Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data

Meichen Yu[1], Kristin A. Linn[2,1], Philip A. Cook[1,3], Mary L. Phillips[4], Melvin McInnis[5], Maurizio Fava[6], Madhukar H. Trivedi[7], Myrna M. Weissman[8,9,10], Russell T. Shinohara[2,1], and Yvette I. Sheline[1,3,11*]

[1] Center for Neuromodulation in Depression and Stress, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States

[2] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, United States

[3] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, United States

[4] Department of Psychiatry, University of Pittsburgh School of Medicine, United States

[5] Department of Psychiatry, University of Michigan School of Medicine, United States

[6] Department of Psychiatry, Massachusetts General Hospital, United States

[7] Department of Psychiatry, University of Texas Southwestern Medical Center, United States

[8] Department of Psychiatry, Columbia University College of Physicians & Surgeons, United States

[9] Division of Epidemiology, New York State Psychiatric Institute, United States

[10] Mailman School of Public Health, Columbia University, United States

[11] Department of Neurology, Perelman School of Medicine, University of Pennsylvania, United States

*Please address correspondence to:

Yvette I. Sheline, M.D., M.S

Email: sheline@mail.med.upenn.edu

Phone: 215-573-0082

Fax: 215-349-5067

Words (abstract): 250

Words (main text): 6230 (excluding References and text in tables and figure captions)

Tables: 2

Figures: 5

Supplementary materials: 2

Tables and figures were added in the end of the manuscript; supplementary materials were separated from this manuscript.

Abstract

Acquiring resting-state functional magnetic resonance imaging (fMRI) datasets at multiple MRI scanners and clinical sites can improve statistical power and generalizability of results. However, multi-site neuroimaging studies have reported considerable non-biological variability in fMRI measurements due to different scanner manufacturers and acquisition protocols. These undesirable sources of variability may limit power to detect effects of interest and may even result in erroneous findings. Until now, there has not been an approach that removes unwanted site effects. In this study, using a relatively large multi-site (4 sites) fMRI dataset, we investigated the impact of site effects on functional connectivity and network measures estimated by widely used connectivity metrics and brain parcellations. The protocols and image acquisition of the dataset used in this study had been homogenized using identical MRI phantom acquisitions from each of the neuroimaging sites, however inter-site acquisition effects were not completely eliminated. Indeed, in the current study we found that the magnitude of site effects depended on the choice of connectivity metric and brain atlas. Therefore, to further remove site effects, we applied ComBat, a harmonization technique previously shown to eliminate site effects in multi-site diffusion tensor imaging (DTI) and cortical thickness studies. In the current work, ComBat successfully removed site effects identified in connectivity and network measures and increased the power to detect age associations when using optimal combinations of connectivity metrics and brain atlases. Our proposed ComBat harmonization approach for fMRI-derived connectivity measures facilitates reliable and efficient analysis of retrospective and prospective multi-site fMRI neuroimaging studies.

KEYWORDS

Aging, atlas, ComBat, fMRI, functional connectivity, graph theory, harmonization, multi-site, network efficiency

1 | INTRODUCTION

Functional magnetic resonance imaging (fMRI), a non-invasive neuroimaging modality with high spatial resolution, enables neural activity to be monitored. Functional connectivity and network measures derived from fMRI data have facilitated the study of the brain's function during development, in aging (Fox and Raichle, 2007; Raichle 2015; Bressler and Menon, 2010), and in the context of various neurological disorders (Bullmore and Sporns, 2009, 2012; Stam, 2014; Fornito et al., 2015, 2016).

Over the last decade, multi-site fMRI studies have become increasingly common (Friedman et al., 2006, 2008; Van Horn and Toga, 2009; Biswal et al., 2010; Gradin et al., 2010; Di Martino et al., 2014; Noble et al., 2017). Indeed, pooling fMRI data from multiple sites can accelerate participant recruitment rates and increase the total sample size of the study, thereby increasing statistical power. Pooling fMRI data is often critical when studying rare disorders and subtle effects and when aiming to generalize the study results to a diverse population (Suckling et al., 2010; McGonigle, 2012; Keshavan et al., 2016; Dansereau et al., 2017). Despite these advantages, multi-site studies are often plagued by non-biological variability that can be attributed to differences in scanner manufacturers, non-standardized imaging acquisition parameters, and other intrinsic factors (Shinohara et al., 2017). These additional sources of unwanted variability may decrease statistical power and lead to spurious results. Many multi-site studies have reported considerable site or scanner effects in fMRI data (Friedman et al., 2006, 2008; Suckling et al., 2008, 2010; Van Horn and Toga, 2009; Gountouna et al., 2010; Brown et al., 2011; McGonigle, 2012; Turner et al., 2013; Forsyth et al., 2014; Feis et al., 2015; Rath et al., 2016; Jovicich et al., 2016; Dansereau et al., 2017; Noble et al., 2017; Abraham et al., 2017). However, most of these studies only describe the problem or report the magnitude of site effects in fMRI measurements.

A few studies have attempted to mitigate site effects by standardizing protocols and image acquisition parameters (Friedman et al., 2008; Glover et al., 2012; Shinohara et al., 2017; Oh et al., 2017; Kochunov et al., 2018; Chavez et al., 2018). However, it has been shown that scanner-to-scanner variation arising from the use of scanners from different manufacturers is not eliminated completely by the standardization of acquisition parameters (Jovicich et al., 2016; Noble et al., 2017), for instance, by use of phantom-based imaging acquisitions (Delaparte et al., 2017). To our knowledge, until now, there has been only one attempt to diminish scanner differences in multi-site resting-state fMRI post-acquisition. The authors used an independent component analysis (ICA) based approach that reduced differences across sites in some resting-state network connectivity measures but did not fully eliminate the structured noise arising from different scanners (Feis et al., 2015).

Recently, our group adapted ComBat harmonization (Johnson et al., 2007) to model and remove site effects in multi-site DTI (Fortin et al., 2017) and cortical thickness (Fortin et al., 2018) measurements. ComBat was originally designed to correct so-called "batch effects" in genomic studies (Johnson et al., 2007) that arise due to processing high-throughput genomic data in different laboratories with different equipment at different times. In our previous studies, we demonstrated that the ComBat harmonization technique successfully removed unwanted non-biological variability, while preserving biological associations between participant age and DTI (fractional anisotropy and mean diffusivity), as well as the association between age and cortical thickness measurements.

In this study, we quantified the site effects in functional connectivity and several brain network measures in the multi-site Establishing Moderators and Biosignatures of Antidepressant Response in

Clinical Care (EMBARC) dataset that was acquired at four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX), and the University of Michigan (UM). Our main objectives were to: (1) remove any identified site effects using ComBat harmonization, and (2) preserve the commonly reported negative correlation between age and functional connectivity within the default mode network (DMN; Damoiseaux et al., 2008; Koch et al., 2010; Grady et al., 2010; Tomasi and Volkow, 2012; Ferreira and Busatto, 2013; Damoiseaux, 2017), as well as preserve previously reported negative correlations between age and network efficiency measures (Achard and Bullmore, 2007; Ajilore et al., 2014). Objective (2) was important to demonstrate that the ComBat technique did not remove important, biologically relevant information. A recently published multi-site autism study (Abraham et al., 2017) reported that the magnitude of site effects was influenced by the choice of functional connectivity metrics and brain parcellation. Therefore, we investigated the degree to which widely used functional connectivity and network metrics derived from a number of brain parcellations were affected by scanner-to-scanner variation and how ComBat harmonization performed in each setting. We hypothesized that (1) considerable site effects exist in both functional connectivity and network efficiency measures calculated from non-harmonized multi-site fMRI data; (2) the magnitude of site effects is not constant across different connectivity metrics and brain parcellations; and (3) ComBat harmonization can be used to remove site effects in connectivity and network measures while preserving age-related associations for numerous combinations of connectivity metrics and brain parcellations.

2 | MATERIAL AND METHODS

2.1 Participants

This study considered 200 unmedicated depressed patients with major depressive disorder (MDD) and 40 healthy subjects recruited for EMBARC that have been analyzed in several previous studies (Greenberg et al., 2015; Trivedi et al., 2016; Webb et al., 2016; Fortin et al., 2018). The current study concentrates on the harmonization of multi-site fMRI-based functional connectivity and network measures. The participants were recruited and scans acquired at four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX), and the University of Michigan (UM). All participants provided written informed consent and the institutional review boards from the four clinical sites approved all study procedures. For both patients and healthy individuals, the Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P; First et al., 2002) was used as inclusion criteria to diagnose the presence or absence of depressive symptoms. The Hamilton Depression Rating Scale (HAMD; Hamilton, 1960) and Quick Inventory for Depression Symptomatology (QIDS; Rush et al., 2003) depression scores were used to estimate depressive severity. Anxiety and depressive severity were also assessed using the Mood and Anxiety Symptom Questionnaire (MASQ; Watson and Clark, 1991), including three subscales: general distress (MASQ-GD), anhedonic depression (MASQ-AD), and anxious arousal (MASQ-AA). The individuals were eligible for the study if they met the following inclusion criteria: (1) age 18–65; (2) reported age of depression onset before age 30; (3) fluent in English. Eleven depressed patients and one healthy individual were excluded due to excessive motion (> 4 mm), low slice signal-to-noise ratio (< 80), and severe slice artifacts in MRI data. The final sample included 189 MDD patients and 39 healthy

individuals. The distribution of age, sex, handedness, and education level were matched between the two groups.

2.2 Image acquisition and data preprocessing

All four sites used 3T scanners, however, the manufacturer differed from site to site: CU used a GE SIGNA HDx 3T scanner, MGH used a Siemens TIM Trio 3T scanner, TX used a Philips Achieva 3T scanner, and UM used a Philips Ingenia 3T scanner (Greenberg et al., 2015; Trivedi et al., 2016; Fortin et al., 2018). Imaging parameters at each site are described in Table 1.

Prior to the project's initiation, in close collaboration with MR physics teams at the acquisition sites, a homogenized imaging protocol was developed to minimize acquisition-related site differences. In particular, data were collected using identical MRI phantom acquisitions from each of the neuroimaging sites. Well established routines for using phantoms were employed to perform quality assurance on the scanners used in this study. However, although the phantom-based approach minimized the inconsistency of signal-to-noise acorss scanners over the time and other variability in image acquisition and quality ascross sites, the inter-site acquisition effects were not completely eliminated (Delaparte et al., 2017). Therefore, we employed a post-processing procedure that further harmonized the fMRI functional connectivity matrices of subjects across the 4 sites.

T1-weighted (T1) images were processed using the ANTS Cortical Thickness pipeline available in the antsCorticalThickness.sh script in advanced normalization tools (ANTs ; Avants et al., 2011a; Tustison et al., 2014). The workflow is sketched out as follows: (1) N4 bias correction to minimize field inhomogeneity (Tustison et al., 2010); (2) brain extraction using an optimal population-specific template created by a Symmetric Group Normalization framework (Avants et al., 2010); (3) Atropos probabilistic six-tissue segmentation (Avants et al., 2011b); (4) DiReCT-based cortical thickness estimation (Das et al., 2009); (5) SyN deformable spatial registration to the population-specific template (Klein et al., 2009).

Resting-state time series data from each participant were processed using the XCP Engine (Ciric et al., 2017), which uses an optimized confound regression procedure to reduce the influence of subject motion (Satterthwaite et al., 2017). Each subject contributed time series data from two resting-state fMRI sessions. The workflow of functional data preprocessing is summarized as follows: (1) removal of the four initial volumes of the Blood-oxygen-level Dependent (BOLD) signals to achieve signal stabilization; (2) realignment of functional images using MCFLIRT (Jenkinson et al., 2002); (3) removal of nine confounding signals (six motion parameters+global/white matter/cerebral spinal fluid) as well as the temporal derivative, quadratic term, and temporal derivatives of each quadratic term (36 regressors total) (Satterthwaite et al., 2017); (4) co-registration of functional images to the T1 image using boundary-based registration (Greve and Fischl, 2009); (5) alignment of the co-registered images to template space using the ANTs-transform for the T1 image as above; and (6) temporal filtering of time series between 0.01 and 0.08 Hz as in previous studies (Biswal et al., 1995) using a first-order Butterworth filter. In this study, all regressors, including motion parameters and confound time courses, were band-pass filtered to the same frequency range as the time series data to prevent frequency-dependent mismatch during confound regression (Hallquist et al., 2013). Functional images were smoothed using a Gaussian convolution at 6mm full width at half maximum.

2.3 Parcellation

To investigate the influence of different parcellations on functional connectivity measures across sites and subsequent harmonization, we partitioned the brain of each participant into cortical and subcortical ROIs using the following three different whole-brain atlases (one anatomical and two functional): (1) 78 cortical and 12 subcortical ROIs identified by automated anatomical labeling (AAL) (Tzourio-Mazoyer et al., 2002); (2) 264 cortical and subcortical ROIs of the widely-used functional Power atlas (Power et al., 2011); and (3) 333 cortical and subcortical ROIs from the functional Gordon atlas (Gordon et al., 2016). The ROIs and MNI-space centroids of the AAL, Power, and Gordon atlases can be found in the Supplementary figure S1 and Supplementary Materials 2.

2.4 Functional connectivity

For each participant, whole-brain functional connectivity between all brain regions was constructed pairwise from the preprocessed fMRI data. The fMRI time series were extracted from each voxel and averaged within each ROI of the three atlases (AAL, Power, and Gordon). The functional connectivity between time series for all pair-wise ROIs was estimated by calculating two commonly used connectivity metrics: Pearson correlation and wavelet coherence. For Pearson correlation, the correlation coefficients were Fisher-transformed in order to draw more statistically interpretable conclusions about the magnitude of the correlations (Cohen and Esposito, 2016; Doucet et al., 2017). Due to poor signal quality and signal dropout, we excluded 61 ROIs from the Power atlas and 26 ROIs from the Gordon atlas, which resulted in 203 and 307 ROIs for the Power and Gordon atlases, respectively. All subsequent analyses were performed using the 90×90 AAL-atlas, 203×203 Power-atlas, and 307×307 Gordon-atlas connectivity matrices based on both Fisher-transformed Pearson correlation coefficients and raw wavelet coherence values from all participants.

2.5 Model for functional connectivity matrix harmonization

Based on the literature (Friedman et al., 2008; Feis et al., 2015; Rath et al., 2016; Dansereau et al., 2017), we speculated that measurements such as DTI fractional anisotropy (Fortin et al., 2017), MRI cortical thickness (Fortin et al., 2018), and fMRI functional connectivity (the present study) would differ among the four sites (CU, MGH, TX and UM) due to systematic bias and non-biological variability attributable to the use of different scanners and different imaging parameters.

In this study, we used ComBat (Johnson et al., 2007) to reduce potential biases and non-biological variability induced by site and scanner effects. ComBat uses a multivariate linear mixed effects regression with terms for biological variables and scanner to model imaging feature measurements. The method uses empirical Bayes to improve the estimation of the model parameters for studies with small sample sizes. Here, we reformulate the ComBat model so that it can be applied to functional connectivity matrices estimated using Pearson correlation and wavelet coherence in combination with the AAL, Power, and Gordon atlases (i.e., six combinations: Correlation-AAL, Coherence-AAL, Correlation-Power, Coherence-Power, Correlation-Gordon, and Coherence-Gordon). Since all

connectivity matrices are symmetric, we applied ComBat to connectivity values in the upper triangles of the matrices. Let y_{ijv} represent the connectivity values of imaging site i ($i \in \{1, …, 4\}$), participant j ($j \in \{1, …, 228\}$), and connectivity value v ($v \in \{1, …, 4,005\}$ for the AAL atlas, $v \in \{1, …, 20,503\}$ for the Power atlas, and $v \in \{1, …, 46,971\}$ for the Gordon atlas) between two ROIs. Then, the ComBat model can be written as

$$y_{ijv} = \alpha_v + X_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv}\epsilon_{ijv}$$

where α_v is the average connectivity value for a particular connectivity value v between two ROIs, X_{ij}^T is a design matrix for the covariates of interest (age , gender, and group), and β_v is a vector of regression coefficients corresponding to X. As in Fortin et al. (2018), we further assume that the residual terms ϵ_{ijv} arise from a normal distribution with zero mean and variance σ_v^2. The terms γ_{iv} and δ_{iv} represent the additive (or location parameter) and multiplicative (or scale parameter) site effects of site i for connectivity value v, respectively. The ComBat-harmonized functional connectivity values were then defined as

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \widehat{\alpha_v} - X_{ij}\widehat{\beta_v} - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha_v} + X_{ij}\widehat{\beta_v},$$

where γ_{iv}^* and δ_{iv}^* are the empirical Bayes estimates of γ_{iv} and δ_{iv}, respectively. Thus, ComBat simultaneously models and estimates biological and non-biological terms and algebraically removes the estimated additive and multiplicative site effects. Of note, in the ComBat model, we included age, sex, and group as covariates to preserve important biological trends in the data and avoid overcorrection.

In this study, we performed the ComBat harmonization analyses for the six-combinations of connectivity matrices in two sessions (S1 and S2), separately. ComBat harmonization analyses were performed using a publicly available MATLAB package hosted at https://github.com/Jfortin1/ComBatHarmonization/tree/master/Matlab.

2.6 Visualization and evaluation of functional connectivity harmonization

We used Kruskal-Wallis tests to quantify the magnitude of site effects in functional connectivity between all pairwise ROIs before and after applying ComBat harmonization to each of the six metric-atlas combinations (Correlation-AAL, Coherence-AAL, Correlation-Power, Coherence-Power, Correlation-Gordon, and Coherence-Gordon). The p-values were adjusted for multiple comparisons by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at 5%, separately for each combination (AAL: 4,005 comparisons; Power: 20,503 comparisons; Gordon: 46,971 comparisons). The numbers and percentages of connectivity values that were significantly different (after FDR correction) across the 4 sites for the six combinations are summarized in Table 2. The FDR-corrected p-values can be found in

supplementary Figure 1. We visualized site effects using boxplots of connectivity values between signals of two randomly selected ROIs for each atlas across the four sites (Figure 2; Supplementary Figures S3-S6, subplot A). The selected ROIs were consistent for the same atlas using the two connectivity measures.

We also performed a principal component analysis (PCA) on the functional connectivity values in the upper triangle of the connectivity matrices for the six metric-atlas combinations before and after ComBat harmonization. In Supplementary Figure S3A-S6A, subplot A and S7, we plotted three-dimensional scatter plots of the first three PC scores from the PCA. If the connectivity values were significantly different (Kruskal-Wallis tests; FDR corrections) across sites before or after ComBat harmonization, the corresponding PC scores are likely to be associated with site, and we would expect to see data from the same scanner roughly clustered together in the scatter plots.

To evaluate whether the assumed empirical Bayes priors for the location (γ) and scale (δ) parameters in the ComBat harmonization model reasonably reflect the observed data, we overlayed the empirical and prior distributions of γ and δ in Figure 3 and Supplementary Figures S3-S6, subplot C.

We applied ComBat harmonization to the connectivity matrices from each fMRI session, separately. We present visualizations of the site effects and plots of ComBat model parameters for the first session. Plots generated from the second session were similar and therefore not included. Figure 2 demonstrates differences in the distribution of functional connectivity across sites for the first session. Figure 3 provides a visualization of the goodness of fit of the ComBat model's prior assumptions to the observed data for the first session. Following ComBat, we extracted four network measures from the harmonized connectivity matrices and averaged these measures across the two sessions. Henceforth, we focus on analyzing the average network measures, which included weigthed DMN connectivity, nodal strength, local efficiency, and global efficiency. We formally define these measures in Section 2.7.

2.7 Calculation of network measures

In order to ensure that our post-processing harmonization did not remove meaningful biological variability along with the undesireable site effects we conducted an additional analysis. As the default mode network (DMN) has been found to have larger negative associations between age and functional connectivity metrics than other resting-state networks (Tomasi and Volkow, 2012; Ferreira and Busatto, 2013; Damoiseaux, 2017), we selected it to conduct analysis of age-related effects. In this study functional connectivity and local network metrics (quantified by weighted nodal strength and nodal efficiency) were thus calculated in the DMN. Global network topology was characterized by weighted global efficiency. The computation details of these connectivity and network metrics are described in the following paragraphs.

For each atlas, the DMN network connectivity was defined by the summation of the functional connectivity values within the DMN ROIs normalized by the number of DMN ROIs for each atlas. The DMN nodal strength was computed by first summing the functional connectivity values (link weights) for each pair of ROIs and then summing up the nodal level connectivity values within the DMN ROIs of each specific atlas. The weighted local and global efficiency (Latora and Marchiori, 2001) were computed using the weighted shortest path length (L^W; Dijkstra, 1959), which is the shortest sum of connection

length (inverse of the connectivity values or link weights) between two nodes (or ROIs; Rubinov and Sporns, 2010). The weighted nodal efficiency (E^W_{nodal}) was calculated as the inverse of the harmonic mean of L^W from one node to all other nodes, as follows:

$$E^W_{nodal} = \frac{1}{N-1} \sum_{j \epsilon G} \frac{1}{L^W_{i,j}}$$

where N is the number of nodes in graph G (represented by the AAL, Power and Gordon connectivity matrices in this study) and $L^W_{i,j}$ is the weighted shortest path length between node i and j.

Weighted local efficiency (E^W_{local}) for a node is defined as the average weighted nodal efficiency among the neighboring nodes of that node (excluding the reference node), as follows:

$$E^W_{local} = \frac{1}{N_{G_i}(N_{G_i} - 1)} \sum_{j,k \epsilon G_i} \frac{1}{L^W_{j,k}}$$

where N_{G_i} is the number of nodes in subgraph G_i that consists of all neighboring nodes of node i, but excluding node i. For the weighted DMN local efficiency, weighted local efficiency values were computed for each ROI and then summed up within the DMN ROIs of each specific atlas.

The weighted global efficiency (E^W_{global}) was calculated as the average weighted nodal efficiency of nodes in a graph G, as follows:

$$E^W_{global} = \frac{1}{N(N-1)} \sum_{i \neq j \epsilon G} \frac{1}{L^W_{i,j}}$$

All the network efficiency measures were computed using the Brain Connectivity Toolbox (BCT) (Rubinov and Sporns, 2010).

For each participant, we first computed DMN network connectivity, DMN nodal strength, weighted DMN local efficiency, and weighted global efficiency for each of the six combinations (3 atlases × 2 connectivity metrics) before and after applying ComBat harmonization. Next, we averaged the values of each participant's network connectivity or efficiency measures from the two sessions. Then, we tested the global null hypothesis of no differences across sites in the network connectivity or efficiency measures using Kruskal-Wallis tests (in total, 2 conditions (before and after ComBat) × 2 connectivity metrics × 4 network measures × 3 atlases = 48 comparisons) with a separate FDR correction at 5% within each condition (2 connectivity metrics × 4 network measures × 3 atlases = 24 comparisons), separately.

2.8 Preservation of biological variability

An optimal harmonization technique should be able to remove most or all non-biological sources of variability caused by site and scanner, yet preserve or increase statistical power to detect biological associations. In this study, there was a broad participant age range (18 to 65 years), enabling investigation of age-related associations. Therefore, we investigated whether negative associations

between age and DMN network connectivity as well as associations between age and network efficiency measures were preserved or made stronger when estimated using ComBat-harmonized data.

We computed the Spearman correlation between each network (or connectivity) measure and age. The p-values were adjusted for multiple comparisons (in total, 2 conditions (before and after ComBat) × 2 connectivity metrics × 4 network measures × 3 atlases = 48 comparisons) by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995). As before, the FDR corrections were applied separately within condition (2 connectivity metrics × 4 network measures × 3 atlases = 24 comparisons). A significance level of p < .05 was used for these tests. Note that, for the Power and Gordon atlases, we used the original definitions of DMN ROIs from Power et al., 2011 and Gordon et al., 2011, respectively; for the AAL atlas, we defined DMN ROIs according to a review article, Rosazza and Minati, 2011. For details of the definition of the DMN ROIs for each atlas, refer to Supplementary Table 1, Figure S1 and Supplementary materials 2. Figure S1, subplot b, c were visualized with BrainNet Viewer (version 1.5, Xia et al., 2013, http://www.nitrc.org/projects/bnv/).

2.9 Statistical analysis of demographic characteristics

Statistical analyses for demographic characteristics of participants were performed using MATLAB (R2017a). Age and educational level were compared among the four sites using Kruskal-Wallis tests followed by Mann Whitney-U tests when appropriate. All p-values from the Mann Whitney-U tests were adjusted for multiple comparisons by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at 5%. We tested for differences in the gender and clinical group distribution among the four sites using Pearson's chi-squared (χ^2) tests.

3 | RESULTS

3.1 Demographic characteristics

The distribution of demographic characteristics across imaging sites is shown in Figure 1. The age distribution (p = .001) was imbalanced across sites; subjects in the TX site were older than the other sites (TX-CU: p < .001; TX-MGH: p = .04; TX-UM: p = .04; FDR correction). There were also weak site effects (p = .03) in the level of education, but no differences were found regarding educational level between pairwise sites after FDR correction. Gender (p = .14) and depressed/control (p = .34) distributions were equally distributed across sites.

3.2 Visualization and evaluation of ComBat harmonization

Functional connectivity values estimated by Pearson correlation showed much stronger site effects than those by wavelet coherence for analyses using the AAL and Power atlases as well as the Gordon atlas (Table 2). Moreover, the AAL atlas had a much larger percentage of connectivity values that differed significantly across the four sites than the Power and Gordon atlases (Table 2). Following ComBat harmonization, there were no statistically significant site effects in the functional connectivity values of the six metric-atlas combinations.

Figure 2 displays boxplots of functional connectivity values between two randomly selected ROIs (for AAL: TPOmid.R and ACG.R; for Power: two regions in the visual cortex) for the Pearson correlation-AAL atlas and Wavelet coherence-Power atlas combinations. The connectivity values between these two ROIs showed consistent patterns across the four sites: for the AAL atlas, MGH values were generally higher than the other three sites; for the Power atlas, MGH values were generally lower than the other three sites. For the Gordon atlas, MGH and UM values were generally lower than the other two sites (see Supplementary Figures S5-S6, subplot A, for the Gordon atlas results). After ComBat harmonization, these significant site effects were dramatically reduced in all six metric-atlas combinations. See Supplementary Figures S3-S6, subplot A for boxplots of functional connectivity for other metric-atlas combinations, before and after ComBat.

To further visualize site effects, we generated three-dimensional scatter plots of the first three principal component (PC) scores obtained from the functional connectivity matrices (see Supplementary Figures S3-S6, subplot B; Figure S7). For all three atlases, the second PC scores from CU and/or MGH patients showed distinct separation from those of TX and UM, particularly when using Pearson correlation. These visual site effects are much less noticable in the ComBat-harmonized data: the first three PC scores were not clearly associated with site by visual inspection for any of the metric-atlas combinations (Supplementary Figures S3-S6, subplot B; Figure S7).

For all metric-atlas combinations, the ComBat-harmonized prior distributions appear to fit the empirical distributions of both the location (γ) and scale (δ) parameters well (Figure 3 and Supplementary Figures S3-S6, subplot C). Visual inspection of these overlaid distributions suggests that the ComBat model used appropriate prior information to capture the underlying site effects in the functional connectivity matrices. Furthermore, for each of the three atlases, the distributions of γ and δ reflected the observed lower magnitudes in the distribution location and variability of MGH values compared with the values from other three sites, i.e., for MGH, $\delta < 1$ on average using Pearson correlation and $\delta < 1$ and $\gamma < 0$ using wavelet coherence.

Before performing ComBat harmonization on the functional connectivity matrices, all the network connectivity and efficiency measures estimated by Pearson correlation and a majority of the measures estimated by wavelet coherence displayed statistically significant site effects with similar patterns across sites (MGH values were generally lower than all the other sites). Figure 5A displays log-transformed p-values from the global site effect tests for each metric-atlas combination before and after ComBat. The p-values for network connectivity and the efficiency measures estimated by Pearson correlation were considerably more significant than those estimated by wavelet coherence (see Supplementary Figures S8-S10, subplot A, for corresponding boxplot visualizations of the site effects). After ComBat harmonization, there were no remaining statistically significant site effects for any metric-atlas combination.

Prior to functional connectivity harmonization using ComBat, there were statistically significant site effects across all network connectivity and efficiency measures estimated by both Pearson correlation and wavelet coherence when using the Gordon atlas (Figure 5A). In contrast, prior to harmonization using Combat, when using the AAL or Power atlases, a number of the measures estimated by wavelet coherence did not display significant differences across sites (Figure 5A, Supplementary Figures S9-S10, subplot A).

3.3 Preservation of biological variability

3.3.1 By connectivity metric

ComBat harmonization preserved or strengthened the anti-correlations between age and DMN functional connectivity as well as between age and network efficiency measures. The p-values and correlation values for each metric-atlas combination are displayed in Figures 5B and 5C, respectively, where we see more significant p-values and stronger correlations post-ComBat. This result was true for both Pearson correlation and wavelet coherence connectivity, with wavelet coherence identifying the strongest anti-correlations both before and after ComBat harmonization. Supplementary Figures S8-S9, subplot B, display scatter plots associated with each correlation value.

3.3.2 By atlas

Using the original data without ComBat harmonization, the Gordon atlas showed significant site effects in all network connectivity and efficiency measures estimated by both Pearson correlation and wavelet coherence (Figure 5A). However, for AAL and Power atlases, there were no site effects even in some non-harmonized measures estimated by wavelet coherence (Figure 5A).

   As shown in Figure 5B, ComBat harmonization strengthened the estimated anti-correlations between age and network measures across all three atlases. In particular, for the AAL and Power atlases, ComBat harmonization uncovered significant anti-correlations that were not detected when using the non-harmonized data (Figure 4B, Figure 5B). Among the three atlases, the AAL atlas identified the fewest significant anti-correlations both before and after ComBat harmonization and the magnitudes were generally smaller than those identified by other atlases. The Power atlas identified stonger anti-correlations than the other two atlases post-ComBat. Moreover, a majority of the network measures estimated by the correlation-AAL combination were not negatively associated with age, even after performing ComBat harmonization (see Supplementary Figures S8B-S10B, for scatter plots associated with the p and correlation values in Figure 5B,C).

   Overall, ComBat harmonization not only removed unwanted site effects in network connectivity and efficiency measures calculated from functional connectivity matrices but also preserved or increased the estimated underlying correlations with age. Some specific combinations of atlases and connectivity metrics appear to be better than others with respect to revealing significant relationships with age. When considering both site effect removal and correlation with age, we found that the coherence-Power combination performed optimally.

4 | DISCUSSION

In this study, we investigated the degree to which combining data from different scanners in a multi-site study could affect downstream analyses of fMRI-based functional connectivity and network efficiency measures. We implemented several visualization techniques and statistical tests to visualize and quantify the scanner effects. We performed ComBat harmonization on fMRI-based functional connectivity matrices to remove site effects before extracting DMN connectivity and network measures. We quantified the site effects and the performance of ComBat harmonization using two different metrics to compute connectivity and three different brain atlases. We demonstrated that ComBat

harmonization can successfully remove site effects in the functional connectivity matrices, thereby leading to network connectivity and efficiency measures that are also not different across sites for any choice of connectivity metric and atlas. Moreover, we found that using wavelet coherence with the Power atlas resulted in the best power to detect anti-correlations between age and DMN functional connectivity as well as network efficiency measures following ComBat harmonization, suggesting the best preservation of underlying biological signal with this combination.

4.1 ComBat harmonization removes site effects

As previous studies (Van Horn and Toga, 2009; Dansereau et al., 2017) have consistently reported the existence of considerable site effects in multi-site fMRI measurements that cannot be removed by performing ICA-based approaches (Feis et al., 2015), we tested whether ComBat harmonization could eliminate site effects in several fMRI-based functional connectivity and network measures. Of note, we only performed ComBat harmonization on the original functional connectivity matrices and then subsequently calculated network connectivity and efficiency measures from the harmonized connectivity matrices. Notably, we did not find statistically significant site effects in the downstream network measures.

Given the excellent performance of ComBat in DTI (Fortin et al., 2017), MRI-based cortical thickness (Fortin et al., 2018), and fMRI (current study) measurements, we conclude that this harmonization method is a reliable and powerful technique that can be widely applied to different neuroimaging modalities and summary measurements.

4.2 Wavelet coherence outperforms Pearson correlation

In this study, to investigate the effects of connectivity metrics on multi-site fMRI measurements and the performance of ComBat harmonization, we used both Pearson correlation and Wavelet coherence to estimate the fMRI functional connectivity. Previous studies have shown that wavelet coherence outperforms Pearson correlation with respect to sensitivity to outliers caused by motion artifact (Huber, 2004; Achard et al., 2006). Additionally, using coherence avoids the need to remove negative correlation coefficients to calculate network measures (Achard and Bullmore, 2007; Bassett et al., 2011) and robustly extracts frequency-specific information from the time series without picking up on edge effects of band-pass filtering (Percival and Walden, 2000; Zhang et al., 2016). However, currently there is no study comparing the sensitivity to scanner differences of the two connectivity metrics applied to fMRI data. Our results indicate that ComBat harmonization can remove scanner effects from the data, regardless of the choice of connectivity metric. However, wavelet coherence-based measures showed weaker differences across sites than Pearson correlation-based measures in non-harmonized data. Moreover, wavelet coherence measures generally resulted in stronger anti-correlations between age and the connectivity and network measures across all the three atlases (AAL, Power and Gordon) both before and after harmonization. For multi-site fMRI studies, this result suggests that wavelet coherence may be preferable to Pearson correlation when extracting connectivity and network summary outcomes.

4.3 Power atlas outperforms AAL and Gordon atlases

We also studied the effects of three atlases (AAL, Power and Gordon) on multi-site fMRI measurements and the performance of ComBat harmonization. A larger percentage of connections between ROIs were significantly affected by site in the AAL atlas than in the Power and Gordan atlases. These results are consistent with previous findings that in multi-site fMRI studies, functional atlases extracted from large resting-state fMRI datasets outperform traditional anatomical atlases (Abraham et al., 2017). For all three atlases, site effects in the functional connectivity and network measures were successfully removed by ComBat harmonization. However, all the network connectivity and efficiency measures using the AAL atlas were less correlated with age than those using the Power and Gordan atlases, suggesting that the AAL atlas may not be as sensitive to underlying biological variability (assessed using age in this study) when using multi-site fMRI data. Interestingly, we did not find significant site effects using the Power atlas among non-ComBat-harmonized network efficiency measures estimated by wavelet coherence. In contrast, the AAL and Gordon atlases demonstrated strong site effects in these non-ComBat-harmonized network measures. Overall, we concluded that the Power atlas outperforms the AAL and Gordon atlases with respect to post-ComBat analyses of biological variability.

4.4 Strengths, limitations and future direction

Our current study has several strengths: (1) we investigated six combinations of two connectivity metrics and three atlases, and thus were able to explore the ability of ComBat harmonization to remove site effects and to identify combinations of connectivity metrics (wavelet coherence) and atlases (Power and Gordon) that best preserved age-related anti-correlations after harmonization; (2) we used a relatively large sample (228 participants), therefore providing relatively reliable and convincing results; (3) by using the ComBat model, which is generic in its formulations and thus could easily be generalized to additional imaging modalities, our findings may have implications for multi-site electroencephalography, magnetoencephalography and other neurophysiological and neuroimaging datasets; (4) ComBat has been implemented the in MATLAB and R (https://github.com/Jfortin1/ComBatHarmonization) and in Python (https://github.com/ncullen93/neuroCombat) making the technique available and largely applicable to analysts using a variety of different software packages for image processing.

   There are also several limitations that should be considered and improved in future studies. First, several previous fMRI studies (Friedman et al., 2006; Brown et al., 2011; Forsyth et al., 2014; Keshavan et al., 2016; Rath et al., 2016; Noble et al., 2017; Shinohara et al., 2017) used traveling-subject datasets in which the same participants were scanned across sites to reduce the subject effects. One recent study (Noble et al., 2017) using a small dataset (8 subjects scanned at each of 8 sites) found that the subject differences were stronger than potential site effects. Although, our ComBat harmonization technique was tested on different participants scanned across different sites with heterogeneous protocols, we speculate that ComBat harmonization will also have excellent performance on removing any non-biological variations when applied to traveling-subject datasets, however this remains to be proven. Second, in some longitudinal datasets, the same participants may be scanned on different scanners over multiple time points. However, the current ComBat harmonization model cannot be directly applied to this type of longitudinal data. Therefore, in the future, we plan to develop a time-dependent ComBat

algorithm to study longitudinal fMRI connectivity and network properties. Third, in the present study, we tested the performance of ComBat harmonization on two functional connectivity metrics and three atlases (six combinations). Although the ComBat model does not require assumptions on connectivity metrics and atlases, we found that the choices of connectivity metrics and atlases had a strong influence on the magnitude of site effects in fMRI measurements and on preserving biological variability (age in this study). Therefore, future work exploring the performance of ComBat harmonization in other combinations of connectivity metrics (e.g. partial correlation) and atlases (anatomical atlas: Brodmann, 1909; Desikan et al., 2006; functional atlas: Yeo et al., 2011; Wig et al., 2014; Glasser et al., 2016; Schaefer et al., 2017) is warranted. Finally, in this project we focused on the ability of ComBat harmonization to preserve age-related associations with several network connectivity and efficiency measures. However, previous studies (Bullmore and Sporns, 2012; Stam, 2014; Fornito et al., 2015; Yu et al., 2016, 2017) have shown that functional brain network organizations are highly correlated with other demographic (e.g. gender, educational level), clinical phenotypes (e.g. disease severity for neurological disorders), and pathological biomarkers (e.g. amyloid-β_{42} and tau proteins in Alzheimer's disease). In particular, the EMBARC functional dataset was originally designed to study the potential differences on fMRI measurements between MDD patients and healthy controls (Greenberg et al., 2015; Trivedi et al., 2016; Webb et al., 2016). Future studies will focus on whether the ComBat-harmonized fMRI data preserve functional brain networks (Kaiser et al., 2015; Gong and He, 2015) associated with depression, and whether the abnormal network attributes in MDD after ComBat harmonization are associated with patients' symptoms (Sheline et al., 2009, 2010; Otte et al., 2016; Williams, 2016).

5 | CONCLUSION

ComBat harmonization is a powerful technique for removing site effects in functional connectivity matrices, network connectivity, and efficiency measures. In addition, it preserves or strengthens the power to detect age-related anti-correlations in network connectivity and efficiency measures. In the current multi-site fMRI study, the optimal performance of ComBat harmonization was obtained by using wavelet coherence to extract functional connectivity from the Power atlas segmentation of functional brain images.

FINANCIAL DISCLOSURES

Except for federal grants all authors report no biomedical financial interests or potential conflicts

of interest.

REFERENCES

Abraham A, Milham MP, Di Martino A, Cameron Craddock R, Samaras D, Thirion B, Varoquaux G (2017): Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147:736–745.

Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E (2006): A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J Neurosci 26(1):63–72.

Achard S, Bullmore E (2007): Efficiency and cost of economical brain brain functional networks. PLoS Comput Biol 3:e17.

Achard S, Delon-Martin C, Vértes PE, Renard F, Schenck M, Schneider F, et al. (2012): Hubs of brain functional networks are radically reorganized in comatose patients. Proc Natl Acad Sci USA 109:20608–13.

Ajilore O, Lamar M, Kumar A (2014): Association of brain network efficiency with aging, depression, and cognition. Am J Geriatr Psychiatry 22(2):102–10.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010): The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49(3):2457–2466.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011a): A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage 54:2033–2044.

Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC (2011b): An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics; 9:381–400.

Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST (2011): Dynamic reconfiguration of human brain networks during learning. Proc Natl Acad Sci U S A 108(18):7641–6.

Benjamini Y, Hochberg Y (1995): Controling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B Met 57:289–300.

Biswal B1, Yetkin FZ, Haughton VM, Hyde JS (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34(4):537–41.

Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski AM, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kötter R, Li SJ, Lin CP, Lowe MJ, Mackay C, Madden DJ, Madsen KH, Margulies DS, Mayberg HS, McMahon K, Monk CS, Mostofsky SH, Nagel BJ, Pekar JJ, Peltier SJ, Petersen SE, Riedl V, Rombouts SA, Rypma B, Schlaggar BL, Schmidt S, Seidler RD, Siegle GJ, Sorg C, Teng GJ, Veijola J, Villringer A, Walter M, Wang L, Weng XC, Whitfield-Gabrieli S, Williamson P, Windischberger C, Zang YF, Zhang HY, Castellanos FX, Milham MP (2010): Toward discovery science of human brain function. Proc Natl Acad Sci U S A 107(10):4734–9.

Bressler SL, Menon V (2010): Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn Sci 14(6):277–90.

Brodmann K (1909/2006): Localization in the Cerebral Cortex, translated by Garey LJ. New York: Springer.

Brown GG, Mathalon DH, Stern H, Ford J, Mueller B, Greve DN, McCarthy G, Voyvodic J, Glover G, Diaz M, Yetter E, Ozyurt IB, Jorgensen KW, Wible CG, Turner JA, Thompson WK, Potkin SG; Function Biomedical Informatics Research Network (2011): Multisite reliability of cognitive BOLD data. Neuroimage 54(3):2163–75.

Bullmore E, Sporns O (2009): Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10:186–98.

Bullmore E, Sporns O (2012): The economy of brain network organization. Nat Rev Neurosci 13:336–49.

Chavez S, Viviano J, Zamyadi M, Kingsley PB, Kochunov P, Strother S, Voineskos A (2018): A novel DTI-QA tool: Automated metric extraction exploiting the sphericity of an agar filled phantom. Magn Reson Imaging 46:28–39.

Ciric R, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, Shinohara RT, Elliott MA, Eickhoff SB, Davatzikos C, Gur RC, Gur RE, Bassett DS, Satterthwaite TD (2017): Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage. 154:174–187.

Cohen JR, D'Esposito M (2016): The Segregation and Integration of Distinct Brain Networks and Their Relationship to Cognition. J Neurosci 36(48):12083–12094.

Damoiseaux JS, Beckmann CF, Arigita EJ, Barkhof F, Scheltens P, Stam CJ, Smith SM, Rombouts SA (2008): Reduced resting-state brain activity in the "default network" in normal aging. Cereb Cortex 18(8):1856–64.

Damoiseaux JS (2017): Effects of aging on functional and structural brain connectivity. Neuroimage 160:32–40.

Dansereau C, Benhajali Y, Risterucci C, Pich EM, Orban P, Arnold D, Bellec P (2017) Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. Neuroimage 149:220–232.

Das SR, Avants BB, Grossman M, Gee JC (2009): Registration based cortical thickness measurement. Neuroimage 45(3):867–79.

Delaparte L, Yeh FC, Adams P, Malchow A, Trivedi MH, Oquendo MA, Deckersbach T, Ogden T, Pizzagalli DA, Fava M, Cooper C, McInnis M, Kurian BT, Weissman MM, McGrath PJ, Klein DN, Parsey RV, DeLorenzo C (2017): A comparison of structural connectivity in anxious depression versus non-anxious depression. J Psychiatr Res 89:38–47.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006): An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968.

Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Müller RA, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP (2014): The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19 (6):659–667.

Dijkstra EW (1959): A note on two problems in connection with graphs. Numerische Math 1:269–271.

Doucet GE, Bassett DS, Yao N, Glahn DC, Frangou S (2017): The Role of Intrinsic Brain Functional Connectivity in Vulnerability and Resilience to Bipolar Disorder. Am J Psychiatry 174(12):1214–1222.

Feis RA, Smith SM, Filippini N, Douaud G, Dopper EG, Heise V, Trachtenberg AJ, van Swieten JC, van Buchem MA, Rombouts SA, Mackay CE (2015): ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. Front Neurosci 9:395.

Ferreira LK, Busatto GF (2013): Resting-state functional connectivity in normal brain aging. Neurosci Biobehav Rev 37(3):384–400.

First MB, Spitzer RL, Gibbon M, et al. (2002): Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P). New York, Biometrics Research, New York State Psychiatric Institute.

Fornito A, Zalesky A, Breakspear M (2015): The connectomics of brain disorders. Nat Rev Neurosci 16:159–72.

Fornito A, Zalesky A, Bullmore E (2016): Fundamentals of Brain Network Analysis. Academic Press, Cambridge.

Forsyth JK, McEwen SC, Gee DG, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos HW, Tsuang MT, van Erp TG, Walker EF, Hamann S, Woods SW, Qiu M, Cannon TD (2014): Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome Longitudinal Study. Neuroimage 97:41–52.

Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur RC, Gur RE, Schultz RT, Verma R, Shinohara RT (2017): Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161:149–170.

Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT (2018): Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120.

Fox MD, Raichle ME (2007): Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8:700–11.

Friedman L, Glover GH; Fbirn Consortium (2006): Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage 33(2):471–81.

Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, Greve DN, Bockholt HJ, Belger A, Mueller B, Doty MJ, He J, Wells W, Smyth P, Pieper S, Kim S, Kubicki M, Vangel M, Potkin SG (2008): Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp 29(8):958–72.

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

Glover GH, Mueller BA, Turner JA, van Erp TG, Liu TT, Greve DN, Voyvodic JT, Rasmussen J, Brown GG, Keator DB, Calhoun VD, Lee HJ, Ford JM, Mathalon DH, Diaz M, O'Leary DS, Gadde S, Preda A, Lim KO, Wible CG, Stern HS, Belger A, McCarthy G, Ozyurt B, Potkin SG (2012): Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J Magn Reson Imaging 36(1):39–54.

Gong Q, He Y (2015): Depression, neuroimaging and connectomics: a selective overview. Biol Psychiatry 77(3):223–235.

Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE (2016): Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb Cortex 26(1):288–303.

Gountouna VE, Job DE, McIntosh AM, Moorhead TW, Lymer GK, Whalley HC, Hall J, Waiter GD, Brennan D, McGonigle DJ, Ahearn TS, Cavanagh J, Condon B, Hadley DM, Marshall I, Murray AD, Steele JD, Wardlaw JM, Lawrie SM (2010): Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49(1):552–60.

Gradin V, Gountouna VE, Waiter G, Ahearn TS, Brennan D, Condon B, Marshall I, McGonigle DJ, Murray AD, Whalley H, Cavanagh J, Hadley D, Lymer K, McIntosh A, Moorhead TW, Job D, Wardlaw J, Lawrie SM, Steele JD (2010): Between- and within-scanner variability in the CaliBrain study n-back cognitive task. Psychiatry Res 184(2):86–95.

Grady CL, Protzner AB, Kovacevic N, Strother SC, Afshin-Pour B, Wojtowicz M, Anderson JA, Churchill N, McIntosh AR (2010): A multivariate analysis of age-related differences in default mode and task-positive networks across multiple cognitive domains. Cereb Cortex 20(6):1432–47.

Greenberg T, Chase HW, Almeida JR, Stiffler R, Zevallos CR, Aslam HA, Deckersbach T, Weyandt S, Cooper C, Toups M, Carmody T, Kurian B, Peltier S, Adams P, McInnis MG, Oquendo MA, McGrath PJ1,

Fava M, Weissman M, Parsey R, Trivedi MH, Phillips ML1 (2015): Moderation of the Relationship Between Reward Expectancy and Prediction Error-Related Ventral Striatal Reactivity by Anhedonia in Unmedicated Major Depressive Disorder: Findings From the EMBARC Study. Am J Psychiatry 172(9):881–91.

Greve DN, Fischl B (2009): Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63–72.

Hallquist MN, Hwang K, Luna B (2013): The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fmri preprocessing reintroduces noise and obscures functional connectivity. Neuroimage 2013; 82C: 208–225.

Hamilton M (1960): A rating scale for depression. J Neurol Neurosurg Psychiatry 23(1):56–62.

Huber PJ (2004): Robust Statistics. Wiley.

Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Johnson WE, Li C, Rabinovic A (2007): Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–27.

Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartrés-Faz D, Arnold J, Benninghoff J, Fiedler U, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Sein J, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Ferretti A, Caulo M, Aiello M, Cavaliere C, Soricelli A, Parnetti L, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Schönknecht P, Hensch T, Hoffmann KT, Kuijer JP, Visser PJ, Barkhof F, Frisoni GB; PharmaCog Consortium (2016): Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: A multicentric resting-state fMRI study. Neuroimage 124(Pt A):442–454.

Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA (2015): Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. JAMA Psychiatry 72(6):603–11.

Keshavan A, Paul F, Beyer MK, Zhu AH, Papinutto N, Shinohara RT, Stern W, Amann M, Bakshi R, Bischof A, Carriero A, Comabella M, Crane JC, D'Alfonso S, Demaerel P, Dubois B, Filippi M, Fleischer V, Fontaine B, Gaetano L, Goris A, Graetz C, Gröger A, Groppa S, Hafler DA, Harbo HF, Hemmer B, Jordan K, Kappos L, Kirkish G, Llufriu S, Magon S, Martinelli-Boneschi F, McCauley JL, Montalban X, Mühlau M, Pelletier D, Pattany PM, Pericak-Vance M, Cournu-Rebeix I, Rocca MA, Rovira A, Schlaeger R, Saiz A, Sprenger T, Stecco A, Uitdehaag BMJ, Villoslada P, Wattjes MP, Weiner H, Wuerfel J, Zimmer C, Zipp F; International Multiple Sclerosis Genetics Consortium. Electronic address: AIVINSON@PARTNERS.ORG, Hauser SL, Oksenberg JR, Henry RG (2016): Power estimation for non-standardized multisite studies. Neuroimage 134:281–294.

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C et al. (2009): Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46:786–802.

Koch W, Teipel S, Mueller S, Buerger K, Bokde AL, Hampel H, Coates U, Reiser M, Meindl T (2010): Effects of aging on default mode network activity in resting state fMRI: does the method of analysis matter? Neuroimage 51(1):280–7.

Kochunov P, Dickie EW, Viviano JD, Turner J, Kingsley PB, Jahanshad N, Thompson PM, Ryan MC, Fieremans E, Novikov D, Veraart J, Hong EL, Malhotra AK, Buchanan RW, Chavez S, Voineskos AN (2018): Integration of routine QA data into mega-analysis may improve quality and sensitivity of multisite diffusion tensor imaging studies. Hum Brain Mapp 39(2):1015–1023.

Latora V, Marchiori M (2001): Efficient behavior of small-world networks. Phys Rev Lett 87(19):198701.

McGonigle DJ (2012): Test-retest reliability in fMRI: Or how I learned to stop worrying and love the variability. Neuroimage 62(2):1116–20.

Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos H, Tsuang MT, van Erp TGM, Walker EF, Hamann S, Woods SW, Cannon TD, Constable RT (2017): Multisite reliability of MR-based functional connectivity. Neuroimage 146:959–970.

Oh J, Bakshi R, Calabresi PA, Crainiceanu C, Henry RG, Nair G, Papinutto N, Constable RT, Reich DS5, Pelletier D, Rooney W, Schwartz D, Tagge I, Shinohara RT, Simon JH, Sicotte NL; NAIMS Cooperative Steering Committee (2017): The NAIMS cooperative pilot project: Design, implementation and future directions. Mult Scler 1352458517739990.

Otte C, Gold SM, Penninx BW, Pariante CM, Etkin A, Fava M, Mohr DC, Schatzberg AF (2016): Major depressive disorder. Nat Rev Dis Primers 2:16065.

Percival DB, Walden AT (2000): Wavelet methods for time series analysis. Cambridge University Press.

Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, Petersen SE (2011): Functional network organization of the human brain. Neuron 72:665–678.

Raichle ME (2015): The brain's default mode network. Annu Rev Neurosci 38:433–47.

Rath J, Wurnig M, Fischmeister F, Klinger N, Höllinger I, Geißler A, Aichhorn M, Foki T, Kronbichler M, Nickel J, Siedentopf C, Staffen W, Verius M, Golaszewski S, Koppelstaetter F, Auff E, Felber S, Seitz RJ, Beisteiner R (2016): Between- and within-site variability of fMRI localizations. Hum Brain Mapp 37(6):2151–60.

Rubinov M, Sporns O (2010): Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52:1059–1069.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB (2003): The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 54(5):573–83.

Satterthwaite TD, Ciric R, Roalf DR, Davatzikos C, Bassett DS, Wolf DH (2017): Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum Brain Mapp doi: 10.1002/hbm.23665. [Epub ahead of print]

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT (2017): Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb Cortex 18:1-20.

Sheline YI, Barch DM, Price JL, Rundle MM, Vaishnavi SN, Snyder AZ et al. (2009): The default mode network and self-referential processes in depression. Proc Natl Acad Sci USA 106:1942–1947.

Sheline YI, Price JL, Yan Z, Mintun MA (2010): Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. Proc Natl Acad Sci USA 107:11020–5.

Shinohara RT, Oh J, Nair G, Calabresi PA, Davatzikos C, Doshi J, Henry RG, Kim G, Linn KA, Papinutto N, Pelletier D, Pham DL, Reich DS, Rooney W, Roy S, Stern W, Tummala S, Yousuf F, Zhu A, Sicotte NL, Bakshi R; NAIMS Cooperative (2017): Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. AJNR Am J Neuroradiol 38(8):1501–1509.

Stam CJ (2014): Modern network science of neurological disorders. Nat Rev Neurosci 15:683–95.

Suckling J, Ohlssen D, Andrew C, Johnson G, Williams SC, Graves M, Chen CH, Spiegelhalter D, Bullmore E (2008): Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp 29(10):1111–22.

Suckling J, Barnes A, Job D, Brenan D, Lymer K, Dazzan P, Marques TR, MacKay C, McKie S, Williams SR, Williams SC, Lawrie S, Deakin B (2010): Power calculations for multicenter imaging studies controlled by the false discovery rate. Hum Brain Mapp 31(8):1183–95.

Tomasi D, Volkow ND (2012): Aging and functional brain networks. Mol Psychiatry 17(5):471, 549–58.

Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, Oquendo MA, Bruder G, Pizzagalli D, Toups M, Cooper C, Adams P, Weyandt S, Morris DW, Grannemann BD, Ogden RT, Buckner R, McInnis M, Kraemer HC, Petkova E, Carmody TJ, Weissman MM (2016): Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. J Psychiatr Res 78:11–23.

Turner JA, Damaraju E, van Erp TG, Mathalon DH, Ford JM, Voyvodic J, Mueller BA, Belger A, Bustillo J, McEwen S, Potkin SG, Fbirn, Calhoun VD (2013): A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia. Front Neurosci 7:137.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA et al. (2010): N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310–1320.

Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT et al. (2014): Large-scale eval- uation of ants and freesurfer cortical thickness measurements. Neuroimage 99:166–179.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–89.

Van Horn JD, Toga AW (2009): Multi-Site Neuroimaging Trials. Curr Opin Neurol 22(4):370–378.

Watson D, Clark LA (1991): The mood and anxiety symptom questionnaire. Iowa City Iowa, University of Iowa.

Webb CA, Dillon DG, Pechtel P, Goer FK, Murray L, Huys QJ, Fava M, McGrath PJ, Weissman M, Parsey R, Kurian BT, Adams P, Weyandt S, Trombello JM, Grannemann B, Cooper CM, Deldin P, Tenke C, Trivedi M, Bruder G, Pizzagalli DA (2016): Neural Correlates of Three Promising Endophenotypes of Depression: Evidence from the EMBARC Study. Neuropsychopharmacology 41(2):454–63.

Wig GS, Laumann TO, Petersen SE (2014): An approach for parcellating human cortical areas using resting-state correlations. Neuroimage 93:276–291.

Williams LM (2016): Precision psychiatry: a neural circuit taxonomy for depression and anxiety. Lancet Psychiatry May;3(5):472–80.

Xia M, Wang J, He Y (2013): BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One 8:e68910.

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, et al. (2011): The organization of the human cerebral cortex estimated by intrinsic functionalconnectivity. J Neurophysiol 106:1125–1165.

Yu M, Gouw AA, Hillebrand A, Tijms BM, Stam CJ, van Straaten EC, et al. (2016): Different functional connectivity and network topology in behavioral variant of frontotemporal dementia and Alzheimer's disease: an EEG study. Neurobiol Aging 42:150–62.

Yu M, Engels MMA, Hillebrand A, van Straaten ECW, Gouw AA, Teunissen C, van der Flier WM, Scheltens P, Stam CJ (2017): Selective impairment of hippocampus and posterior hub areas in Alzheimer's disease: an MEG-based multiplex network study. Brain 140(5):1466–1485.

Zhang Z, Telesford QK, Giusti C, Lim KO, Bassett DS (2016): Choosing wavelet methods, filters, and lengths for functional brain network construction. PLoS One 11:e0157243.

TABLES

TABLE 1 Imaging parameters for the four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX) and the University of Michigan (UM).

	CU	**MGH**	**TX**	**UM**
Scanner	General Electric 3T	Siemens 3T	Phillips 3T	Phillips 3T
Structural	FSPGR TR = 6.0 ms TE = 2.4 ms TI = 900 ms Flip Angle = 9° FOV= 256×256 mm Slice Thickness = 1 mm Matrix = 256×256 178 continuous slices (4 discarded)	MPRAGE TR = 2300 ms TE = 2.54 ms TI = 900 ms Flip Angle = 9° FOV= 256×256 mm Slice Thickness = 1 mm Matrix = 256×256 176 continuous slices	Turbo Field Echo (TFE) TR = 8.2 ms TE = 3.7 ms TI = 1100 ms Flip Angle = 12° FOV= 256×256 mm Slice Thickness = 1mm Matrix = 256×256 178 continuous slices	MPRAGE TR = 2100 ms TE = 3.7 ms TI = 1100 ms Flip Angle = 12° FOV= 256×256 mm Slice Thickness = 1mm Matrix = 256×256 178 continuous slices
Functional	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV=205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64

TABLE 2 Site effects in functional connectivity values for different connectivity measures (Pearson correlation and wavelet coherence) using different atlases (AAL, Power and Gordon).

Six combinations of two connectivity measures and three atlases		Number of ROIs	Original	ComBat
Pearson correlation	AAL	90	1610 (40.2%)	0 (0%)
	Power	203	515 (2.5%)	0 (0%)
	Gordon	307	603 (1.3%)	0 (0%)
Wavelet coherence	AAL	90	7 (0.2%)	0 (0%)
	Power	203	12 (0.06%)	0 (0%)
	Gordon	307	17 (0.04%)	0 (0%)

Note that, the numbers and percentages in the round brackets represent the numbers and percentages of connectivity values that were significantly different across 4 sites for six combinations of two connectivity measures and three atlases, respectively. Significant effects for each combination were computed using Kruskal-Wallis tests with FDR corrections. Here we only presented the results of the first-session data, as the two sessions showed extremely similar results.

FIGURE LEGENDS

FIGURE 1 Distribution of subject demographic characteristics across 4 sites. Abbreviations: MDD = major depressive disorder; Control = healthy controls; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan. Note that whiskers in the boxplots represent variability outside the upper and lower quartiles.

FIGURE 2 Site effects in functional connectivity estimated by "Correlation-AAL" (A) and "Coherence-Power" (B) before and after ComBat harmonization. Note that the functional connectivity values in (A) and (B) were computed from the time series of two randomly selected ROIs: for AAL, the TPOmid.R and ACG.R were selected; for Power, two regions in the visual cortex were selected; whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan; TPOmid.R = right Temporal pole: middle temporal gyrus; ACG.R = right Anterior cingulate and paracingulate gyri.

FIGURE 3 Empirical (dashed lines) and ComBat-estimated (solid lines) prior distributions for the site-specific location (γ) and scale (δ) parameters for "Correlation-AAL" (A) and "Coherence-Power" (B). Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan.

FIGURE 4 DMN connectivity estimated by "Coherence-Power" and "Correlation-AAL" across sites (**A**) and their anti-correlations with age (**B**). Note that DMN connectivity was computed by first summing the functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (Power and AAL); the significant site effects in DMN connectivity were removed by ComBat harmonization (**A**). The strong anti-correlation between age and DMN connectivity estimated by "Coherence-Power" was preserved in ComBat-harmonized data; for "Correlation-AAL", ComBat harmonization increased the detection power of the anti-correlation (**B**). Note that whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: DMN = default mode network.

FIGURE 5 Negative log-transformed p-value heat maps for Kruskal-Wallis tests (e.g. Figure 4A; Supplementary Figure S8-S10, subplot A) of network connectivity and efficiency measures (A), and correlation coefficient (B) heat maps for correlation analyses between the network measures and age (e.g. Figure 4B; Supplementary Figure S8-S10, subplot B). Note that the asterisks (blue or white colors) represent $p < .05$, which corresponds to a negative $\log_{10}$ transformed p-value greater than 1.301 in the color bar of (A). Abbreviations: DC = DMN connectivity; DNS = DMN nodal strength; DLE = DMN local efficiency; GE = global efficiency. Note that DMN connectivity was computed by first summing the

functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (AAL, Power and Gordon); please see the details of the computation of DNS, DLE and GE in Section 2.7 Calculation of network measures.

Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data

Meichen Yu[1], Kristin A. Linn[2,1], Philip A. Cook[1,3], Mary L. Phillips[4], Melvin McInnis[5], Maurizio Fava[6], Madhukar H. Trivedi[7], Myrna M. Weissman[8,9,10], Russell T. Shinohara[2,1], and Yvette I. Sheline[1,3,11*]

[1] Center for Neuromodulation in Depression and Stress, Department of Psychiatry, Perelman School of Medicine, University of Pennsylvania, United States

[2] Department of Biostatistics, Epidemiology, and Informatics, Perelman School of Medicine, University of Pennsylvania, United States

[3] Department of Radiology, Perelman School of Medicine, University of Pennsylvania, United States

[4] Department of Psychiatry, University of Pittsburgh School of Medicine, United States

[5] Department of Psychiatry, University of Michigan School of Medicine, United States

[6] Department of Psychiatry, Massachusetts General Hospital, United States

[7] Department of Psychiatry, University of Texas Southwestern Medical Center, United States

[8] Department of Psychiatry, Columbia University College of Physicians & Surgeons, United States

[9] Division of Epidemiology, New York State Psychiatric Institute, United States

[10] Mailman School of Public Health, Columbia University, United States

[11] Department of Neurology, Perelman School of Medicine, University of Pennsylvania, United States

*Please address correspondence to:

Yvette I. Sheline, M.D., M.S

Email: sheline@mail.med.upenn.edu

Phone: 215-573-0082

Fax: 215-349-5067

Words (abstract): 250

Words (main text): 6230 (excluding References and text in tables and figure captions)

Tables: 2

Figures: 5

Supplementary materials: 2

Tables and figures were added in the end of the manuscript; supplementary materials were separated from this manuscript.

Abstract

Acquiring resting-state functional magnetic resonance imaging (fMRI) datasets at multiple MRI scanners and clinical sites can improve statistical power and generalizability of results. However, multi-site neuroimaging studies have reported considerable non-biological variability in fMRI measurements due to different scanner manufacturers and acquisition protocols. These undesirable sources of variability may limit power to detect effects of interest and may even result in erroneous findings. Until now, there has not been an approach that removes unwanted site effects. In this study, using a relatively large multi-site (4 sites) fMRI dataset, we investigated the impact of site effects on functional connectivity and network measures estimated by widely used connectivity metrics and brain parcellations. The protocols and image acquisition of the dataset used in this study had been homogenized using identical MRI phantom acquisitions from each of the neuroimaging sites, however inter-site acquisition effects were not completely eliminated. Indeed, in the current study we found that the magnitude of site effects depended on the choice of connectivity metric and brain atlas. Therefore, to further remove site effects, we applied ComBat, a harmonization technique previously shown to eliminate site effects in multi-site diffusion tensor imaging (DTI) and cortical thickness studies. In the current work, ComBat successfully removed site effects identified in connectivity and network measures and increased the power to detect age associations when using optimal combinations of connectivity metrics and brain atlases. Our proposed ComBat harmonization approach for fMRI-derived connectivity measures facilitates reliable and efficient analysis of retrospective and prospective multi-site fMRI neuroimaging studies.

KEYWORDS

Aging, atlas, ComBat, fMRI, functional connectivity, graph theory, harmonization, multi-site, network efficiency

1 | INTRODUCTION

Functional magnetic resonance imaging (fMRI), a non-invasive neuroimaging modality with high spatial resolution, enables neural activity to be monitored. Functional connectivity and network measures derived from fMRI data have facilitated the study of the brain's function during development, in aging (Fox and Raichle, 2007; Raichle 2015; Bressler and Menon, 2010), and in the context of various neurological disorders (Bullmore and Sporns, 2009, 2012; Stam, 2014; Fornito et al., 2015, 2016).

Over the last decade, multi-site fMRI studies have become increasingly common (Friedman et al., 2006, 2008; Van Horn and Toga, 2009; Biswal et al., 2010; Gradin et al., 2010; Di Martino et al., 2014; Noble et al., 2017). Indeed, pooling fMRI data from multiple sites can accelerate participant recruitment rates and increase the total sample size of the study, thereby increasing statistical power. Pooling fMRI data is often critical when studying rare disorders and subtle effects and when aiming to generalize the study results to a diverse population (Suckling et al., 2010; McGonigle, 2012; Keshavan et al., 2016; Dansereau et al., 2017). Despite these advantages, multi-site studies are often plagued by non-biological variability that can be attributed to differences in scanner manufacturers, non-standardized imaging acquisition parameters, and other intrinsic factors (Shinohara et al., 2017). These additional sources of unwanted variability may decrease statistical power and lead to spurious results. Many multi-site studies have reported considerable site or scanner effects in fMRI data (Friedman et al., 2006, 2008; Suckling et al., 2008, 2010; Van Horn and Toga, 2009; Gountouna et al., 2010; Brown et al., 2011; McGonigle, 2012; Turner et al., 2013; Forsyth et al., 2014; Feis et al., 2015; Rath et al., 2016; Jovicich et al., 2016; Dansereau et al., 2017; Noble et al., 2017; Abraham et al., 2017). However, most of these studies only describe the problem or report the magnitude of site effects in fMRI measurements.

A few studies have attempted to mitigate site effects by standardizing protocols and image acquisition parameters (Friedman et al., 2008; Glover et al., 2012; Shinohara et al., 2017; Oh et al., 2017; Kochunov et al., 2018; Chavez et al., 2018). However, it has been shown that scanner-to-scanner variation arising from the use of scanners from different manufacturers is not eliminated completely by the standardization of acquisition parameters (Jovicich et al., 2016; Noble et al., 2017), for instance, by use of phantom-based imaging acquisitions (Delaparte et al., 2017). To our knowledge, until now, there has been only one attempt to diminish scanner differences in multi-site resting-state fMRI post-acquisition. The authors used an independent component analysis (ICA) based approach that reduced differences across sites in some resting-state network connectivity measures but did not fully eliminate the structured noise arising from different scanners (Feis et al., 2015).

Recently, our group adapted ComBat harmonization (Johnson et al., 2007) to model and remove site effects in multi-site DTI (Fortin et al., 2017) and cortical thickness (Fortin et al., 2018) measurements. ComBat was originally designed to correct so-called "batch effects" in genomic studies (Johnson et al., 2007) that arise due to processing high-throughput genomic data in different laboratories with different equipment at different times. In our previous studies, we demonstrated that the ComBat harmonization technique successfully removed unwanted non-biological variability, while preserving biological associations between participant age and DTI (fractional anisotropy and mean diffusivity), as well as the association between age and cortical thickness measurements.

In this study, we quantified the site effects in functional connectivity and several brain network measures in the multi-site Establishing Moderators and Biosignatures of Antidepressant Response in

Clinical Care (EMBARC) dataset that was acquired at four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX), and the University of Michigan (UM). Our main objectives were to: (1) remove any identified site effects using ComBat harmonization, and (2) preserve the commonly reported negative correlation between age and functional connectivity within the default mode network (DMN; Damoiseaux et al., 2008; Koch et al., 2010; Grady et al., 2010; Tomasi and Volkow, 2012; Ferreira and Busatto, 2013; Damoiseaux, 2017), as well as preserve previously reported negative correlations between age and network efficiency measures (Achard and Bullmore, 2007; Ajilore et al., 2014). Objective (2) was important to demonstrate that the ComBat technique did not remove important, biologically relevant information. A recently published multi-site autism study (Abraham et al., 2017) reported that the magnitude of site effects was influenced by the choice of functional connectivity metrics and brain parcellation. Therefore, we investigated the degree to which widely used functional connectivity and network metrics derived from a number of brain parcellations were affected by scanner-to-scanner variation and how ComBat harmonization performed in each setting. We hypothesized that (1) considerable site effects exist in both functional connectivity and network efficiency measures calculated from non-harmonized multi-site fMRI data; (2) the magnitude of site effects is not constant across different connectivity metrics and brain parcellations; and (3) ComBat harmonization can be used to remove site effects in connectivity and network measures while preserving age-related associations for numerous combinations of connectivity metrics and brain parcellations.

2 | MATERIAL AND METHODS

2.1 Participants

This study considered 200 unmedicated depressed patients with major depressive disorder (MDD) and 40 healthy subjects recruited for EMBARC that have been analyzed in several previous studies (Greenberg et al., 2015; Trivedi et al., 2016; Webb et al., 2016; Fortin et al., 2018). The current study concentrates on the harmonization of multi-site fMRI-based functional connectivity and network measures. The participants were recruited and scans acquired at four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX), and the University of Michigan (UM). All participants provided written informed consent and the institutional review boards from the four clinical sites approved all study procedures. For both patients and healthy individuals, the Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P; First et al., 2002) was used as inclusion criteria to diagnose the presence or absence of depressive symptoms. The Hamilton Depression Rating Scale (HAMD; Hamilton, 1960) and Quick Inventory for Depression Symptomatology (QIDS; Rush et al., 2003) depression scores were used to estimate depressive severity. Anxiety and depressive severity were also assessed using the Mood and Anxiety Symptom Questionnaire (MASQ; Watson and Clark, 1991), including three subscales: general distress (MASQ-GD), anhedonic depression (MASQ-AD), and anxious arousal (MASQ-AA). The individuals were eligible for the study if they met the following inclusion criteria: (1) age 18–65; (2) reported age of depression onset before age 30; (3) fluent in English. Eleven depressed patients and one healthy individual were excluded due to excessive motion (> 4 mm), low slice signal-to-noise ratio (< 80), and severe slice artifacts in MRI data. The final sample included 189 MDD patients and 39 healthy

individuals. The distribution of age, sex, handedness, and education level were matched between the two groups.

2.2 Image acquisition and data preprocessing

All four sites used 3T scanners, however, the manufacturer differed from site to site: CU used a GE SIGNA HDx 3T scanner, MGH used a Siemens TIM Trio 3T scanner, TX used a Philips Achieva 3T scanner, and UM used a Philips Ingenia 3T scanner (Greenberg et al., 2015; Trivedi et al., 2016; Fortin et al., 2018). Imaging parameters at each site are described in Table 1.

Prior to the project's initiation, in close collaboration with MR physics teams at the acquisition sites, a homogenized imaging protocol was developed to minimize acquisition-related site differences. In particular, data were collected using identical MRI phantom acquisitions from each of the neuroimaging sites. Well established routines for using phantoms were employed to perform quality assurance on the scanners used in this study. However, although the phantom-based approach minimized the inconsistency of signal-to-noise acorss scanners over the time and other variability in image acquisition and quality ascross sites, the inter-site acquisition effects were not completely eliminated (Delaparte et al., 2017). Therefore, we employed a post-processing procedure that further harmonized the fMRI functional connectivity matrices of subjects across the 4 sites.

T1-weighted (T1) images were processed using the ANTS Cortical Thickness pipeline available in the antsCorticalThickness.sh script in advanced normalization tools (ANTs ; Avants et al., 2011a; Tustison et al., 2014). The workflow is sketched out as follows: (1) N4 bias correction to minimize field inhomogeneity (Tustison et al., 2010); (2) brain extraction using an optimal population-specific template created by a Symmetric Group Normalization framework (Avants et al., 2010); (3) Atropos probabilistic six-tissue segmentation (Avants et al., 2011b); (4) DiReCT-based cortical thickness estimation (Das et al., 2009); (5) SyN deformable spatial registration to the population-specific template (Klein et al., 2009).

Resting-state time series data from each participant were processed using the XCP Engine (Ciric et al., 2017), which uses an optimized confound regression procedure to reduce the influence of subject motion (Satterthwaite et al., 2017). Each subject contributed time series data from two resting-state fMRI sessions. The workflow of functional data preprocessing is summarized as follows: (1) removal of the four initial volumes of the Blood-oxygen-level Dependent (BOLD) signals to achieve signal stabilization; (2) realignment of functional images using MCFLIRT (Jenkinson et al., 2002); (3) removal of nine confounding signals (six motion parameters+global/white matter/cerebral spinal fluid) as well as the temporal derivative, quadratic term, and temporal derivatives of each quadratic term (36 regressors total) (Satterthwaite et al., 2017); (4) co-registration of functional images to the T1 image using boundary-based registration (Greve and Fischl, 2009); (5) alignment of the co-registered images to template space using the ANTs-transform for the T1 image as above; and (6) temporal filtering of time series between 0.01 and 0.08 Hz as in previous studies (Biswal et al., 1995) using a first-order Butterworth filter. In this study, all regressors, including motion parameters and confound time courses, were band-pass filtered to the same frequency range as the time series data to prevent frequency-dependent mismatch during confound regression (Hallquist et al., 2013). Functional images were smoothed using a Gaussian convolution at 6mm full width at half maximum.

2.3 Parcellation

To investigate the influence of different parcellations on functional connectivity measures across sites and subsequent harmonization, we partitioned the brain of each participant into cortical and subcortical ROIs using the following three different whole-brain atlases (one anatomical and two functional): (1) 78 cortical and 12 subcortical ROIs identified by automated anatomical labeling (AAL) (Tzourio-Mazoyer et al., 2002); (2) 264 cortical and subcortical ROIs of the widely-used functional Power atlas (Power et al., 2011); and (3) 333 cortical and subcortical ROIs from the functional Gordon atlas (Gordon et al., 2016). The ROIs and MNI-space centroids of the AAL, Power, and Gordon atlases can be found in the Supplementary figure S1 and Supplementary Materials 2.

2.4 Functional connectivity

For each participant, whole-brain functional connectivity between all brain regions was constructed pairwise from the preprocessed fMRI data. The fMRI time series were extracted from each voxel and averaged within each ROI of the three atlases (AAL, Power, and Gordon). The functional connectivity between time series for all pair-wise ROIs was estimated by calculating two commonly used connectivity metrics: Pearson correlation and wavelet coherence. For Pearson correlation, the correlation coefficients were Fisher-transformed in order to draw more statistically interpretable conclusions about the magnitude of the correlations (Cohen and Esposito, 2016; Doucet et al., 2017). Due to poor signal quality and signal dropout, we excluded 61 ROIs from the Power atlas and 26 ROIs from the Gordon atlas, which resulted in 203 and 307 ROIs for the Power and Gordon atlases, respectively. All subsequent analyses were performed using the 90×90 AAL-atlas, 203×203 Power-atlas, and 307×307 Gordon-atlas connectivity matrices based on both Fisher-transformed Pearson correlation coefficients and raw wavelet coherence values from all participants.

2.5 Model for functional connectivity matrix harmonization

Based on the literature (Friedman et al., 2008; Feis et al., 2015; Rath et al., 2016; Dansereau et al., 2017), we speculated that measurements such as DTI fractional anisotropy (Fortin et al., 2017), MRI cortical thickness (Fortin et al., 2018), and fMRI functional connectivity (the present study) would differ among the four sites (CU, MGH, TX and UM) due to systematic bias and non-biological variability attributable to the use of different scanners and different imaging parameters.

   In this study, we used ComBat (Johnson et al., 2007) to reduce potential biases and non-biological variability induced by site and scanner effects. ComBat uses a multivariate linear mixed effects regression with terms for biological variables and scanner to model imaging feature measurements. The method uses empirical Bayes to improve the estimation of the model parameters for studies with small sample sizes. Here, we reformulate the ComBat model so that it can be applied to functional connectivity matrices estimated using Pearson correlation and wavelet coherence in combination with the AAL, Power, and Gordon atlases (i.e., six combinations: Correlation-AAL, Coherence-AAL, Correlation-Power, Coherence-Power, Correlation-Gordon, and Coherence-Gordon). Since all

connectivity matrices are symmetric, we applied ComBat to connectivity values in the upper triangles of the matrices. Let y_{ijv} represent the connectivity values of imaging site i ($i \in \{1, \dots, 4\}$), participant j ($j \in \{1, \dots, 228\}$), and connectivity value v ($v \in \{1, \dots, 4,005\}$ for the AAL atlas, $v \in \{1, \dots, 20,503\}$ for the Power atlas, and $v \in \{1, \dots, 46,971\}$ for the Gordon atlas) between two ROIs. Then, the ComBat model can be written as

$$y_{ijv} = \alpha_v + X_{ij}^T \beta_v + \gamma_{iv} + \delta_{iv} \epsilon_{ijv}$$

where α_v is the average connectivity value for a particular connectivity value v between two ROIs, X_{ij}^T is a design matrix for the covariates of interest (age , gender, and group), and β_v is a vector of regression coefficients corresponding to X. As in Fortin et al. (2018), we further assume that the residual terms ϵ_{ijv} arise from a normal distribution with zero mean and variance σ_v^2. The terms γ_{iv} and δ_{iv} represent the additive (or location parameter) and multiplicative (or scale parameter) site effects of site i for connectivity value v, respectively. The ComBat-harmonized functional connectivity values were then defined as

$$y_{ijv}^{ComBat} = \frac{y_{ijv} - \widehat{\alpha_v} - X_{ij}\widehat{\beta_v} - \gamma_{iv}^*}{\delta_{iv}^*} + \widehat{\alpha_v} + X_{ij}\widehat{\beta_v},$$

where γ_{iv}^* and δ_{iv}^* are the empirical Bayes estimates of γ_{iv} and δ_{iv}, respectively. Thus, ComBat simultaneously models and estimates biological and non-biological terms and algebraically removes the estimated additive and multiplicative site effects. Of note, in the ComBat model, we included age, sex, and group as covariates to preserve important biological trends in the data and avoid overcorrection.

In this study, we performed the ComBat harmonization analyses for the six-combinations of connectivity matrices in two sessions (S1 and S2), separately. ComBat harmonization analyses were performed using a publicly available MATLAB package hosted at https://github.com/Jfortin1/ComBatHarmonization/tree/master/Matlab.

2.6 Visualization and evaluation of functional connectivity harmonization

We used Kruskal-Wallis tests to quantify the magnitude of site effects in functional connectivity between all pairwise ROIs before and after applying ComBat harmonization to each of the six metric-atlas combinations (Correlation-AAL, Coherence-AAL, Correlation-Power, Coherence-Power, Correlation-Gordon, and Coherence-Gordon). The p-values were adjusted for multiple comparisons by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at 5%, separately for each combination (AAL: 4,005 comparisons; Power: 20,503 comparisons; Gordon: 46,971 comparisons). The numbers and percentages of connectivity values that were significantly different (after FDR correction) across the 4 sites for the six combinations are summarized in Table 2. The FDR-corrected p-values can be found in

supplementary Figure 1. We visualized site effects using boxplots of connectivity values between signals of two randomly selected ROIs for each atlas across the four sites (Figure 2; Supplementary Figures S3-S6, subplot A). The selected ROIs were consistent for the same atlas using the two connectivity measures.

We also performed a principal component analysis (PCA) on the functional connectivity values in the upper triangle of the connectivity matrices for the six metric-atlas combinations before and after ComBat harmonization. In Supplementary Figure S3A-S6A, subplot A and S7, we plotted three-dimensional scatter plots of the first three PC scores from the PCA. If the connectivity values were significantly different (Kruskal-Wallis tests; FDR corrections) across sites before or after ComBat harmonization, the corresponding PC scores are likely to be associated with site, and we would expect to see data from the same scanner roughly clustered together in the scatter plots.

To evaluate whether the assumed empirical Bayes priors for the location (γ) and scale (δ) parameters in the ComBat harmonization model reasonably reflect the observed data, we overlayed the empirical and prior distributions of γ and δ in Figure 3 and Supplementary Figures S3-S6, subplot C.

We applied ComBat harmonization to the connectivity matrices from each fMRI session, separately. We present visualizations of the site effects and plots of ComBat model parameters for the first session. Plots generated from the second session were similar and therefore not included. Figure 2 demonstrates differences in the distribution of functional connectivity across sites for the first session. Figure 3 provides a visualization of the goodness of fit of the ComBat model's prior assumptions to the observed data for the first session. Following ComBat, we extracted four network measures from the harmonized connectivity matrices and averaged these measures across the two sessions. Henceforth, we focus on analyzing the average network measures, which included weigthed DMN connectivity, nodal strength, local efficiency, and global efficiency. We formally define these measures in Section 2.7.

2.7 Calculation of network measures

In order to ensure that our post-processing harmonization did not remove meaningful biological variability along with the undesireable site effects we conducted an additional analysis. As the default mode network (DMN) has been found to have larger negative associations between age and functional connectivity metrics than other resting-state networks (Tomasi and Volkow, 2012; Ferreira and Busatto, 2013; Damoiseaux, 2017), we selected it to conduct analysis of age-related effects. In this study functional connectivity and local network metrics (quantified by weighted nodal strength and nodal efficiency) were thus calculated in the DMN. Global network topology was characterized by weighted global efficiency. The computation details of these connectivity and network metrics are described in the following paragraphs.

For each atlas, the DMN network connectivity was defined by the summation of the functional connectivity values within the DMN ROIs normalized by the number of DMN ROIs for each atlas. The DMN nodal strength was computed by first summing the functional connectivity values (link weights) for each pair of ROIs and then summing up the nodal level connectivity values within the DMN ROIs of each specific atlas. The weighted local and global efficiency (Latora and Marchiori, 2001) were computed using the weighted shortest path length (L^W; Dijkstra, 1959), which is the shortest sum of connection

length (inverse of the connectivity values or link weights) between two nodes (or ROIs; Rubinov and Sporns, 2010). The weighted nodal efficiency (E^W_{nodal}) was calculated as the inverse of the harmonic mean of L^W from one node to all other nodes, as follows:

$$E^W_{nodal} = \frac{1}{N-1} \sum_{j \epsilon G} \frac{1}{L^W_{i,j}}$$

where N is the number of nodes in graph G (represented by the AAL, Power and Gordon connectivity matrices in this study) and $L^W_{i,j}$ is the weighted shortest path length between node i and j.

Weighted local efficiency (E^W_{local}) for a node is defined as the average weighted nodal efficiency among the neighboring nodes of that node (excluding the reference node), as follows:

$$E^W_{local} = \frac{1}{N_{G_i}(N_{G_i} - 1)} \sum_{j,k \epsilon G_i} \frac{1}{L^W_{j,k}}$$

where N_{G_i} is the number of nodes in subgraph G_i that consists of all neighboring nodes of node i, but excluding node i. For the weighted DMN local efficiency, weighted local efficiency values were computed for each ROI and then summed up within the DMN ROIs of each specific atlas.

The weighted global efficiency (E^W_{global}) was calculated as the average weighted nodal efficiency of nodes in a graph G, as follows:

$$E^W_{global} = \frac{1}{N(N-1)} \sum_{i \neq j \epsilon G} \frac{1}{L^W_{i,j}}$$

All the network efficiency measures were computed using the Brain Connectivity Toolbox (BCT) (Rubinov and Sporns, 2010).

For each participant, we first computed DMN network connectivity, DMN nodal strength, weighted DMN local efficiency, and weighted global efficiency for each of the six combinations (3 atlases × 2 connectivity metrics) before and after applying ComBat harmonization. Next, we averaged the values of each participant's network connectivity or efficiency measures from the two sessions. Then, we tested the global null hypothesis of no differences across sites in the network connectivity or efficiency measures using Kruskal-Wallis tests (in total, 2 conditions (before and after ComBat) × 2 connectivity metrics × 4 network measures × 3 atlases = 48 comparisons) with a separate FDR correction at 5% within each condition (2 connectivity metrics × 4 network measures × 3 atlases = 24 comparisons), separately.

2.8 Preservation of biological variability

An optimal harmonization technique should be able to remove most or all non-biological sources of variability caused by site and scanner, yet preserve or increase statistical power to detect biological associations. In this study, there was a broad participant age range (18 to 65 years), enabling investigation of age-related associations. Therefore, we investigated whether negative associations

between age and DMN network connectivity as well as associations between age and network efficiency measures were preserved or made stronger when estimated using ComBat-harmonized data.

We computed the Spearman correlation between each network (or connectivity) measure and age. The p-values were adjusted for multiple comparisons (in total, 2 conditions (before and after ComBat) × 2 connectivity metrics × 4 network measures × 3 atlases = 48 comparisons) by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995). As before, the FDR corrections were applied separately within condition (2 connectivity metrics × 4 network measures × 3 atlases = 24 comparisons). A significance level of p < .05 was used for these tests. Note that, for the Power and Gordon atlases, we used the original definitions of DMN ROIs from Power et al., 2011 and Gordon et al., 2011, respectively; for the AAL atlas, we defined DMN ROIs according to a review article, Rosazza and Minati, 2011. For details of the definition of the DMN ROIs for each atlas, refer to Supplementary Table 1, Figure S1 and Supplementary materials 2. Figure S1, subplot b, c were visualized with BrainNet Viewer (version 1.5, Xia et al., 2013, http://www.nitrc.org/projects/bnv/).

2.9 Statistical analysis of demographic characteristics

Statistical analyses for demographic characteristics of participants were performed using MATLAB (R2017a). Age and educational level were compared among the four sites using Kruskal-Wallis tests followed by Mann Whitney-U tests when appropriate. All p-values from the Mann Whitney-U tests were adjusted for multiple comparisons by controlling the false discovery rate (FDR) (Benjamini and Hochberg, 1995) at 5%. We tested for differences in the gender and clinical group distribution among the four sites using Pearson's chi-squared (χ^2) tests.

3 | RESULTS

3.1 Demographic characteristics

The distribution of demographic characteristics across imaging sites is shown in Figure 1. The age distribution (p = .001) was imbalanced across sites; subjects in the TX site were older than the other sites (TX-CU: p < .001; TX-MGH: p = .04; TX-UM: p = .04; FDR correction). There were also weak site effects (p = .03) in the level of education, but no differences were found regarding educational level between pairwise sites after FDR correction. Gender (p = .14) and depressed/control (p = .34) distributions were equally distributed across sites.

3.2 Visualization and evaluation of ComBat harmonization

Functional connectivity values estimated by Pearson correlation showed much stronger site effects than those by wavelet coherence for analyses using the AAL and Power atlases as well as the Gordon atlas (Table 2). Moreover, the AAL atlas had a much larger percentage of connectivity values that differed significantly across the four sites than the Power and Gordon atlases (Table 2). Following ComBat harmonization, there were no statistically significant site effects in the functional connectivity values of the six metric-atlas combinations.

Figure 2 displays boxplots of functional connectivity values between two randomly selected ROIs (for AAL: TPOmid.R and ACG.R; for Power: two regions in the visual cortex) for the Pearson correlation-AAL atlas and Wavelet coherence-Power atlas combinations. The connectivity values between these two ROIs showed consistent patterns across the four sites: for the AAL atlas, MGH values were generally higher than the other three sites; for the Power atlas, MGH values were generally lower than the other three sites. For the Gordon atlas, MGH and UM values were generally lower than the other two sites (see Supplementary Figures S5-S6, subplot A, for the Gordon atlas results). After ComBat harmonization, these significant site effects were dramatically reduced in all six metric-atlas combinations. See Supplementary Figures S3-S6, subplot A for boxplots of functional connectivity for other metric-atlas combinations, before and after ComBat.

To further visualize site effects, we generated three-dimensional scatter plots of the first three principal component (PC) scores obtained from the functional connectivity matrices (see Supplementary Figures S3-S6, subplot B; Figure S7). For all three atlases, the second PC scores from CU and/or MGH patients showed distinct separation from those of TX and UM, particularly when using Pearson correlation. These visual site effects are much less noticeable in the ComBat-harmonized data: the first three PC scores were not clearly associated with site by visual inspection for any of the metric-atlas combinations (Supplementary Figures S3-S6, subplot B; Figure S7).

For all metric-atlas combinations, the ComBat-harmonized prior distributions appear to fit the empirical distributions of both the location (γ) and scale (δ) parameters well (Figure 3 and Supplementary Figures S3-S6, subplot C). Visual inspection of these overlaid distributions suggests that the ComBat model used appropriate prior information to capture the underlying site effects in the functional connectivity matrices. Furthermore, for each of the three atlases, the distributions of γ and δ reflected the observed lower magnitudes in the distribution location and variability of MGH values compared with the values from other three sites, i.e., for MGH, $\delta < 1$ on average using Pearson correlation and $\delta < 1$ and $\gamma < 0$ using wavelet coherence.

Before performing ComBat harmonization on the functional connectivity matrices, all the network connectivity and efficiency measures estimated by Pearson correlation and a majority of the measures estimated by wavelet coherence displayed statistically significant site effects with similar patterns across sites (MGH values were generally lower than all the other sites). Figure 5A displays log-transformed p-values from the global site effect tests for each metric-atlas combination before and after ComBat. The p-values for network connectivity and the efficiency measures estimated by Pearson correlation were considerably more significant than those estimated by wavelet coherence (see Supplementary Figures S8-S10, subplot A, for corresponding boxplot visualizations of the site effects). After ComBat harmonization, there were no remaining statistically significant site effects for any metric-atlas combination.

Prior to functional connectivity harmonization using ComBat, there were statistically significant site effects across all network connectivity and efficiency measures estimated by both Pearson correlation and wavelet coherence when using the Gordon atlas (Figure 5A). In contrast, prior to harmonization using Combat, when using the AAL or Power atlases, a number of the measures estimated by wavelet coherence did not display significant differences across sites (Figure 5A, Supplementary Figures S9-S10, subplot A).

3.3 Preservation of biological variability

3.3.1 By connectivity metric

ComBat harmonization preserved or strengthened the anti-correlations between age and DMN functional connectivity as well as between age and network efficiency measures. The p-values and correlation values for each metric-atlas combination are displayed in Figures 5B and 5C, respectively, where we see more significant p-values and stronger correlations post-ComBat. This result was true for both Pearson correlation and wavelet coherence connectivity, with wavelet coherence identifying the strongest anti-correlations both before and after ComBat harmonization. Supplementary Figures S8-S9, subplot B, display scatter plots associated with each correlation value.

3.3.2 By atlas

Using the original data without ComBat harmonization, the Gordon atlas showed significant site effects in all network connectivity and efficiency measures estimated by both Pearson correlation and wavelet coherence (Figure 5A). However, for AAL and Power atlases, there were no site effects even in some non-harmonized measures estimated by wavelet coherence (Figure 5A).

   As shown in Figure 5B, ComBat harmonization strengthened the estimated anti-correlations between age and network measures across all three atlases. In particular, for the AAL and Power atlases, ComBat harmonization uncovered significant anti-correlations that were not detected when using the non-harmonized data (Figure 4B, Figure 5B). Among the three atlases, the AAL atlas identified the fewest significant anti-correlations both before and after ComBat harmonization and the magnitudes were generally smaller than those identified by other atlases. The Power atlas identified stonger anti-correlations than the other two atlases post-ComBat. Moreover, a majority of the network measures estimated by the correlation-AAL combination were not negatively associated with age, even after performing ComBat harmonization (see Supplementary Figures S8B-S10B, for scatter plots associated with the p and correlation values in Figure 5B,C).

   Overall, ComBat harmonization not only removed unwanted site effects in network connectivity and efficiency measures calculated from functional connectivity matrices but also preserved or increased the estimated underlying correlations with age. Some specific combinations of atlases and connectivity metrics appear to be better than others with respect to revealing significant relationships with age. When considering both site effect removal and correlation with age, we found that the coherence-Power combination performed optimally.

4 | DISCUSSION

In this study, we investigated the degree to which combining data from different scanners in a multi-site study could affect downstream analyses of fMRI-based functional connectivity and network efficiency measures. We implemented several visualization techniques and statistical tests to visualize and quantify the scanner effects. We performed ComBat harmonization on fMRI-based functional connectivity matrices to remove site effects before extracting DMN connectivity and network measures. We quantified the site effects and the performance of ComBat harmonization using two different metrics to compute connectivity and three different brain atlases. We demonstrated that ComBat

harmonization can successfully remove site effects in the functional connectivity matrices, thereby leading to network connectivity and efficiency measures that are also not different across sites for any choice of connectivity metric and atlas. Moreover, we found that using wavelet coherence with the Power atlas resulted in the best power to detect anti-correlations between age and DMN functional connectivity as well as network efficiency measures following ComBat harmonization, suggesting the best preservation of underlying biological signal with this combination.

4.1 ComBat harmonization removes site effects

As previous studies (Van Horn and Toga, 2009; Dansereau et al., 2017) have consistently reported the existence of considerable site effects in multi-site fMRI measurements that cannot be removed by performing ICA-based approaches (Feis et al., 2015), we tested whether ComBat harmonization could eliminate site effects in several fMRI-based functional connectivity and network measures. Of note, we only performed ComBat harmonization on the original functional connectivity matrices and then subsequently calculated network connectivity and efficiency measures from the harmonized connectivity matrices. Notably, we did not find statistically significant site effects in the downstream network measures.

   Given the excellent performance of ComBat in DTI (Fortin et al., 2017), MRI-based cortical thickness (Fortin et al., 2018), and fMRI (current study) measurements, we conclude that this harmonization method is a reliable and powerful technique that can be widely applied to different neuroimaging modalities and summary measurements.

4.2 Wavelet coherence outperforms Pearson correlation

In this study, to investigate the effects of connectivity metrics on multi-site fMRI measurements and the performance of ComBat harmonization, we used both Pearson correlation and Wavelet coherence to estimate the fMRI functional connectivity. Previous studies have shown that wavelet coherence outperforms Pearson correlation with respect to sensitivity to outliers caused by motion artifact (Huber, 2004; Achard et al., 2006). Additionally, using coherence avoids the need to remove negative correlation coefficients to calculate network measures (Achard and Bullmore, 2007; Bassett et al., 2011) and robustly extracts frequency-specific information from the time series without picking up on edge effects of band-pass filtering (Percival and Walden, 2000; Zhang et al., 2016). However, currently there is no study comparing the sensitivity to scanner differences of the two connectivity metrics applied to fMRI data. Our results indicate that ComBat harmonization can remove scanner effects from the data, regardless of the choice of connectivity metric. However, wavelet coherence-based measures showed weaker differences across sites than Pearson correlation-based measures in non-harmonized data. Moreover, wavelet coherence measures generally resulted in stronger anti-correlations between age and the connectivity and network measures across all the three atlases (AAL, Power and Gordon) both before and after harmonization. For multi-site fMRI studies, this result suggests that wavelet coherence may be preferable to Pearson correlation when extracting connectivity and network summary outcomes.

4.3 Power atlas outperforms AAL and Gordon atlases

We also studied the effects of three atlases (AAL, Power and Gordon) on multi-site fMRI measurements and the performance of ComBat harmonization. A larger percentage of connections between ROIs were significantly affected by site in the AAL atlas than in the Power and Gordan atlases. These results are consistent with previous findings that in multi-site fMRI studies, functional atlases extracted from large resting-state fMRI datasets outperform traditional anatomical atlases (Abraham et al., 2017). For all three atlases, site effects in the functional connectivity and network measures were successfully removed by ComBat harmonization. However, all the network connectivity and efficiency measures using the AAL atlas were less correlated with age than those using the Power and Gordan atlases, suggesting that the AAL atlas may not be as sensitive to underlying biological variability (assessed using age in this study) when using multi-site fMRI data. Interestingly, we did not find significant site effects using the Power atlas among non-ComBat-harmonized network efficiency measures estimated by wavelet coherence. In contrast, the AAL and Gordon atlases demonstrated strong site effects in these non-ComBat-harmonized network measures. Overall, we concluded that the Power atlas outperforms the AAL and Gordon atlases with respect to post-ComBat analyses of biological variability.

4.4 Strengths, limitations and future direction

Our current study has several strengths: (1) we investigated six combinations of two connectivity metrics and three atlases, and thus were able to explore the ability of ComBat harmonization to remove site effects and to identify combinations of connectivity metrics (wavelet coherence) and atlases (Power and Gordon) that best preserved age-related anti-correlations after harmonization; (2) we used a relatively large sample (228 participants), therefore providing relatively reliable and convincing results; (3) by using the ComBat model, which is generic in its formulations and thus could easily be generalized to additional imaging modalities, our findings may have implications for multi-site electroencephalography, magnetoencephalography and other neurophysiological and neuroimaging datasets; (4) ComBat has been implemented the in MATLAB and R (https://github.com/Jfortin1/ComBatHarmonization) and in Python (https://github.com/ncullen93/neuroCombat) making the technique available and largely applicable to analysts using a variety of different software packages for image processing.

There are also several limitations that should be considered and improved in future studies. First, several previous fMRI studies (Friedman et al., 2006; Brown et al., 2011; Forsyth et al., 2014; Keshavan et al., 2016; Rath et al., 2016; Noble et al., 2017; Shinohara et al., 2017) used traveling-subject datasets in which the same participants were scanned across sites to reduce the subject effects. One recent study (Noble et al., 2017) using a small dataset (8 subjects scanned at each of 8 sites) found that the subject differences were stronger than potential site effects. Although, our ComBat harmonization technique was tested on different participants scanned across different sites with heterogeneous protocols, we speculate that ComBat harmonization will also have excellent performance on removing any non-biological variations when applied to traveling-subject datasets, however this remains to be proven. Second, in some longitudinal datasets, the same participants may be scanned on different scanners over multiple time points. However, the current ComBat harmonization model cannot be directly applied to this type of longitudinal data. Therefore, in the future, we plan to develop a time-dependent ComBat

algorithm to study longitudinal fMRI connectivity and network properties. Third, in the present study, we tested the performance of ComBat harmonization on two functional connectivity metrics and three atlases (six combinations). Although the ComBat model does not require assumptions on connectivity metrics and atlases, we found that the choices of connectivity metrics and atlases had a strong influence on the magnitude of site effects in fMRI measurements and on preserving biological variability (age in this study). Therefore, future work exploring the performance of ComBat harmonization in other combinations of connectivity metrics (e.g. partial correlation) and atlases (anatomical atlas: Brodmann, 1909; Desikan et al., 2006; functional atlas: Yeo et al., 2011; Wig et al., 2014; Glasser et al., 2016; Schaefer et al., 2017) is warranted. Finally, in this project we focused on the ability of ComBat harmonization to preserve age-related associations with several network connectivity and efficiency measures. However, previous studies (Bullmore and Sporns, 2012; Stam, 2014; Fornito et al., 2015; Yu et al., 2016, 2017) have shown that functional brain network organizations are highly correlated with other demographic (e.g. gender, educational level), clinical phenotypes (e.g. disease severity for neurological disorders), and pathological biomarkers (e.g. amyloid-β_{42} and tau proteins in Alzheimer's disease). In particular, the EMBARC functional dataset was originally designed to study the potential differences on fMRI measurements between MDD patients and healthy controls (Greenberg et al., 2015; Trivedi et al., 2016; Webb et al., 2016). Future studies will focus on whether the ComBat-harmonized fMRI data preserve functional brain networks (Kaiser et al., 2015; Gong and He, 2015) associated with depression, and whether the abnormal network attributes in MDD after ComBat harmonization are associated with patients' symptoms (Sheline et al., 2009, 2010; Otte et al., 2016; Williams, 2016).

5 | CONCLUSION

ComBat harmonization is a powerful technique for removing site effects in functional connectivity matrices, network connectivity, and efficiency measures. In addition, it preserves or strengthens the power to detect age-related anti-correlations in network connectivity and efficiency measures. In the current multi-site fMRI study, the optimal performance of ComBat harmonization was obtained by using wavelet coherence to extract functional connectivity from the Power atlas segmentation of functional brain images.

FINANCIAL DISCLOSURES

REFERENCES

Abraham A, Milham MP, Di Martino A, Cameron Craddock R, Samaras D, Thirion B, Varoquaux G (2017): Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example. Neuroimage 147:736–745.

Achard S, Salvador R, Whitcher B, Suckling J, Bullmore E (2006): A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. J Neurosci 26(1):63–72.

Achard S, Bullmore E (2007): Efficiency and cost of economical brain brain functional networks. PLoS Comput Biol 3:e17.

Achard S, Delon-Martin C, Vértes PE, Renard F, Schenck M, Schneider F, et al. (2012): Hubs of brain functional networks are radically reorganized in comatose patients. Proc Natl Acad Sci USA 109:20608–13.

Ajilore O, Lamar M, Kumar A (2014): Association of brain network efficiency with aging, depression, and cognition. Am J Geriatr Psychiatry 22(2):102–10.

Avants BB, Yushkevich P, Pluta J, Minkoff D, Korczykowski M, Detre J, Gee JC (2010): The optimal template effect in hippocampus studies of diseased populations. Neuroimage 49(3):2457–2466.

Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC (2011a): A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage 54:2033–2044.

Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC (2011b): An open source multivariate framework for n-tissue segmentation with evaluation on public data. Neuroinformatics; 9:381–400.

Bassett DS, Wymbs NF, Porter MA, Mucha PJ, Carlson JM, Grafton ST (2011): Dynamic reconfiguration of human brain networks during learning. Proc Natl Acad Sci U S A 108(18):7641–6.

Benjamini Y, Hochberg Y (1995): Controling the false discovery rate: a practical and powerful approach to multiple testing. J Roy Stat Soc B Met 57:289–300.

Biswal B1, Yetkin FZ, Haughton VM, Hyde JS (1995): Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34(4):537–41.

Biswal BB, Mennes M, Zuo XN, Gohel S, Kelly C, Smith SM, Beckmann CF, Adelstein JS, Buckner RL, Colcombe S, Dogonowski AM, Ernst M, Fair D, Hampson M, Hoptman MJ, Hyde JS, Kiviniemi VJ, Kötter R, Li SJ, Lin CP, Lowe MJ, Mackay C, Madden DJ, Madsen KH, Margulies DS, Mayberg HS, McMahon K, Monk CS, Mostofsky SH, Nagel BJ, Pekar JJ, Peltier SJ, Petersen SE, Riedl V, Rombouts SA, Rypma B, Schlaggar BL, Schmidt S, Seidler RD, Siegle GJ, Sorg C, Teng GJ, Veijola J, Villringer A, Walter M, Wang L, Weng XC, Whitfield-Gabrieli S, Williamson P, Windischberger C, Zang YF, Zhang HY, Castellanos FX, Milham MP (2010): Toward discovery science of human brain function. Proc Natl Acad Sci U S A 107(10):4734–9.

Bressler SL, Menon V (2010): Large-scale brain networks in cognition: emerging methods and principles. Trends Cogn Sci 14(6):277–90.

Brodmann K (1909/2006): Localization in the Cerebral Cortex, translated by Garey LJ. New York: Springer.

Brown GG, Mathalon DH, Stern H, Ford J, Mueller B, Greve DN, McCarthy G, Voyvodic J, Glover G, Diaz M, Yetter E, Ozyurt IB, Jorgensen KW, Wible CG, Turner JA, Thompson WK, Potkin SG; Function Biomedical Informatics Research Network (2011): Multisite reliability of cognitive BOLD data. Neuroimage 54(3):2163–75.

Bullmore E, Sporns O (2009): Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10:186–98.

Bullmore E, Sporns O (2012): The economy of brain network organization. Nat Rev Neurosci 13:336–49.

Chavez S, Viviano J, Zamyadi M, Kingsley PB, Kochunov P, Strother S, Voineskos A (2018): A novel DTI-QA tool: Automated metric extraction exploiting the sphericity of an agar filled phantom. Magn Reson Imaging 46:28–39.

Ciric R, Wolf DH, Power JD, Roalf DR, Baum GL, Ruparel K, Shinohara RT, Elliott MA, Eickhoff SB, Davatzikos C, Gur RC, Gur RE, Bassett DS, Satterthwaite TD (2017): Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. Neuroimage. 154:174–187.

Cohen JR, D'Esposito M (2016): The Segregation and Integration of Distinct Brain Networks and Their Relationship to Cognition. J Neurosci 36(48):12083–12094.

Damoiseaux JS, Beckmann CF, Arigita EJ, Barkhof F, Scheltens P, Stam CJ, Smith SM, Rombouts SA (2008): Reduced resting-state brain activity in the "default network" in normal aging. Cereb Cortex 18(8):1856–64.

Damoiseaux JS (2017): Effects of aging on functional and structural brain connectivity. Neuroimage 160:32–40.

Dansereau C, Benhajali Y, Risterucci C, Pich EM, Orban P, Arnold D, Bellec P (2017) Statistical power and prediction accuracy in multisite resting-state fMRI connectivity. Neuroimage 149:220–232.

Das SR, Avants BB, Grossman M, Gee JC (2009): Registration based cortical thickness measurement. Neuroimage 45(3):867–79.

Delaparte L, Yeh FC, Adams P, Malchow A, Trivedi MH, Oquendo MA, Deckersbach T, Ogden T, Pizzagalli DA, Fava M, Cooper C, McInnis M, Kurian BT, Weissman MM, McGrath PJ, Klein DN, Parsey RV, DeLorenzo C (2017): A comparison of structural connectivity in anxious depression versus non-anxious depression. J Psychiatr Res 89:38–47.

Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006): An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. Neuroimage 31, 968.

Di Martino A, Yan CG, Li Q, Denio E, Castellanos FX, Alaerts K, Anderson JS, Assaf M, Bookheimer SY, Dapretto M, Deen B, Delmonte S, Dinstein I, Ertl-Wagner B, Fair DA, Gallagher L, Kennedy DP, Keown CL, Keysers C, Lainhart JE, Lord C, Luna B, Menon V, Minshew NJ, Monk CS, Mueller S, Müller RA, Nebel MB, Nigg JT, O'Hearn K, Pelphrey KA, Peltier SJ, Rudie JD, Sunaert S, Thioux M, Tyszka JM, Uddin LQ, Verhoeven JS, Wenderoth N, Wiggins JL, Mostofsky SH, Milham MP (2014): The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. Mol Psychiatry 19 (6):659–667.

Dijkstra EW (1959): A note on two problems in connection with graphs. Numerische Math 1:269–271.

Doucet GE, Bassett DS, Yao N, Glahn DC, Frangou S (2017): The Role of Intrinsic Brain Functional Connectivity in Vulnerability and Resilience to Bipolar Disorder. Am J Psychiatry 174(12):1214–1222.

Feis RA, Smith SM, Filippini N, Douaud G, Dopper EG, Heise V, Trachtenberg AJ, van Swieten JC, van Buchem MA, Rombouts SA, Mackay CE (2015): ICA-based artifact removal diminishes scan site differences in multi-center resting-state fMRI. Front Neurosci 9:395.

Ferreira LK, Busatto GF (2013): Resting-state functional connectivity in normal brain aging. Neurosci Biobehav Rev 37(3):384–400.

First MB, Spitzer RL, Gibbon M, et al. (2002): Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Research Version, Patient Edition (SCID-I/P). New York, Biometrics Research, New York State Psychiatric Institute.

Fornito A, Zalesky A, Breakspear M (2015): The connectomics of brain disorders. Nat Rev Neurosci 16:159–72.

Fornito A, Zalesky A, Bullmore E (2016): Fundamentals of Brain Network Analysis. Academic Press, Cambridge.

Forsyth JK, McEwen SC, Gee DG, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos HW, Tsuang MT, van Erp TG, Walker EF, Hamann S, Woods SW, Qiu M, Cannon TD (2014): Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome Longitudinal Study. Neuroimage 97:41–52.

Fortin JP, Parker D, Tunç B, Watanabe T, Elliott MA, Ruparel K, Roalf DR, Satterthwaite TD, Gur RC, Gur RE, Schultz RT, Verma R, Shinohara RT (2017): Harmonization of multi-site diffusion tensor imaging data. Neuroimage 161:149–170.

Fortin JP, Cullen N, Sheline YI, Taylor WD, Aselcioglu I, Cook PA, Adams P, Cooper C, Fava M, McGrath PJ, McInnis M, Phillips ML, Trivedi MH, Weissman MM, Shinohara RT (2018): Harmonization of cortical thickness measurements across scanners and sites. Neuroimage 167:104–120.

Fox MD, Raichle ME (2007): Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8:700–11.

Friedman L, Glover GH; Fbirn Consortium (2006): Reducing interscanner variability of activation in a multicenter fMRI study: controlling for signal-to-fluctuation-noise-ratio (SFNR) differences. Neuroimage 33(2):471–81.

Friedman L, Stern H, Brown GG, Mathalon DH, Turner J, Glover GH, Gollub RL, Lauriello J, Lim KO, Cannon T, Greve DN, Bockholt HJ, Belger A, Mueller B, Doty MJ, He J, Wells W, Smyth P, Pieper S, Kim S, Kubicki M, Vangel M, Potkin SG (2008): Test-retest and between-site reliability in a multicenter fMRI study. Hum Brain Mapp 29(8):958–72.

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

Glover GH, Mueller BA, Turner JA, van Erp TG, Liu TT, Greve DN, Voyvodic JT, Rasmussen J, Brown GG, Keator DB, Calhoun VD, Lee HJ, Ford JM, Mathalon DH, Diaz M, O'Leary DS, Gadde S, Preda A, Lim KO, Wible CG, Stern HS, Belger A, McCarthy G, Ozyurt B, Potkin SG (2012): Function biomedical informatics research network recommendations for prospective multicenter functional MRI studies. J Magn Reson Imaging 36(1):39–54.

Gong Q, He Y (2015): Depression, neuroimaging and connectomics: a selective overview. Biol Psychiatry 77(3):223–235.

Gordon EM, Laumann TO, Adeyemo B, Huckins JF, Kelley WM, Petersen SE (2016): Generation and evaluation of a cortical area parcellation from resting-state correlations. Cereb Cortex 26(1):288–303.

Gountouna VE, Job DE, McIntosh AM, Moorhead TW, Lymer GK, Whalley HC, Hall J, Waiter GD, Brennan D, McGonigle DJ, Ahearn TS, Cavanagh J, Condon B, Hadley DM, Marshall I, Murray AD, Steele JD, Wardlaw JM, Lawrie SM (2010): Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. Neuroimage 49(1):552–60.

Gradin V, Gountouna VE, Waiter G, Ahearn TS, Brennan D, Condon B, Marshall I, McGonigle DJ, Murray AD, Whalley H, Cavanagh J, Hadley D, Lymer K, McIntosh A, Moorhead TW, Job D, Wardlaw J, Lawrie SM, Steele JD (2010): Between- and within-scanner variability in the CaliBrain study n-back cognitive task. Psychiatry Res 184(2):86–95.

Grady CL, Protzner AB, Kovacevic N, Strother SC, Afshin-Pour B, Wojtowicz M, Anderson JA, Churchill N, McIntosh AR (2010): A multivariate analysis of age-related differences in default mode and task-positive networks across multiple cognitive domains. Cereb Cortex 20(6):1432–47.

Greenberg T, Chase HW, Almeida JR, Stiffler R, Zevallos CR, Aslam HA, Deckersbach T, Weyandt S, Cooper C, Toups M, Carmody T, Kurian B, Peltier S, Adams P, McInnis MG, Oquendo MA, McGrath PJ1,

Fava M, Weissman M, Parsey R, Trivedi MH, Phillips ML1 (2015): Moderation of the Relationship Between Reward Expectancy and Prediction Error-Related Ventral Striatal Reactivity by Anhedonia in Unmedicated Major Depressive Disorder: Findings From the EMBARC Study. Am J Psychiatry 172(9):881–91.

Greve DN, Fischl B (2009): Accurate and robust brain image alignment using boundary-based registration. Neuroimage 48:63–72.

Hallquist MN, Hwang K, Luna B (2013): The nuisance of nuisance regression: spectral misspecification in a common approach to resting-state fmri preprocessing reintroduces noise and obscures functional connectivity. Neuroimage 2013; 82C: 208–225.

Hamilton M (1960): A rating scale for depression. J Neurol Neurosurg Psychiatry 23(1):56–62.

Huber PJ (2004): Robust Statistics. Wiley.

Jenkinson M, Bannister P, Brady M, Smith S (2002): Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17:825–841.

Johnson WE, Li C, Rabinovic A (2007): Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8(1):118–27.

Jovicich J, Minati L, Marizzoni M, Marchitelli R, Sala-Llonch R, Bartrés-Faz D, Arnold J, Benninghoff J, Fiedler U, Roccatagliata L, Picco A, Nobili F, Blin O, Bombois S, Lopes R, Bordet R, Sein J, Ranjeva JP, Didic M, Gros-Dagnac H, Payoux P, Zoccatelli G, Alessandrini F, Beltramello A, Bargalló N, Ferretti A, Caulo M, Aiello M, Cavaliere C, Soricelli A, Parnetti L, Tarducci R, Floridi P, Tsolaki M, Constantinidis M, Drevelegas A, Rossini PM, Marra C, Schönknecht P, Hensch T, Hoffmann KT, Kuijer JP, Visser PJ, Barkhof F, Frisoni GB; PharmaCog Consortium (2016): Longitudinal reproducibility of default-mode network connectivity in healthy elderly participants: A multicentric resting-state fMRI study. Neuroimage 124(Pt A):442–454.

Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA (2015): Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. JAMA Psychiatry 72(6):603–11.

Keshavan A, Paul F, Beyer MK, Zhu AH, Papinutto N, Shinohara RT, Stern W, Amann M, Bakshi R, Bischof A, Carriero A, Comabella M, Crane JC, D'Alfonso S, Demaerel P, Dubois B, Filippi M, Fleischer V, Fontaine B, Gaetano L, Goris A, Graetz C, Gröger A, Groppa S, Hafler DA, Harbo HF, Hemmer B, Jordan K, Kappos L, Kirkish G, Llufriu S, Magon S, Martinelli-Boneschi F, McCauley JL, Montalban X, Mühlau M, Pelletier D, Pattany PM, Pericak-Vance M, Cournu-Rebeix I, Rocca MA, Rovira A, Schlaeger R, Saiz A, Sprenger T, Stecco A, Uitdehaag BMJ, Villoslada P, Wattjes MP, Weiner H, Wuerfel J, Zimmer C, Zipp F; International Multiple Sclerosis Genetics Consortium. Electronic address: AIVINSON@PARTNERS.ORG, Hauser SL, Oksenberg JR, Henry RG (2016): Power estimation for non-standardized multisite studies. Neuroimage 134:281–294.

Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang M-C et al. (2009): Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. Neuroimage 46:786–802.

Koch W, Teipel S, Mueller S, Buerger K, Bokde AL, Hampel H, Coates U, Reiser M, Meindl T (2010): Effects of aging on default mode network activity in resting state fMRI: does the method of analysis matter? Neuroimage 51(1):280–7.

Kochunov P, Dickie EW, Viviano JD, Turner J, Kingsley PB, Jahanshad N, Thompson PM, Ryan MC, Fieremans E, Novikov D, Veraart J, Hong EL, Malhotra AK, Buchanan RW, Chavez S, Voineskos AN (2018): Integration of routine QA data into mega-analysis may improve quality and sensitivity of multisite diffusion tensor imaging studies. Hum Brain Mapp 39(2):1015–1023.

Latora V, Marchiori M (2001): Efficient behavior of small-world networks. Phys Rev Lett 87(19):198701.

McGonigle DJ (2012): Test-retest reliability in fMRI: Or how I learned to stop worrying and love the variability. Neuroimage 62(2):1116–20.

Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM, Mathalon DH, McGlashan TH, Perkins DO, Belger A, Seidman LJ, Thermenos H, Tsuang MT, van Erp TGM, Walker EF, Hamann S, Woods SW, Cannon TD, Constable RT (2017): Multisite reliability of MR-based functional connectivity. Neuroimage 146:959–970.

Oh J, Bakshi R, Calabresi PA, Crainiceanu C, Henry RG, Nair G, Papinutto N, Constable RT, Reich DS5, Pelletier D, Rooney W, Schwartz D, Tagge I, Shinohara RT, Simon JH, Sicotte NL; NAIMS Cooperative Steering Committee (2017): The NAIMS cooperative pilot project: Design, implementation and future directions. Mult Scler 1352458517739990.

Otte C, Gold SM, Penninx BW, Pariante CM, Etkin A, Fava M, Mohr DC, Schatzberg AF (2016): Major depressive disorder. Nat Rev Dis Primers 2:16065.

Percival DB, Walden AT (2000): Wavelet methods for time series analysis. Cambridge University Press.

Power JD, Cohen AL, Nelson SM, Wig GS, Barnes KA, Church JA, Vogel AC, Laumann TO, Miezin FM, Schlaggar BL, Petersen SE (2011): Functional network organization of the human brain. Neuron 72:665–678.

Raichle ME (2015): The brain's default mode network. Annu Rev Neurosci 38:433–47.

Rath J, Wurnig M, Fischmeister F, Klinger N, Höllinger I, Geißler A, Aichhorn M, Foki T, Kronbichler M, Nickel J, Siedentopf C, Staffen W, Verius M, Golaszewski S, Koppelstaetter F, Auff E, Felber S, Seitz RJ, Beisteiner R (2016): Between- and within-site variability of fMRI localizations. Hum Brain Mapp 37(6):2151–60.

Rubinov M, Sporns O (2010): Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52:1059–1069.

Rush AJ, Trivedi MH, Ibrahim HM, Carmody TJ, Arnow B, Klein DN, Markowitz JC, Ninan PT, Kornstein S, Manber R, Thase ME, Kocsis JH, Keller MB (2003): The 16-Item Quick Inventory of Depressive Symptomatology (QIDS), clinician rating (QIDS-C), and self-report (QIDS-SR): a psychometric evaluation in patients with chronic major depression. Biol Psychiatry 54(5):573–83.

Satterthwaite TD, Ciric R, Roalf DR, Davatzikos C, Bassett DS, Wolf DH (2017): Motion artifact in studies of functional connectivity: Characteristics and mitigation strategies. Hum Brain Mapp doi: 10.1002/hbm.23665. [Epub ahead of print]

Schaefer A, Kong R, Gordon EM, Laumann TO, Zuo XN, Holmes AJ, Eickhoff SB, Yeo BTT (2017): Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. Cereb Cortex 18:1-20.

Sheline YI, Barch DM, Price JL, Rundle MM, Vaishnavi SN, Snyder AZ et al. (2009): The default mode network and self-referential processes in depression. Proc Natl Acad Sci USA 106:1942–1947.

Sheline YI, Price JL, Yan Z, Mintun MA (2010): Resting-state functional MRI in depression unmasks increased connectivity between networks via the dorsal nexus. Proc Natl Acad Sci USA 107:11020–5.

Shinohara RT, Oh J, Nair G, Calabresi PA, Davatzikos C, Doshi J, Henry RG, Kim G, Linn KA, Papinutto N, Pelletier D, Pham DL, Reich DS, Rooney W, Roy S, Stern W, Tummala S, Yousuf F, Zhu A, Sicotte NL, Bakshi R; NAIMS Cooperative (2017): Volumetric Analysis from a Harmonized Multisite Brain MRI Study of a Single Subject with Multiple Sclerosis. AJNR Am J Neuroradiol 38(8):1501–1509.

Stam CJ (2014): Modern network science of neurological disorders. Nat Rev Neurosci 15:683–95.

Suckling J, Ohlssen D, Andrew C, Johnson G, Williams SC, Graves M, Chen CH, Spiegelhalter D, Bullmore E (2008): Components of variance in a multicentre functional MRI study and implications for calculation of statistical power. Hum Brain Mapp 29(10):1111–22.

Suckling J, Barnes A, Job D, Brenan D, Lymer K, Dazzan P, Marques TR, MacKay C, McKie S, Williams SR, Williams SC, Lawrie S, Deakin B (2010): Power calculations for multicenter imaging studies controlled by the false discovery rate. Hum Brain Mapp 31(8):1183–95.

Tomasi D, Volkow ND (2012): Aging and functional brain networks. Mol Psychiatry 17(5):471, 549–58.

Trivedi MH, McGrath PJ, Fava M, Parsey RV, Kurian BT, Phillips ML, Oquendo MA, Bruder G, Pizzagalli D, Toups M, Cooper C, Adams P, Weyandt S, Morris DW, Grannemann BD, Ogden RT, Buckner R, McInnis M, Kraemer HC, Petkova E, Carmody TJ, Weissman MM (2016): Establishing moderators and biosignatures of antidepressant response in clinical care (EMBARC): Rationale and design. J Psychiatr Res 78:11–23.

Turner JA, Damaraju E, van Erp TG, Mathalon DH, Ford JM, Voyvodic J, Mueller BA, Belger A, Bustillo J, McEwen S, Potkin SG, Fbirn, Calhoun VD (2013): A multi-site resting state fMRI study on the amplitude of low frequency fluctuations in schizophrenia. Front Neurosci 7:137.

Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA et al. (2010): N4ITK: improved N3 bias correction. IEEE Trans Med Imaging 29:1310–1320.

Tustison NJ, Cook PA, Klein A, Song G, Das SR, Duda JT et al. (2014): Large-scale eval- uation of ants and freesurfer cortical thickness measurements. Neuroimage 99:166–179.

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, et al. (2002): Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. Neuroimage 15:273–89.

Van Horn JD, Toga AW (2009): Multi-Site Neuroimaging Trials. Curr Opin Neurol 22(4):370–378.

Watson D, Clark LA (1991): The mood and anxiety symptom questionnaire. Iowa City Iowa, University of Iowa.

Webb CA, Dillon DG, Pechtel P, Goer FK, Murray L, Huys QJ, Fava M, McGrath PJ, Weissman M, Parsey R, Kurian BT, Adams P, Weyandt S, Trombello JM, Grannemann B, Cooper CM, Deldin P, Tenke C, Trivedi M, Bruder G, Pizzagalli DA (2016): Neural Correlates of Three Promising Endophenotypes of Depression: Evidence from the EMBARC Study. Neuropsychopharmacology 41(2):454–63.

Wig GS, Laumann TO, Petersen SE (2014): An approach for parcellating human cortical areas using resting-state correlations. Neuroimage 93:276–291.

Williams LM (2016): Precision psychiatry: a neural circuit taxonomy for depression and anxiety. Lancet Psychiatry May;3(5):472–80.

Xia M, Wang J, He Y (2013): BrainNet Viewer: a network visualization tool for human brain connectomics. PLoS One 8:e68910.

Yeo BTT, Krienen FM, Sepulcre J, Sabuncu MR, Lashkari D, Hollinshead M, Roffman JL, Smoller JW, Zöllei L, Polimeni JR, et al. (2011): The organization of the human cerebral cortex estimated by intrinsic functionalconnectivity. J Neurophysiol 106:1125–1165.

Yu M, Gouw AA, Hillebrand A, Tijms BM, Stam CJ, van Straaten EC, et al. (2016): Different functional connectivity and network topology in behavioral variant of frontotemporal dementia and Alzheimer's disease: an EEG study. Neurobiol Aging 42:150–62.

Yu M, Engels MMA, Hillebrand A, van Straaten ECW, Gouw AA, Teunissen C, van der Flier WM, Scheltens P, Stam CJ (2017): Selective impairment of hippocampus and posterior hub areas in Alzheimer's disease: an MEG-based multiplex network study. Brain 140(5):1466–1485.

Zhang Z, Telesford QK, Giusti C, Lim KO, Bassett DS (2016): Choosing wavelet methods, filters, and lengths for functional brain network construction. PLoS One 11:e0157243.

TABLES

TABLE 1 Imaging parameters for the four clinical sites: Columbia University (CU), Massachusetts General Hospital (MGH), the University of Texas Southwestern Medical Center (TX) and the University of Michigan (UM).

	CU	MGH	TX	UM
Scanner	General Electric 3T	Siemens 3T	Phillips 3T	Phillips 3T
Structural	FSPGR TR = 6.0 ms TE = 2.4 ms TI = 900 ms Flip Angle = 9° FOV= 256×256 mm Slice Thickness = 1 mm Matrix = 256×256 178 continuous slices (4 discarded)	MPRAGE TR = 2300 ms TE = 2.54 ms TI = 900 ms Flip Angle = 9° FOV= 256×256 mm Slice Thickness = 1 mm Matrix = 256×256 176 continuous slices	Turbo Field Echo (TFE) TR = 8.2 ms TE = 3.7 ms TI = 1100 ms Flip Angle = 12° FOV= 256×256 mm Slice Thickness = 1mm Matrix = 256×256 178 continuous slices	MPRAGE TR = 2100 ms TE = 3.7 ms TI = 1100 ms Flip Angle = 12° FOV= 256×256 mm Slice Thickness = 1mm Matrix = 256×256 178 continuous slices
Functional	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV=205×205 mm Slice thickness: 3.1 mm Matrix 64×64	TR/TE = 2000/28msec Flip Angle 90° FOV = 205×205 mm Slice thickness: 3.1 mm Matrix 64×64

TABLE 2 Site effects in functional connectivity values for different connectivity measures (Pearson correlation and wavelet coherence) using different atlases (AAL, Power and Gordon).

Six combinations of two connectivity measures and three atlases		Number of ROIs	Original	ComBat
Pearson correlation	AAL	90	1610 (40.2%)	0 (0%)
	Power	203	515 (2.5%)	0 (0%)
	Gordon	307	603 (1.3%)	0 (0%)
Wavelet coherence	AAL	90	7 (0.2%)	0 (0%)
	Power	203	12 (0.06%)	0 (0%)
	Gordon	307	17 (0.04%)	0 (0%)

Note that, the numbers and percentages in the round brackets represent the numbers and percentages of connectivity values that were significantly different across 4 sites for six combinations of two connectivity measures and three atlases, respectively. Significant effects for each combination were computed using Kruskal-Wallis tests with FDR corrections. Here we only presented the results of the first-session data, as the two sessions showed extremely similar results.

FIGURE LEGENDS

FIGURE 1 Distribution of subject demographic characteristics across 4 sites. Abbreviations: MDD = major depressive disorder; Control = healthy controls; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan. Note that whiskers in the boxplots represent variability outside the upper and lower quartiles.

FIGURE 2 Site effects in functional connectivity estimated by "Correlation-AAL" (A) and "Coherence-Power" (B) before and after ComBat harmonization. Note that the functional connectivity values in (A) and (B) were computed from the time series of two randomly selected ROIs: for AAL, the TPOmid.R and ACG.R were selected; for Power, two regions in the visual cortex were selected; whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan; TPOmid.R = right Temporal pole: middle temporal gyrus; ACG.R = right Anterior cingulate and paracingulate gyri.
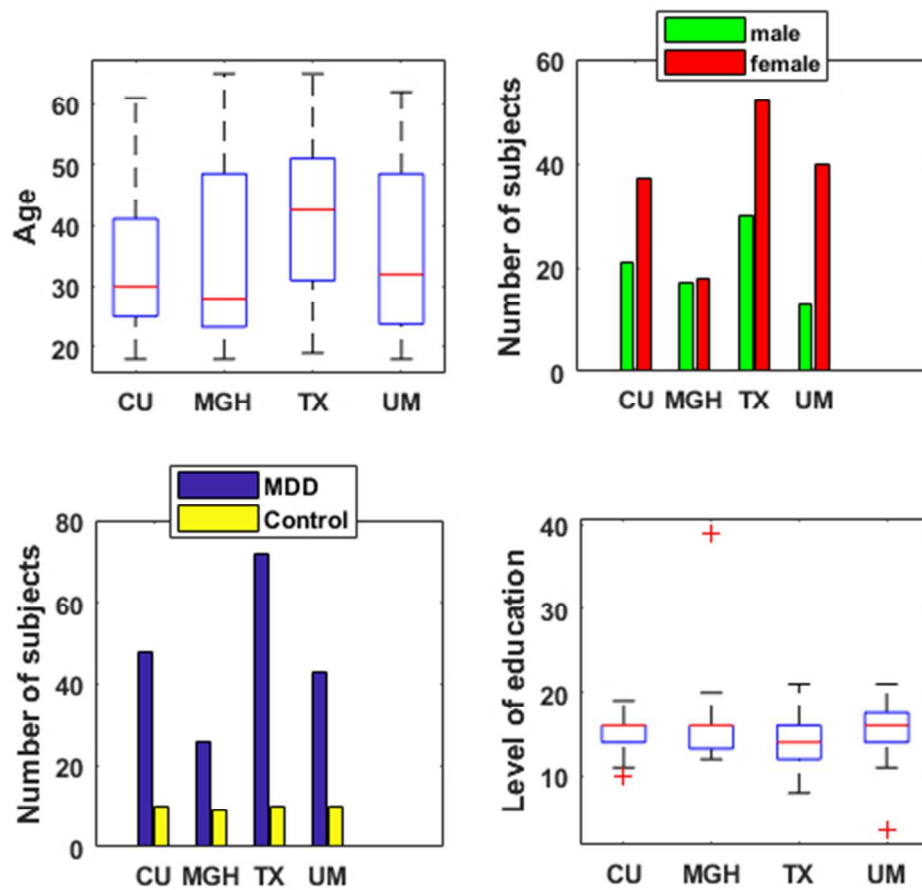
FIGURE 3 Empirical (dashed lines) and ComBat-estimated (solid lines) prior distributions for the site-specific location (γ) and scale (δ) parameters for "Correlation-AAL" (A) and "Coherence-Power" (B). Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan.

FIGURE 4 DMN connectivity estimated by "Coherence-Power" and "Correlation-AAL" across sites (**A**) and their anti-correlations with age (**B**). Note that DMN connectivity was computed by first summing the functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (Power and AAL); the significant site effects in DMN connectivity were removed by ComBat harmonization (**A**). The strong anti-correlation between age and DMN connectivity estimated by "Coherence-Power" was preserved in ComBat-harmonized data; for "Correlation-AAL", ComBat harmonization increased the detection power of the anti-correlation (**B**). Note that whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: DMN = default mode network.

FIGURE 5 Negative log-transformed p-value heat maps for Kruskal-Wallis tests (e.g. Figure 4A; Supplementary Figure S8-S10, subplot A) of network connectivity and efficiency measures (A), and correlation coefficient (B) heat maps for correlation analyses between the network measures and age (e.g. Figure 4B; Supplementary Figure S8-S10, subplot B). Note that the asterisks (blue or white colors) represent $p < .05$, which corresponds to a negative $\log_{10}$ transformed p-value greater than 1.301 in the color bar of (A). Abbreviations: DC = DMN connectivity; DNS = DMN nodal strength; DLE = DMN local efficiency; GE = global efficiency. Note that DMN connectivity was computed by first summing the

functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (AAL, Power and Gordon); please see the details of the computation of DNS, DLE and GE in Section 2.7 Calculation of network measures.
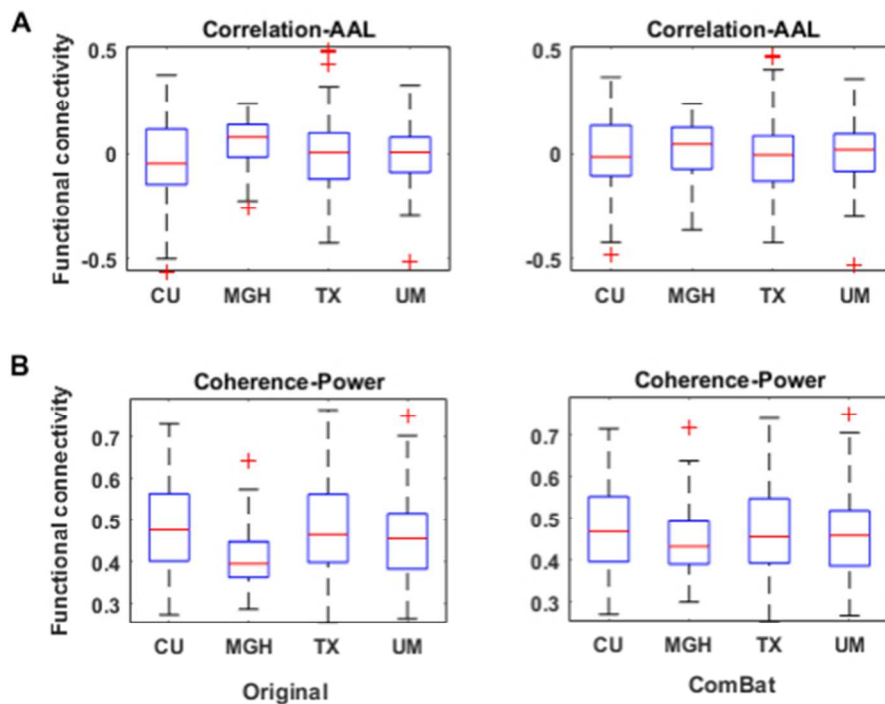
Distribution of subject demographic characteristics across 4 sites. Abbreviations: MDD = major depressive disorder; Control = healthy controls; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan. Note that whiskers in the boxplots represent variability outside the upper and lower quartiles.

127x120mm (96 x 96 DPI)

Site effects in functional connectivity estimated by "Correlation-AAL" (A) and "Coherence-Power" (B) before and after ComBat harmonization. Note that the functional connectivity values in (A) and (B) were computed from the time series of two randomly selected ROIs: for AAL, the TPOmid.R and ACG.R were selected; for Power, two regions in the visual cortex were selected; whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan; TPOmid.R = right Temporal pole: middle temporal gyrus; ACG.R = right Anterior cingulate and paracingulate gyri.
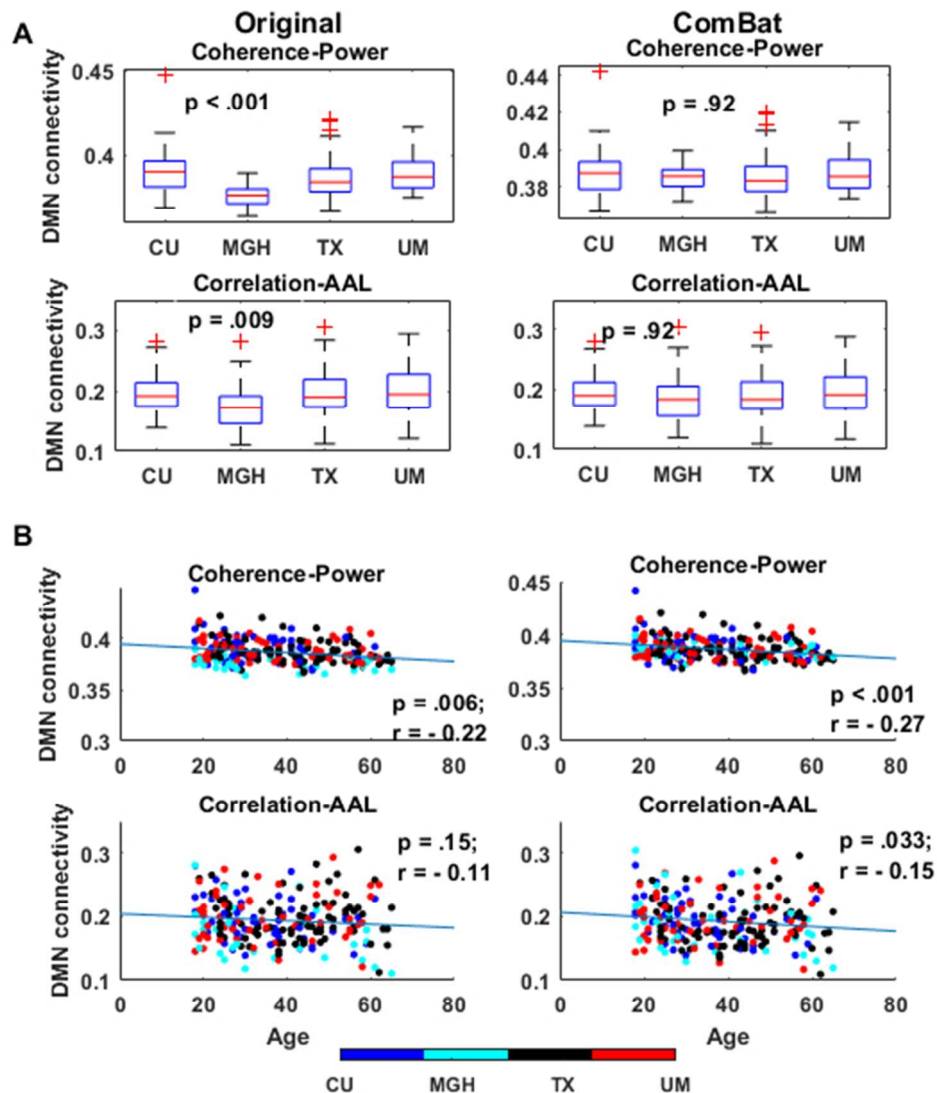
152x110mm (96 x 96 DPI)
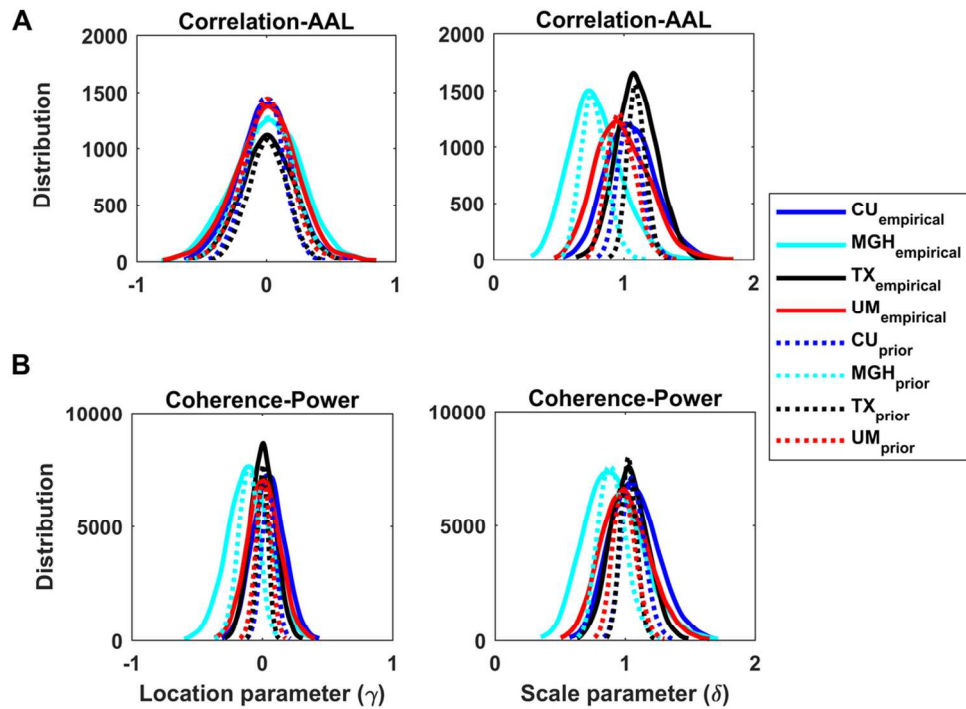
Empirical (dashed lines) and ComBat-estimated (solid lines) prior distributions for the site-specific location (γ) and scale (δ) parameters for "Correlation-AAL" (A) and "Coherence-Power" (B). Abbreviations: CU = Columbia University; MGH = Massachusetts General Hospital; TX = University of Texas Southwestern Medical Center; UM = University of Michigan.

111x81mm (300 x 300 DPI)

DMN connectivity estimated by "Coherence-Power" and "Correlation-AAL" across sites (A) and their anti-correlations with age (B). Note that DMN connectivity was computed by first summing the functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (Power and AAL); the significant site effects in DMN connectivity were removed by ComBat harmonization (A). The strong anti-correlation between age and DMN connectivity estimated by "Coherence-Power" was preserved in ComBat-harmonized data; for "Correlation-AAL", ComBat harmonization increased the detection power of the anti-correlation (B). Note that whiskers in the boxplots represent variability outside the upper and lower quartiles. Abbreviations: DMN = default mode network.

148x161mm (96 x 96 DPI)
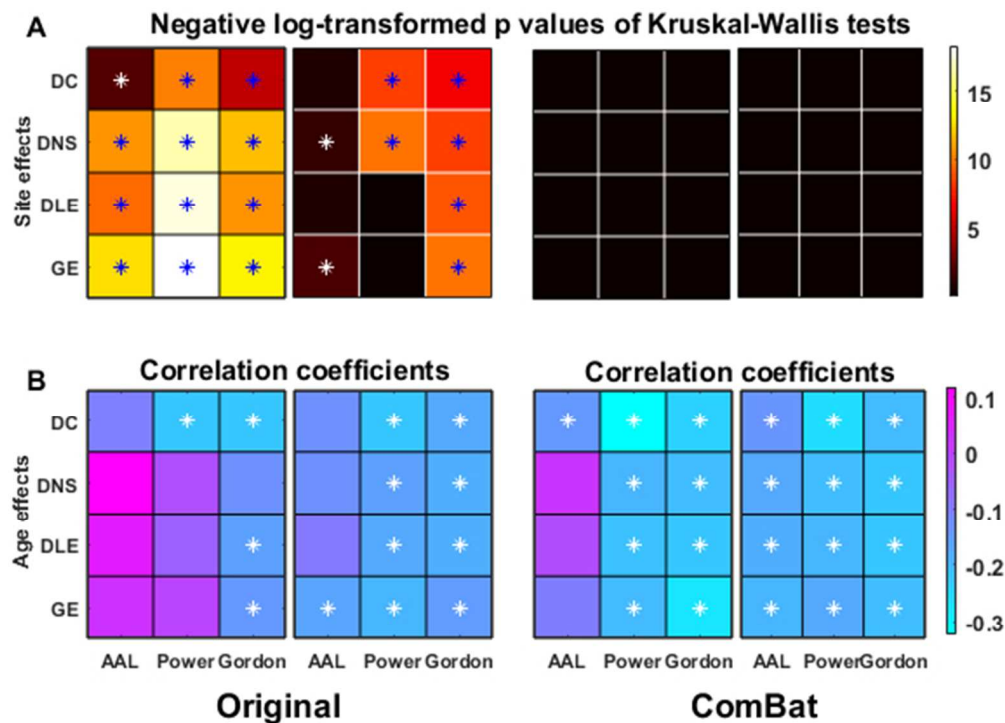
5 Negative log-transformed p-value heat maps for Kruskal-Wallis tests (e.g. Figure 4A; Supplementary Figure S8-S10, subplot A) of network connectivity and efficiency measures (A), and correlation coefficient (B) heat maps for correlation analyses between the network measures and age (e.g. Figure 4B; Supplementary Figure S8-S10, subplot B). Note that the asterisks (blue or white colors) represent p < .05, which corresponds to a negative log10 transformed p-value greater than 1.301 in the color bar of (A). Abbreviations: DC = DMN connectivity; DNS = DMN nodal strength; DLE = DMN local efficiency; GE = global efficiency. Note that DMN connectivity was computed by first summing the functional connectivity values within the DMN ROIs, and then normalizing by the number of DMN ROIs corresponding to each atlas (AAL, Power and Gordon); please see the details of the computation of DNS, DLE and GE in Section 2.7 Calculation of network measures.

152x111mm (96 x 96 DPI)