Distribution Agreement

In presenting this thesis or dissertation as a partial fulfillment of the requirements for an advanced degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis or dissertation in whole or in part in all forms of media, now or hereafter known, including display on the world wide web. I understand that I may select some access restrictions as part of the online submission of this thesis or dissertation. I retain all ownership rights to the copyright of the thesis or dissertation. I also retain the right to use in future works (such as articles or books) all or part of this thesis or dissertation.

Signature:

_____ _____

Liangkang Wang Date

Compensating for selection bias when detecting altered functional connectivity in autism

By

Liangkang Wang
Master of Science in Public Health

Department of Biostatistics and Bioinformatics

Benjamin Risk, Ph.D.
(Thesis Advisor)

Raphiel Murden, Ph.D.
(Reader)

Ying Guo, Ph.D.
(Reader)

Date

Compensating for selection bias when detecting altered functional connectivity in autism

By

Liangkang Wang
B.S., Wuhan University, 2021

Advisor: Benjamin Risk, Ph.D.

An abstract of
A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2023

Abstract

Compensating for selection bias when detecting altered functional connectivity in autism

By Liangkang Wang

The exclusion of high-motion participants in functional Magnetic Resonance Imaging (fMRI) studies is a common practice to reduce motion-related artifacts. However, this exclusion can introduce biases by altering the distribution of clinically relevant variables, leading to a non-representative study sample. This paper aims to introduce a framework that employs the Average Inverse Probability Weighted Estimator (AIPWE) method to address these biases by treating excluded scans as missing data. Using simulated datasets, we tested the AIPWE method on single-region and multi-region scenarios with varying block correlations to evaluate its effectiveness in addressing selection bias. Our results demonstrate that the AIPWE method effectively mitigates the impact of the selection bias in these simulations, providing more accurate estimates of functional connectivity. We applied the AIPWE method to real-world data from 396 children aged 8-13 (144 with autism spectrum disorder and 252 typically developing) from the Autism Brain Imaging Data Exchange (ABIDE) datasets. Our findings reveal that autistic children are more likely to be excluded compared to typically developing children, suggesting that the generalizability of previous studies may be limited due to the selection of older children with less severe clinical profiles. To address data loss and resulting biases, we adapted the AIPWE method in conjunction with an ensemble of machine learning algorithms. The proposed approach identified more edges with differing functional connectivity between autistic and typically developing children compared to the standard approach, highlighting the potential of our framework to improve the study of heterogeneous populations where motion is prevalent. Overall, this study underscores the importance of addressing selection bias in fMRI studies and demonstrates the utility of the AIPWE method in enhancing the reliability and validity of functional connectivity analyses.

Compensating for selection bias when detecting altered functional connectivity in autism

By

Liangkang Wang
B.S., Wuhan University, 2021

Advisor: Benjamin Risk, Ph.D.

A dissertation submitted to the Faculty of the
James T. Laney School of Graduate Studies of Emory University
in partial fulfillment of the requirements for the degree of
Master of Science in Public Health
in Department of Biostatistics and Bioinformatics
2023

Acknowledgments

I am deeply grateful to the Rollins School of Public Health at Emory University for their unwavering support throughout my academic journey. The two-year study in Biostatistics at RSPH provided me with a comprehensive knowledge system and helped me discover my passion for bioinformatics, setting my career direction for the future. I would like to extend my heartfelt thanks to my thesis advisor, Benjamin Risk, for providing me with the opportunity to conduct research in neuroimaging and for his consistent support throughout the thesis writing process. His expertise, sincere guidance, and encouragement have been invaluable, especially during times when I felt disheartened. I am also grateful to Jialu Ran and Zihang Wang, members of the research team, for providing me with the preprocessed data, which were the cornerstone of my study. Their willingness to help and friendly demeanor made my thesis writing journey much more manageable. Lastly, I would like to express my gratitude to my family for their unwavering support, especially during my time studying abroad. Their love and encouragement have been a constant source of comfort, making me feel not alone even when we are miles apart.

Contents

1 Introduction

In brain research, resting-state functional magnetic resonance imaging (rs-fMRI) is commonly used to investigate functional connectivity(Bednarz et al., 2017), which is characterized by spontaneous, interregional correlations in blood-oxygen-level-dependent signal fluctuations(Biswal et al., 1995). rs-fMRI research faces the major challenge of separating signals reflecting neural activity from thermal noise and irrelevant structured signals. Participant head motion may result in spurious pattern of functional connectivity (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). Motion quality control (QC) consists of two steps: eliminating scans with major motion and minimizing artifacts from tolerable motion. Both post-acquisition cleaning techniques and guidelines for removing motion-affected rs-fMRI data have been developed (Power et al., 2014, 2017; Satterthwaite et al., 2013a; Parkes et al., 2018; Muschelli et al., 2014; Pruim et al., 2015; Inoue et al., 1988). The conventional post-acquisition cleaning procedures ignore the impact of scan exclusion on study samples and selection bias (Nebel et al., 2022).

Motion is frequent in pediatric and clinical populations (Fassbender et al., 2017; Greene et al., 2018). The fear that motion artifacts may create erroneous functional connectivity differences between groups of interest if they are not carefully removed from the data has led to the focus on increasing rs-fMRI data quality. According to the "connectivity hypothesis" of autism, the brain's long-range connections suffer while short-range connections are strengthened. However, this pattern is regularly observed in distortions caused by sub-millimeter motion. High-motion subjects still show stronger connections between adjacent brain regions and worse correlations between distant ones as compared to low-motion participants, even after motion has been modified through many modeling steps (Power et al., 2012; Satterthwaite et al., 2012; Van Dijk et al., 2012). According to research on the functional connectivity of autism spectrum disorder (ASD), there are various patterns of significant hypoconnectivity, hyperconnectivity, and mixtures of the two (Di Martino et al., 2011a; Keown et al., 2013; Lombardo et al., 2019; Rudie and Dapretto, 2013; Supekar et al., 2013; Dajani and Uddin, 2016). Studies applying stricter motion quality control have found typical functional connectivity patterns (Dajani and Uddin, 2016), evidence that motion artifacts may have contributed to discrepancies in the literature (Deen and Pelphrey, 2012).

Excluding people with high levels of movement may assist in reducing motion artifacts in estimates of functional connectivity, but doing so may create new issues by systematically changing the research group. Implementing scan exclusion parameters may result in significant sample size decreases. Bias may occur in the mean difference between two groups calculated from observed outcomes in studies where some participants' results are excluded non-randomly (Hernan and Robins, 2020). In the studies of ASD excluding high-motion participants, functional connectivity differences in autistic vs typical children and the severity of motor and social skill deficits are found (D'Souza et al., 2021; Lake et al., 2019; Uddin et al., 2013; Wymbs et al., 2021). In Nebel et al. (2022), we examined the impact of scan exclusion on the composition of the study sample with usable data. To counterbalance the impact of scan exclusion, we employed a doubly robust targeted

minimum loss-based estimation approach (Benkeser et al., 2017) combined with a collection of machine learning techniques (Polley et al., 2019). Inference was based on asymptotic normality, but in practice, finite samples can lead to anti-conservative p-values. On the other hand, corrections for multiplicity using Bonferroni, Holm's, or False Discovery Rate (FDR) are overly conservative because the edge-wise analysis does not incorporate dependencies between edge-locations. Using the max statistic across edges from a permutation test can lead to more powerful corrections for multiple comparisons (Nichols and Holmes, 2002). In this paper, we address both possible issues with smaller sample sizes and propose more powerful tests for accounting for multiple comparisons.

Our research contributes to statistical analysis in several ways. Firstly, we introduce a more effective approach for correcting multiple comparisons that accounts for correlations between edges, which is especially relevant for analyzing complex networks where the edges are often interdependent. Secondly, we develop a finite sample inference method suitable for studies with moderate sample sizes commonly found in neuroimaging research. This method can provide more precise and accurate results than traditional methods of inference that may be unreliable with small sample sizes.

Additionally, we evaluate the augmented inverse propensity weighted estimator (AIPWE) as an alternative to the double robust targeted maximum likelihood estimator (DRTMLE) and find that AIPWE exhibits better finite sample performance with lower Monte Carlo error. Moreover, we propose a permutation test based on AIPWE that is computationally scalable, reduces type-1 errors, and enhances statistical power when p-values are highly correlated.

Finally, we apply our methodology to analyze connectivity with a seed in the default mode network in a dataset of 400 children from ABIDE I/ABIDE II (Di Martino et al., 2014, 2017). This analysis demonstrates the utility of our approach in identifying meaningful connectivity patterns and highlights the potential of our method to advance our understanding of brain function in various populations.

2 Statistical Methods

2.1 Parameter of Interest and Target Parameter

We aim to identify the differences in average functional connectivity between children with autism without intellectual disabilities and typically developing children. To achieve this, we employ a causally informed method to address any possible selection bias that may have occurred due to the exclusion of observations in the motion quality control.

Assume that Y is a random variable representing the functional connectivity between two locations defined in the brain. To indicate the presence or absence of autism spectrum disorder (ASD) in a participant, we will use the variable A. Specifically, A takes on a value of one if the participant has ASD and zero otherwise. We will use the variable W to represent the covariates, which consist of measures that may be associated with both functional connectivity and the severity

of autism spectrum disorder (ASD). It is worth noting that W and A are not independent, given that the distribution of certain behavioral variables varies by diagnosis group. Therefore, the values of W are likely to differ between individuals with ASD and typically developing individuals. Children may often move excessively during their rs-fMRI scan, resulting in unusable data. However, we can still gather significant behavioral and sociodemographic covariates from these children. For the purpose of our analysis, we consider data that fail motion quality control as "missing data." Let Δ denote a binary random variable capturing the missing data mechanism that is equal to one if the data are usable and zero otherwise. Then data are realizations of the random vector $\{ Y, A, W, \Delta \}$.

Let $Y(1)$ represent the counterfactual outcome indicating that a participant's scan is usable. Define the counterfactual parameter $\psi^* = E^*[Y(1)|A = 1] - E^[Y(1)|A = 0]$ as the difference in expected counterfactual outcomes between two groups. Similarly, define the usable parameter $\psi = E[Y|\Delta = 1, A = 1] - E[Y|\Delta = 1, A = 0]$ as the difference in expected outcomes between the two groups of usable scans. There exists selection bias $\psi^* = \psi$ when $\Delta \leftrightarrow W$, $W \leftrightarrow Y$ due to the lack of exchangeability between usable and unusable data (Hernan and Robins, 2020).

Identifying the parameter of interest ψ^* from the target parameter ψ requires three assumptions:

(A1.1) *Mean exchangeability* : for $a = 0, 1$, $E^*\{Y(1)|A = a, W\} = E^*\{Y(1)|\Delta = 1, A = a, W\}$.

(A1.2) *Positivity* : for $a = 0, 1$ and all possible w, $P(\Delta = 1|A = a, W = w) > 0$.

(A1.3) *Causal Consistency* : for all i such that $\Delta_i = 1$, $Y_i(1) = Y_i$.

Then we proposed the target parameter in Nebel et al. (2022),

$$\begin{aligned} \psi^* = \psi &= E[Y|\Delta = 1, A = 1] - E[Y|\Delta = 1, A = 0] \\ &= E\{E(Y \mid \Delta = 1, A = 1, W) \mid A = 1\} - E\{E(Y \mid \Delta = 1, A = 1, W) \mid A = 0\}. \end{aligned} \tag{1}$$

Put simply, ψ represents the difference between the average functional connectivity of individuals with autism spectrum disorder (ASD) and those without (TD) across the range of behavioral phenotypes in each group. The inclusion of covariates W, which includes measures of autism severity, is important to maintain the distribution of autism severity within each diagnosis group.

Remark: The definition of W is an important scientific decision. In our application, we distinguish between covariates whose distribution we want to be balanced in the ASD and TD, which we call to as diagnosis-independent covariates, and those that we consider to be part of ASD, which we call diagnosis-dependent covariates or W. Diagnosis-independent covariates that are not balanced between groups are a potential source of bias. These are generally variables that are conventionally included as regressors in multiple regression, for example, age, sex, handedness, and gross motion measures (Di Martino et al., 2013). Here, we account for these effects in an initial processing step, see Section 4.4.3.

2.2 AIPWE

The augmented inverse propensity weighted estimator (Glynn and Quinn, 2010) can be adapted to the missing data problem with diagnosis-specific distributions. Let n_1 be the number of children with ASD, and let $\mathcal{S}_1$ denote the set of indices for ASD children. Similarly, define n_0 and $\mathcal{S}_0$ for the TD children.

First, we define the propensity model as $g_n(A, W) = P(\Delta = 1|A = a, W = w)$, which estimates the probability of data inclusion in the motion control, derived by logistic regression in Section 4.6. Moreover, we introduce the outcome model $\bar{Q}_n(A, W) = E(Y|\Delta = 1, A = a, W = w)$, aiming to predict $Y(1)|A, W$ for $\Delta = 0$ or 1 (refer to Section 4.6). The subscript n highlights the use of all observations within the study (encompassing both ASD and TD children) for fitting the propensity and outcome regression models. In this context, let $g_n(A_i, W_i)$ represent the propensity, i.e., the predicted probability of data inclusion in the motion control step, and let $\bar{Q}_n(A_i, W_i)$ denote the predicted functional connectivity from the outcome model.

Subsequently, we estimate the mean connectivity in the ASD group.

$$\psi_{n,AIP,1} = \frac{1}{n_1} \sum_{i \in \mathcal{S}_1} \left[\frac{I(\Delta_i = 1)}{g_n(A_i, W_i)} \right] Y_i + \frac{1}{n_1} \sum_{i \in \mathcal{S}_1} \left[1 - \frac{I(\Delta_i = 1)}{g_n(A_i, W_i)} \right] \bar{Q}_n(A_i, W_i). \tag{2}$$

The asymptotic variance when both the propensity and outcome model are correctly specified is

$$Var(\psi_{n,AIP,1}) = \frac{1}{n_1(n_1 - 1)} \sum_{i \in \mathcal{S}_1} \left[Z_i - \frac{1}{n_1} \sum_{i \in \mathcal{S}_1} Z_i \right]^2, \tag{3}$$

where $Z_i = \frac{I(\Delta_i = 1)}{g_n(A_i, W_i)} \left[Y_i * \bar{Q}_n(A_i, W_i) \right] + \bar{Q}_n(A_i, W_i) - \psi_{n,AIP,1}$.

In a similar manner, we estimate the mean connectivity within the TD group, denoted as $\psi_{n,AIP,0}$, and its corresponding variance. As a result, we obtain $\psi_{n,AIP} = \psi_{n,AIP,1} - \psi_{n,AIP,0}$ and $var(\psi_{n,AIP}) = \sqrt{var(\psi_{n,AIP,1}) + var(\psi_{n,AIP,0})}$, under the assumption that the autistic and typically developing groups are independent.

2.3 Permutation Test

We propose a novel permutation test that is computationally scalable. Note that a classic permutation test would involve refitting the propensity and outcome models thousands of times. In our practical application, we examine 418 edges, which represent the functional connectivity between the 14th parcel and the remaining parcels. We employ 20 random seeds for each location in the outcome model during a 10-fold cross-validation process, as cross-validation predictions are notably sensitive to the random seed. A naive permutation test with 1,000 permutations would require more than 80,000,000 outcome regressions, as calculated by $1000 \times 418 \times 20 \times 10$. The computationally scalable permutation test treats the propensity-adjusted AIPWE terms as data points, as described below.

For each edge, denoted as j, we permute membership in $\mathcal{S}_1$ and $\mathcal{S}_0$, call these permuted memberships $\mathcal{S}_1^{(k)}$ and $\mathcal{S}_0^{(k)}$.

$$\begin{aligned}\hat{\psi}_{j,AIP}^{(k)} = \frac{1}{n_1}\sum_{i\in\mathcal{S}_1^{(k)}}\left(\left[\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]Y_{ij}+\left[1-\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]\bar{Q}_j(A_i,W_i)\right)\\ -\frac{1}{n_0}\sum_{i\in\mathcal{S}_0^{(k)}}\left(\left[\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]Y_{ij}+\left[1-\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]\bar{Q}_j(A_i,W_i)\right).\end{aligned} \tag{4}$$

Standardize by the asymptotic standard error(3) to generate the statistic of each edge within every permutation:

$$z_j^{(k)} = \frac{\hat{\psi}_{j,AIP}^{(k)}}{\sqrt{Var(\hat{\psi}_{j,AIP,1}^{(k)}) + Var(\hat{\psi}_{j,AIP,0}^{(k)})}} \tag{5}$$

Family-wise error rate control: Take max of absolute value across all edges and compare to our original z statistic, K means the permutation times:

$$p_{j,fwer} = \frac{1}{K}\sum_{k=1}^{K} I\left(\left\{\max_j |z_j^{(k)}|\right\} > |z_j|\right) \tag{6}$$

Note that under the null hypothesis:

$$E\{E(Y_j \mid \Delta=1, A=1, W) \mid A=1\} - E\{E(Y_j \mid \Delta=1, A=0, W) \mid A=0\} = 0.$$

This permutation test preserves the inner conditional expectation $E(Y_j|\Delta=1,A,W)$ by using the estimates $\left(\left[\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]Y_{ij}+\left[1-\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]\bar{Q}_n(A_i,W_i)\right)$ for $i\in\{1,\ldots,n\}$. Each term of the permutation test is an estimate of $E\{E(Y_j \mid \Delta=1, A=1, W) \mid A=1\} - E\{E(Y_j \mid \Delta=1, A=0, W)|A=0\}$. We obtain a null distribution for our finite sample since $E\{\hat{\psi}_{j,AIP}^{(k)}\} = E\{E(Y_j \mid \Delta=1, A=1, W)\} - E\{E(Y_j \mid \Delta=1, A=0, W)\}$.

In practice, using predicted values in both the propensity and outcome models can lead to violations of the exchangeabililty assumption, which can lead to inflated type-1 errors. This will be investigated in simulations. We also evaluate a sandwich estimator treating $\left(\left[\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]Y_i+\left[1-\frac{I(\Delta_i=1)}{g_n(A_i,W_i)}\right]\bar{Q}_n(A_i,W_i)\right)$ as data in an attempt to address possible issues with heteroscedasticity (Zeileis, 2006).

3 Simulations

In this section, we conducted a series of simulation studies to assess the performance of the Augmented Inverse Probability Weighted (AIPW) estimator (Glynn and Quinn, 2010), and the permutation test. Our real data analysis concentrated on the functional connectivity between the 14th parcel and the other parcels within the brain (see in Section 4.3.2). The simulations were designed to replicate the functional connectivity between a single region and the seed region, as presented in Simulation 3.1, as well as the functional connectivity between multiple regions and the seed region, as illustrated in Simulation 3.2. These simulations aimed to showcase the application of the AIPW

estimator for both single-edge and multi-edge scenarios.

3.1 Single region

In this single-region simulation, we illustrate the efficacy of DRTMLE and AIPWE in mitigating selection bias and demonstrate enhancements in Type I error and power as sample size increases.

We used the same severity of ASD and probability to pass the movement control, as in Nebel et al. (2022), to simulate biased datasets and estimate the deconfounded group difference. The simulated sample includes around 35% of individuals with ASD, which is similar to the proportion observed in real data (around 36% ASD). Additionally, we generated a covariate, denoted as W_c, to represent ASD severity, which is equal to zero for typically developing children and is generated from a log-normal distribution in the ASD group with a log mean of 2 and a standard deviation of 0.4. In addition, we generated nine standard normal variables that have no relationship with the diagnosis. The propensity model's data usability was generated from a logistic regression model, $logit(E[\Delta = 1|W_c = w_c]) = 2 - 0.2 * w_c$. The simulation setup led to roughly 88% and 60% usable data in the typically developing and ASD groups, respectively, in contrast to 85% and 69% in the real data when using Ciric criteria outlined as ($\Delta = 1$ if relative RMS displacement < 0.2, and $>$ 5 minutes of data remain after removing > 0.25 framewise relative RMS displacement(Ciric et al., 2017)).

We used a linear model to specify the outcome model, with a slope of -0.2 for W_c, 0 for the remaining nine covariates, and intercepts of 0 and 1.4 for typically developing individuals and those with Autism Spectrum Disorder (ASD), respectively. Through the implementation of this simulation design, we successfully attained a correlation of -0.58 between W_c and Y, in addition to a "true" functional connectivity, denoted by $E[Y(1)|A = a]$, approximating -0.20 in the ASD group and 0 in the typically developing group. Consequently, this produced a $Cohen'sd$ value of 0.39 when comparing the two groups. We generated a random sample of 500 participants and employed it to estimate the deconfounded group difference, as depicted in fig. 1.

We also designed a simulation setting with all covariates of linear regression set to 0. This new setting is used to assess the type 1 error, as it conforms to the null hypothesis that there is no difference in functional connectivity between autistic and typically developing children. We then modified the sample size while maintaining the covariates of the two simulation settings to assess the variation of type 1 error and power of various tests across different sample sizes. Table **??** shows that the type 1 error for the permutation-based Augmented Inverse Probability Weighting Estimator (AIPWE) approaches 0.05 when the sample size exceeds 500. Conversely, the Doubly Robust Targeted Maximum Likelihood Estimator (drtmle) method exhibits a considerably higher type 1 error when the sample size is below 500. Regarding rejection capability, Table **??** demonstrates that drtmle has higher power for smaller sample sizes, such as 100, and comparable power when the sample size exceeds 500. Our simulations suggest that AIPWE is a superior technique for controlling type 1 error, particularly in smaller sample sizes, as drtmle utilizes an increased type 1

error rate to achieve enhanced power.

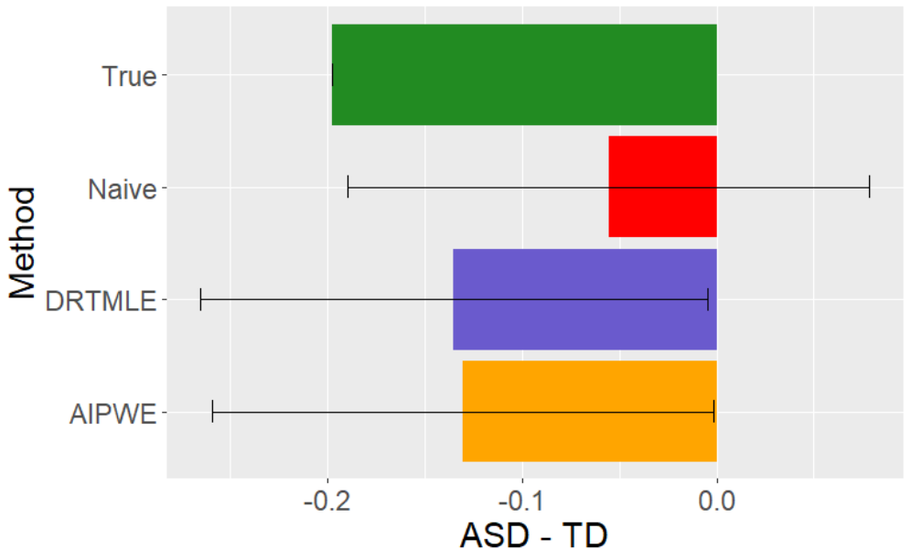
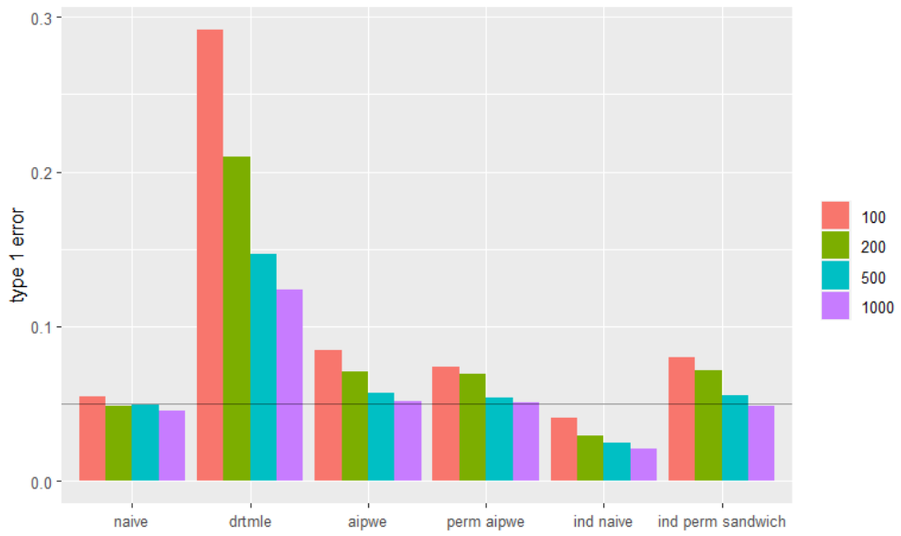


Figure 1: **An illustration of the improvement in functional connectivity from DRTMLE and AIPWE compared to the naive approach from a single simulated dataset.** The authentic mean difference in functional connectivity between Autism Spectrum Disorder (ASD) and typically developing (TD) groups is negative (illustrated by the green bar), with the true mean in the ASD group exhibiting a negative value and the true mean in the TD group presenting a slightly positive value. The estimate of the mean ASD-TD difference derived from the naïve approach (depicted by the red bar) is also negative, albeit closer to zero. Furthermore, the 95% confidence interval encompasses zero. By employing Doubly Robust Targeted Maximum Likelihood Estimator (DRTMLE) and Augmented Inverse Probability Weighting Estimator (AIPWE), the deconfounded group differences (represented by the purple and yellow bars) are more proximate to the true values, and the 95% confidence intervals do not include zero.

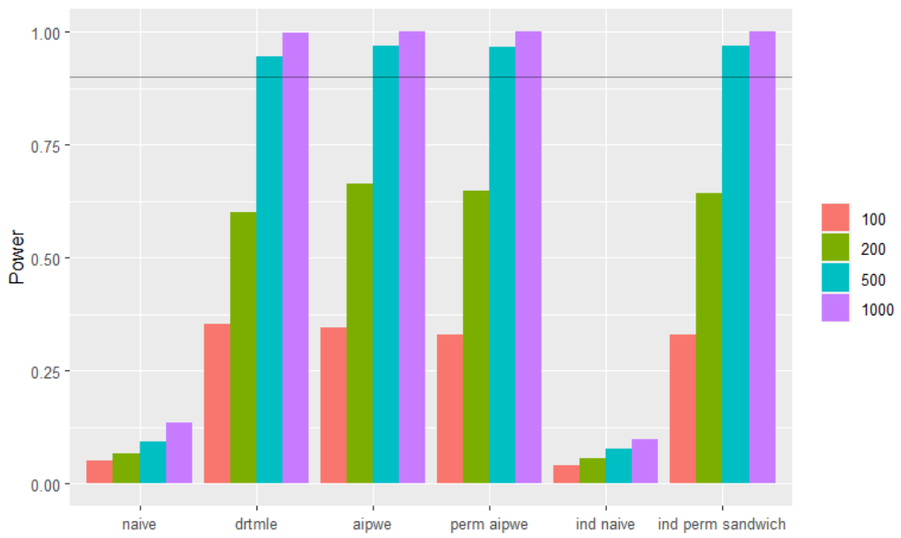
3.2 Multiple regions

In this section, our objective is to utilize the Monte Carlo method to simulate and compare the performance of various tests on multiple tests. Specifically, we concentrate on the tests' capacity to control the family-wise error rate and their power in effectively distinguishing regions of interest. Our aim is to identify the most appropriate approach for our real data analysis.

In this section, we retain the same settings for the variables W_c, A, and usability Δ as outlined in section 3.1. This approach ensures that the proportion of usable and unusable data conforms to our actual data. Furthermore, we incorporate a block structure across multiple subjects to simulate the interdependencies among various time series, mirroring the characteristics of functional magnetic resonance imaging (fMRI) data. As a result, this simulation provides a trustworthy representation of the properties observed in fMRI data. We specified the outcome model using a linear model, setting the slope for W_c at 0.05, and 0 for the remaining nine covariates. The intercepts for typically developing individuals and those with Autism Spectrum Disorder (ASD) were 0 and 0.15, respectively. This linear model was simulated based on real data, specifically the correlation between the 14th and 273rd region (17networks_RH_ContC_IPL_1) (Schaefer et al., 2018), which was identified as an effective area exhibiting significantly altered functional connectivity to our seed region.



(a) Type 1 Error for single region tests



(b) Power for single region tests

Figure 2: **Performance of tests in simulation setting 1 (single region).** a) Type 1 Error for single region tests. b) Power for single region tests.

It is crucial to note that the correlation between the 14th and 273rd region represents the most effective functional connectivity in our real data analysis, with the smallest p-value in multiple tests. The signal of the difference between ASD and typically developing (TD) groups is considerably stronger than in Simulation 3.1. We employed $Y = 0.15 * I(ASD) + 0.05 * W_c$, resulting in a mean Y value of 0.55 in the ASD group and 0 in the TD group. In Simulation 3.1, the mean Y value is -0.2 in the ASD group and 0 in the TD group.

3.2.1 Strong block-wise correlation

In this section, we generate stochastic data that includes three pronounced block structures, consisting of one effective edge and 80 non-effective edges. Each block comprises 27 edges, and the within-block correlation is set to 0.9. The correlation matrix displayed in fig.3(a) exhibits the distinct block structure. Our aim is to evaluate and compare the family-wise error rate (FWER),

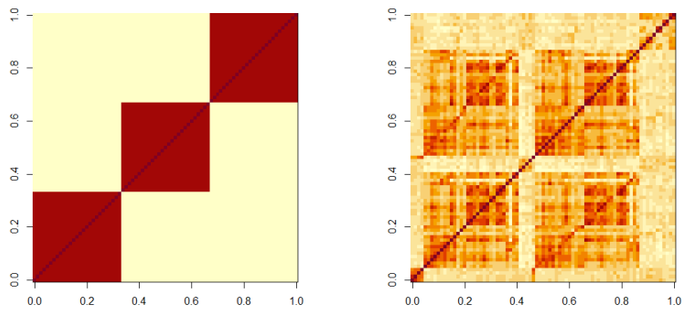
false discovery rate (FDR), and power of Bonferroni-adjusted AIPWE, Benjamini-Hochberg (BH)-adjusted AIPWE, permutation-based AIPWE, and permutation-based sandwich estimator (refer to Section 2.3) under multiple edges with a strong block structure.

As illustrated in fig.4(b), the maximum statistic permutation-based Augmented Inverse Probability Weighting Estimator (max perm AIPWE) reveals significantly higher power with a sample size of 200 compared to the naïve and AIPWE tests employing Bonferroni or Benjamini-Hochberg (BH) adjustment. Moreover, in fig.4(a), the max perm AIPWE showcases a relatively acceptable FWER when the sample size reaches 500, similar to most real data analyses. In fact, AIPWE with BH adjustment effectively controls the FWER when the sample size is 500. If the primary goal is not to achieve the most powerful test under a small sample size and stringent block-wise correlation conditions, we recommend using AIPWE with BH adjustment.

3.2.2 Correlation from a seed-based analysis

We employ the same parameters for the propensity and outcome models as in section 3.2.1 but reduce the block structure to a more realistic level. We generate simulation data with multivariate normal errors utilizing a subset of the correlation matrix between 81 edges from the data analysis discussed in section 4. The weak block structure is represented by the correlation matrix in fig.3(b). To assess the performance of AIPWE, we examine the false discovery rate (FDR) and power using three methods: AIPWE with Bonferroni correction, AIPWE with permutation test, and AIPWE with max statistics.

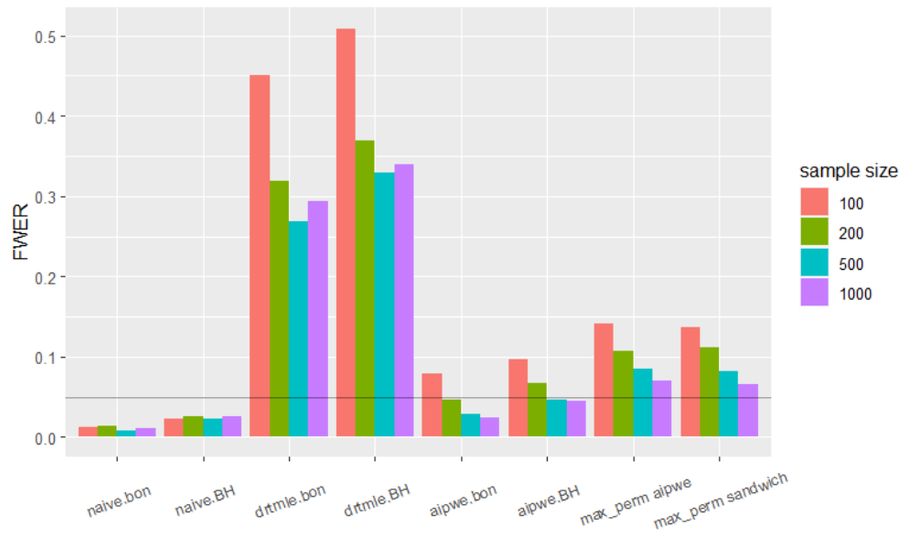
In the setting with a weaker correlation, which more accurately reflects real data analysis, we draw a similar conclusion regarding power as in Simulation 3.2.1. The maximum permutation-based Augmented Inverse Probability Weighting Estimator (max perm AIPWE) exhibits superior performance in small sample sizes, as shown in fig.5(b). It is crucial to note that in the weaker correlation setting, the Family-Wise Error Rate (FWER) of AIPWE with Bonferroni and Benjamini-Hochberg (BH) adjustment increases and exceeds that of max perm AIPWE, as depicted in fig.5(a). This result emphasized the effectiveness of max perm AIPWE in controlling FWER in datasets with weak inter-correlations, ultimately reducing the false discovery rate and the number of incorrectly identified significant areas in real data analysis.



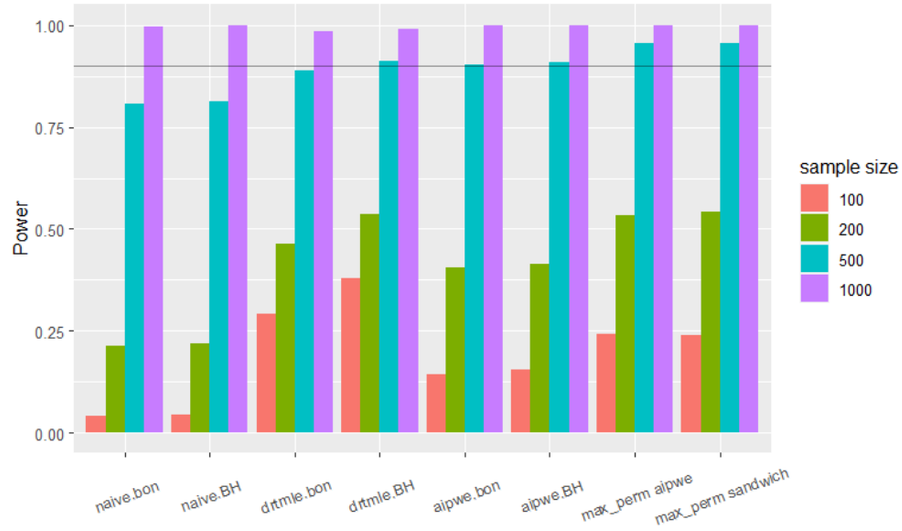
(a) Strong block-wise correlation

(b) Correlation from a seed-based analysis

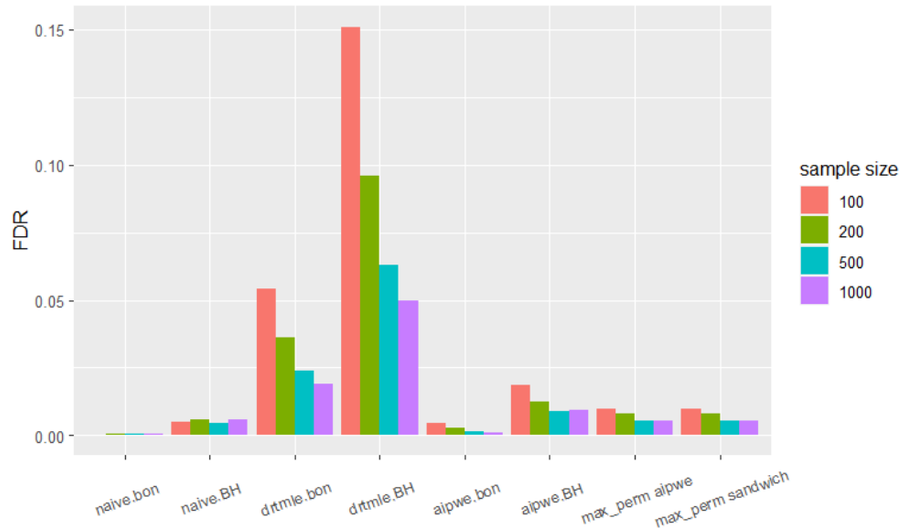
Figure 3: **Block structure used in simulation.**



(a) Family-wise error rate for multiple regions tests

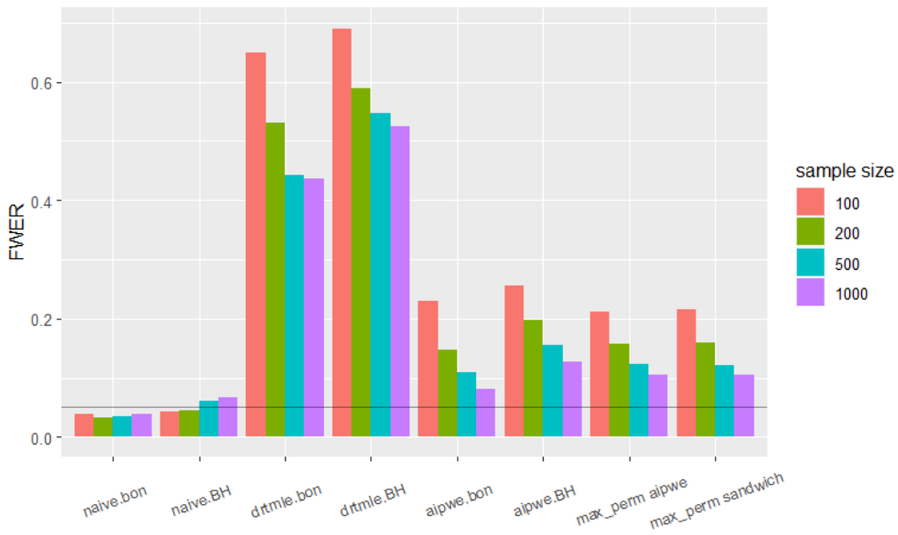


(b) Power for multiple regions tests

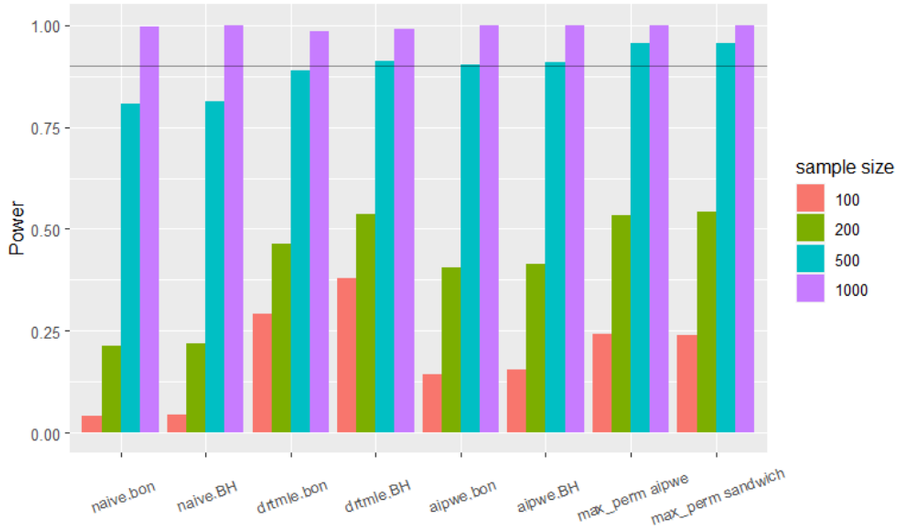


(c) False discovery rate for multiple regions tests

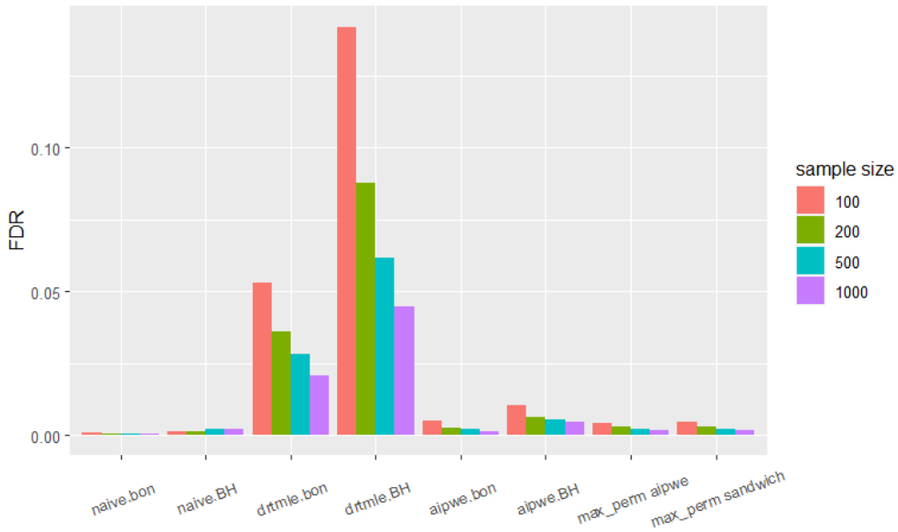
Figure 4: **Performance of tests simulation setting 2 (strong block-wise correlation between 81 regions).** a)Family-wise error rate for multiple regions tests. b)Power for multiple regions tests. c) False discovery rate for multiple regions tests.



(a) Family-wise error rate for multiple regions tests



(b) Power for multiple regions tests



(c) False discovery rate for multiple regions tests

Figure 5: **Performance of tests for simulation setting 3 (correlations between regions based on real data analysis, weaker correlation).** a) Family-wise error rate for multiple regions tests. b) Power for multiple regions tests. c) False discovery rate for multiple regions tests.

4 Data and Methods

4.1 Dataset

4.1.1 Study population

Our cohort comprised 396 children aged 8-13 years old, 144 autistic children without intellectual disabilities (119 boys), and 252 typically developing children (174 boys). We used rs-fMRI scans and phenotypic data for these children from the Autism Brain Imaging Data Exchange (ABIDE)(Di Martino et al., 2014, 2017). Table 1 summarizes the socio-demographic traits for all predictor cases. The demographic traits related to passing movement control are presented in separate summaries for both autistic individuals (see Table 2) and typically developing individuals (see Table 3). Section 4.3.1 will describe how to calculate patients' movements and how to decide to pass quality control.

4.1.2 Phenotypic Assessment

To evaluate the severity of the primary symptoms of Autism Spectrum Disorder (ASD) in the ASD group, we employed scores from either the Autism Diagnostic Observation Schedule (ADOS) (Lord et al., 2000) or the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) (Lord and Jones, 2012). These scores were calibrated to ensure comparability across different versions of the instrument (Hus et al., 2014). We focused on the ADOS/ADOS-2 Comparable Total Score, hereinafter referred to as ADOS. Higher ADOS scores indicate more severe symptoms for individuals with autism spectrum disorder (ASD), and this score measures the severity of those symptoms. These semi-structured ASD observation schedules are not commonly used with participants in control groups. Scores for typically developing children are often zero or very close to zero because they were not designed to capture significant variance in individuals who do not have ASD. However, ASD-like traits in non-clinical individuals can vary, and those who meet the criteria for an ASD diagnosis are at one end of a continuum that also includes the general population.

Intellectual ability was assessed using the Full-Scale Intelligence Quotient (FIQ). It is a measure of general intelligence obtained from a standardized IQ test that evaluates various cognitive abilities, such as verbal comprehension, perceptual reasoning, working memory, and processing speed. The Full-Scale IQ score is derived by combining scores from these individual subtests, and it is used to provide an overall measure of cognitive ability and intellectual functioning. The Wechsler Intelligence Scale for Children (WISC) is the most commonly used IQ test that generates an FIQ score (Wechsler, 2003).

4.2 rs-fMRI acquisition and preprocessing

KKI acquisitions included 2 dummy scans, and visual assessment of the global signal indicated it was stable. NYU data use 82 or 90 degree flip angles, and discarding two TRs suffices for achieving steady-state. Results included in this section come from preprocessing performed using **fMRIPrep**

	TD (N=252)	ASD (N=144)	Total (N=396)	p value
Age				0.680
Mean (SD)	10.400 (1.347)	10.338 (1.594)	10.378 (1.440)	
Range	8.010 - 13.720	8.014 - 13.950	8.010 - 13.950	
Gender				0.003
Male	174 (69.0%)	119 (82.6%)	293 (74.0%)	
Female	78 (31.0%)	25 (17.4%)	103 (26.0%)	
FIQ				< 0.001
Mean (SD)	114.627 (11.507)	103.625 (17.477)	110.626 (14.926)	
Range	80.000 - 144.000	63.000 - 148.000	63.000 - 148.000	
Handedness				0.289
Right	235 (93.3%)	130 (90.3%)	365 (92.2%)	
Left	17 (6.7%)	14 (9.7%)	31 (7.8%)	
ADOS				< 0.001
Mean (SD)	0.000 (0.000)	13.403 (5.245)	4.874 (7.186)	
Range	0.000 - 0.000	6.000 - 35.000	0.000 - 35.000	
Currently on Stimulants				< 0.001
No	252 (100.0%)	116 (80.6%)	368 (92.9%)	
Yes	0 (0.0%)	28 (19.4%)	28 (7.1%)	
Currently on NonStimulants				< 0.001
No	251 (99.6%)	118 (81.9%)	369 (93.2%)	
Yes	1 (0.4%)	26 (18.1%)	27 (6.8%)	
Site ID				< 0.001
ABIDEI-KKI	33 (13.1%)	22 (15.3%)	55 (13.9%)	
ABIDEI-NYU	44 (17.5%)	43 (29.9%)	87 (22.0%)	
ABIDEII-KKI_1	155 (61.5%)	56 (38.9%)	211 (53.3%)	
ABIDEII-NYU_1	20 (7.9%)	23 (16.0%)	43 (10.9%)	
ciric				< 0.001
Unusable	39 (15.5%)	45 (31.2%)	84 (21.2%)	
Usable	213 (84.5%)	99 (68.8%)	312 (78.8%)	
powerpt2				< 0.001
Unusable	126 (50.0%)	110 (76.4%)	236 (59.6%)	
Usable	126 (50.0%)	34 (23.6%)	160 (40.4%)	

Table 1: **Socio-demographic characteristics** Mean and standard deviation (SD) are presented for continuous variables, and Kruskal-Wallis rank-sum tests were used to compare diagnosis groups. Frequencies and percentages summarize binary and categorical variables, and differences between diagnosis groups were analyzed using either the Chi-square test or Fisher's exact test. Although data were pooled from multiple studies, age, and handedness were balanced across diagnosis groups, while gender and FIQ were not. ASD = autism spectrum disorder. TD = typically developing. SD = standard deviation.

21.0.2 (Esteban et al., 2019) which is based on **Nipype 1.6.1** (Gorgolewski et al., 2011)

4.2.1 Anatomical Data Preprocessing

The input BIDS dataset contained a single T1-weighted (T1w) image. This image underwent intensity non-uniformity (INU) correction using *N4BiasFieldCorrection* (Tustison et al., 2010), which is included in **ANTs 2.3.3** (Avants et al., 2009). The corrected T1w image served as the T1w-reference throughout the workflow. Subsequently, the T1w-reference was skull-stripped using a Nipype implementation of the *antsBrainExtraction.sh* workflow from ANTs, with OASIS 30ANTs as the target template. *fast* (Woolrich et al., 2009; Abramian et al., 2022) was employed to perform brain tissue segmentation of cerebrospinal fluid (CSF), white matter (WM), and gray matter (GM) on the brain-extracted T1w. Brain surfaces were reconstructed using *recon-all* (Dale et al., 1999a; Fischl, 2012). A custom adaptation of the Mindboggle method (Klein et al., 2017) was utilized to reconcile ANTs-derived and FreeSurfer-derived segmentations of cortical gray matter, which refined

	Unusable (N=45)	Usable (N=99)	Total (N=144)	p value
Age				0.311
Mean (SD)	10.138 (1.755)	10.429 (1.516)	10.338 (1.594)	
Range	8.014 - 13.630	8.030 - 13.950	8.014 - 13.950	
Gender				0.130
Male	34 (75.6%)	85 (85.9%)	119 (82.6%)	
Female	11 (24.4%)	14 (14.1%)	25 (17.4%)	
FIQ				0.567
Mean (SD)	104.867 (18.145)	103.061 (17.229)	103.625 (17.477)	
Range	63.000 - 136.000	67.000 - 148.000	63.000 - 148.000	
Handedness				0.704
Right	40 (88.9%)	90 (90.9%)	130 (90.3%)	
Left	5 (11.1%)	9 (9.1%)	14 (9.7%)	
ADOS				0.391
Mean (SD)	12.844 (4.311)	13.657 (5.621)	13.403 (5.245)	
Range	7.000 - 28.000	6.000 - 35.000	6.000 - 35.000	
Currently on Stimulants				0.733
No	37 (82.2%)	79 (79.8%)	116 (80.6%)	
Yes	8 (17.8%)	20 (20.2%)	28 (19.4%)	
Currently on NonStimulants				0.683
No	36 (80.0%)	82 (82.8%)	118 (81.9%)	
Yes	9 (20.0%)	17 (17.2%)	26 (18.1%)	
Site ID				0.007
ABIDEI-KKI	10 (22.2%)	12 (12.1%)	22 (15.3%)	
ABIDEI-NYU	7 (15.6%)	36 (36.4%)	43 (29.9%)	
ABIDEII-KKI_1	24 (53.3%)	32 (32.3%)	56 (38.9%)	
ABIDEII-NYU_1	4 (8.9%)	19 (19.2%)	23 (16.0%)	

Table 2: **Demographic for autistic group** Unusable = the patient was excluded in Ciric criteria. Usable = the patient passed the Ciric criteria.

the previously estimated brain mask. The T1w reference and T1w template (both brain-extracted) were used for volume-based spatial normalization to two standard spaces, MNI152NLin2009cAsym and MNI152NLin6Asym, through nonlinear registration with antsRegistration (ANTs 2.3.3). The chosen templates for spatial normalization included the ICBM 152 Nonlinear Asymmetrical template version 2009c (Fonov et al., 2009) and FSL's MNI ICBM 152 non-linear 6th Generation Asymmetric Average Brain Stereotaxic Registration Model (Jenkinson et al., 2012).

4.2.2 Functional and Anatomical Data Preprocessing

The preprocessing steps for each subject's single BOLD run, encompassing all tasks and sessions, consisted of the following procedures. Initially, fMRIPrep's custom methodology was utilized to generate a reference volume and its skull-stripped counterpart. Subsequently, head-motion parameters were estimated with respect to the BOLD reference using transformation matrices and six corresponding rotation and translation parameters before performing any spatiotemporal filtering, which was accomplished through the use of *mcflirt* (Jenkinson et al., 2002). After applying head-motion correction transforms, the BOLD time-series (with slice-timing correction if applicable) were resampled to their original, native space. These resampled BOLD time series are referred to as preprocessed BOLD in original space or simply preprocessed BOLD. To align the functional data with anatomical data, the BOLD reference was registered to the T1w reference using a boundary-based registration tool called *bbregister* from FreeSurfer (Dale et al., 1999b), as described in (Tobyne

	Unusable (N=39)	Usable (N=213)	Total (N=252)	p value
Age				0.361
Mean (SD)	10.219 (1.286)	10.434 (1.359)	10.400 (1.347)	
Range	8.025 - 12.899	8.010 - 13.720	8.010 - 13.720	
Gender				0.435
Male	29 (74.4%)	145 (68.1%)	174 (69.0%)	
Female	10 (25.6%)	68 (31.9%)	78 (31.0%)	
FIQ				0.474
Mean (SD)	113.410 (10.166)	114.850 (11.744)	114.627 (11.507)	
Range	93.000 - 136.000	80.000 - 144.000	80.000 - 144.000	
Handedness				0.798
Right	36 (92.3%)	199 (93.4%)	235 (93.3%)	
Left	3 (7.7%)	14 (6.6%)	17 (6.7%)	
ADOS				NaN
Mean (SD)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	
Range	0.000 - 0.000	0.000 - 0.000	0.000 - 0.000	
Currently on Stimulants				
No	39 (100.0%)	213 (100.0%)	252 (100.0%)	
Yes	0 (0.0%)	0 (0.0%)	0 (0.0%)	
Currently on NonStimulants				0.019
No	38 (97.4%)	213 (100.0%)	251 (99.6%)	
Yes	1 (2.6%)	0 (0.0%)	1 (0.4%)	
Site ID				0.090
ABIDEI-KKI	2 (5.1%)	31 (14.6%)	33 (13.1%)	
ABIDEI-NYU	4 (10.3%)	40 (18.8%)	44 (17.5%)	
ABIDEII-KKI_1	31 (79.5%)	124 (58.2%)	155 (61.5%)	
ABIDEII-NYU_1	2 (5.1%)	18 (8.5%)	20 (7.9%)	

Table 3: **Demographic for typically developing group** Unusable = the patient was excluded in Ciric criteria. Usable = the patient passed the Ciric criteria.

et al., 2016).

The co-registration was configured with six degrees of freedom. After preprocessing the BOLD data, several confounding time-series were calculated, including framewise displacement (FD), DVARS, and three region-wise global signals. Two formulations were used to compute FD: the absolute sum of relative motions as described by Power et al. (2013), and the relative root mean square displacement between affines using Jenkinson et al. (2002). These computations were conducted using the *power_fd_dvars* and *mcflirt* functions, respectively. For each functional run, framewise displacement (FD) and DVARS were computed using their respective implementations in **Nipype** [following the definitions by Power et al. (2013)]. Three global signals were extracted from the cerebrospinal fluid (CSF), white matter (WM), and whole-brain masks. Additionally, a set of physiological regressors was extracted to enable component-based noise correction using *CompCor* [as described in Behzadi et al. (2007)]. To estimate principal components, the preprocessed BOLD time series were high-pass filtered using a discrete cosine filter with a 128-second cut-off for the two CompCor variants: temporal (*tCompCor*) and anatomical (*aCompCor*). The top 2% of variable voxels within the brain mask were used to calculate *tCompCor* components. For *aCompCor*, three probabilistic masks were generated in anatomical space: CSF, WM, and a combination of CSF+WM. However, the implementation differs from that of Behzadi et al. (2007) because instead of eroding the masks by 2 pixels in BOLD space, the aCompCor masks are subtracted from a mask of pixels that likely contain a volume fraction of gray matter (GM).

To ensure that components are not extracted from voxels containing a minimal fraction of gray matter (GM), a GM mask extracted from FreeSurfer's *aseg* segmentation is dilated and subtracted from the aCompCor masks. These masks are then resampled into BOLD space and binarized by thresholding at 0.99, as in the original implementation. Components are calculated separately within the WM and CSF masks. For each CompCor decomposition, the k components with the largest singular values are retained, such that the retained components' time series are sufficient to explain 50% of variance across the nuisance mask (CSF, WM, combined, or temporal). The remaining components are dropped from consideration. The head-motion estimates calculated in the correction step are also included in the corresponding confounds file. The confound time series derived from head motion estimates and global signals are expanded by including temporal derivatives and quadratic terms for each, following Satterthwaite et al. (2013b).

Frames with motion exceeding a threshold of 0.5 mm FD or 1.5 standardized DVARS were identified as motion outliers. Subsequently, the BOLD time series were resampled to standard space to generate a preprocessed BOLD run in MNI152NLin2009cAsym space. The custom methodology of fMRIPrep was employed to generate a reference volume and its skull-stripped version. Resampling of the BOLD time series onto surfaces was carried out using the FreeSurfer reconstruction nomenclature. Additionally, Grayordinates files (Glasser et al., 2013) containing 91k samples were generated using the highest-resolution *fsaverage* as an intermediate standardized surface space.

The resampling process was optimized by utilizing a single interpolation step that combines all relevant transformations, including head-motion correction transform matrices, susceptibility distortion correction (if applied), and co-registrations to anatomical and output spaces. Gridded (volumetric) resamplings were performed using *antsApplyTransforms* from **ANTs** software, which was configured to use Lanczos interpolation to minimize the smoothing effects of other kernels (Lanczos, 1956). Non-gridded (surface) resamplings were performed using *mri_vol2surf* from FreeSurfer.

Many internal operations of fMRIPrep rely on Nilearn 0.8.1 (), primarily within the functional processing workflow. For more information on the pipeline, refer to the workflows section in the fMRIPrep documentation (https://fmriprep.readthedocs.io/en/latest/workflows.html "fMRIPrep documentation").

4.3 Seed correlations

4.3.1 Motion quality control

We employed two criteria for the exclusion of gross motion:

Ciric Criteria: Scans were deemed unusable and excluded if the mean RMSD exceeded 0.2 mm or if they contained less than five minutes of data without frames where RMSD exceeded 0.25 mm (Ciric et al., 2017).

Powerpt2 Criteria: Scans were excluded if the participant had less than 5 minutes of data remaining after the removal of frames in which the FD exceeded 0.2 mm (Power et al., 2014).

The outcome of the two criteria are summarized in table 1. Table 2 and 3 display the demographic statistics for children who either passed or did not pass the motion control based on Ciric criteria.

4.3.2 Parcel correlation

We utilized Schaefer400 to divide the cortex into 400 regions and chose 17networks_LH_DefaultA_pCun_1 as our target parcel (Schaefer et al., 2018), located between the precuneus and posterior cingulate cortex. This specific parcel was used as the seed region for the default mode network in previous studies and corresponds to region 14 in the R package ciftiTools (Pham et al., 2022). After Fisher z transforming the parcel correlation matrices, we extracted the lower triangle for statistical analysis.

4.4 Data normalization

4.4.1 Covariates description

In this study, ten covariates related to motion parameters, sociodemographic factors, and disease-specific elements were employed, as detailed in table 4. These covariates were divided into four categories for the purpose of site harmonization and covariate balancing. The first category, represented by M, includes *PropRMSD025* and *Mean RMSD*, which pertain to subject movement in the context of the Ciric criteria. In contrast, for the Powerpt2 criteria, *PropFD02* and *Mean FD* serve as alternative motion covariates. The second category, denoted as W_1, encompasses *AGE*, *SEX*, and *HANDEDNESS*; these demographic variables differ between autistic and typically developing children but are not directly associated with the diagnosis. The third category, represented by W_2, consists of *Stimulant*, *NonStimulant*, *ADOS*, and *FIQ*, which are covariates directly linked to autism diagnosis. Lastly, the fourth category, denoted as A, solely includes the covariate *ASD*.

PropRMSD025 denotes the frequency of occurrences where the relative root mean square displacement is less than 0.25 mm during fMRI testing. *Mean RMSD* signifies the average value of the root mean square displacement throughout the fMRI testing. *PropFD02* denotes the frequency of occurrences where the framewise displacement is less than 0.2 mm during fMRI testing. *Mean FD* signifies the average value of the framewise displacement throughout the fMRI testing. *AGE* indicates the age of the child at the time of fMRI testing. *SEX* indicates the gender of the child. *HANDEDNESS* indicates the dominant hand of the child. *Stimulant* indicates whether the child is taking stimulant medication during fMRI testing. *NonStimulant* indicates whether the child is taking non-stimulant medication during fMRI testing. *ADOS*, *FIQ*, *ASD* were defined in section 4.1.2.

4.4.2 Site harmonization

We utilized ABIDE data from four different sites collected by the Kennedy Krieger Institute (Di Martino et al., 2014) and New York University (Di Martino et al., 2011b). To minimize the potential impact of data acquisition and processing differences between different imaging sites

Group	Covariates
$M(Criteria)$	PropRMSD025, Mean RMSD
$M(Powerpt2)$	PropFD02, Mean FD
W_1	AGE, SEX, HANDEDNESS
W_2	Stimulant, NonStimulant, ADOS, FIQ
A	ASD

Table 4: **Covariates Groups** M denotes the covariates related to motion control in different criteria, W_1 denotes the covariates not related to disease diagnosis, W_2 denotes the covariates related to disease diagnosis, and A denotes the disease diagnosis. In Powerpt2 criteria, we need to change M to PropRMSD02 and Mean FD.

or scanners on our study results, we utilized all the covariates described in Section 4.4.1 to fit a mixed-effects model employed for site harmonization using NeuroCombat (Fortin et al., 2017).

4.4.3 Balancing diagnosis-independent variables

Diagnosis-independent covariates, denoted as $W_1 \nleftrightarrow A$, that are not balanced between groups, can potentially introduce confounding bias. We know $W_1 \to Y$, and its imbalance across the two ASD and TD groups causes the difference in Y between the groups to be confounded. Our aim is to make $W_1 \nrightarrow Y$, allowing us to obtain a more effective outcome model focusing on diagnosis-dependent covariates W_2.

To address potential confounding effects and minimize the impact of motion on functional connectivity, we adjusted the correlations using the following approach. We fit a linear model for each edge that incorporated predictors M, W_1, and A. The model included age, sex, and handedness, as these factors differed between autistic and typically developing children (refer to Section 4.1.1). After fitting the model, we extracted the residuals and added the estimated intercept and effect of primary diagnosis. This method assisted in controlling for mean effects that may vary between the two groups (ASD versus typically developing).

4.5 Impact of motion QC on the sample size and composition

4.5.1 Impact of motion QC on group sample size

Pearson's chi-squared tests were employed to evaluate whether the ASD and typically developing groups differed in the proportion of excluded children when considering the motion exclusion based on RMSD.

4.5.2 rs-fMRI exclusion probability as a function of phenotypes

We employed univariate generalized additive models (GAMs) to examine the association between the log odds of exclusion and phenotype covariates, including ADOS (for the ASD group), FIQ, and age. To ensure adherence to the Ciric motion exclusion criteria, our analysis focused on the children within the final study sample (see Table 1, containing 312 usable and 74 unusable participants). Automatic smoothing through random effects with restricted maximum likelihood estimation (REML) was used to determine the smoothing parameters (Wood, 2017). We opted

for univariate models instead of considering all covariates simultaneously, as some variables were correlated, making it difficult to estimate the impact of each variable on rs-fMRI usability. These models are related to the propensity models used to estimate deconfounded group differences (see Section 4.6). Controlling for these variables yielded highly similar results (not shown). Our focus for this analysis was on interpretable models, although the propensity models would employ an ensemble of machine learning models to predict usability from multiple predictors. We controlled for multiple comparisons using the false discovery rate (FDR) for the three univariate models (Benjamini and Hochberg, 1995). Although FDR correction is typically employed in high-throughput studies in computational biology, Benjamini and Hochberg (1995) originally demonstrated its utility for controlling the expected number of falsely rejected null hypotheses in a study involving a moderate number of tests (15), which is similar to our analysis.

We also conducted univariate analyses using generalized additive models (GAMs) with Gaussian errors to investigate the association between phenotypes and mean RMSD. Separate analyses were performed for the complete study sample (including both usable and unusable cases) and for the subgroups of children (ASD or TD) who passed the Ciric exclusion criteria, respectively. We used the false discovery rate (FDR) correction for the three comparisons within each sample to control for multiple comparisons. Potential sex differences in mean RMSD were also examined using Mann-Whitney U-tests for the three samples.

4.5.3 Impact of motion QC on distributions of phenotypes among children with usable data

We examined potential differences in the distribution of various covariates, such as ADOS, age, and FIQ, between participants included and excluded in the study. To gain further insights into the impact of scan exclusion on autistic versus typically developing children, we stratified the analysis by diagnosis. Kernel density estimation with default bandwidths in ggplot2 was used to visualize the distribution of factors (Wickham, 2016). To test for differences between included and excluded participants, we conducted two-sided Mann-Whitney U tests for each measure, stratified by diagnosis. Additionally, we calculated effect sizes as $Z/\sqrt{N}$. To control for multiple comparisons, we separately applied the false discovery rate to the thirteen tests (3 for the ASD group and 2 for the typically developing group).

4.5.4 Functional connectivity as a function of phenotypes

We explored the association between phenotypes and functional connectivity using univariate generalized additive models (GAMs). For each edge of signal-to-signal components in the partial correlation matrix, we examined the relationship between each phenotypic measure and the adjusted residuals. These residuals were calculated from a linear model that included sex, socioeconomic status, and diagnosis as covariates, with the effect of diagnosis added back in, as explained in Section 4.4.3. To determine smoothing, we used the random effects formulation of spline coefficients

with restricted maximum likelihood estimation (REML) (Wood, 2017).

By examining these associations, we aimed to better understand the potential influence of phenotypic factors on functional connectivity. This approach allows us to identify patterns and relationships between phenotypic measures and connectivity while accounting for confounding variables. The results from these analyses provide valuable insights into the complex interplay between autism, typically developing children, and brain connectivity, ultimately contributing to a more comprehensive understanding of the factors that influence rs-fMRI measurements in these populations.

4.6 Application of AIPWE in abide data

4.6.1 Procedure flow

Our approach for estimating our target involves three steps.

In step 1, we fit a propensity model, denoted as $P(\Delta|A, W_1, W_2)$, to estimate the probability that the rs-fMRI data meet the motion quality control criteria, training the model on all available data.

In step 2, we fit an outcome model, denoted as $E(Y|\Delta = 1, A, W_1, W_2)$, to estimate functional connectivity for participants with usable rs-fMRI data based on their covariates, allowing us to predict and identify functional connectivity for both usable and unusable participants.

In step 3, We utilize the augmented inverse propensity weighted estimator (AIPWE) with the Benjamin-Hochberg correction Glynn and Quinn (2010)).

The super learner technique used in steps 1 and 2 combines multiple regression models and selects weights for each model by minimizing cross-validated risk (Polley et al., 2019). In step 3, We used AIPWE to integrate functional connectivity data from usable subjects weighted by the inverse probability of usability with functional connectivity predictions for all subjects (both usable and unusable), separately for each diagnosis group. The mean functional connectivity is then calculated by integrating across the diagnosis-specific distribution of covariates for both usable and non-usable participants. AIPWE is advantageous as it provides statistically consistent estimates of the deconfounded group difference and its variance, even if the propensity or outcome model is inconsistently estimated (Bang and Robins, 2005).

4.6.2 Procedure details

For step 1 and 2 of our analysis, we employed several learners and R packages with the super learner technique, an ensemble machine learning method. The learners and packages included multivariate adaptive regression splines from the R package earth (Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani., 2011), lasso from glmnet (Friedman et al., 2008), generalized additive models from gam (Hastie and Tibshirani, 2010), generalized linear models from glm, random forests with ranger (Wright and Ziegler, 2017), step-wise regression from step, stepwise regression with interactions, xgboost from (Chen and Guestrin, 2016), and the intercept only (mean) model. For

the outcome model with continuous response, we additionally used ridge from MASS (Venables and Ripley, 2002). The parameters were set to their defaults, except for the family set to binomial (logistic link) in the propensity model, and the method set to minimize the negative log-likelihood. The method was set to minimize the squared error loss in the outcome models. Note that the outcome model is fitted separately for each of the 418 edges, while the same propensities are used for all edges. The propensity model is fit using the complete predictor cases, and the outcome model is fit using the complete usable cases.

The same predictors are used in both the propensity and outcome models, including age at scan, sex, handedness, primary diagnosis, indicator variables for a current prescription for stimulants and non-stimulants, FIQ, and ADOS (only for children in the ASD group). FIQ is missing in one observation, and we use the average of FIQ in the specific subgroup to implement. To balance confounding variables W_1 depicted in section 4.4.3, we fit a linear model including M, W_1, and A, and extracted the residuals and added the estimated intercept and effect of the primary diagnosis.

For each edge, AIPWE is utilized first for the ASD group, then for the TD group, employing both propensities and predicted outcomes to determine the deconfounded mean for each group as well as their respective variances. A $z-$ statistic is computed based on their difference, assuming independent groups, and utilized to test the null hypothesis that there is no difference in functional connectivity between autistic and typically developing children.

As the super learner employs cross-validation, its outcomes may vary due to the random seeds. To address this, we computed the average value of propensity model prediction $g_n = P(\Delta = 1|A, W)$ across two hundred different random seeds. Additionally, we computed the average value of the outcome models prediction $\bar{Q}_n$ across twenty random seeds for all the 418 edges. Then we calculated the AIPWE-based z-statistic for the difference in functional connectivity for each edge from our average prediction of g_n and $\bar{Q}_n$.

4.7 Data and code availability

All the data used in this research can be obtained through ABIDE datasets (Autism Brain Imaging Data Exchange) (Di Martino et al., 2014, 2017). g The code for recreating this study's analyses, tables, and figures is available at https://github.com/thebrisklab/LiangkangThesis.

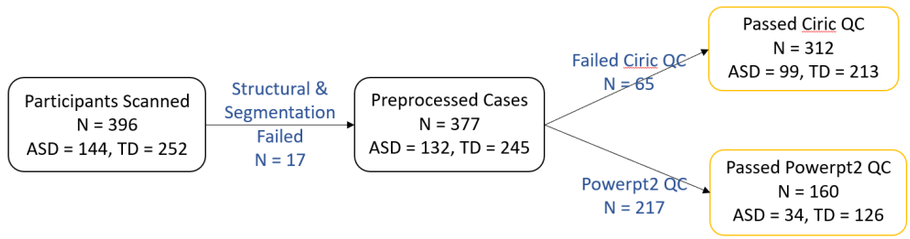
5 Results

5.1 Impact of motion QC on the study sample and sample bias

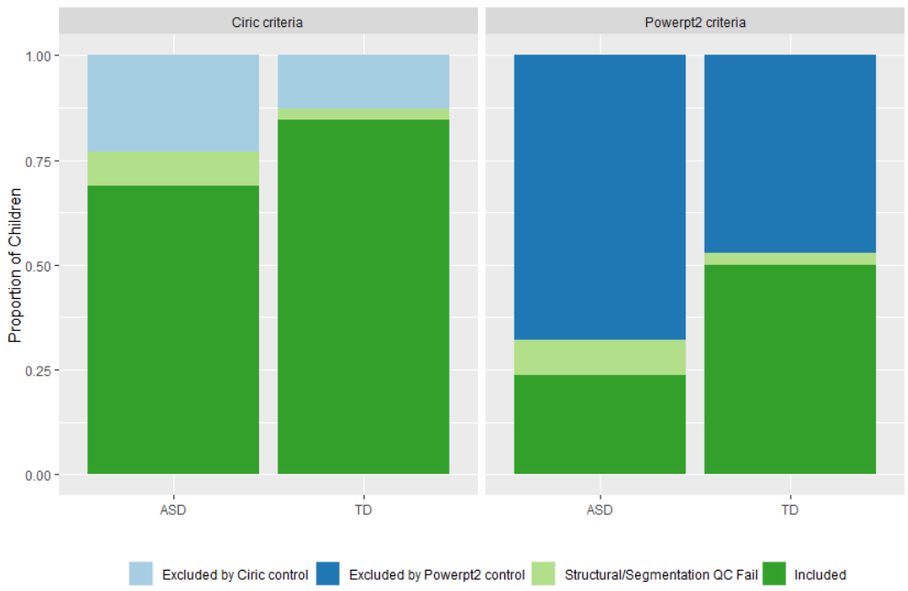
5.1.1 The impact of motion QC on sample size can be dramatic and differs by diagnosis group

Figure 6 presents the criteria employed for the inclusion of participants in our analyses and enumerates the participants excluded at each stage. Out of the total scanned participants, 4.8% (19 participants) were excluded due to preprocessing failures during structural acquisition. The Ciric

criteria resulted in the exclusion of 17.2% of the preprocessed cases, while the Powerpt2 criteria excluded 57.6% of the preprocessed cases. Moreover, upon examining the proportion of excluded participants by diagnosis group using both levels of motion quality control (QC) as depicted in fig.6(b), we discovered that 13.1% of typically developing children were excluded, whereas 25% of children in the autism spectrum disorder (ASD) group were excluded within Ciric criteria ($\chi^2 = 12.7$, df = 1, p = 0.00036). Utilizing Powerpt2 criteria, 48.6% of children in the ASD group were excluded, in contrast to 74.2% of children in the ASD group who were excluded (p=4.6e-07). These findings indicate that the commonly employed Ciric and Powerpt2 motion QC procedures resulted in substantial data loss, particularly for the ASD group.



(a)



(b)

Figure 6: **Motion quality control significantly reduces sample size.** a) Flow chart of inclusion criteria for this study, illustrating the number of participants remaining after each exclusion step. In total, 19 participants (4.8% of the overall number of scanned participants) were excluded due to preprocessing failures during structural acquisition. The Ciric criteria resulted in the exclusion of 17.2% of the preprocessed cases, whereas the Powerpt2 criteria excluded 57.6% of the preprocessed cases. b) The proportion of children in each diagnosis group with scans that were included (green) and excluded (slate blue) across different diagnostic groups. A higher proportion of children in the autism spectrum disorder (ASD) group were excluded compared to typically developing (TD) children when using Ciric criteria ($\chi^2 = 12.715$, df = 1, $p < 0.001$) and Powerpt2 criteria ($p < 4.6e\text{-}07$).

5.1.2 The relationship between rs-fMRI exclusion probability and phenotype and age

In fig. 7, the top panel displays the outcomes of our univariate analyses concerning the relationship between exclusion probability and phenotypes. Utilizing the Benjamini-Hochberg (BH) adjustment for each criterion, no significant association is observed between the rs-fMRI exclusion probability and the three phenotypes (FDR-adjusted p-values > 0.2), which contrasts with the findings of Nebel et al. (2022).

The lower panel of fig. 7 presents the distribution of covariates for each diagnostic group, encompassing both included and excluded participants. Children with autism exhibit a concentrated range of 10-15 for ADOS scores, whereas the age distribution is similar for both groups of children. In comparison to children with autism, typically developing children display a more evenly distributed density curve for FIQ values.

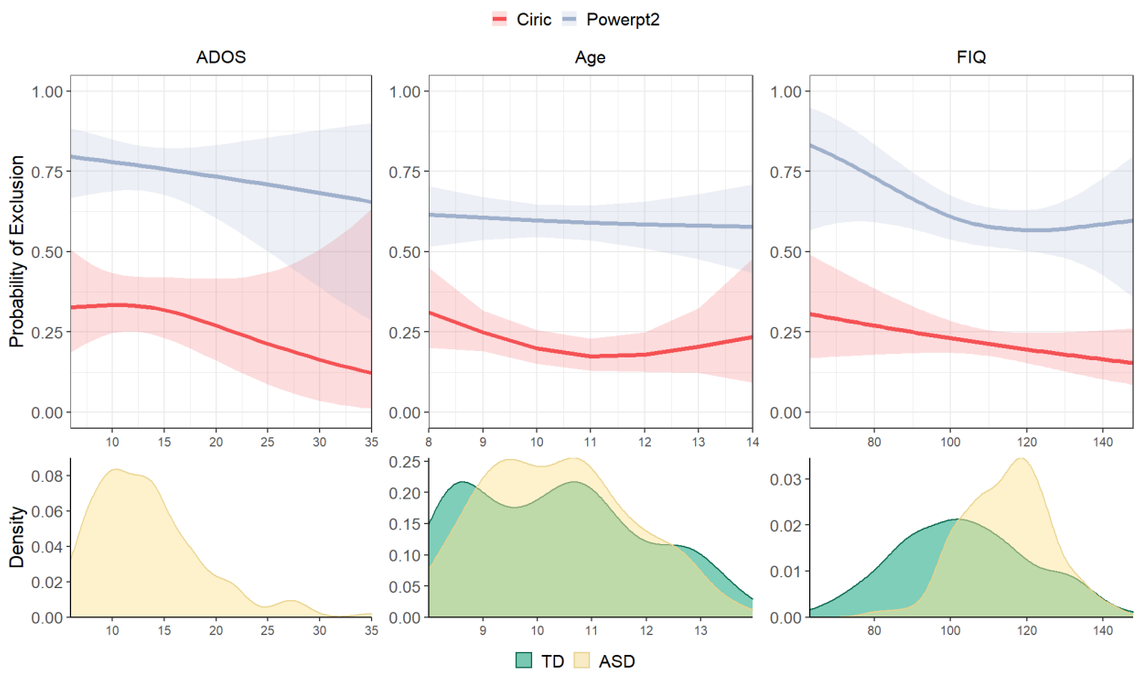


Figure 7: The univariate analysis of rs-fMRI exclusion probability in relation to participant characteristics investigates three variables from left to right: Autism Diagnostic Observation Schedule (ADOS) total scores, age, and full-scale IQ (FIQ). The bottom panel illustrates the variable distributions for each diagnostic group, encompassing participants with both included and excluded scans (TD = typically developing, blue; ASD = autism spectrum disorder, red).

5.1.3 Phenotype representations do not differ between included and excluded children

Figure 8 depicts the covariate distributions for included and excluded participants, stratified by diagnostic group and motion QC level. For both Ciric and Powerpt2 motion QC, median values, effect sizes, and FDR-adjusted p-values for each measure and diagnostic group are provided in the Web Supplement Tables S1 and S2. With the Ciric criteria, no differences were observed in ADOS, age, and FIQ between included and excluded participants. Consequently, we conclude that selection bias in our datasets of phenotypes is negligible.

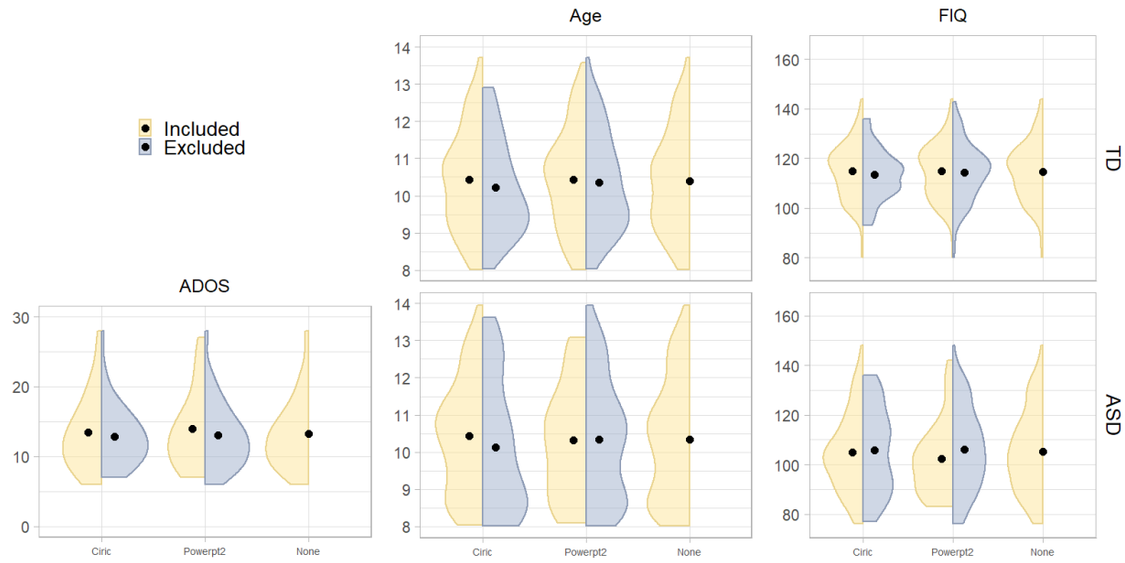


Figure 8: **Comparison of participants with usable and unusable rs-fMRI data.** This figure compares Autism Diagnostic Observation Schedule (ADOS) scores, age, and Full-Scale Intelligence Quotient (FIQ) for included (yellow) and excluded (slate blue) participants, stratified by diagnostic group and motion exclusion level. TD is at the top, and ASD is at the bottom. The deconfounded mean integrates across the diagnosis-specific distribution of usable and unusable covariates for the variables described in Section 4.6, which is labeled as "None" in this figure. Mean values are indicated by a black dot. The R code to produce these split violin plots was adapted from DeBruine (2018).

5.1.4 Phenotypes are also related to functional connectivity

The relationships identified between rs-fMRI data usability and the covariates explored in the earlier analyses might impact our parameter of interest if those variables are also connected to functional connectivity. Fig. 9 displays histograms of p-values for GAMs investigating the association between edgewise functional connectivity (adjusted for sex, age, and motion, as described in Section 4.4.3) and ADOS, along with FIQ, for participants with usable rs-fMRI data using Powerpt2 motion QC (slate blue bins) and Ciric motion QC (red bins). This examination is pertinent to the outcome model applied in the deconfounded group difference, as it provides an understanding of whether the sampling bias would influence the average difference in functional connectivity between the groups. Since we focus on a single phenotype in each GAM for the sake of interpretability, we have not included the analysis of stimulant and nonstimulant drug usage during the fMRI test, as they represent categorical variables in our dataset. In the context of a particular phenotype, clustering of p-values close to zero suggests that a covariate has a stronger association with functional connectivity for a greater number of edges. If no such relationship exists between the covariate and functional connectivity, we would anticipate a more uniform distribution of p-values. For total ADOS, we notice considerable clustering of p-values close to zero among participants with usable rs-fMRI data under both Ciric and Powerpt2 motion QC.

5.2 Application: Deconfounded Group Differences in the KKI and NYU Dataset

We assessed the average propensity scores' consistency. Propensities near zero can elevate both the bias and variance of causal effects (Petersen et al., 2010) and may suggest a potential breach of the

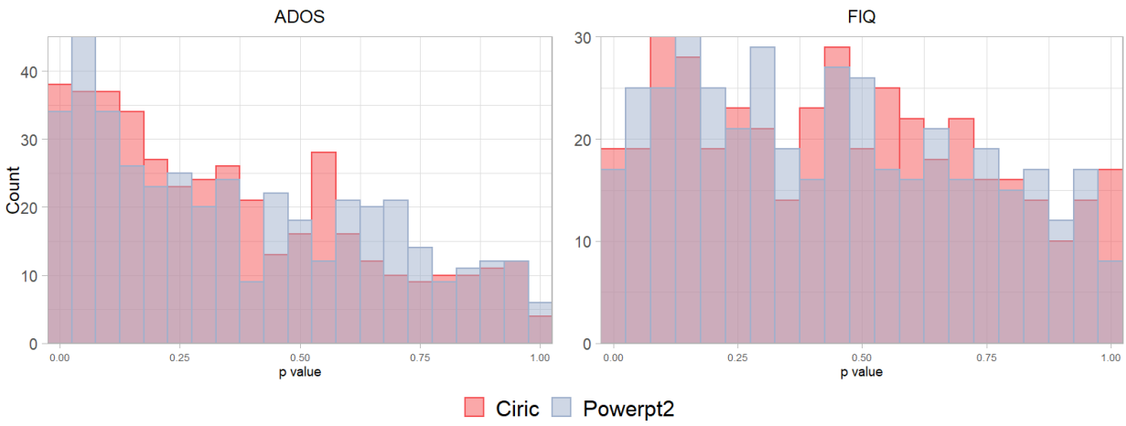


Figure 9: **Some covariates related to rs-fMRI exclusion probability are also related to functional connectivity.** Histograms of p values for generalized additive models of the relationship between edgewise functional connectivity in participants with usable rs-fMRI data and (from left to right) he total scores of the Autism Diagnostic Observation Schedule (ADOS), and full-scale IQ (FIQ). For a given covariate, a clustering of p values near zero suggests that covariate is associated with functional connectivity for a greater number of edges. ADOS appear to be related to functional connectivity using both the Powerpt2 motion quality control (lavender bins) and the Ciric motion quality control (red bins).

positivity assumption (A1.2). The propensity spanned 0.23-0.60 (Powerpt2 criteria) and 0.61-0.88 (Ciric criteria), indicating a reasonable likelihood of data inclusion across the range of A, W and suggesting that Assumption (A1.2) is likely sufficiently met.

Using the framework (see in Section 4.6) to analyze group differences in the KKI and NYU dataset, we set significance levels at $\alpha = 0.05$ and $\alpha = 0.2$ and employed the naive test (participant exclusion with Welch t-tests and BH adjustment for FWER control), max statistic permutation-based AIPWE (detailed in section 2.3), and AIPWE with BH adjustment. These methods were also simulated and discussed in section 4.3.2. The effectiveness of these tests when applied to real data is evident in the analysis of the ABIDE I/ABIDE II datasets (Di Martino et al., 2014, 2017).

Figures 10, 11, and 12 present the differential functional connectivity linked to our seed region between autistic children and typically developing children. The left column plots display all z-statistics in the brain maps, while the right two columns show the significant regions identified by different tests.

Figure 10 indicates that, without participant removal, the naive test identifies the most significant regions. Nonetheless, this test includes false significant regions due to insufficient motion control during analysis, increasing its false discovery rate. When applying the Ciric criteria, the naive test exhibits moderate power at both α levels. Conversely, the naive test using Powerpt2 criteria detects only one significant region at both $\alpha = 0.05$ and $\alpha = 0.2$. As a stringent criterion, Powerpt2 excludes 57.6% of preprocessed cases (refer to Section 5.1.1), resulting in fewer significant regions.

Figure 11 demonstrates that max perm AIPWE uncovers two more significant regions than the naive tests at $\alpha = 0.2$. Considering the simulation findings with weak block correlation in section 3.2, AIPWE max perm outperforms the naive tests for sample sizes between 200-500, aligning with our empirical data analysis. BH-adjusted AIPWE reveals more significant regions than max

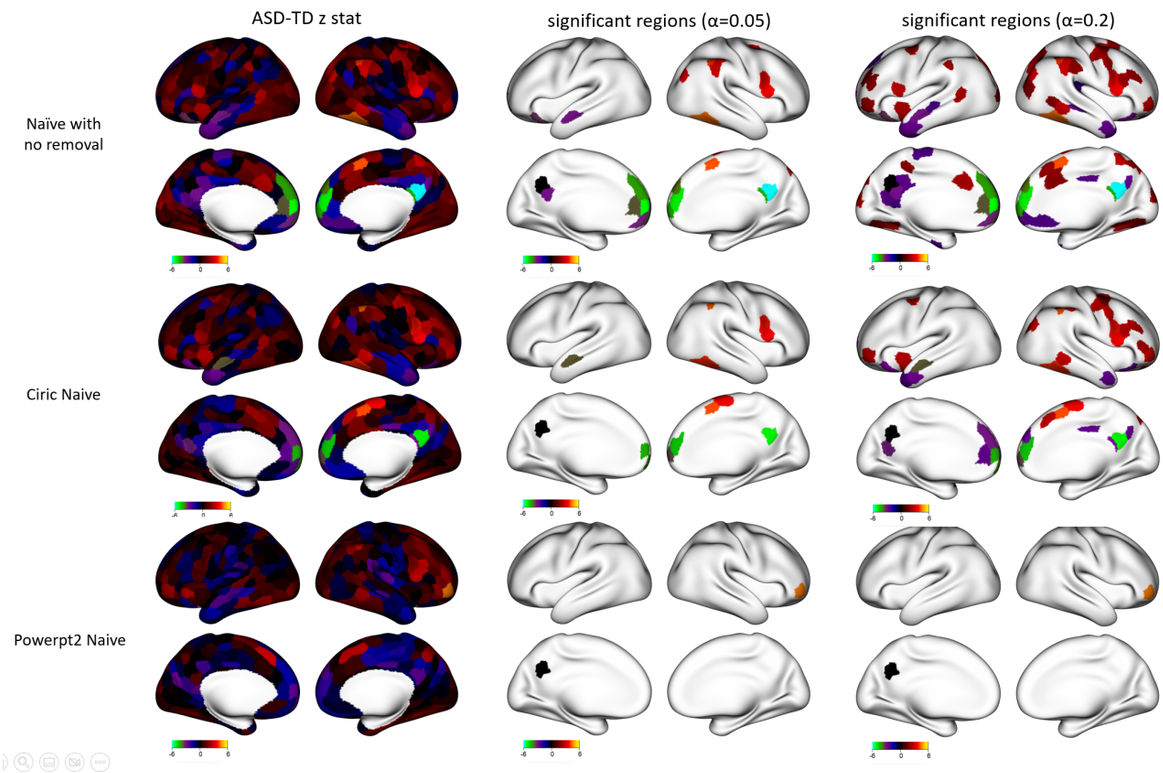


Figure 10: **Naive test with different motion control** The black area in the left hemisphere serves as our seed region, described in Section 3.2. Naive tests with no motion control has more false significant areas than the naive tests with Ciric and Powerpt2 criteria.

perm AIPWE due to its higher family-wise error rate, as depicted in fig. 5(a). We can infer that BH-adjusted AIPWE is a more powerful method, although it has a higher family-wise error rate. If better FWER control is desired, max perm AIPWE can be used instead.

Figure 12 yields similar conclusions to those found in the analysis with Powerpt2 motion QC. However, the naive test uncovers more significant regions than max perm AIPWE. The Ciric criteria is a lenient motion QC that only reduces 17.2% of preprocessed cases, retaining a larger sample size in the naive tests. Consequently, the utility of AIPWE is diminished, leading to fewer significant regions at $\alpha = 0.2$. When $\alpha = 0.05$, BH-adjusted AIPWE is a considerably more powerful test than max perm and naive tests. Furthermore, given the outcomes in Section 5.1.2, no significant selection bias is present in our datasets, which explains the superior performance of naive tests in real data compared to our simulations.

When selecting a method for an unfamiliar dataset with unknown selection bias, we recommend using BH-adjusted AIPWE because it offers greater power and a comparatively smaller FWER based on our real data analysis. This approach can provide more reliable results and minimize the impact of selection bias. However, it is essential to consider the specific context of the dataset and research question when deciding which method to employ, as the optimal choice may vary depending on the dataset characteristics and objectives of the study. By evaluating the methods presented in this paper, researchers can make informed decisions and achieve more accurate and reliable outcomes in their analyses of group differences in functional connectivity.

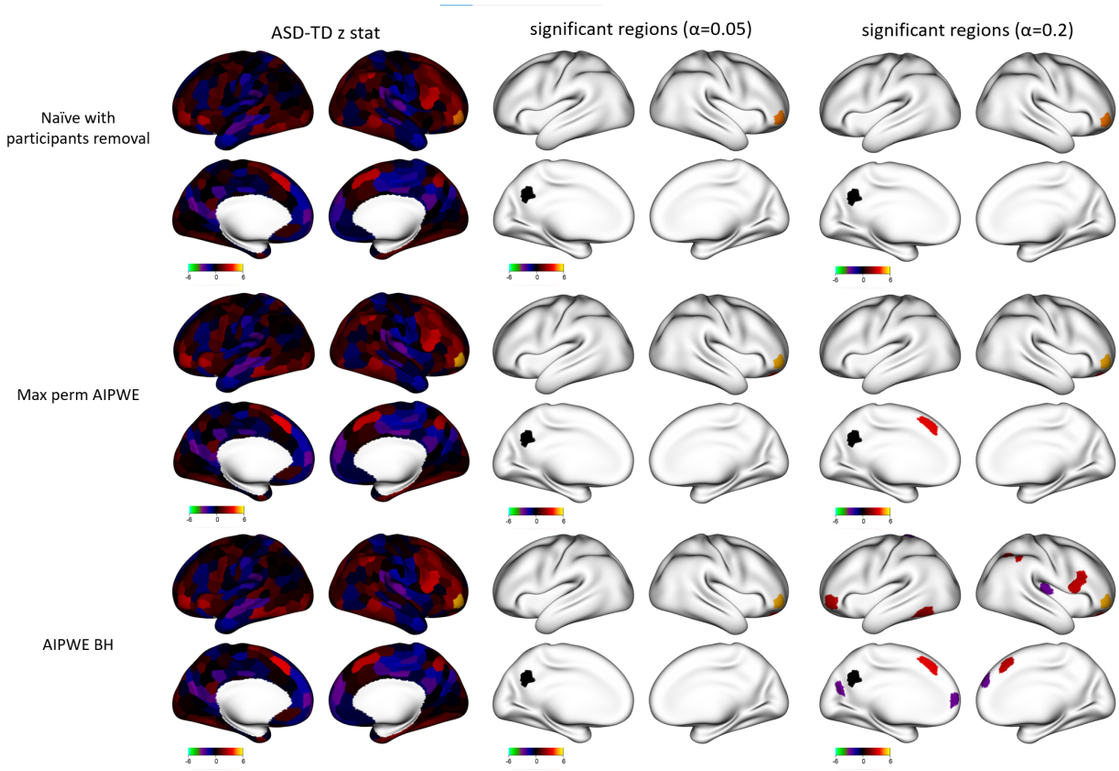


Figure 11: **Different tests with powerpt2 criteria** The black area in the left hemisphere serves as our seed region, described in Section 3.2. BH-adjusted AIPWE is more powerful when more than 60% of observations were excluded by motion control.

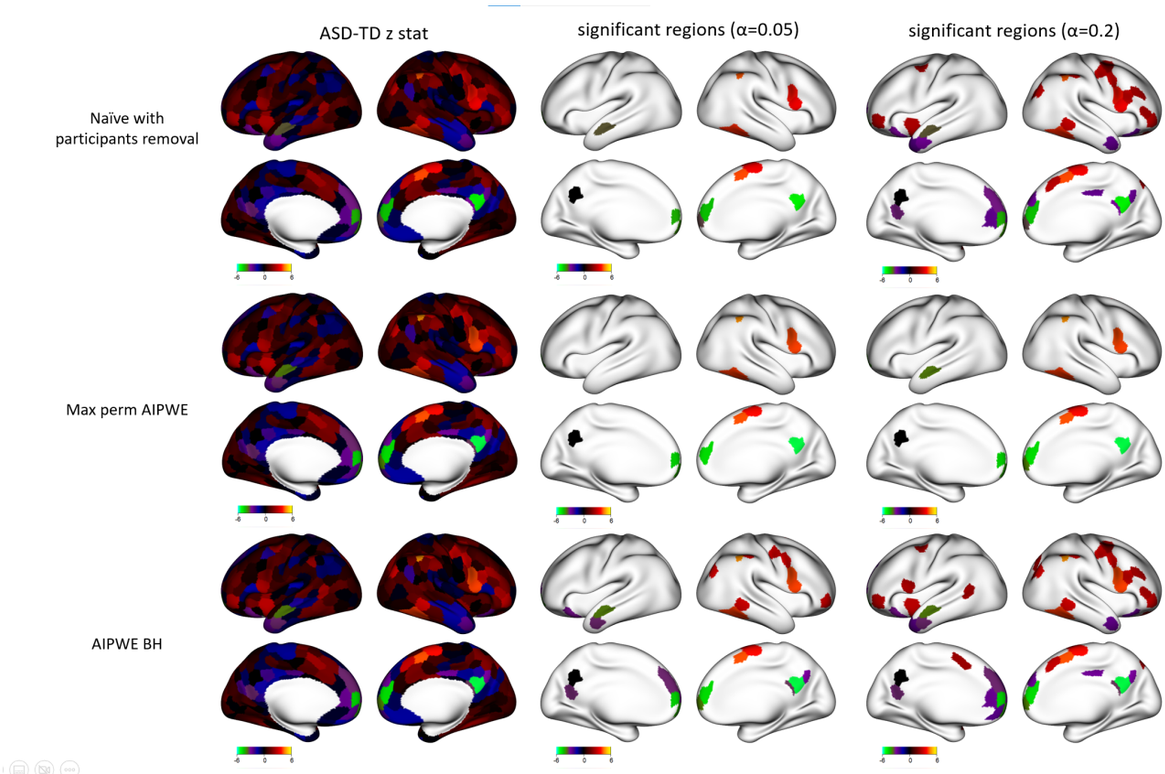


Figure 12: **Different tests with ciric criteria** The black area in the left hemisphere serves as our seed region, described in Section 3.2. Max statistic permutation-based AIPWE is more conservative.

6 Discussion

6.1 Differences between simulation setting 1 and real data

We aimed to use simulations to replicate real data, focusing on the impact of the diagnostic variable A and the $ADOS$ covariate. We employed a linear model to specify the outcome model, as described in Section 3.1. By implementing this simulation design, we successfully achieved a correlation of -0.58 between W_c and Y, along with a "true" functional connectivity, denoted by $E[Y(1)|A = a]$, approximating -0.20 in the ASD group and 0 in the typically developing group. As a result, this yielded a $Cohen'sd$ value of 0.39 when comparing the two groups.

However, our analysis revealed a weak correlation between Autism Diagnostic Observation Schedule (ADOS) scores and partial correlations within our dataset, exhibiting minimum and maximum values of -0.01 and 0.06, respectively, across 418 edges for children with usable data under our motion quality control (QC) measures. Furthermore, we discovered that the largest naïve $Cohen'sd$ equated to 0.51. In a study examining sleeping autistic toddlers, correlations between functional connectivity and ADOS reached as high as -0.78 in specific subgroups, and certain effect sizes were substantial (exceeding 0.8) (Lombardo et al., 2019). This comparison highlights the differences between our simulation setting and real-world observations.

It is important to emphasize that the naïve difference and deconfounded group difference in the actual data analysis exhibit greater similarity than illustrated in this simplified example 2(b). Nonetheless, this example serves to highlight the potential influence of selection bias within a realistic experimental context.

6.2 Motion quality control bias

In our investigation, the influence of motion quality control (QC) on sample size was substantial and varied across diagnostic groups. Although comprehensive reporting of participant exclusion due to excessive head movement is not standard practice, we observed that motion QC led to a higher proportion of autistic children being removed compared to typically developing children. This observation aligns with the findings reported in Redcay et al. (2013), and Jones et al. (2010). Contrasting with the results from Nebel et al. (2022), our dataset revealed no selection bias, as diagnosis-dependent covariates (ADOS, FIQ), and age did not significantly impact the likelihood of a child being excluded during motion control. The distribution of these three covariates was consistent across excluded and included groups (see Section 5.1.2).

The average functional connectivity estimation should represent all children participating in the study, assuming that the participants accurately reflect the target population. Given the varying definitions of usability across resting-state fMRI (rs-fMRI) studies, our findings imply that the differing representation of symptom severity among children with usable data post-motion QC may have contributed to inconsistencies in ASD-related functional connectivity findings in the literature. To enhance comparability across studies, it is essential for rs-fMRI researchers to transparently

evaluate the information loss after motion QC (fig.6), and examine if participant characteristics associated with usability are also related to the effect of interest, and address the potential bias and power loss, if applicable.

We have addressed this issue by employing techniques from missing data and causal inference literature, combined with a collection of machine learning algorithms. Our framework treats missingness due to motion QC as a source of bias, and we define a target parameter known as the deconfounded group difference. This parameter uses the distribution of diagnosis-specific behavioral variables across usable and unusable scans. The framework's general concept is to acknowledge that children with usable data may not accurately represent all enrolled children within each diagnostic group. AIPWE combines the results of inverse propensity weighting and G-computation, enhancing robustness compared to either approach individually. Inverse propensity weighting assigns more weight to children more likely to be missing, as functional connectivity is related to symptom severity. Consequently, we need these children to represent all those with more severe symptoms who were excluded due to data quality issues. The outcome model estimates functional connectivity for all children, including those with greater symptom severity, thus accounting for children with unusable data. We employ a set of machine learning methods to model potential non-linear relationships between phenotypic traits and data usability (the propensity model) and between phenotypic traits and functional connectivity (the outcome model) flexibly. We include a comprehensive set of variables in both the propensity and outcome models that may be associated with rs-fMRI usability, functional connectivity, or both. Incorporating variables that contribute to both rs-fMRI usability and functional connectivity provides an opportunity to reduce bias. Including variables that contribute to functional connectivity but not necessarily to rs-fMRI usability offers a chance to decrease the variance of our estimate without increasing bias. AIPWE is then used to combine the propensity and outcome models, resulting in statistically consistent estimation of the deconfounded group difference and its variances under the assumptions in Section 2.1 and as discussed in Section 6.3.

6.3 Assumptions and Potential Violations

To estimate the difference in functional connectivity between autistic and typically developing children in a counterfactual scenario where all data is usable, we rely on three assumptions: mean exchangeability, positivity, and consistency between the counterfactual and observed outcomes (causal consistency) (Section 2.1).

Regarding mean exchangeability, or the assumption of no unmeasured confounders, we presume that functional connectivity is independent of the missingness mechanism, given our variables W, A. As stated earlier, the missingness mechanism is deterministic based on head motion; however, we substitute it with a stochastic model that estimates missingness from W, A. In our study, it is crucial that we do not include summary measures of head motion in the propensity and outcome models. This is because children who nearly fail motion QC may still have motion impacts on

their functional connectivity signal. The deconfounded group difference assumes that Y represents the signal of interest, i.e., neural sources of variation that are not influenced by motion. We took several measures to account for potential motion impacts on functional connectivity in children who almost fail; we used the residuals from a linear model including motion, as described in Sections 4.4.2, resulting in a Y that more accurately reflects neural sources of variation. However, if we then incorporated summary motion measures into our propensity and outcome models, the propensity model would up-weight these children who almost failed, and the outcome model, integrating over the full range of head motion, would potentially reintroduce the motion impacts we aimed to remove. Moreover, our statistical estimator possesses the double robustness property: if at least one of the propensity or outcome models is correctly specified, we obtain a statistically consistent estimator of the deconfounded group difference (Glynn and Quinn, 2010). We include a comprehensive set of predictors and a collection of machine learning algorithms to help address the assumption of no unmeasured confounding.

Positivity assumes that no values of W, A will always render the data unusable. Violations of positivity assumptions result in out-of-sample prediction of functional connectivity in the outcome model and instabilities in the propensity model, potentially leading to increased variance and bias (Petersen et al., 2010). In fig. 8, we observe that for the Ciric criteria, the range of behavioral traits generally overlap between included and excluded participants, although the most severe ADOS score is not present among the excluded children.

The maximum value of ADOS in excluded data (Ciric criteria) is 28, while the maximum value of ADOS in included data (Ciric criteria) is 35. The positivity assumption (see in Section 2.1) is confirmed. As reported in Section 5.2, the lowest average propensity score is 0.23. The absence of propensities close to zero for children with usable or unusable data suggests that the assumption of positivity is reasonable in our study. Concerning the final assumption, causal consistency is a technical assumption that presumes that $Y(1)$ is the same as Y when a child has usable data, which generally cannot be tested but appears reasonable.

In the permutation test (Section 2.3), our objective is to maintain all connections between $(A, W) \to \Delta$, while disrupting the direct and indirect paths of A: $(A \leftrightarrow W) \nrightarrow Y$ in the permutation test. However, by permuting the A label and using the original $\hat{Q}_n(A_i, W_i)$ and $g_n(A_i, W_i)$ to calculate the $z_j^{(k)}$ as the outcome of the k-th permutation, we violate the exchangeability of the permutation test. Consequently, we observe inflated type 1 error and family-wise error rate for the permutation-based AIPWE in the simulation for single region (section 3.1) and multiple regions (section 3.2).

6.4 Overview and outlook

In this paper, we have presented an approach to account for selection bias and improve statistical power in fMRI studies. By employing DRTMLE and AIPWE, we demonstrated the potential to reduce sample bias and enhance power in functional connectivity analysis. In the context of smaller

sample sizes simulation, we compared the performance of AIPWE to DRTMLE to determine the relative efficacy of each method.

Our simulations revealed that AIPWE exhibited better type-1 error control than DRTMLE in most settings, though some inflation was observed. To address this issue, we proposed a computationally scalable permutation test, which demonstrated mixed results in controlling type-1 error while maintaining power.

We applied our methods to real data from school-age children in the Autism Brain Imaging Data Exchange. With 34 usable scans from autistic children, we identified significant differences between ASD and typically developing (TD) children using AIPWE. However, further research is necessary to disentangle false positives from true positives in these findings.

Our preliminary results using bootstraps show promise, and we are currently working on enhancing computational scalability to facilitate broader application of these methods. In future work, we plan to examine other applications where selection bias is expected, such as ABCD developmental differences between boys and girls, cortical thickness studies in Alzheimer's disease, ADHD, and a prospective autism study in the brisklab with richer phenotyping.

These additional analyses will provide further insights into the effectiveness of our proposed methods in different contexts and contribute to the development of more accurate and reliable approaches for addressing selection bias and improving statistical power in fMRI studies. By refining these techniques and extending their applications, we aim to advance the field of neuroimaging research and facilitate a better understanding of the brain's functional connectivity and its implications in various disorders and conditions.

7 Acknowledgement

This research was supported by the National Institute of Mental Health of the National Institutes of Health under award number R01 MH129855. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Abraham, A. et al. Nilearn: Machine learning for neuro-imaging in Python.

Abramian, D., Blystad, I., and Eklund, A. (2022). Evaluation of inverse treatment planning for gamma knife radiosurgery using fmri brain activation maps as organs at risk. *medRxiv*.

Avants, B. B., Tustison, N., Song, G., et al. (2009). Advanced normalization tools (ants). *Insight j*, 2(365):1–35.

Bang, H. and Robins, J. M. (2005). Doubly Robust Estimation in Missing Data and Causal Inference Models. *Biometrics*, 61(4):962–973.

Bednarz, H. M., Maximo, J. O., Murdaugh, D. L., O'Kelley, S., and Kana, R. K. (2017). "decoding versus comprehension": Brain responses underlying reading comprehension in children with autism. *Brain and language*, 169:39–47.

Behzadi, Y., Restom, K., Liau, J., and Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *Neuroimage*, 37(1):90–101.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, pages 289–300.

Benkeser, D., Carone, M., Laan, M. J. V. D., and Gilbert, P. B. (2017). Doubly robust nonparametric inference on the average treatment effect. *Biometrika*, 104(4):863–880.

Biswal, B., Yetkin, F. Z., Haughton, V. M., and Hyde, J. S. (1995). Functional connectivity in the motor cortex of resting human brain using echo-planar mri. *Magn Reson Med*, 34(4):537–41.

Chen, T. and Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 785–794, New York, NY, USA. ACM.

Ciric, R., Wolf, D. H., Power, J. D., Roalf, D. R., Baum, G. L., Ruparel, K., Shinohara, R. T., Elliott, M. A., Eickhoff, S. B., Davatzikos, C., Gur, R. C., Gur, R. E., Bassett, D. S., and Satterthwaite, T. D. (2017). Benchmarking of participant-level confound regression strategies for the control of motion artifact in studies of functional connectivity. *NeuroImage*, 154:174–187.

Dajani, D. R. and Uddin, L. Q. (2016). Local brain connectivity across development in autism spectrum disorder: A cross-sectional investigation. *Autism Res*, 9(1):43–54.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999a). Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

Dale, A. M., Fischl, B., and Sereno, M. I. (1999b). Cortical surface-based analysis. i. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–194.

DeBruine, L. (2018). Plot comparison. `https://www.debruine.github.io/post/plot-comparison`. [blog post].

Deen, B. and Pelphrey, K. (2012). Perspective: Brain scans need a rethink.

Di Martino, A., Kelly, C., Grzadzinski, R., Zuo, X. N., Mennes, M., Mairena, M. A., Lord, C., Castellanos, F. X., and Milham, M. P. (2011a). Aberrant striatal functional connectivity in children with autism. *Biological Psychiatry*, 69(9):847–856.

Di Martino, A., Kelly, C., Grzadzinski, R., Zuo, X.-N., Mennes, M., Mairena, M. A., Lord, C., Castellanos, F. X., and Milham, M. P. (2011b). Aberrant striatal functional connectivity in children with autism. *Biol. Psychiatry*, 69(9):847–856.

Di Martino, A., O'Connor, D., Chen, B., Alaerts, K., Anderson, J. S., Assaf, M., Balsters, J. H., Baxter, L., Beggiato, A., Bernaerts, S., Blanken, L. M. E., Bookheimer, S. Y., Braden, B. B., Byrge, L., Castellanos, F. X., Dapretto, M., Delorme, R., Fair, D. A., Fishman, I., Fitzgerald, J., Gallagher, L., Keehn, R. J. J., Kennedy, D. P., Lainhart, J. E., Luna, B., Mostofsky, S. H., Müller, R.-A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Solomon, M., Toro, R., Vaidya, C. J., Wenderoth, N., White, T., Craddock, R. C., Lord, C., Leventhal, B., and Milham, M. P. (2017). Enhancing studies of the connectome in autism using the autism brain imaging data exchange II. *Sci. Data*, 4:170010.

Di Martino, A., Yan, C.-G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Müller, R.-A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H., and Milham, M. P. (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.

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., Anderson, J. S., Assaf, M., Bookheimer, S. Y., Dapretto, M., Deen, B., Delmonte, S., Dinstein, I., Ertl-Wagner, B., Fair, D. A., Gallagher, L., Kennedy, D. P., Keown, C. L., Keysers, C., Lainhart, J. E., Lord, C., Luna, B., Menon, V., Minshew, N. J., Monk, C. S., Mueller, S., Müller, R. A., Nebel, M. B., Nigg, J. T., O'Hearn, K., Pelphrey, K. A., Peltier, S. J., Rudie, J. D., Sunaert, S., Thioux, M., Tyszka, J. M., Uddin, L. Q., Verhoeven, J. S., Wenderoth, N., Wiggins, J. L., Mostofsky, S. H., and Milham, M. P. (2013). The autism brain imaging data exchange: towards a large-scale evaluation of the intrinsic brain architecture in autism. *Molecular Psychiatry 2014 19:6*, 19(6):659–667.

D'Souza, N. S., Nebel, M. B., Crocetti, D., Robinson, J., Wymbs, N., Mostofsky, S. H., and Venkataraman, A. (2021). Deep sr-DDL: Deep structurally regularized dynamic dictionary learn-

ing to integrate multimodal and dynamic functional connectomics data for multidimensional clinical characterizations. *Neuroimage*, 241:118388.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., and Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nat. Methods*, 16(1):111–116.

Fassbender, C., Mukherjee, P., and Schweitzer, J. B. (2017). Reprint of: Minimizing noise in pediatric task-based functional MRI; Adolescents with developmental disabilities and typical development. *NeuroImage*, 154:230–239.

Fischl, B. (2012). FreeSurfer. *Neuroimage*, 62(2):774–781.

Fonov, V., Evans, A., McKinstry, R., Almli, C., and Collins, D. (2009). Unbiased nonlinear average age-appropriate brain templates from birth to adulthood. *NeuroImage*, 47:S102. Organization for Human Brain Mapping 2009 Annual Meeting.

Fortin, J.-P., Parker, D., Tunç, B., Watanabe, T., Elliott, M. A., Ruparel, K., Roalf, D. R., Satterthwaite, T. D., Gur, R. C., Gur, R. E., Schultz, R. T., Verma, R., and Shinohara, R. T. (2017). Harmonization of multi-site diffusion tensor imaging data. *Neuroimage*, 161:149–170.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441.

Glasser, M. F., Sotiropoulos, S. N., Wilson, J. A., Coalson, T. S., Fischl, B., Andersson, J. L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J. R., Van Essen, D. C., Jenkinson, M., and WU-Minn HCP Consortium (2013). The minimal preprocessing pipelines for the human connectome project. *Neuroimage*, 80:105–124.

Glynn, A. N. and Quinn, K. M. (2010). An introduction to the augmented inverse propensity weighted estimator. *Political Analysis*, 18(1):36–56.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., and Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Front Neuroinform*, 5.

Greene, D. J., Koller, J. M., Hampton, J. M., Wesevich, V., Van, A. N., Nguyen, A. L., Hoyt, C. R., McIntyre, L., Earl, E. A., Klein, R. L., Shimony, J. S., Petersen, S. E., Schlaggar, B. L., Fair, D. A., and Dosenbach, N. U. F. (2018). Behavioral interventions for reducing head motion during MRI scans in children. *Neuroimage*, 171:234–245.

Hastie, T. and Tibshirani, R. (2010). ProDenICA: Product Density Estimation for ICA using tilted Gaussian density estimates.

Hernan, M. and Robins, J. (2020). *Causal Inference: What If*. Chapman Hall/CRC, Boca Raton.

Hus, V., Gotham, K., and Lord, C. (2014). Standardizing ADOS domain scores: separating severity of social affect and restricted and repetitive behaviors. *J Autism Dev Disord*, 44(10):2400–2412.

Inoue, M., Honma, T., Saitoh, T., Suyama, T., Hamada, M., Matsuki, K., and Hasegawa, S. (1988). [Lung function and morphometry in a canine model of papain-induced emphysema]. *Nihon Kyobu Shikkan Gakkai Zasshi*, 26(11):1161–1169.

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, 17(2):825–841.

Jenkinson, M., Beckmann, C. F., Behrens, T. E. J., Woolrich, M. W., and Smith, S. M. (2012). FSL. *Neuroimage*, 62(2):782–790.

Jones, T. B., Bandettini, P. A., Kenworthy, L., Case, L. K., Milleville, S. C., Martin, A., and Birn, R. M. (2010). Sources of group differences in functional connectivity: an investigation applied to autism spectrum disorder. *Neuroimage*, 49(1):401–414.

Keown, C. L., Shih, P., Nair, A., Peterson, N., Mulvey, M. E., and Müller, R. A. (2013). Local functional overconnectivity in posterior brain regions is associated with symptom severity in autism spectrum disorders. *Cell Reports*, 5(3):567–572.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., me, Y., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., and Keshavan, A. (2017). Mindboggling morphometry of human brains. *PLoS Comput Biol*, 13(2):e1005350.

Lake, E. M. R., Finn, E. S., Noble, S. M., Vanderwal, T., Shen, X., Rosenberg, M. D., Spann, M. N., Chun, M. M., Scheinost, D., and Constable, R. T. (2019). The Functional Brain Organization of an Individual Allows Prediction of Measures of Social Abilities Transdiagnostically in Autism and Attention-Deficit/Hyperactivity Disorder. *Biol Psychiatry*, 86(4):315–326.

Lanczos, C. (1956). Evaluation of noisy data. *Journal of the Society for Industrial and Applied Mathematics*, 4(1):76–85.

Lombardo, M. V., Eyler, L., Moore, A., Datko, M., Barnes, C. C., Cha, D., Courchesne, E., and Pierce, K. (2019). Default mode-visual network hypoconnectivity in an autism subtype with pronounced social visual engagement difficulties. *eLife*, 8.

Lord, C. and Jones, R. M. (2012). Annual research review: re-thinking the classification of autism spectrum disorders. *J Child Psychol Psychiatry*, 53(5):490–509.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., and Rutter, M. (2000). The autism diagnostic observation schedule-generic: a standard measure of social and communication deficits associated with the spectrum of autism. *J Autism Dev Disord*, 30(3):205–223.

Milborrow. Derived from mda:mars by T. Hastie and R. Tibshirani., S. (2011). *earth: Multivariate Adaptive Regression Splines*. R package.

Muschelli, J., Nebel, M. B., Caffo, B. S., Barber, A. D., Pekar, J. J., and Mostofsky, S. H. (2014). Reduction of motion-related artifacts in resting state fMRI using aCompCor. *NeuroImage*, 96:22–35.

Nebel, M. B., Lidstone, D. E., Wang, L., Benkeser, D., Mostofsky, S. H., and Risk, B. B. (2022). Accounting for motion in resting-state fMRI: What part of the spectrum are we characterizing in autism spectrum disorder? *NeuroImage*, 257:119296.

Nichols, T. E. and Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum Brain Mapp*, 15(1):1–25.

Parkes, L., Fulcher, B., Yücel, M., and Fornito, A. (2018). An evaluation of the efficacy, reliability, and sensitivity of motion correction strategies for resting-state functional MRI. *NeuroImage*, 171.

Petersen, M. L., Porter, K. E., Gruber, S., Wang, Y., and Laan, M. J. v. d. (2010). Diagnosing and responding to violations in the positivity assumption:. *Statistical Methods in Medical Research*, 21(1):31–54.

Pham, D. D., Muschelli, J., and Mejia, A. F. (2022). ciftitools: A package for reading, writing, visualizing, and manipulating CIFTI files in R. *Neuroimage*, 250(118877):118877.

Polley, E., LeDell, E., Kennedy, C., and van der Laan, M. (2019). SuperLearner: Super Learner Prediction. *R package v. 2.0-26*.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, 59(3):2142–2154.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2013). Steps toward optimizing motion artifact removal in functional connectivity MRI; a reply to carp. *Neuroimage*, 76:439–441.

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., and Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, 84:320–341.

Power, J. D., Plitt, M., Laumann, T. O., and Martin, A. (2017). Sources and implications of whole-brain fMRI signals in humans. *NeuroImage*, 146:609–625.

Pruim, R. H. R., Mennes, M., Buitelaar, J. K., and Beckmann, C. F. (2015). Evaluation of ICA-AROMA and alternative strategies for motion artifact removal in resting state fMRI. *Neuroimage*, 112:278–287.

Redcay, E., Moran, J. M., Mavros, P. L., Tager-Flusberg, H., Gabrieli, J. D. E., and Whitfield-Gabrieli, S. (2013). Intrinsic functional network organization in high-functioning adolescents with autism spectrum disorder. *Front. Hum. Neurosci.*, 7:573.

Rudie, J. D. and Dapretto, M. (2013). Convergent evidence of brain overconnectivity in children with autism?

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and others (2013a). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, 64:240–256.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., Eickhoff, S. B., Hakonarson, H., Gur, R. C., Gur, R. E., and Wolf, D. H. (2013b). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, 64:240–256.

Satterthwaite, T. D., Wolf, D. H., Loughead, J., Ruparel, K., Elliott, M. A., Hakonarson, H., Gur, R. C., and Gur, R. E. (2012). Impact of in-scanner head motion on multiple measures of functional connectivity: relevance for studies of neurodevelopment in youth. *Neuroimage*, 60(1):623–632.

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., Eickhoff, S. B., and Yeo, B. T. T. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cereb. Cortex*, 28(9):3095–3114.

Supekar, K., Uddin, L. Q., Khouzam, A., Phillips, J., Gaillard, W. D., Kenworthy, L. E., Yerys, B. E., Vaidya, C. J., and Menon, V. (2013). Brain Hyperconnectivity in Children with Autism and its Links to Social Deficits. *Cell Reports*, 5(3):738–747.

Tobyne, S. M., Boratyn, D., Johnson, J. A., Greve, D. N., Mainero, C., and Klawiter, E. C. (2016). A surface-based technique for mapping homotopic interhemispheric connectivity: Development, characterization, and clinical application. *Hum. Brain Mapp.*, 37(8):2849–2868.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., and Gee, J. C. (2010). N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*, 29(6):1310–1320.

Uddin, L. Q., Supekar, K., Lynch, C. J., Khouzam, A., Phillips, J., Feinstein, C., Ryali, S., and Menon, V. (2013). Salience network-based classification and prediction of symptom severity in children with autism. *JAMA Psychiatry*, 70(8):869–879.

Van Dijk, K. R., Sabuncu, M. R., and Buckner, R. L. (2012). The influence of head motion on intrinsic functional connectivity MRI. *Neuroimage*, 59(1):431–438.

Venables, W. and Ripley, B. (2002). *Modern Applied Statistics with S*. Springer.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children-WISC-IV*. Psychological Corporation.

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wood, S. (2017). *Generalized additive models: an introduction with R*. CRC Press.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., Beckmann, C., Jenkinson, M., and Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, 45(1 Suppl):S173–86.

Wright, M. N. and Ziegler, A. (2017). ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17.

Wymbs, N. F., Nebel, M. B., Ewen, J. B., and Mostofsky, S. H. (2021). Altered Inferior Parietal Functional Connectivity is Correlated with Praxis and Social Skill Performance in Children with Autism Spectrum Disorder. *Cereb Cortex*, 31(5):2639–2652.

Zeileis, A. (2006). Object-oriented computation of sandwich estimators. *Journal of Statistical Software*, 16(9):1–16.