Harmonization of Structural Brain Networks for Multi-site Studies

Rui Sherry Shen[1,2*], Drew Parker[2], Ragini Verma[1,2]

[1]*Department of Bioengineering, University of Pennsylvania, Philadelphia, United States*
[2]*Perelman School of Medicine, University of Pennsylvania, Philadelphia, United States*

Abstract

Research on structural networks often suffers from limited sample sizes and inherent selection biases in individual studies, which restrict their ability to address complex questions regarding human brain organization. Pooling data across studies is crucial for achieving a more comprehensive representation of the population and effectively managing individual heterogeneity; however, structural networks acquired from multiple sites are susceptible to significant site-related differences. This necessitates harmonization to mitigate biases and reveal true biological variability in multi-site analyses. Our work marks the first effort to develop and evaluate harmonization frameworks specifically for structural networks. We adapt several statistical approaches for harmonizing structural networks and provide a comprehensive evaluation to rigorously test their effectiveness. Our findings demonstrate that the adaptation of the Gamma Generalized Linear Model (gamma-GLM) outperforms other methods in modeling structural network data, effectively eliminating site-related effects in structural connectivity matrices and downstream graph-based analyses while preserving biological variability. Additionally, we highlight gamma-GLM's superiority in addressing confounding factors between site and age. Two practical applications further illustrate the utility of our harmonization framework in tackling common challenges in multi-site structural network studies. Specifically, harmonization using gamma-GLM enhances the transferability of brain age predictors to new datasets and facilitates data integration in patient studies, ultimately improving statistical power. Together, our work offers essential guidelines for harmonizing and integrating multi-site structural network studies, paving the way for more robust discoveries through collaborative research in the era of big data.
Keywords: Strucutral network, Harmonization, Diffusion MRI, Gamma Generalized Linear Model, Multi-site Analysis, Connectome, Big Data

1. Introduction

Investigating brain disorders through structural brain networks faces major challenges due to the high individual heterogeneity and small sample sizes of single studies [1-3]. This often leads to inconclusive or divergent findings within the field [4, 5]. To tackle these issues, the scientific community has initiated efforts to pool and maintain imaging data from various clinical sites, aiming to address sample size limitations and reduce selection biases inherent in individual studies, thereby achieving a more comprehensive representation of the population. As a result, several large-scale structural network datasets have been developed, including the UK Biobank [6], Autism Brain Imaging Data Exchange (ABIDE) [7], and Action-Based Conversations Dataset (ABCD) [8]. However, integrating structural network data from multiple sources and utilizing large datasets retrospectively necessitates effective harmonization methods to eliminate site-specific differences that are unrelated to the underlying biology.

Structural networks are typically mapped non-invasively using diffusion-weighted MRI (dMRI) [9] and represented as connectivity matrices, with nodes denoting brain regions and edges representing their axonal connections. Data collected from different sites often exhibits substantial variability due to technical differences in scanner manufacturers, acquisition protocols, preprocessing pipelines, tractography algorithms, and definition of networks [10-12]. Although several studies have attempted to reduce site-related effects by standardizing acquisition protocols and parameters [13-15], significant multi-site variations persist. These discrepancies can introduce site-related bias in brain network analysis, compromise data comparability in collaborative research, obscure biologically meaningful associations, and even lead to spurious associations [16-18].

Current efforts to mitigate site-related variations primarily focus on harmonizing dMRI signals. Mirzaalian et al. [19] introduced a technique that transforms dMRI signals into spherical harmonic space and learns site-to-site mapping by matching their rotation-invariant spherical harmonic (RISH) features. This approach has proven effective in harmonizing dMRI signals from similar acquisitions. Additionally, deep learning methods based on

* Corresponding author. *E-mail address:* ruishen@seas.upenn.edu

RISH features have been developed and evaluated [20, 21]. Huynh et al. [22] employed the method of moments (MoM), which directly harmonizes dMRI signals without transforming them into other domains, offering greater flexibility for integrating dMRI data with varying numbers of gradient directions.

However, these approaches address site effects only at the dMRI signal level, assuming that downstream outcomes, such as structural networks, will be harmonized accordingly. Kurokawa et al. [23], using a traveling-subject design, observed that even after the harmonization of dMRI signals, cross-scanner differences in structural networks and their graph metrics still persisted. This highlights the need to address inter-site variability at the connectivity matrix level before conducting network analyses or downstream tasks, to avoid biases in multi-site studies.

The complexity of structural network data—marked by non-Gaussian distributions, high dimensionality, and graph-structured organization— demands specialized methods and rigorous evaluation to develop effective harmonization tools. Although various statistical models have been proposed for harmonizing other neuroimaging modalities, their suitability for structural networks remains unexplored. ComBat, originally designed to address 'batch effects' in genomic data [24], uses parametric empirical Bayes [25] to mitigate site-related effects by adjusting the data for mean and variance while accounting for biological covariates such as age and sex. Fortin et al. [26] extended ComBat to harmonize dMRI scalar maps such as fractional anisotropy (FA) and mean diffusivity (MD), and Yu et al. [27] applied ComBat to resting-state functional networks. In addition to ComBat, methods such as the generalized linear model (GLM) [28, 29] and CovBat [30] have been proposed for neuroimaging data harmonization. However, most of these methods are limited to harmonizing measures that follow normal distribution. To date, no harmonization tools have been specifically designed and validated for structural networks.

In this work, we fill the gap by extending harmonization framework to the structural network data. We adapt several statistical harmonization models specifically for structural networks and provide a comprehensive evaluation of these methods. Each harmonization method is rigorously tested to eliminate site-specific effects in structural connectivity matrices and downstream graph-based analyses [31], while preserving biological variability and ensuring robustness against confounding factors across different sites. Furthermore, we present two practical use cases that demonstrate how these harmonization approaches can overcome common challenges in multi-site structural network studies: first, we evaluate their application in enhancing the transferability of machine learning models to new datasets, particularly in the context of brain age prediction; second, recognizing that most structural network studies involve both control and patient groups, we illustrate the effectiveness of harmonization methods for both diagnostic groups, facilitating data integration in patient studies and enhancing their statistical power. Collectively, our work provides valuable guidelines for the harmonization and integration of multi-site structural network studies, supporting more robust findings in collaborative research and contributing to innovative diagnostic and therapeutic strategies.

2. Methods

2.1. Overview of Datasets

We assess several harmonization approaches on diffusion MRI datasets collected from six separate sites, including the Philadelphia Neurodevelopmental Cohort (PNC) [32], the Center for Autism Research (CAR) [33], TimRobert [34], and three clinical sites from Autism Brain Imaging Data Exchange II (ABIDEII-NYU, ABIDEII-SDSU, ABIDEII-TCD) [7]. A total of 1503 participants (890 males, 613 females) are included in this study. The detailed demographic information and acquisition parameters of each dataset are described in Table 1 .

Table 1 Demographic information and acquisition parameters for six datasets.

Dataset	TDC			ASD			Scanner	dMRI					T1-w MP-RAGE				
	Age range (years)	Sex (M, F)	Sample size	Age range (years)	Sex (M, F)	Sample size		b values (s/mm^2)	Gradient directions	TR (ms)	TE (ms)	Resolution (mm)	TR (ms)	TE (ms)	TI (ms)	Flip angle	Resolution (mm)
PNC	8.17-21.00 (15.06±3.33)	M=421 F=533	954	-	-	-	Siemen Trio 3T	0, 1000	64	8100	82	1.875× 1.875 × 2	1810	3.5	1100	9°	0.9375 × 0.9375 × 1
CAR	6.26-17.90 (11.81±2.85)	M=112 F=30	142	6.37-18.82 (12.05±2.73)	M=129 F=22	151	Siemens Verio 3T	0, 1000	30	11000	76	2 × 2 × 2	1900	2.54	900	9°	0.8 × 0.8 × 0.9

TimRobert	6.17-13.83 (9.97±2.04)	M=32 F=6	38	6.17-16.92 (10.65±2.18)	M=61 F=9	70	Siemens Verio 3T	0, 1000	30	11000	76.4	2 × 2 × 2	1900	2.87	1050	9°	1 × 1 × 1
ABIDEII-NYU	6.66-12.90 (9.55±1.77)	M=19 F=0	19	6.05-17.93 (8.33±2.17)	M=35 F=4	39	Siemens Allegra	0, 1000	64	5200	78	3 × 3 × 3	2530	3.25	1100	7°	1.3 × 1 × 1.33
ABIDEII-SDSU	8.10-17.70 (13.50±3.05)	M=21 F=2	23	7.40-18.00 (12.95±3.35)	M=24 F=7	31	GE MR750	0, 1000	61	8500	84.9	1.875 × 1.875 × 2	8.136	3.172	600	8°	1 × 1 × 1
ABIDEII-TCD	10.25-20.00 (16.18±2.97)	M=18 F=0	18	10.00-19.50 (14.47±3.32)	M=18 F=0	18	Philips Achieva	0, 1500	61	20244	79	1.94 × 1.94 × 2	8.4	3.9	1150	8°	0.9 × 0.9 × 0.9

[1]Mean and standard deviation of age are shown in parentheses.

[2]Abbreviations: M, male; F, female, TR, repetition time; TE, echo time; TI, inversion time

To evaluate the effectiveness of harmonization methods in different scenarios, we create four cohorts using several subsets of these datasets.

2.1.1. Paired PNC-CAR cohort

We carefully select typically developing children from our two largest datasets, the PNC and CAR sites, ensuring they are closely matched in age and sex. This approach aims to mimic the traveling-subject research paradigm, minimizing confounding effects that could impact the harmonization results. As a result, we form two groups from different sites, each consisting of 100 participants (76 males, 24 females). The age of participants from the PNC site ranges from 8.17 to 18.08 years, with a mean of 12.31 years and a standard deviation of 2.70 years at enrollment. Participants from the CAR site range in age from 8.08 to 17.83 years, with a mean of 12.27 years and a standard deviation of 2.73 years. This cohort will be used to assess whether harmonization methods effectively remove scanner effects and preserve biological variability with minimal confounding biases involved.

2.1.2. Confounded NYU-TCD cohort

We use all typically developing children from the ABIDEII-NYU and ABIDEII-TCD datasets to create a cohort where age and site are confounded. The ABIDEII-NYU dataset includes the youngest participants among the six datasets, with an age range of 6.66 to 12.90 years, a mean of 9.55 years, and a standard deviation of 1.77 years. In contrast, the ABIDEII-TCD dataset consists of the oldest participants, with an age range of 10.25 to 20.00 years, a mean of 16.18 years, and a standard deviation of 2.97 years. Since no female subjects were included in these datasets, no sex-related biases are present. This cohort will be used to validate different harmonization techniques in datasets that involves confounding between age and site.

2.1.3. TDC cohort

We pool all TDC data from our datasets, resulting in a total of 1194 participants (623 males, 571 females). The age of participants ranges from 6.17 to 21.00 years, with a mean age of 14.41 years and a standard deviation of 3.53 years. This cohort represents a more generic case in which age and sex are unbalanced with respect to site, potentially introducing statistical confounding. It will be used to evaluate the robustness of harmonization methods in the presence of such confounding factors. Additionally, the cohort will be applied in brain age prediction to assess whether harmonization approaches can effectively remove site-related variations and enhance the transferability of machine learning-based prediction models to new datasets.

2.1.4. ASD cohort

All datasets, except for the PNC site, also contain neuroimaging data from children with autism. We pool the ASD data from these five datasets to create a cohort of 309 participants (267 males, 42 females) aged from 6.05 to 19.50 years, with a mean age of 11.49 years and a standard deviation of 3.05 years. The structural networks in the ASD cohort will be harmonized by adjusting for site effects, which are estimated from the corresponding TDC cohort. This cohort will be used to evaluate the effectiveness of harmonization methods in removing site-related variations while preserving biological and diagnostic features, with the goal of improving the statistical power of analyses through data integration.

2.2. Image preprocessing

The dMRI data underwent a series of preprocessing steps, including denoising using joint local principal component analysis (PCA), correction for eddy currents and movements, and skull stripping with BET2. The constrained

spherical deconvolution (CSD) of dMRI is then conducted using DIPY, with the fiber response function estimated using dhollander algorithm [35, 36] and the fiber orientation distribution (FOD) fitted using the method proposed by Jeurissen and others [37]. Probabilistic fiber tracking is performed using the iFOD2 algorithm and anatomically constrained tractography (ACT) [38], with 500 seeds placed at random inside each voxel of the grey-matter white-matter interface (GMWMI). The high-resolution T1-w images are first preprocessed using FreeSurfer recon-all pipeline (http://surfer.nmr.mgh.harvard.edu) [39], followed by registration to the dMRI data of each subject using restricted deformable SyN algorithm in ANTs [40]. Quality control is conducted manually, and the ones with poor quality are excluded from the analysis.

2.3. Construction of structural networks

We parcellate the brain into 86 regions of interest (ROIs) using the Desikan-Killiany atlas [41] and calculate the number of streamlines intersecting each pairwise combination of these ROIs, yielding an 86x86 connectivity matrix. Post-processing of the structural networks involves removing self-connections and symmetrizing the connectivity matrix. Additionally, we normalize the structural networks by dividing edge weights by each subject's total WMGMI volume. Since structural networks derived from probabilistic tractography may include numerous spurious connections that could affect downstream analysis, we apply consistency-based thresholding to filter out spurious connections and retain only the most consistent ones. For our primary analyses, the consistency-based thresholding at 40% density has been applied to retain edges that are highly consistent across the population. Results using other thresholding levels are reported in the Supplementary Materials. All harmonization methods are performed with sex, age, and age^2 included as covariates.

2.4. Harmonization methods

We propose to use and adapt five statistical approaches for the harmonization of structural networks at the level of connectivity matrices. Table 2 presents a comparison of these harmonization methods. Consider structural network data gathered from M imaging sites, where $m = 1, 2, \dots, M$ denotes the index of the site, and each site comprises N_m samples, with $n = 1, 2, \dots, N_m$ representing the index of the subject. Let Y_{mne} represent the observed edge value before harmonization at a given edge index $e \in E(G)$ within the structural network of subject n from site m. In the following sections, we provide detailed descriptions of each technique and explain their implementation in the context of structural network data.

Table 2 Comparison of mainstream harmonization methodologies.

Harmonization method	Addresses mean site effects	Addresses variance site effects	Addresses covariance site effects	Address skewness	Assure non-negative edge values	Assumption for data distribution
ComBat	Yes	Yes				Normal
CovBat	Yes	Yes	Yes			Normal
log-ComBat	Yes	Yes		Yes	Yes	Log-normal
log-CovBat	Yes	Yes	Yes	Yes	Yes	Log-normal
gamma-GLM	Yes			Yes	Yes	Gamma

2.4.1. ComBat

The ComBat model, originally developed by Johnson et al. [24] to mitigate batch effects when integrating multiple microarray datasets for gene expression analysis, has been adapted for harmonizing multiple scalar metrics in neuroimaging studies [26, 27, 42-45]. ComBat formulates each edge value as:

$$Y_{mne} = \alpha_e + \boldsymbol{X_{mn}^T \beta_e} + \gamma_{me} + \delta_{me}\epsilon_{mne}$$

In this model, α_e represents intercept of the connectivity value at edge e, the vector $\boldsymbol{X_{mn}}$ incorporates covariates of interest, such as age and sex, with $\boldsymbol{\beta_e}$ representing the vector of regression coefficients corresponding to $\boldsymbol{X_{mn}}$ at edge e. γ_{me} and δ_{me} denote the mean and variance of the site effect for site m at edge e. Combat assumes that the error term ϵ_{mne} follows an independent normal distribution $\epsilon_{mne} \sim N(0, \sigma_e^2)$. It first estimates $\hat{\alpha}_e$, $\hat{\beta}_e$ using ordinary least-squares by constraining $\Sigma_m N_m \hat{\gamma}_{me} = 0$ for all edges. To estimate empirical statistical distributions of γ_{me} and δ_{me}, ComBat employs an empirical Bayes approach with the assumption that all edges share a common yet site-

specific prior distribution, with hyperparameters estimated empirically using method of moments. This allows information from all edges to be leveraged to infer the statistical properties of the site effect. Specifically, γ_{me} is assumed to have an independent normal prior distribution, and δ_{me} follows an independent inverse gamma prior distribution. The estimates γ_{me}^* and δ_{me}^* are obtained by computing the conditional posterior means. Finally, ComBat adjusts for the estimated site effects to obtain the harmonized connectivity edge:

$$ComBat(Y_{mne}) = \frac{Y_{mne} - \hat{\alpha}_e - \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}} - \gamma_{me}^*}{\delta_{me}^*} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}}$$

ComBat harmonization may introduce spurious connections for those originally with 0 edge values (indicating no connection between two nodes), altering the physical interpretation of the connectivity. To correct this, we follow the approach of Onicas et al. [46] by reassigning zeros to those connections. Specifically, we apply a binary connectivity matrix derived pre-harmonization to the harmonized weighted connectivity matrix for each subject.

2.4.2. CovBat

ComBat treats each edge in the structural networks independently, without accounting for the covariance between edges. However, in structural networks, site effects might be edgewise correlated and exhibit varying covariances across sites. To address these potential covariance effects, Chen et al. [30] adapted the ComBat framework into CovBat. CovBat posits that the error vector $\boldsymbol{\epsilon_{mn}} = (\epsilon_{mn1}, \epsilon_{mn2}, \dots, \epsilon_{mne}, \dots, \epsilon_{mnE})^T \sim N(\mathbf{0}, \boldsymbol{\Sigma_m})$, where E denotes total number of edges. CovBat initially applies ComBat to remove the mean and variance of the sites and obtain ComBat-adjusted residuals:

$$\epsilon_{mne}^{ComBat} = \frac{Y_{mne} - \hat{\alpha}_e - \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}} - \gamma_{me}^*}{\delta_{me}^*}$$

It then performs principal component analysis (PCA) on the residuals. These residuals can be projected onto the principal components, expressed as:

$$\boldsymbol{\epsilon_{mn}^{ComBat}} = \sum_{l=1}^{L} \xi_{mnl} \boldsymbol{\Phi_l}$$

Here, $\boldsymbol{\Phi_l}$ represent the eigenvectors of the covariance of the residual data, and $L = \min\left(\sum_{m=1}^{M} N_m,\ E\right)$ denotes the rank. ξ_{mnl} indicates the principal component scores. Assuming that the covariance site effect is captured within these principal component scores, we can harmonize them to mitigate the covariance site effects. Suppose:

$$\xi_{mnl} = \mu_{ml} + \rho_{ml}\epsilon_{mnl}$$

Where μ_{ml} and ρ_{ml} are location and scale shifts on principal component scores for site m at principal component l. The CovBat adjusts these site-related shifts in principal component scores:

$$\xi_{mnl}^{CovBat} = (\xi_{mnl} - \widehat{\mu_{ml}})/\widehat{\rho_{ml}}$$

The covariance-adjusted residuals can be obtained by projecting the adjusted principal component scores back into the original space, with a hyperparameter K controlling the desired portion of principal component scores to be adjusted:

$$\boldsymbol{e_{mn}^{CovBat}} = \sum_{k=1}^{K} \xi_{mnk}^{CovBat} \boldsymbol{\Phi_k} + \sum_{l=K+1}^{L} \xi_{mnl}^{CovBat} \boldsymbol{\Phi_l}$$

In this study, we select K such that the principal components explain 95% of the variation. The final harmonized connectivity edge values can be represented by replacing the residuals:

$$CovBat(Y_{mne}) = e_{mne}^{CovBat} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T}\widehat{\boldsymbol{\beta}}_{\boldsymbol{e}}$$

Similar to the post-processing in ComBat, we reassign zeros to connections that had zero edge values prior to harmonization.

2.4.3. log-ComBat

The ComBat algorithm relies on the assumption of data conforming to a normal distribution. However, this premise may not hold true for structural network data, where edge values frequently demonstrate notable skewness. In biomedical research, the log-scaling technique is commonly employed to transform skewed data to approximate a normal distribution [47, 48]. The log-scaling approach assumes that edgewise structural connectivity strength follows a log-normal distribution or closely approximate it. Consequently, the ComBat algorithm can be reformulated as follows:

$$\ln(Y_{mne}) = \alpha_e + \boldsymbol{X_{mn}^T \beta_e} + \gamma_{me} + \delta_{me}\epsilon_{mne}$$

and the harmonized connectivity values can be obtained from the inverse of a logarithm:

$$logComBat(Y_{mne}) = \exp\left(\frac{\ln(Y_{mne}) - \hat{\alpha}_e - \boldsymbol{X_{mn}^T \hat{\beta}_e} - \gamma_{me}^*}{\delta_{me}^*} + \hat{\alpha}_e + \boldsymbol{X_{mn}^T \hat{\beta}_e}\right)$$

log-ComBat ensures that harmonized connectivity values remain non-negative. To handle numerical errors with the logarithm when original connectivity values Y_{mne} are zero, we add a small positive constant c to Y_{mne}. Those connections with zero edge values prior to harmonization will be reassigned to zeros to ensure no spurious connections are introduced by the harmonization process

2.4.4. log-CovBat

We also adapt the CovBat algorithm by incorporating log-scaling to tackle the skewness of the edge weights in structural networks. First, we apply CovBat to log-transformed connectivity values, and then we apply the inverse of logarithm to obtain the final harmonized connectivity values:

$$logCovBat(Y_{mne}) = \exp\left(CovBat(\ln(Y_{mne}))\right)$$

log-CovBat ensures that connectivity edges remain non-negative after harmonization. Similarly, we add a small positive constant c to original connectivity value Y_{mne} during log-CovBat harmonization and restore connections that originally had zero edge values back to zeros.

2.4.5. gamma-GLM with log-link

The generalized linear model (GLM) provides an alternative for regression when the assumptions of normality and homoskedasticity are not met [49]. By employing a gamma distribution with a log-link function, GLM offers greater flexibility in modeling a wide range of data shapes, particularly well-suited for positive and right-skewed data. We model the effects of site and other covariates on expected structural networks with a log-link function, expressed as:

$$\ln\left(E_{mn}[Y_{mne}]\right) = \alpha_e + \boldsymbol{X_{mn}^T \beta_e} + \gamma_{me}$$

In the gamma-GLM, the edge values for each connection are assumed to follow a gamma distribution, $Y_{mne} \sim \Gamma(k_e, \theta_e)$, where k_e and θ_e denote the shape and scale parameters, respectively. The expected value for edge e is given by:

$$E_{mn}[Y_{mne}] = k_e\theta_e$$

The parameters k_e and θ_e can be fitted using GLM regression, enabling the estimation of the mean site effects γ_{me}^* for site m at edge e can be then estimated. To obtain the final harmonized connectivity values, the estimated mean site effects γ_{me}^* are regressed out, followed by applying the inverse logarithm transformation:

$$gammaGLM(Y_{mne}) = \exp(\ln(Y_{mne}) - \gamma_{me}^*) = Y_{mne}/\exp\left(\gamma_{me}^*\right)$$

The gamma-GLM with a log-link function ensures that the harmonized edge values remain non-negative. As gamma fitting also requires positive edge values, we add a small positive constant c to original connectivity value Y_{mne} and restore connections with zero edge values back to zeros after harmonization.

2.5. Test for distributional properties of structural network connections

To assess the validity of the distributional assumptions underlying different statistical harmonization methods, we conduct goodness-of-fit tests to evaluate the distributional characteristics of the structural network data. Specifically, we fit a normal distribution (as required by ComBat and CovBat), a log-normal distribution (for log-ComBat and log-CovBat), and a gamma distribution (for gamma-GLM) to the observed structural networks at each site. We employ the Kolmogorov–Smirnov (KS) test to quantify the distance between the observed data distribution and the fitted distributions.

2.6. Evaluation of harmonization methods

2.6.1. Tests on edgewise site effects

We first examine whether each harmonization method can eliminate site-related biases in the mean, variance, and covariance of edgewise connectivity strength. To evaluate mean site effects, we perform a two-sample t-test to assess edgewise site differences for the paired PNC-CAR cohort, with sex, age, and age^2 included as covariates. For the TDC cohort, where multiple sites are involved, we conduct a one-way analysis of variance (ANOVA) to test for mean site effects, followed by post-hoc analysis to identify specific site differences when the ANOVA results are significant. To evaluate variance site effects, we use Brown-Forsythe test on the residuals of edgewise connectivity strength, after regressing out sex, age, and age^2. Brown-Forsyth test is the non-parametric version of Levene's test to examine whether two comparing groups have equal variances. For site-specific covariance effects, we first calculate the empirical covariance for each site and then compute the Frobenius distance between the within-site covariance matrices for each pair of sites, both before and after harmonization. Additional Box-M test is conducted to evaluate whether the site-specific covariance matrices are equal.

2.6.2. Evaluation on derived graph topological measures

Since structural networks are represented as graphs with specific topological properties rather than merely edgewise connections, it is essential that harmonizing edgewise connectivity also ensures the derived graph topological measures to be harmonized. We calculate graph topological measures at two levels for the brain networks: node-level features (node strength, betweenness centrality, local efficiency, clustering coefficient) and global graph measures (global strength, intra- and inter-hemisphere strength, characteristic path length, global efficiency, modularity). We then examine site-related differences in these metrics before and after applying ComBat, CovBat, log-ComBat, log-CovBat, and gamma-GLM harmonization methods. We perform a two-sample t-test to assess site differences for each metric for the paired PNC-CAR cohort and use a one-way ANOVA test for the multi-site TDC cohort. Effect sizes are calculated using Cohen's d for the paired PNC-CAR cohort and Cohen's f for the multi-site TDC cohort.

2.7. Preservation of biological variability

An ideal harmonization method should eliminate non-biological variability introduced by different sites and scanners while retaining the statistical power to detect biologically meaningful associations To validate this, we calculate the Pearson correlation between edgewise connectivity strength and subjects' chronological age within each site, controlling for sex covariates. We expect that the harmonized structural networks to recover the age-connectivity relationships observed in the original data. The restoration is quantified by the true positive rate (sensitivity) of detected connections exhibiting significant age correlations. Additionally, we rank the connections by the absolute values of their age correlations and compute the overlap between the top k correlations in the harmonized and original data across all values of k, visualizing the results using concordance at the top (CAT) curves [50]. A CAT curve closer to one indicates better overlap between the two sets of correlation relationships.

2.8. Harmonization with confounding factors

Confounding between age and site poses a major challenge for harmonization, as removing site-related variation can inadvertently change age-related variation if not handled carefully. To assess different harmonization methods in the

presence of such confounding, we perform a Pearson's correlation analysis, testing the association between edgewise connections and subjects' chronological age within each site for the confounded NYU-TCD cohort. CAT plots are used to visualize the replicability of these age associations before and after applying different harmonization methods, and the true positive rate is employed to quantify the recovery of connections with significant age correlations. Additionally, we compute the Frobenius distance between the edgewise age correlation matrices of the two sites. We expect this distance to remain similar before and after harmonization, indicating that age-related site differences are preserved. Lastly, we jointly test the replicability of age associations with two sites combined after harmonization. We also include the scenario where the two sites are combined without proper harmonization. By comparing the results from the paired PNC-CAR cohort and the confounded NYU-TCD cohort, we highlight the critical role of harmonization in addressing confounding factors for data integration.

2.9. Application of harmonization for multi-site analysis

To further demonstrate the importance of harmonization in multi-site data analysis, we assess the effectiveness of these harmonization methods through two practical use cases: 1) improving the transferability of prediction models to new datasets. We use brain age prediction as an example. 2) enhancing statistical power for clinical studies through data integration. Here we focus on detecting group differences and clinical associations in ASD.

2.9.1. Brain age prediction on new datasets

Brain development is accompanied by structural changes in brain networks, and the gap between predicted brain age from neuroimaging data and chronological age has been linked to various health conditions [51-55]. While many researchers have used machine learning to model neurotypical brain development trajectories and predict brain age, these predictors often struggle with transferability across datasets due to site and scanner variability. We use brain age prediction as a case study to assess whether harmonization methods can eliminate site-related effects while preserving age-related biological variability in structural networks, thereby improving the prediction performance of machine learning models on new datasets.

We train a support vector regression (SVR) model to predict brain age from structural networks, using data from typically developing children at the PNC site. Features are extracted from the upper triangular portion of the structural connectivity matrices, with PCA applied for dimensionality reduction. We randomly split the PNC dataset into 50:50 for the training set and the testing set. To optimize the model, we apply 5-fold cross-validation on the training set to tune hyperparameters of the model. The final SVR model is trained on all subjects in the training set using the optimal hyperparameters. The performance of the model is reported using the testing set of the PNC dataset. We then assess the model's transferability by testing it on data from other sites in the TDC cohort, evaluating prediction performance through root mean square error (RMSE), mean absolute error (MAE), and Pearson correlation coefficient (R).

2.9.2. Enhancing statistical power for ASD studies

Previous research has identified several connectivity profiles in individuals with ASD [56-60]. However, most studies are conducted at single sites and often suffer from small sample sizes, which limit their statistical power and make it difficult to draw generalizable conclusions. Pooling data from multiple sources may introduce site-related biases which can obscure the detection of clinical associations. Thus, harmonizing structural networks becomes crucial for revealing reliable ASD-related characteristics with multi-site datasets.

We validate the effectiveness of harmonization algorithms in this context by estimating site-related effects using the TDC cohort and applying the resulting site-specific coefficients to both the TDC and ASD cohorts. We calculate global and nodewise graph topological features of structural networks before and after harmonization, assessing age-related differences in these features between the ASD and TDC groups. Due to pronounced sex differences in autistic children and the small sample size of autistic females, we exclude female subjects from this analysis to minimize potential sex bias. Additionally, to detect clinical associations for those graph topological features before and after harmonization, we perform correlation analyses to test their relationship to various clinical measures for ASD, including the Autism Diagnostic Observation Schedule (ADOS) [61] and the Social Responsiveness Scale (SRS-2) [62], while controlling for age and age^2.

3. Results

The results are organized as follows: Section 3.1 examines the distributional properties of structural network connections before harmonization, validating the data assumptions underlying different harmonization methods. Section 3.2 presents evidence of substantial mean, variance and covariance site effects in edgewise connectivity strength, and evaluates how effectively harmonization methods reduce site variability. Section 3.3 explores the site effects on derived graph measures of structural networks. Section 3.4 assesses the replicability of biological patterns such as age associations. Section 3.5 tests the robustness of harmonization methods in a more generic context involving confounding factors. Sections 3.6 and 3.7 highlight two practical applications of harmonization: enhance the transferability of machine learning-based brain age predictors to new datasets, and improve the statistical power of detection for clinical associations.

3.1. Distributional properties of structural network connections

For each pairwise connection in the structural connectivity matrices, we fit the edge values across the population to three hypothesized probability distributions: normal, log-normal, and gamma. The edge distributions are fitted separately for each site and diagnostic group (TDC or ASD). Figure 1A shows the KS distances between the observed distributions and three hypothesized probability distributions for structural network connections in the PNC dataset, our largest site for the TDC group. A larger KS distance indicates that the observed data deviate more from the hypothesized distribution. The gamma distribution demonstrates the best fit, with significantly lower KS distances compared to the other two hypothesized distributions ($p<0.0001$, paired t-test), followed by the log-normal distribution. Figure 1B illustrates the heatmaps of KS distances for each connection. For each hypothesized model, only connections with significant differences from the observed distributions are shown in the heatmaps ($p<0.05$, FDR-adjusted). Among the top 40% of highly consistent connections (1480 edges in total), 97% do not follow a normal distribution, while 31% have distributions that are significantly different from the log-normal distribution, and fewer than 5% of these connections show significant differences from the gamma distribution fitting.

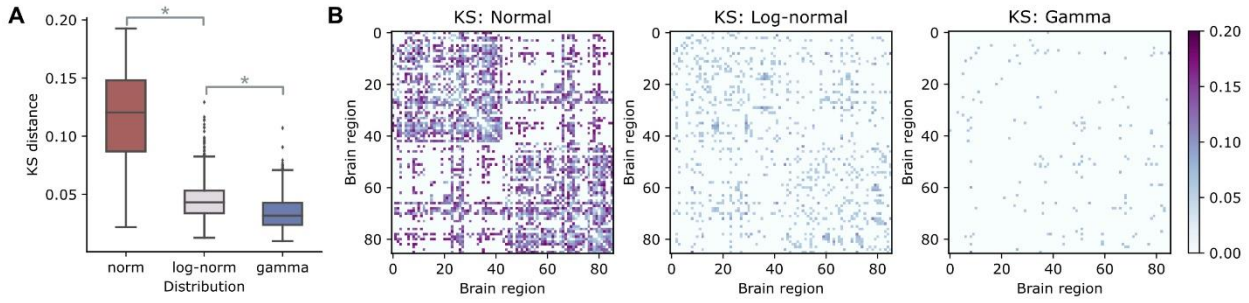

Figure 1 Goodness of fit test for the PNC site. A) Edgewise KS distances between the observed distributions of structural connectivity strengths and three hypothesized distributions (normal, log-normal, and gamma). Asterisks denote significant group differences ($p < 0.0001$, paired t-test). The gamma distribution provides a significantly better fit than the other two hypothesized models. B) Heatmaps of edgewise KS distances for each hypothesized model, showing only connections with significant differences from the observed distributions in structural networks ($p < 0.05$, FDR-adjusted).

Similar distributional properties of structural network connections are also observed in the TDC group across the other five sites and hold true for the ASD group as well, as shown in Figure S1. Notably, the normality assumptions underlying ComBat and CovBat are not met for most structural network connections. These connections are more likely to follow a gamma distribution, regardless of site or diagnostic group, while the log-normal distribution may also provide a satisfactory fit in some cases.

3.2. Evaluation of site effects on edgewise connectivity strength

We utilize the paired PNC-CAR cohort to assess the site effects on edgewise connectivity strength. The MA-plot [63] is a common method for visualizing differences in measurements taken from two sites. In this plot, the

edgewise mean connectivity strength is calculated for each site, and the between-site differences in log-transformed means are displayed as a function of the average log-transformed site means. Figure 2A presents the MA-plots comparing the PNC and CAR sites for participants paired by age and sex. If the structural network data were free of site-related effects, the scatterplot would align along the horizontal line at zero, as indicated by the dashed lines. Pre-harmonization, we observe striking site differences in edgewise connectivity strength between the two sites. These site differences are largely reduced after applying any of those tested harmonization approaches.

Figure 2B illustrates site-related biases in the mean and variance of edgewise connectivity strength, with t-statistics shown for mean site effects (first row) and F* statistics from the Brown-Forsythe test for variance site effects (second row). The log-CovBat and gamma-GLM approaches are favorable in eliminating mean site effects, and log-ComBat and log-CovBat show superiority in harmonizing site effects in variance. Specifically, among the 1480 most consistent connections in the structural networks, 1035 (70%) show significant differences in site means before harmonization ($p<0.05$, two-sample t-test). After ComBat harmonization, this number drops to just 11 connections (0.74%), while logComBat reduces the number to 5 (0.34%). CovBat harmonization leaves only 3 connections (0.2%) with significant site differences in means, and no significant site differences are observed after log-CovBat and gamma-GLM harmonization. For variance site effects, 741 connections (50%) exhibit significant site differences in the unharmonized data ($p<0.05$, Brown-Forsythe test). After ComBat harmonization, 110 connections (7.4%) still show significant differences, with 104 (7%) remaining significant after CovBat, and 7 (0.47%) after gamma-GLM. Following log-ComBat and log-CovBat harmonization, all connections are free from site-related effects on variance.

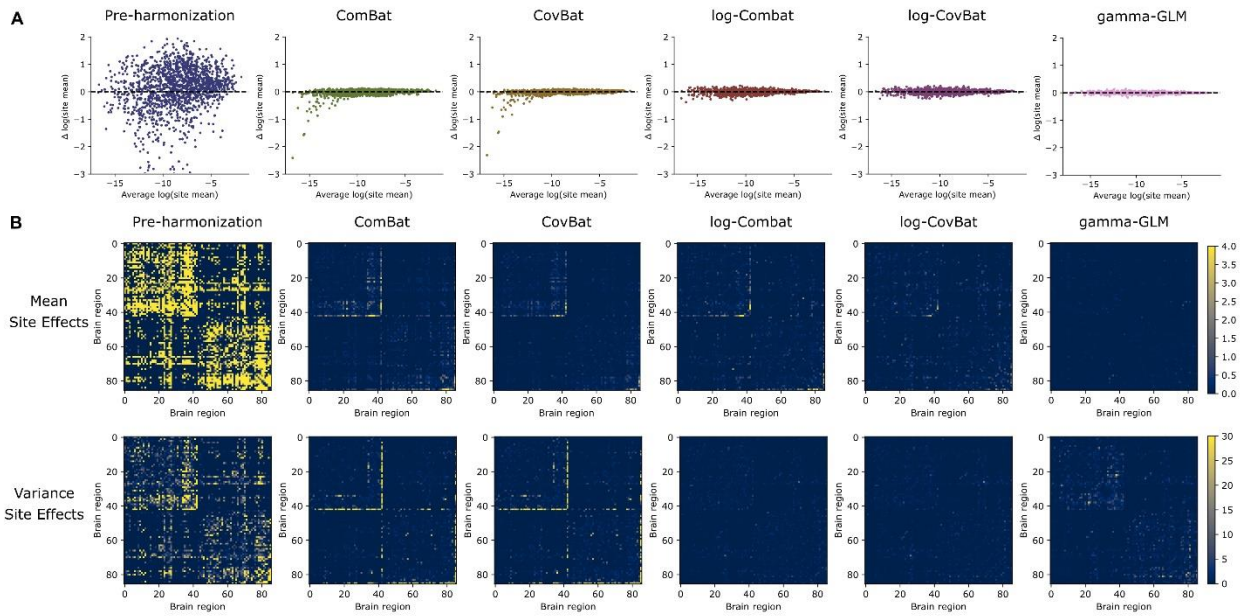

Figure 2 Site-related effects on mean and variance of edgewise connectivity strength. A) MA-plots for visualization of site differences between the PNC site and the CAR site with paired subjects. The x-axis represents the averaged log-transformed means across sites and the y-axis represents between-site differences in log-transformed means. The horizontal line at zero indicates no site-related effects. B) Edgewise site effects on mean (first row) and variance (second row) connectivity strength. The t-statistics are shown for mean site effects and F statistics from the Brown-Forsythe test for variance site effects.*

Table 3 shows the site-specific covariance effects on edgewise connectivity strength before and after harmonization, quantified by the Frobenius distance between covariance matrices of two sites. We regress out the effects of site demographics, including sex, age, and age^2, jointly across both sites using a linear model for the connectivity strength at each edge. The log-CovBat harmonization approach performs best in addressing site effects on covariance, followed by log-ComBat and gamma-GLM. However, the results of the Box's M test indicate that

covariance site effects on structural networks in the paired PNC-CAR cohort are not significant across all tested cases, including the pre-harmonization data.

Table 3 Frobenius distances between site-specific covariance matrices

Pairwise Site differences	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
Paired PNC-CAR	0.1022	0.0924	0.0912	0.0728	0.0726	0.0786

3.3. Evaluation of site effects on derived graph topological measures

We then examine the derived graph topological measures of structural networks to determine whether harmonizing edgewise connectivity strength can lead to harmonized downstream outcomes. Figure 3 illustrates the Cohen's d effect sizes of site differences across six global graph topological measures, with asterisks denoting significant site effects. All six global measures show significant site effects before harmonization ($p<0.05$, two-sample t-test), with three strength-based graph measures exhibiting large effect sizes (Cohen's $d > 1.0$), characteristic path length displaying a moderate effect size (Cohen's $d=0.68$), and global efficiency showing a smaller effect size (Cohen's $d=0.30$). After ComBat harmonization of edgewise connections, site effects on four global measures are eliminated. However, site effects on characteristic path length and global efficiency remain significant ($p<0.05$, two-sample t-test), though their effect sizes are reduced to small (Cohen's $d = 0.20$ for characteristic path length, Cohen's $d = 0.27$ for global efficiency). Among CovBat, log-ComBat, log-CovBat, and gamma-GLM, all of which address site effects on each global measure, gamma-GLM demonstrates the best overall harmonization performance, achieving the smallest effect sizes for most measures.

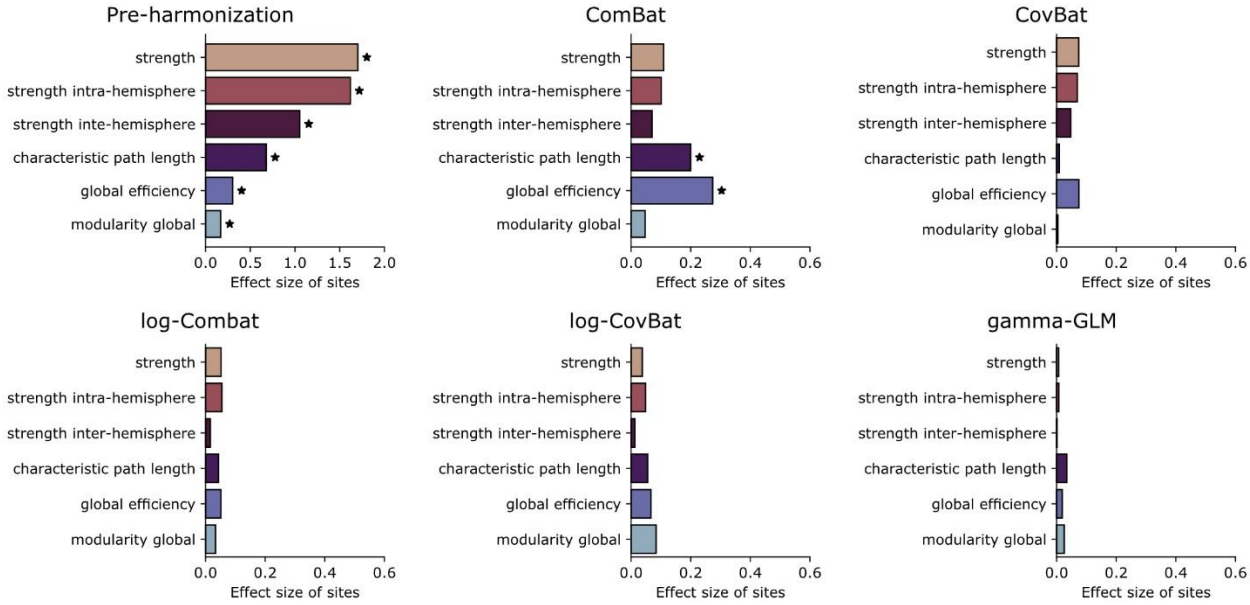

Figure 3 Effect size of site differences on global graph topological measures. For each harmonization method, the Cohen's d effect sizes of sites are evaluated on six global graph topological measures (global strength, intra- and inter-hemisphere strength, characteristic path length, global efficiency, modularity). Significant site effects are indicated by asterisks ($p < 0.05$, two-sample t-test).

We then explore the site effects on nodewise graph topological features. Figure 4 presents the effect sizes of sites on four nodewise graph topological measures, with significant site effects marked by asterisks. The number of nodes with significant site effects is indicated in the top left corner of each plot. While all tested harmonization approaches address most site effects on node strength and clustering coefficient, ComBat and CovBat fail to correct site effects

on betweenness centrality and local efficiency, leaving a substantial number of nodes with significant site effects ($p<0.05$, two-sample t-test).

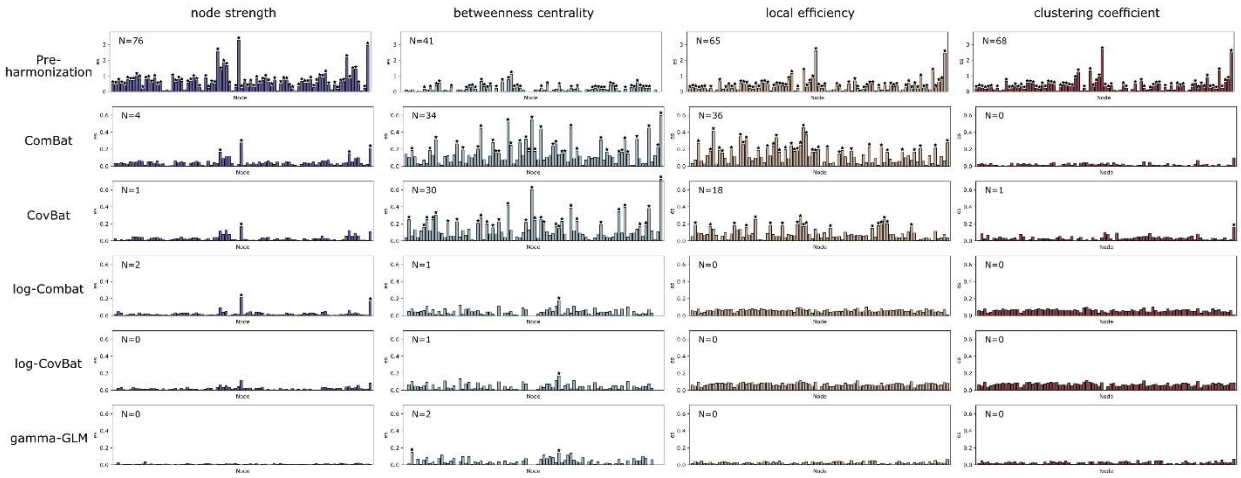

Figure 4 Effect size of site differences on nodal graph topological measures. Four nodewise topological measures (node strength, betweenness centrality, local efficiency and clustering coefficient) are evaluated. Significant site effects ($p < 0.05$, two-sample t-test) are indicated by asterisks. The number of nodes with significant site effects is noted in the top left corner of each plot.

3.4. Preservation of biological variability

We then assess the replicability of age-related biological variability after applying different harmonization methods. For each site, Pearson correlations between edgewise connectivity strength and subjects' chronological age are calculated before and after harmonization, controlling for sex as a covariate. Figure 5A shows the edgewise correlations for the PNC site (first row) and the CAR site (second row) before and after harmonization with paired subjects, displaying only connections with significant age associations ($p < 0.05$, Pearson's R). Notably, the gamma-GLM method most closely resembles the pre-harmonization patterns of age associations. Table 4 quantifies the replicability of these age associations using the true positive rate (TPR) of connections with significant correlations, compared to the pre-harmonization data. Our results show that gamma-GLM outperforms other methods by recovering all the positive and negative correlations in the original structural networks across both sites in the paired PNC-CAR cohort, followed by log-ComBat and log-CovBat.

We visualize the replicability of age associations with CAT plots, where each curve represents the concordance of rankings for all edgewise age correlations before and after harmonization, again controlling for sex (Figure 5A). In the absence of confounding, gamma-GLM performs the best at both PNC and CAR sites, showing an almost flat CAT curve near 1. The log-ComBat and log-CovBat methods exhibit good overlap, performing slightly behind gamma-GLM, while ComBat and CovBat rank last.

Table 4 Recovery of connections with significant age correlation

Cohort	Site	# positive correlations	TPR of positive correlations					# negative correlations	TPR of negative correlations				
		pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM	pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
Paired PNC-CAR	PNC	228	94.30%	94.30%	99.56%	95.61%	100%	143	82.51%	82.51%	98.60%	97.90%	100%
	CAR	227	72.69%	74.89%	94.27%	92.95%	100%	253	90.12%	92.09%	92.49%	92.09%	100%
Confounded NYU-TCD	ABIDEII-NYU	19	84.21%	78.95%	94.74%	84.21%	100%	78	80.76%	78.21%	82.05%	73.08%	100%
	ABIDEII-TCD	36	86.11%	58.33%	94.44%	66.67%	100%	95	84.21%	70.53%	85.26%	80.00%	100%

Abbreviations: TPR, true positive rate; # positive (or negative) correlations, number of connections with positive (or negative) age correlations

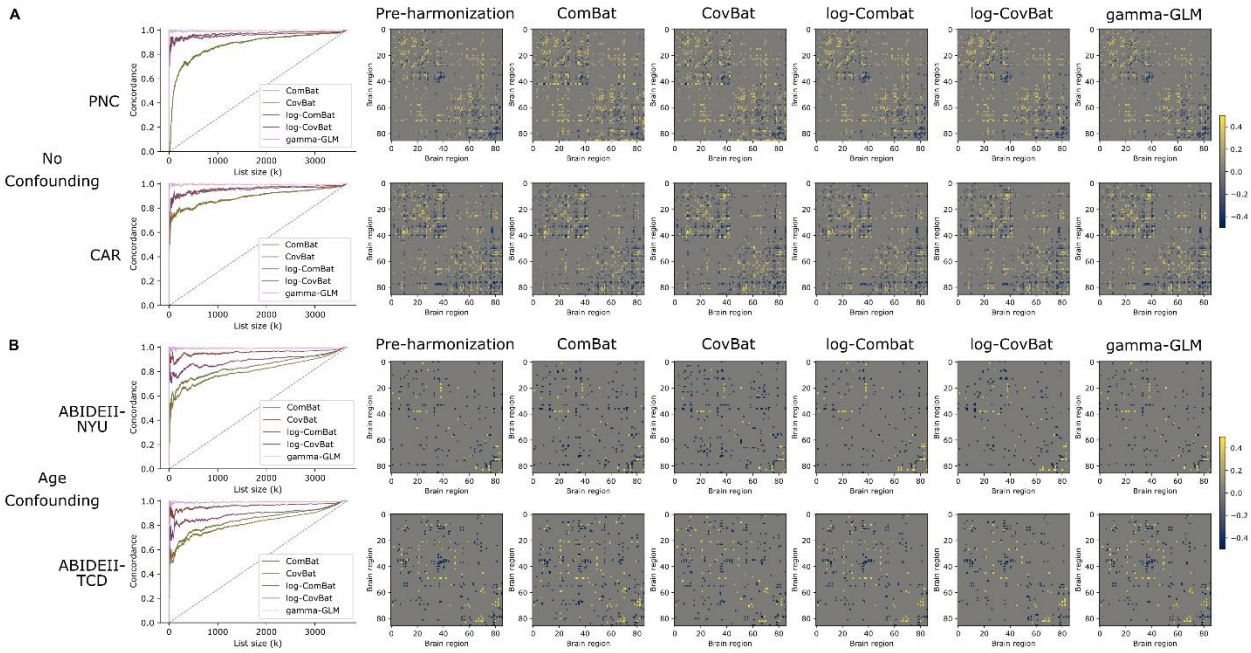

Figure 5 Replicability of age associations with edgewise connectivity after applying different harmonization methods. Two scenarios are tested: (A) the paired PNC-CAR cohort and (B) the confounded NYU-TCD cohort. For each scenario, edgewise correlations between connectivity strength and age are shown for each site and harmonization method, displaying only significant age-associated connections ($p < 0.05$, Pearson's R). The CAT curves visualize the concordance of edgewise age associations before and after harmonization for each method. A CAT curve closer to one indicates better overlap in age associations.

3.5. Harmonization with confounding factors

The previous results are obtained by harmonizing two sites that paired for age and sex to minimize potential confounding of those variables across sites. However, this kind of pairing is not always feasible in multi-site studies. In the confounded NYU-TCD cohort where there is a poor overlap between age across sites, enough paired samples can be used to estimate site-related effects. A better harmonization method should be applied on all available data across sites without the pairing process and should be robust to confounding. Figure 5B shows the edgewise within-site correlations with age for ABIDEII-NYU (first row) and ABIDEII-TCD (second row) before and after harmonization. We note that, compared to the paired cohort, the confounded cohort demonstrates different edgewise patterns between two sites, which may be due to the effects of age on structural networks are different across different age groups. The CAT plots show that the site-specific age-related patterns are best preserved for each site after using gamma-GLM harmonization, followed by log-ComBat, log-CovBat, ComBat, and CovBat. The gamma-GLM harmonization can recover all connections with significant positive and negative correlations in presence of confounding between age and sites (Table 4).

To assess the age-effects between sites, we calculate the Frobenius distances between the edgewise correlation patterns of two sites, as shown in Table 5. Ideally, we expect the between-site distance for age-related patterns should remain the same before and after harmonization. The gamma-GLM most successfully perseveres the age-related differences across sites, while log-ComBat and log-CovBat inadvertently reduce these differences. The ComBat and CovBat increase the age-related differences across sites, which may be due to the poor fitting using the normal distribution for edges.

Table 5 Frobenius distances between site-specific edgewise patterns of age associations

Pairwise Site differences	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM

Confounded NYU-TCD	11.7748	13.6246	12.8325	11.3402	11.4245	11.7764

Lastly, we evaluate the replicability of age associations when combining two sites. Figure S2 shows the CAT curves for the paired PNC-CAR cohort and the confounded NYU-TCD cohort. Our results demonstrate that, in the presence of confounding, the replicability of age associations drops dramatically compared to the paired cohort if sites are combined without proper harmonization, whereas properly harmonized ones maintain high replicability.

3.6. Application 1: Brain age prediction on new datasets

In the following sessions, we evaluate the effectiveness of different harmonization methods in addressing several common challenges that encountered in multi-site structural network studies.

Many studies face limitations in the transferability of prediction models when applied to new datasets, as these datasets may exhibit substantial differences in data distribution compared to the training cohort. Harmonization is one approach to mitigate these data differences, which are often caused by scanner variations and other site-specific factors. For this analysis, we use the entire TDC cohort. The site effects before and after harmonization are reported in the Supplementary Materials.

We train a brain age predictor using an SVR model on the PNC site's training set and apply it to the remaining sites before and after different harmonization methods. Table 6 shows the model's prediction performance at each new site, as well as for the combined data from all new sites. For reference, the original performance of the trained brain age predictor on the PNC site's test set is provided (underlined). Our results demonstrate that using proper harmonization to remove site-specific variations between the PNC and new sites can largely improve the transferability of the brain age predictor to new datasets. The gamma-GLM-harmonized data achieve the lowest MAE and RMSE, along with the highest Pearson R correlation, across most new sites with unseen data and for the overall results. Although log-ComBat attains the lowest MAE and RMSE in two sites (TimRobert, ABIDEII-TCD), it does not match gamma-GLM's performance in terms of Pearson R correlation.

Table 6 Impact of harmonization on the transferability of brain age predictors to new datasets

Dataset	Metric	Pre-harmonization	ComBat	CovBat	log-ComBat	log-CovBat	gamma-GLM
PNC	MAE	1.7321	1.6752	1.6433	1.7078	1.7001	1.7281
	RMSE	2.1681	2.0683	2.0400	2.1396	2.1329	2.1651
	Pearson R	0.7405	0.7662	0.7734	0.7483	0.7504	0.7414
CAR	MAE	3.7216	2.4807	2.4579	1.9644	1.9739	**1.8536**
	RMSE	4.3141	3.1131	3.0585	2.3928	2.4044	**2.2631**
	Pearson R	0.0548	0.3260	0.3175	0.6400	0.6446	**0.7074**
TimRobert	MAE	5.0619	2.7072	2.5526	**1.9518**	2.0532	2.0260
	RMSE	5.4438	3.2332	3.0941	**2.2854**	2.4712	2.3330
	Pearson R	0.2194	0.2418	0.2436	0.5063	0.4355	**0.5426**
ABIDEII-NYU	MAE	5.5198	3.1685	3.1896	1.8013	2.0526	**1.7870**
	RMSE	5.7817	3.7338	3.9420	2.2315	2.5171	**2.2105**
	Pearson R	-0.5111	-0.2693	-0.3182	0.4385	0.3965	**0.4676**
ABIDEII-SDSU	MAE	2.8504	3.4530	3.4925	2.5960	2.6739	**2.3822**
	RMSE	3.3662	4.1185	4.1908	3.1410	3.2959	**2.9947**
	Pearson R	0.2313	-0.4330	-0.4747	0.2078	0.1278	**0.3514**
ABIDEII-TCD	MAE	2.5921	2.5443	2.9942	**2.0126**	2.1363	2.0621
	RMSE	3.0745	3.2309	3.6377	**2.3909**	2.5667	2.4667
	Pearson R	0.5522	0.0682	-0.1160	0.6647	0.5909	**0.6804**
All New Sites	MAE	3.9080	2.6690	2.6702	2.0137	2.0719	**1.9419**
	RMSE	4.4883	3.3017	3.3100	2.4460	2.5346	**2.3653**
	Pearson R	0.2027	0.3935	0.3433	0.6790	0.6587	**0.7201**

[1]The performance of brain age predictors is reported on the testing set of the PNC site (underlined). The brain age predictors are directly applied to other sites to evaluate the transferability of the model to new datasets.

[2]The best-performing harmonization method for each metric is highlighted in bold.

[3]Abbreviations: MAE, mean absolute error; RMSE, root mean square error; Pearson R, Pearson's correlation.

3.7. Application 2: Enhancing statistical power for ASD studies

4. Discussion

This study addresses a key challenge in multi-site structural network research: how to effectively harmonize data collected from various sources to enable integration of multi-site studies and retrospective use of existing datasets. By doing so, it seeks to expand sample sizes and provide a more comprehensive representation of both neurotypical and disordered populations, supporting more robust findings in the field. While previous studies have proposed various harmonization approaches at the level of dMRI signals [19, 22], they fail to remove site effects at the level of structural networks [23], introducing biases in downstream network-level analyses. Therefore, we believe these site effects should be specifically addressed and evaluated at the network level. To the best of our knowledge, this is the first effort to develop harmonization frameworks and provide comprehensive evaluations specifically tailored to structural networks.

Structural networks have distinct distributional properties in edgewise connectivity strength, requiring careful consideration when modeling and removing site-related effects. Many commonly used statistical harmonization approaches in neuroimaging, such as ComBat, CovBat, or simple linear regression, assume that the data follow a normal distribution. However, most connections in structural networks do not follow a normal distribution but rather a gamma distribution (Section 3.1). Some researchers have applied harmonization methods without validating the underlying data assumptions [46], which can lead to unintended consequences. For instance, ComBat and CovBat may produce negative values, which lack physical meaning in structural networks where edges represent (normalized) numbers of streamlines connecting brain regions. Moreover, inappropriate use of harmonization methods that do not align with the correct data distribution may introduce additional biases. We demonstrate that applying harmonization techniques based on normal or log-normal distributions to structural network data may fail to remove mean site effects at some connections (Section 3.2), potentially resulting in spurious findings in downstream analyses.

Importantly, structural networks are graph-structured, so harmonization should not only remove edgewise site effects but also eliminate site-related variations in the graph properties, commonly measured by graph topological measures [31]. Prior work only evaluated graph measures at the global level [27, 46], but our results indicate that structural networks showing no site differences in global measures may still exhibit site differences at the nodal level. For example, while CovBat effectively removes site effects in all tested global graph measures, significant site effects persist in nodal measures such as betweenness centrality and local efficiency (Section 3.3). Therefore, we argue that the effectiveness of harmonization for structural networks should be evaluated at all three levels—edgewise, nodal, and global graph properties—to ensure that site effects are fully eliminated. Our results show that gamma-GLM performs best across all three levels, followed by log-CovBat and log-ComBat.

Since the major goal of harmonization is to reveal true biological variability through data integration, it is crucial that biologically relevant properties are not inadvertently altered or removed. To assess this, we test the replicability of biological patterns, such as edgewise associations with age, pre- and post-harmonization, as age-related connectivity changes are critical for many studies investigating neurodevelopment or neurodegenerative processes [54, 64]. Our findings demonstrate that gamma-GLM is the most successful in preserving these biologically meaningful associations across sites and remains robust even when age is confounded with site effects (Section 3.4 and Section 3.5).

Our comprehensive evaluation offers practical guidance for researchers looking to harmonize structural network data. Overall, gamma-GLM is our best recommendation, as it outperforms other models in removing site effects in edgewise mean connectivity strength, global and nodal graph measures, and in preserving age-related biological variability, even when confounded with site effects. We also examine variance and covariance site effects (Section 3.2). Our findings indicate that in cases where substantial variance-related site effects persist, log-ComBat or log-CovBat may be more effective at correcting these biases. Although we did not observe significant covariance site

effects in whole-brain analyses, these effects may emerge in subnetwork analyses or when structural networks are modulated by diffusion metrics like fractional anisotropy (FA) or mean diffusivity (MD), where log-CovBat could be more appropriate.

We further demonstrate the practical utility of harmonizing structural networks in two critical applications. One commonly reported issue in collaborative research is that predictive models developed with one dataset often fail to predict unseen data from the new sites [65, 66]. This occurs because these datasets may exhibit substantial differences in scanners, acquisition protocols, and other site-related effects compared to the training cohort. We create machine learning models to predict brain age using structural networks from one site and evaluate the transferability of these predictors on other sites. Our findings indicate that harmonizing structural networks across sites, particularly using gamma-GLM, enhances the transferability of machine learning models, making them more robust to unseen data at new sites (Section 3.6). This approach can further be adapted for a federated learning setup in healthcare [67, 68]. In this context, researchers can first harmonize and standardize decentralized data sources to a central site to ensure comparability, thereby improving the transferability of models when deployed from and to the central site.

Another key application of harmonization is to integrate structural network data from multiple sources, thereby increasing the statistical power of clinical studies. This is especially beneficial for patient studies, where data from a single site is often limited. By pooling patient data from various sources, researchers can access larger and more representative samples, which are essential for detecting subtle differences in structural network patterns associated with different conditions. Therefore, the effectiveness of harmonization should be evaluated not only for control groups but also for patient populations. We illustrate the effectiveness of harmonization methods using the ASD cohort as an example. Our results show that without proper harmonization, the associations between graph topological measures and autistic traits may be obscured by site-related effects. Structural network data after harmonization successfully uncovers the biological features related to disorders in structural networks and enhances statistical power compared to single-site studies (Section 3.7).

While we have demonstrated the effectiveness of our harmonization framework, there are several limitations that warrant attention in future research. For instance, we test only a few distributions from the exponential family and focus on modeling mean, variance, and covariance site effects. It would be valuable to explore additional distributions and statistical metrics, particularly when structural networks are created using more complex methods or modulated by diffusion metrics such as FA and MD. In such cases, more flexible distributional regression models, like the Generalized Additive Model for Location, Scale, and Shape (GAMLSS) [69], might be suitable. Another area for future work lies in expanding demographic variables included during harmonization. Incorporating factors like ethnicity and socioeconomic status could enhance the applicability of findings across diverse populations to promote health equity [70, 71]. Our harmonization and evaluation framework can also be adapted to better address the specific goals of structural network studies in these contexts.

References

1. Button, K.S., et al., *Power failure: why small sample size undermines the reliability of neuroscience.* Nature reviews neuroscience, 2013. **14**(5): p. 365-376.
2. Smith, S.M. and T.E. Nichols, *Statistical challenges in "big data" human neuroimaging.* Neuron, 2018. **97**(2): p. 263-268.
3. Ioannidis, J.P., *Why most published research findings are false.* PLoS medicine, 2005. **2**(8): p. e124.
4. Wheeler, A.L. and A.N. Voineskos, *A review of structural neuroimaging in schizophrenia: from connectivity to connectomics.* Frontiers in human neuroscience, 2014. **8**: p. 653.
5. Picci, G., S.J. Gotts, and K.S. Scherf, *A theoretical rut: revisiting and critically evaluating the generalized under/over-connectivity hypothesis of autism.* Developmental science, 2016. **19**(4): p. 524-549.

6. Sudlow, C., et al., *UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.* PLoS medicine, 2015. **12**(3): p. e1001779.
7. Di Martino, A., et al., *Enhancing studies of the connectome in autism using the autism brain imaging data exchange II.* Scientific data, 2017. **4**(1): p. 1-15.
8. Casey, B.J., et al., *The adolescent brain cognitive development (ABCD) study: imaging acquisition across 21 sites.* Developmental cognitive neuroscience, 2018. **32**: p. 43-54.
9. Iturria-Medina, Y., et al., *Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory.* Neuroimage, 2008. **40**(3): p. 1064-1076.
10. Qi, S., et al., *The influence of construction methodology on structural brain network measures: A review.* Journal of neuroscience methods, 2015. **253**: p. 170-182.
11. Grech-Sollars, M., et al., *Multi-centre reproducibility of diffusion MRI parameters for clinical sequences in the brain.* NMR in Biomedicine, 2015. **28**(4): p. 468-485.
12. Zhu, T., et al., *Quantification of accuracy and precision of multi-center DTI measurements: a diffusion phantom and human brain study.* Neuroimage, 2011. **56**(3): p. 1398-1411.
13. Vollmar, C., et al., *Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners.* Neuroimage, 2010. **51**(4): p. 1384-1394.
14. Magnotta, V.A., et al., *Multicenter reliability of diffusion tensor imaging.* Brain connectivity, 2012. **2**(6): p. 345-355.
15. Jovicich, J., et al., *Multisite longitudinal reliability of tract-based spatial statistics in diffusion tensor imaging of healthy elderly subjects.* Neuroimage, 2014. **101**: p. 390-403.
16. Fornito, A., A. Zalesky, and M. Breakspear, *Graph analysis of the human connectome: promise, progress, and pitfalls.* Neuroimage, 2013. **80**: p. 426-444.
17. Cao, H., et al., *Toward leveraging human connectomic data in large consortia: generalizability of fMRI-based brain graphs across sites, sessions, and paradigms.* Cerebral Cortex, 2019. **29**(3): p. 1263-1279.
18. Abraham, A., et al., *Deriving reproducible biomarkers from multi-site resting-state data: An Autism-based example.* NeuroImage, 2017. **147**: p. 736-745.
19. Mirzaalian, H., et al., *Inter-site and inter-scanner diffusion MRI data harmonization.* NeuroImage, 2016. **135**: p. 311-323.
20. Koppers, S., C. Haarburger, and D. Merhof. *Diffusion MRI signal augmentation: from single shell to multi shell with deep learning*. in *Computational Diffusion MRI: MICCAI Workshop, Athens, Greece, October 2016 19*. 2017. Springer.
21. Koppers, S., et al. *Spherical harmonic residual network for diffusion signal harmonization*. in *Computational Diffusion MRI: International MICCAI Workshop, Granada, Spain, September 2018 22*. 2019. Springer.
22. Huynh, K.M., et al., *Multi-site harmonization of diffusion MRI data via method of moments.* IEEE transactions on medical imaging, 2019. **38**(7): p. 1599-1609.
23. Kurokawa, R., et al., *Cross-scanner reproducibility and harmonization of a diffusion MRI structural brain network: A traveling subject study of multi-b acquisition.* NeuroImage, 2021. **245**: p. 118675.
24. Johnson, W.E., C. Li, and A. Rabinovic, *Adjusting batch effects in microarray expression data using empirical Bayes methods.* Biostatistics, 2007. **8**(1): p. 118-127.
25. Morris, C.N., *Parametric empirical Bayes inference: theory and applications.* Journal of the American statistical Association, 1983. **78**(381): p. 47-55.
26. Fortin, J.-P., et al., *Harmonization of multi-site diffusion tensor imaging data.* Neuroimage, 2017. **161**: p. 149-170.
27. Yu, M., et al., *Statistical harmonization corrects site effects in functional connectivity measurements from multi-site fMRI data.* Human brain mapping, 2018. **39**(11): p. 4213-4227.

28. Yamashita, A., et al., *Harmonization of resting-state functional MRI data across multiple imaging sites via the separation of site differences into sampling bias and measurement bias.* PLoS biology, 2019. **17**(4): p. e3000042.
29. Ma, D., et al., *Quantitative assessment of field strength, total intracranial volume, sex, and age effects on the goodness of harmonization for volumetric analysis on the ADNI database.* Human brain mapping, 2019. **40**(5): p. 1507-1527.
30. Chen, A.A., et al., *Mitigating site effects in covariance for machine learning in neuroimaging data.* Human brain mapping, 2022. **43**(4): p. 1179-1195.
31. Rubinov, M. and O. Sporns, *Complex network measures of brain connectivity: uses and interpretations.* Neuroimage, 2010. **52**(3): p. 1059-1069.
32. Satterthwaite, T.D., et al., *Neuroimaging of the Philadelphia Neurodevelopmental Cohort.* NeuroImage, 2014. **86**: p. 544-553.
33. Tunç, B., et al., *Deviation from normative brain development is associated with symptom severity in autism spectrum disorder.* Molecular autism, 2019. **10**: p. 1-14.
34. Ghanbari, Y., et al., *Identifying group discriminative and age regressive sub-networks from DTI-based connectivity via a unified framework of non-negative matrix factorization and graph embedding.* Medical Image Analysis, 2014. **18**(8): p. 1337-1348.
35. Dhollander, T., et al. *Improved white matter response function estimation for 3-tissue constrained spherical deconvolution*. in *Proc. Intl. Soc. Mag. Reson. Med*. 2019.
36. Dhollander, T., D. Raffelt, and A. Connelly. *Unsupervised 3-tissue response function estimation from single-shell or multi-shell diffusion MR data without a co-registered T1 image*. in *ISMRM workshop on breaking the barriers of diffusion MRI*. 2016. Lisbon, Portugal.
37. Jeurissen, B., et al., *Investigating the prevalence of complex fiber configurations in white matter tissue with diffusion magnetic resonance imaging.* Human brain mapping, 2013. **34**(11): p. 2747-2766.
38. Tournier, J.D., F. Calamante, and A. Connelly. *Improved probabilistic streamlines tractography by 2nd order integration over fibre orientation distributions*. in *Proceedings of the international society for magnetic resonance in medicine*. 2010. John Wiley & Sons, Inc, New Jersey.
39. Fischl, B., *FreeSurfer.* Neuroimage, 2012. **62**(2): p. 774-781.
40. Avants, B.B., et al., *Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain.* Medical image analysis, 2008. **12**(1): p. 26-41.
41. Desikan, R.S., et al., *An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest.* Neuroimage, 2006. **31**(3): p. 968-980.
42. Fortin, J.-P., et al., *Harmonization of cortical thickness measurements across scanners and sites.* Neuroimage, 2018. **167**: p. 104-120.
43. Radua, J., et al., *Increased power by harmonizing structural MRI site differences with the ComBat batch adjustment method in ENIGMA.* Neuroimage, 2020. **218**: p. 116956.
44. Orlhac, F., et al., *How can we combat multicenter variability in MR radiomics? Validation of a correction procedure.* European radiology, 2021. **31**: p. 2272-2280.
45. Wang, Y.-W., X. Chen, and C.-G. Yan, *Comprehensive evaluation of harmonization on functional brain imaging for multisite data-fusion.* NeuroImage, 2023. **274**: p. 120089.
46. Onicas, A.I., et al., *Multisite harmonization of structural DTI networks in children: An A-CAP study.* Frontiers in Neurology, 2022. **13**: p. 850642.
47. Manning, W.G. and J. Mullahy, *Estimating log models: to transform or not to transform?* Journal of health economics, 2001. **20**(4): p. 461-494.
48. Keene, O.N., *The log transformation is special.* Statistics in medicine, 1995. **14**(8): p. 811-819.

49. Nelder, J.A. and R.W. Wedderburn, *Generalized linear models.* Journal of the Royal Statistical Society Series A: Statistics in Society, 1972. **135**(3): p. 370-384.
50. Irizarry, R.A., et al., *Multiple-laboratory comparison of microarray platforms.* Nature methods, 2005. **2**(5): p. 345-350.
51. Franke, K. and C. Gaser, *Ten years of BrainAGE as a neuroimaging biomarker of brain aging: what insights have we gained?* Frontiers in neurology, 2019. **10**: p. 789.
52. Liem, F., et al., *Predicting brain-age from multimodal imaging data captures cognitive impairment.* Neuroimage, 2017. **148**: p. 179-188.
53. Dosenbach, N.U., et al., *Prediction of individual brain maturity using fMRI.* Science, 2010. **329**(5997): p. 1358-1361.
54. Bethlehem, R.A., et al., *Brain charts for the human lifespan.* Nature, 2022. **604**(7906): p. 525-533.
55. Kaufmann, T., et al., *Common brain disorders are associated with heritable patterns of apparent aging of the brain.* Nature neuroscience, 2019. **22**(10): p. 1617-1623.
56. Rudie, J.D., et al., *Altered functional and structural brain network organization in autism.* Neuroimage Clin, 2012. **2**: p. 79-94.
57. Rane, P., et al., *Connectivity in autism: a review of MRI connectivity studies.* Harvard review of psychiatry, 2015. **23**(4): p. 223-244.
58. Ha, S., et al., *Characteristics of brains in autism spectrum disorder: structure, function and connectivity across the lifespan.* Experimental neurobiology, 2015. **24**(4): p. 273.
59. Vissers, M.E., M.X. Cohen, and H.M. Geurts, *Brain connectivity and high functioning autism: a promising path of research that needs refined models, methodological convergence, and stronger behavioral links.* Neuroscience & Biobehavioral Reviews, 2012. **36**(1): p. 604-625.
60. Lewis, J.D., et al., *Network inefficiencies in autism spectrum disorder at 24 months.* Translational psychiatry, 2014. **4**(5): p. e388-e388.
61. Lord, C., et al., *The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism.* Journal of autism and developmental disorders, 2000. **30**: p. 205-223.
62. Constantino, J.N., *Social responsiveness scale*, in *Encyclopedia of autism spectrum disorders*. 2021, Springer. p. 4457-4467.
63. Dudoit, S., et al., *Statistical methods for identifying differentially expressed genes in replicated cDNA microarray experiments.* Statistica sinica, 2002: p. 111-139.
64. Hagmann, P., et al., *White matter maturation reshapes structural connectivity in the late developing human brain.* Proceedings of the National Academy of Sciences, 2010. **107**(44): p. 19067-19072.
65. Goetz, L., et al., *Generalization—a key challenge for responsible AI in patient-facing clinical applications.* npj Digital Medicine, 2024. **7**(1): p. 126.
66. Rajpurkar, P., et al., *AI in health and medicine.* Nature medicine, 2022. **28**(1): p. 31-38.
67. Antunes, R.S., et al., *Federated learning for healthcare: Systematic review and architecture proposal.* ACM Transactions on Intelligent Systems and Technology (TIST), 2022. **13**(4): p. 1-23.
68. Li, X., et al., *Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results.* Medical image analysis, 2020. **65**: p. 101765.
69. Rigby, R.A. and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape.* Journal of the Royal Statistical Society Series C: Applied Statistics, 2005. **54**(3): p. 507-554.
70. Norori, N., et al., *Addressing bias in big data and AI for health care: A call for open science.* Patterns, 2021. **2**(10).

71. Wang, R., P. Chaudhari, and C. Davatzikos, *Bias in machine learning models can be significantly mitigated by careful training: Evidence from neuroimaging studies.* Proceedings of the National Academy of Sciences, 2023. **120**(6): p. e2211613120.