Washington University in St. Louis

Washington University Open Scholarship

Arts & Sciences Electronic Theses and Dissertations

Arts & Sciences

Spring 5-13-2024

Assessing Reproducibility of Brain-Behavior Associations Using Bootstrap Aggregation Methods

ZHETAO CHEN
Washington University in St. Louis

Follow this and additional works at: https://openscholarship.wustl.edu/art_sci_etds

Part of the Biostatistics Commons, Data Science Commons, Longitudinal Data Analysis and Time Series Commons, and the Statistical Methodology Commons

Recommended Citation
CHEN, ZHETAO, "Assessing Reproducibility of Brain-Behavior Associations Using Bootstrap Aggregation Methods" (2024). *Arts & Sciences Electronic Theses and Dissertations*. 3107.
https://openscholarship.wustl.edu/art_sci_etds/3107

This Thesis is brought to you for free and open access by the Arts & Sciences at Washington University Open Scholarship. It has been accepted for inclusion in Arts & Sciences Electronic Theses and Dissertations by an authorized administrator of Washington University Open Scholarship. For more information, please contact digital@wumail.wustl.edu.

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Mathematics and Statistics

Assessing Reproducibility of Brain-Behavior Associations Using Bootstrap Aggregation Methods

by
Zhetao Chen

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

May 2024
St. Louis, Missouri

© 2024, Zhetao Chen

Table of Contents

List of Figures

Acknowledgments

Completing this master's thesis has been a profound journey, not only academically but also personally. I am deeply grateful to a number of individuals whose support and insights have been instrumental in my study and research.

First, my heartfelt thanks go to my advisors, Dr. Muriah D. Wheelock and Dr. Soumendra Lahiri. Their guidance and expertise have been the cornerstone of my research. Their patience, encouragement, and insightful critiques have not only aided in the successful completion of this thesis but have also inspired me to study deeper into the statistical methods of neuroscience.

I would also like to express my sincere appreciation to the committee member, Dr. Robert Lunde, and all faculty members of the Department of Mathematics and Statistics. Their valuable comments and suggestions have contributed to the refinement of my work.

Special thanks are extended to Yumiao Lei, Wei Li, Jiaqi Li, and Donna Dierker for their invaluable advice, guidance, and support throughout this process. Their perspectives and expertise have been a great source of motivation and have contributed to the depth and quality of my research.

I am also grateful to Andy Eck, Ari Segel, Jiaxin (Cindy) Tu, Nilanjan Chakraborty, Rezwana Rosen Razzaque, Xinyang Feng, Hangcen Zou, and everyone in the Wheelock Lab for their assistance and the enriching environment they have provided. The discussions and brainstorming sessions have been a source of inspiration and learning.

Lastly, I want to express my gratitude to my family and friends.

This research is supported by NIH grant R00EB029343 to MW. HCP data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen

and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

Zhetao Chen

Washington University in St. Louis

May 2024

ABSTRACT OF THE THESIS

Assessing reproducibility of brain-behavior

associations using bootstrap aggregation methods

by

Zhetao Chen

Master of Arts in Statistics

Washington University in St. Louis, 2024

Professor Soumendra Lahiri, Chair

Assistant Professor Muriah Wheelock, Co-Chair

In this thesis, amidst growing utilization of resting-state functional connectivity MRI (rs-fcMRI) for linking neural activity to pathological conditions, we confront the prevalent concerns regarding the reliability of such data. Our exploration concentrates on improving the reproducibility of brain-behavior associations within the framework of the Human Connectome Project (HCP) dataset. We employ two distinct bootstrap aggregation approaches to investigate the enhancement of functional connectivity reliability: individual time series bagging using Circular Block Bootstrap (CBB) and subject-level bagging utilizing Linear Support Vector Regression (LSVR) models. Our investigation into individual time series bagging with CBB reveals that this method does not significantly bolster the reproducibility of brain-behavior associations. This finding points to the complexity of achieving reliable functional connectivity measures and the limitations of certain aggregation methods in overcoming this challenge. In contrast, our examination of subject-level bagging through LSVR models presents a more promising outcome. This approach markedly enhances the reliability of model weights between analyses, demonstrating its efficacy in improving data robustness and reproducibility. This differential impact of the two methodologies underscores the critical role of appropriate analytical strategies in enhancing the reliability of

neuroimaging data. By delineating the outcomes of these two methodologies, this thesis contributes to the broader discussion on data reliability in the field of neuroimaging. It underscores the necessity for continued methodological innovations and validations across varied datasets to advance the reliability and interpretability of rs-fcMRI studies.

Chapter 1: Introduction

The field of neuroimaging endeavors to establish robust connections between brain structure and function as indexed by various behavioral or trait manifestations (Woo et al., 2017). This is often achieved by identifying distinctive imaging features that, when integrated into statistical models, can shed light on potential causal relationships or accurately predict observable traits for new participants. The overarching objective is to derive reliable biomarkers for diagnoses or intervention strategies from the wealth of imaging data available (Insel et al., 2010).

Resting-state functional connectivity Magnetic Resonance Imaging (rs-fcMRI) has gained significant traction for its ability to assess correlations in neural activity through monitoring spontaneous fluctuations in the blood oxygen level dependent (BOLD) signal while subjects are at rest. These correlations hold substantial interest within the medical realm as they have been increasingly correlated with various pathological conditions impacting specific brain regions (Seeley et al., 2009). Regrettably, research indicates that the duration of scan time necessary to attain satisfactory levels of reproducibility and reliability often exceeds the scanning time allocated in the majority of clinical and cognitive neuroscience datasets that have been collected or are presently underway (Laumann et al., 2015). Consequently, there is a growing demand for methodologies capable of enhancing the reliability of resting-state functional connectivity. In our study, time-based Pearson correlations between different regions were used as functional connectivity (FC).

Bootstrap aggregation, commonly known as bagging, is a well-regarded technique from statistics and machine learning that holds promise for enhancing the reproducibility of functional connectivity. Both bootstrapping and bagging are versatile procedures, whether parametric or nonparametric, aimed at refining statistical inference across various dimensions, such as improving estimates of effect certainty, estimating p-values, and enhancing accuracy. Bootstrapping involves resampling a dataset with replacement and has found widespread applications across scientific and statistical domains (Efron & Tibshirani, 1994). For instance, it's commonly employed for estimating the accuracy of sample estimates, rooted in the idea that a sample can effectively represent its population through resampling. Building upon the principles of bootstrapping, bootstrap aggregation or bagging was initially devised to resample data for predictive modeling and aggregate predictions across bootstrap samples to enhance predictive performance (Breiman, 1996). Bagging is believed to diminish variability in estimation by averaging prediction labels or cluster memberships across multiple resampled datasets (Dudoit & Fridlyand, 2003). More recently, bagging has emerged as a pivotal technique in ensemble clustering, a methodology for consolidating various cluster solutions into a final clustering outcome(Fischer & Buhmann, 2003). Therefore, it appears to be a feasible approach to diminish sample variability and enhance data reliability by employing bagging to resample individuals or time within individuals.

Previous research has demonstrated the benefits of bagging in enhancing the reproducibility and test-retest reliability of both cortical and subcortical functional parcellations across a range of sites, scanners, samples, scan lengths, clustering algorithms, and clustering parameters (Nikolaidis et al., 2020). In the current study, we aim to assess the efficacy of bagging methodologies in enhancing the reliability of data utilizing Human Connectome Project (HCP) data. Preliminary investigations have indicated the efficacy of circular block bootstrap (CBB) in robustly estimating associations between brain regions (Bellec et al., 2008). Subsequent research has delved into leveraging CBB and bootstrapping techniques at the group level to elucidate the hierarchical organization of brain networks (Bellec et al., 2010). Consequently, our study endeavors to employ CBB to resample individual time points, subsequently computing the average of functional connections, with the aim of ascertaining whether the reliability of the data can be bolstered through network-level analysis.

Additionally, resampling aggregated models, which involve subject-level bootstrapping with replacement, offer improved model performance and generalizability, particularly within the context of the connectome predictive modeling (CPM)(O'Connor et al., 2021). While we want to apply the models to predict behavioral or clinical outcomes using the functional connectome, a challenge introduced by cross-validation procedures in datasets with related individuals is that these related subjects may appear in both the training and hold-out test sets. Given that related individuals and twins tend to have more similar connectome data compared to unrelated individuals (Miranda-Dominguez et al., 2018), the shared variance among families violates the assumption of independent datasets in cross-validation. Consequently, when predicting labels in the test set, the presence of related subjects from the training set with known labels undermines the reliability of using test set prediction accuracy as a measure of model performance. To mitigate this limitation, a random sampling approach that accounts for family structure is implemented (Li et al., 2023). In this study, we intend to integrate the HCP data into a linear support vector regression model using the random sampling approach, thereby evaluating whether averaging model features across a sample of individuals can enhance the reliability of the data.

In our study, we aim to assess the efficacy of the bagging technique in enhancing data reliability from two primary perspectives. Firstly, we employ the CBB method to sample individual time series data. By computing and averaging the functional connectivity matrix derived from the sampled data, we anticipate that this approach will bolster the reliability of rs-fcMRI estimates. Subsequently, we conduct subject-level resampling to aggregate the data while ensuring that familial relationships are preserved within each group. Utilizing these two distinct sets of data for concurrent model training, we anticipate an improvement in the reliability of model weights.

Chapter 2: Methods

2.1 Dataset

The Human Connectome Project (HCP) aimed to study and freely share data from 1200 young adults (ages 22-35) from families with twins and non-twin siblings, using a protocol that includes structural and functional magnetic resonance imaging (fMRI) at 3 Tesla (3T) and behavioral testing (Seitzman et al., 2020; Van Essen et al., 2012). In all parts of the HCP, participants were scanned on the same equipment using the same protocol for every subject. Participants are administered the Mini Mental Status Exam (Folstein et al., 1975) as a broad measure of cognitive status (participants are excluded if they scored below a 27)(Crum et al., 1993).

2.2 Sleep Quality Behavioral Data

The Pittsburgh Sleep Questionnaire (Buysse et al., 1989) as a measure of sleep quality. Respondents are asked to indicate how frequently they have experienced certain sleep difficulties over the past month and to rate their overall sleep quality in this test. Examples of such difficulties include waking up multiple times during the night or experiencing difficulty falling asleep within a specified timeframe. Scores for each question range from 0 to 3, with higher scores indicating more acute sleep disturbances. Then the scores for all questions are added up to get a total score. In this study, we mainly used rs-fMRI data and subjects' total scores in HCP to explore the correlation between them. Given that sleep quality assessments were conducted on the first scan day and corresponded solely to the sleep quality experienced the night before the first scan, any variance observed on the second scan day could be attributed to the fact that the sleep quality data aligned most closely with the brain function observed on the first day. Therefore, we opted to analyze data from the first day to explore this aspect of reliability.

2.3 Data grouping based on Gower’s distance

To ascertain the reliability of the correlation between behavioral scores and functional connectivity observed on the initial day of observation, it is imperative to stratify the subjects into two cohorts. This stratification enables the subsequent computation of correlations between behavioral scores and functional connectivity within each group for the same day, followed by an assessment of the reliability of these correlations across the two cohorts. To accomplish this stratification, we propose utilizing Gower's distance as a metric to quantify the dissimilarity between subjects based on a comprehensive set of variables, including education level, Body Mass Index (BMI), handedness, gender, ethics, race, age, family number, income, smoking history, heavy drinking frequency, and illicit drug usage frequency. In light of the identified limitations associated with the unweighted application of Gower's distance—specifically, its vulnerability to outliers in ratio-scaled variables and the unequal weighting of variables in the overall distance calculation—modifications have been proposed to refine this methodology (D’Orazio, 2021). These modifications are designed to mitigate the impact of outliers and ensure a more equitable contribution of each variable to the composite distance measure. We use the method to compute the distance between each subject's behaviour records and aggregating these distances to all other subjects. Subsequently, a sorting process is undertaken based on the total distances, enabling the systematic allocation of subjects into distinct groups, which are marked as split1 and split2 respectively. This ensures a more balanced distribution, where individuals within each group exhibit comparable distances to all other subjects. Through these refinements, the reliability and robustness of the Gower's distance calculation are enhanced, facilitating more meaningful data grouping and analysis.

2.4 fMRI data preprocessing

The preprocessing steps applied to HCP data in this study, as described by Gordon et al., (2017), involved several key procedures. Initially, the first 29.52 seconds or 41 frames of each resting-state run were discarded to account for magnetization equilibrium and any responses evoked by the scan start (Laumann et al., 2015). Subsequently, the functional data were aligned to the first frame of the first run using rigid body transforms, motion corrected, and whole-brain mode 1000 normalized (Miezin et al., 2000). The resting-state data, comprised of 2x2x2mm voxels, were then registered to the T1-weighted image and a WashU MNI atlas using affine and FSL transforms (Smith et al., 2004).

Further preprocessing steps were implemented to remove artifacts from the resting-state BOLD data (Ciric et al., 2017). Frame-wise displacement (FD), a metric quantifying motion or displacement between consecutive frames in fMRI data (Power et al., 2012), was calculated, and artifact removal (Ciric et al., 2017; Power et al., 2014) was conducted with a low-pass filter at 0.1 Hz to address respiration artifacts affecting the FD estimates (Fair et al., 2020). Frames with FD greater than 0.04 mm were removed (Dworetsky et al., 2021). Nuisance variable regression was performed, including whole-brain mean, ventricular and white matter CSF signals, the temporal derivatives of those regressors, and 24 movement regressors (Yan et al., 2013). Temporal masks of the gray matter, white matter, and ventricles were created using Freesurfer automatic segmentation (Fischl et al., 2002), and segments of data lasting fewer than 5 contiguous frames were excluded (Hocke & Kämpfer, 2009; Power et al., 2014). Data were then bandpass filtered from 0.009 to 0.08 Hz, and censored frames were removed from the time series (Seitzman et al., 2020).

In accordance with the established protocols outlined by Gordon et al., (2016), the preprocessed BOLD time series data underwent surface processing. This involved utilizing the ribbon-constrained sampling procedure within the Connectome Workbench software to project the BOLD volumes onto each subject's individual native surface. During this process, voxels exhibiting a time series coefficient with a variation exceeding 0.5 standard deviations above the mean of nearby voxels were excluded from further analysis, as per the methodology described by Glasser et al., (2013) and Gordon et al., (2016). Subsequently, the sampled data were subjected to deformation, resampling, and Gaussian smoothing (FWHM = 4mm, sigma = 1.7) to enhance spatial coherence and mitigate noise. The Connectome Workbench software was then utilized to integrate these processed surfaces with volumetric subcortical and cerebellar data, resulting in the generation of full brain time courses in the CIFTI format while excluding non-gray matter tissue, as outlined by Glasser et al., (2013).

Following the outlined processing steps, surface-based parcels and canonical functional networks (Gordon et al., 2016) were utilized to partition a predefined set of regions of interest (ROIs) into 12 networks and one unspecified network (Figure 1). This unspecified network comprised parcels that were not strongly connected with any others, as defined in Gordon et al., (2016). The Pearson correlation between the mean time courses of each pair of ROIs was evaluated (55,278 pairs in total, ROIs on the diagonal excluded), which are used to estimate functional connectivity (FC).

2.5 Bootstrap aggregation.

2.5.1 Circular Block Bootstrap on time series data

The Circular Block Bootstrap (CBB) method is a resampling technique employed in statistical inference, particularly in the context of time series data and spatial data with periodic characteristics. It is designed to address the inherent autocorrelation present in such data by preserving the underlying temporal or spatial structure during resampling. The primary principle of it involves dividing the time series or spatial data into blocks of contiguous observations, with the last block potentially overlapping with the first block in a circular fashion. This circular arrangement ensures that the temporal or spatial structure is maintained during the resampling process (Politis & Romano, 1994). At the individual level, bagging is implemented utilizing a CBB approach, with a window size set to the square root of the number of time steps. This method is applied to individual time series data to generate numerous resampled functional connectivity instances for each individual. Subsequently, these resampled instances are aggregated and averaged, following which network-level analyses are conducted to assess the reliability of the data.

2.5.2 Resample Aggregation on participant functional connectivity matrices

The Linear Support Vector Regression (LSVR) model proposed by Li et al., (2023) incorporates a random sampling regime considering family structure. Although maintaining individuals from the same family in either the training or test set led to lower prediction accuracy and reliability, it ensured that predictions in the test set were not influenced by information in the training set, thereby mitigating falsely inflated prediction accuracy. We incorporated the data sampled according to this random sampling system into the model and calculated the average of the final model weights. This averaging process will serve as a crucial step in evaluating the reliability of the model across different samples, thereby enhancing the validity of the model's predictions.

In the LSVR model (Li et al., 2023), we conducted a 1000 repetitions with 5-fold cross-validation procedure, where the data were divided into an 80% training set and a 20% test set (family members were never in both the training and testing sets). Within each iteration, the hyper-parameter of the LSVR model was selected using a 5-fold cross-validation within the training data.

During each iteration within the training set, which contained 80% of the subjects, we further randomly divided the data into an inner-cross-validation training set (80% of subjects) and an inner-cross-validation test set (20% of subjects). Various hyper-parameters were tested within a specified range to fit the LSVR model using the inner-training set. The optimal hyper-parameter was chosen based on the minimized mean square error (MSE) observed in the prediction using the inner-test set.

In the context of investigating the relationship between connectivity matrices and labels(i.e., scores), Pearson correlations are commonly employed as a univariate method, allowing for the examination of individual feature-label associations. However, our utilization of the LSVR model enables simultaneous consideration of the collective impact of multiple variables on the label, offering a more comprehensive analysis. Additionally, in line with insights from Haufe et al., (2014), an inversion process applied to machine learning (ML) results is deemed necessary for enhanced biological interpretation. Haufe et al., (2014) demonstrated that for a linear predictive model, the appropriate transformation can be obtained by computing the covariance of each feature and the predicted target variable in the training set. With a sufficient sample size, feature importance (operationalized as Haufe-transformed weights) have been demonstrated to achieve fair to excellent split-half reliability, which refer to as ICC (Chen et al., 2023). Therefore we will simultaneously calculate the original weights and Haufe-transformed weights for each model. Then, in our study, the effectiveness of bagging methods in improving the reliability of model weights are concurrently evaluated in the three models.

2.6 Reliability

Reliability in the context of statistical analysis refers to the consistency and stability of measurements or observations across different conditions, time points, or raters. It signifies the extent to which the measurements or observations yield consistent results under varying circumstances. In the realm of research, reliability is a fundamental concept as it reflects the degree to which findings can be replicated or trusted. One commonly used metric to assess reliability is the Intraclass Correlation Coefficient (ICC). The ICC is a statistical measure that quantifies the proportion of total variance in measurements or observations that can be attributed to differences between subjects or raters, relative to the total variance (Tian & Zalesky, 2021). For example, if we have two sets of feature importance $(f_1 \, and \, f_2)$,

$$ICC = \frac{1}{Ks^2} \sum_{k=1}^{K} (f_{1,k} - \bar{f})(f_{2,k} - \bar{f})$$

where K is the total number of features and

$$\bar{f} = \frac{1}{2K} \sum_{k=1}^{K} (f_{1,k} + f_{2,k}), s^2 = \frac{1}{2K} \sum_{k=1}^{K} \left((f_{1,k} - \bar{f})^2 + (f_{2,k} - \bar{f})^2 \right)$$

This is referred to as the 1-1 formulation of the ICC. Specifically, ICC(1,1) typically denotes a measure of absolute agreement for single measurements, meaning it evaluates the reliability of single ratings made by evaluators on the same set of subjects..

2.7 Network Level Analysis

2.7.1 Observed Network-Level Enrichment

For the LSVR model, we computed the mean estimated FC-score beta weights across 1000 repetitions to serve as input. Similarly, for the LSVR model with the inversion applied, we calculated the inverse weights in each repetition and averaged them across 1000 replications. Specifically, the covariance between each rsFC and the predicted score was computed as the estimated inverted weight for each ROI pair (Haufe et al., 2014). These weights underwent zscoring, followed by thresholding at $|Z| > 2$ and subsequent binarization. We selected a threshold value of 2 after testing several thresholds, with 2 yielding a balanced distribution of features across the connectome, striking a suitable balance between sparsity and density. This threshold was quantitatively evaluated in a prior study (Li et al., 2023). Weights within each network block that passed this threshold were used to compute the χ^2 values and Welch's t values relative to the distribution of all weights passing the threshold in the rest of the connectome. The 1-degree-of-freedom χ^2 test and Welch's t test compared the observed number of strong (thresholded and binarized) brain-behavior weights within a pair of functional networks to the expected number if such values were uniformly distributed across the full connectome (Wheelock et al., 2021). A large resulting test statistic indicates enrichment of strong associations within a specific network block, implying a greater number of strong weights than expected.

2.7.2 Permutation Test

To evaluate the significance of network-level associations, we employed a permutation test to compute permutation-based p-values. This involved shuffling the sleep disturbance score labels and fitting the same model with the connectome data to generate null FC-score weight matrices. This process was repeated for 1000 repetitions for each model, yielding a null distribution of χ^2 test and Welch's t statistics. Subsequently, the observed (real) χ^2 values and Welch's t values were compared to this null distribution to determine network-level significance, defined as family-wise error rate (FWER) permutation-based p-values < 0.05.

2.7.3 Network-level Reliability

Additionally, we assessed network-level reliability using Matthews Correlation Coefficient (MCC). To compute MCC scores, we constructed a confusion matrix considering split1 and split2 as the observation and prediction sets, respectively. We categorized significant blocks in split1 as "true" and other blocks as "false," while significant blocks in split2 were labeled as "positive" and other blocks as "negative." With this configuration, MCC scores were calculated using the formula

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}$$

Chapter 3: Results

3.1 Bagging individual time series data can not improve the reliability of data significantly

To rigorously assess the reliability of our dataset, we applied Gower's distance as a metric to segment the dataset into two distinct cohorts, encompassing 483 and 482 subjects respectively. Subsequent to this division, we implemented the Composite Bootstrap Bagging (CBB) methodology to perform 1,000-fold bootstrap sampling on the daily functional magnetic resonance imaging (fMRI) data of each subject, subsequently averaging the data to enhance robustness. The investigation into the relationship between resting-state functional connectivity (rsFC) features and sleep-related behavioral scores was conducted through the application of univariate marginal Pearson correlation analysis for each region of interest (ROI) pair in relation to the response variable. The evaluation of the reliability of the weight coefficients derived post-application of the Bagging technique to the univariate model revealed a remarkable consistency across both the χ^2 test and Welch's T test, indicating a negligible disparity in reliability outcomes (Figure 2,3).

Moreover, the ICC scores were computed for these weight coefficients, yielding values of 0.397 and 0.400 in χ^2 test and 0.755 and 0.701 in welch's T test, respectively, which further underscores the similarity in reliability metrics. This finding led us to the conclusion that the employment of the bagging methodology for time-series data does not significantly enhance the reliability of the weight estimates within the Pearson univariate model framework.

3.2 Using bagging methods on subjects does improve the reliability of model weights

In our endeavor to enhance the reliability of time-series data analysis, an alternative approach was adopted by implementing bagging on subjects instead of time points, utilizing both the Linear

Support Vector Regression (LSVR) model and its inverse for analysis. The dataset was meticulously partitioned into an 80% training set and a 20% test set, ensuring no overlap of family members between the training and testing subsets. Within each iteration of the model training process, hyper-parameters for the LSVR model were fine-tuned via 5-fold cross-validation exclusively within the training dataset.

Our analysis aimed to elucidate the relationships between different Region of Interest (ROI) pairs and sleep-related behavioral scores through three distinct methodologies: firstly, by exploring the univariate marginal Pearson correlation between each resting-state functional connectivity (rsFC) feature and the behavioral scores; secondly, by incorporating rsFC as predictive features in the LSVR model; and thirdly, by employing rsFC features in predicting scores utilizing the inverse of the LSVR model's beta weights (figure 4). Concurrently, the application of NLA χ^2 test and Welch's T test across these three analytical outcomes identified specific network blocks most predictive of the behavioral scores.

In the χ^2 test, both Pearson correlations and the inverse weights derived from the LSVR model pinpointed similar significant network blocks, indicating a consistency in network significance (figure 5). Notably, the network-level reliability, as assessed by MCC scores in conjunction with the network-level χ^2 test, exhibited marked improvement when employing the inverse weights of the LSVR model. Despite the raw beta weights from the LSVR model achieving high MCC scores, the networks identified by NLA using these weights as inputs were distinctly disparate, underscoring the interpretational limitations of strong Machine Learning regression raw weights as direct neural predictors of behavioral labels, as highlighted in the literature (Haufe et al., 2014).

The Welch's T test further accentuated the deficiencies associated with the utilization of raw LSVR model weights (figure 6), where the incorporation of weighted, signed data (reflecting mean shifts within the entire network block) resulted in poor MCC scores (-0.028), akin to random prediction accuracy. Conversely, the outcomes from Pearson correlations mirrored those obtained through the inverse weights of the LSVR model, both yielding MCC scores of 0.566 and identifying identical significant network blocks.

In examining ICC scores for raw weights, a notable discrepancy emerges between different statistical tests: a high ICC of 0.788 in the χ^2 test versus a significantly lower ICC of 0.054 in Welch's T-test, both of which align with the MCC, indicating consistent statistical behavior. Despite this statistical consistency within certain network blocks, as evidenced by the χ^2analysis, the raw weights suffer from a lack of biological interpretability, underscoring the need for cautious interpretation of these findings. Moreover, regardless of the statistical approach(χ^2test or Welch's T test), LSVR inversion and simple Pearson correlation yield similar ICC scores, mirroring the MCC results. However, the additional computational cost of LSVR inversion may not be justified when simple linear correlation produces comparable results with similar reliability.

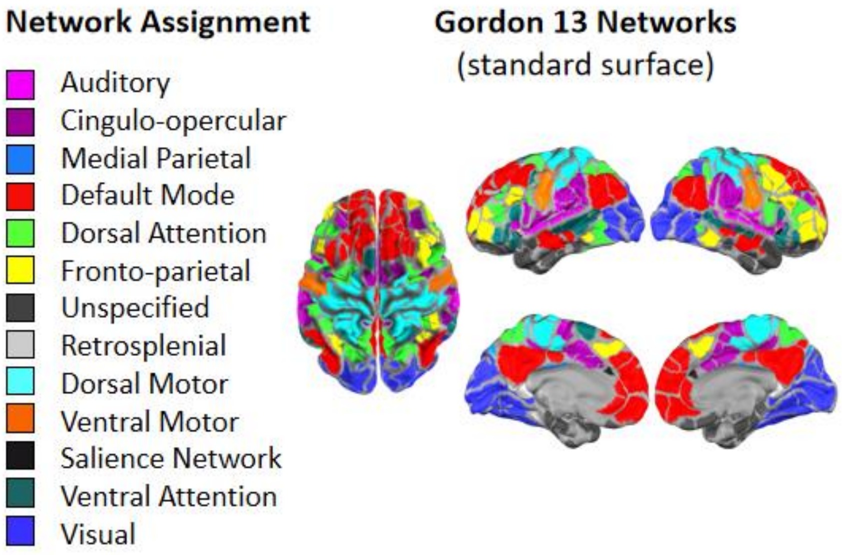

Figure 1 Gordon networks.

333 parcels in the brain were used to extract mean rsFC and were grouped into 12 networks and one unspecified network.

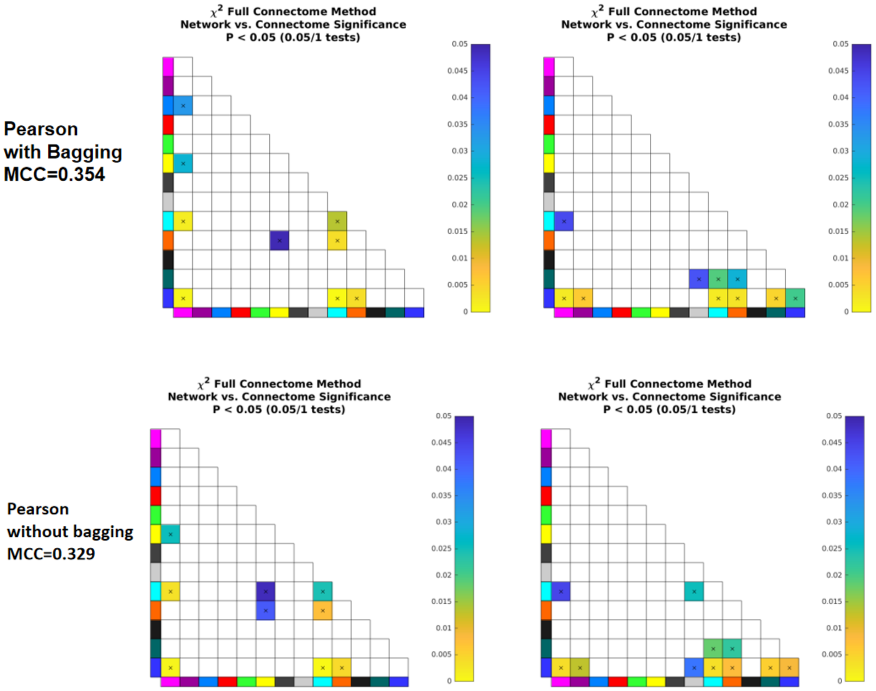

Figure 2 Significant network blocks selected by χ^2 test of two groups on day 1.

(Left) split1, (Right) split2

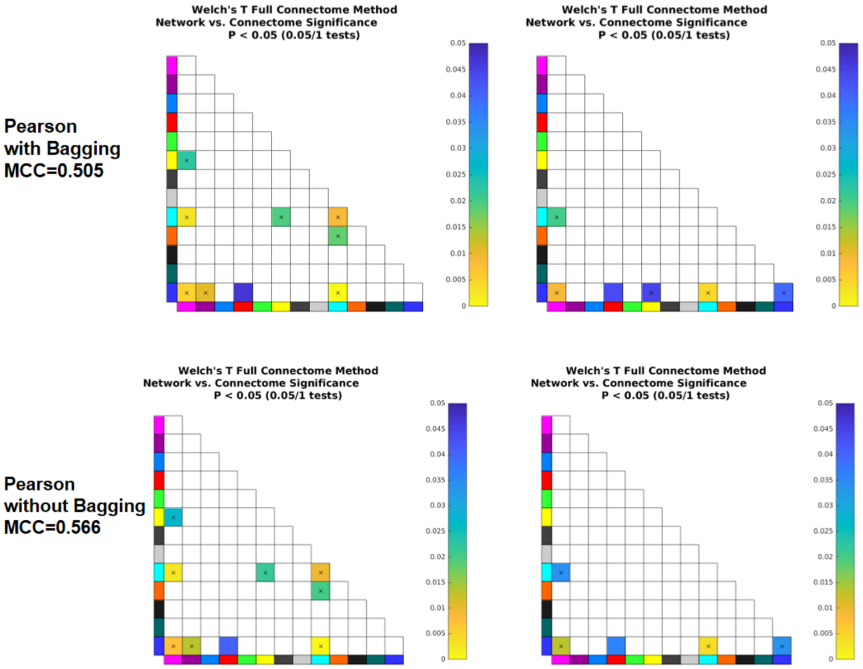

Figure 3 Significant network blocks selected by Welch's T test of two groups on day 1.

(Left) split1, (Right) split2

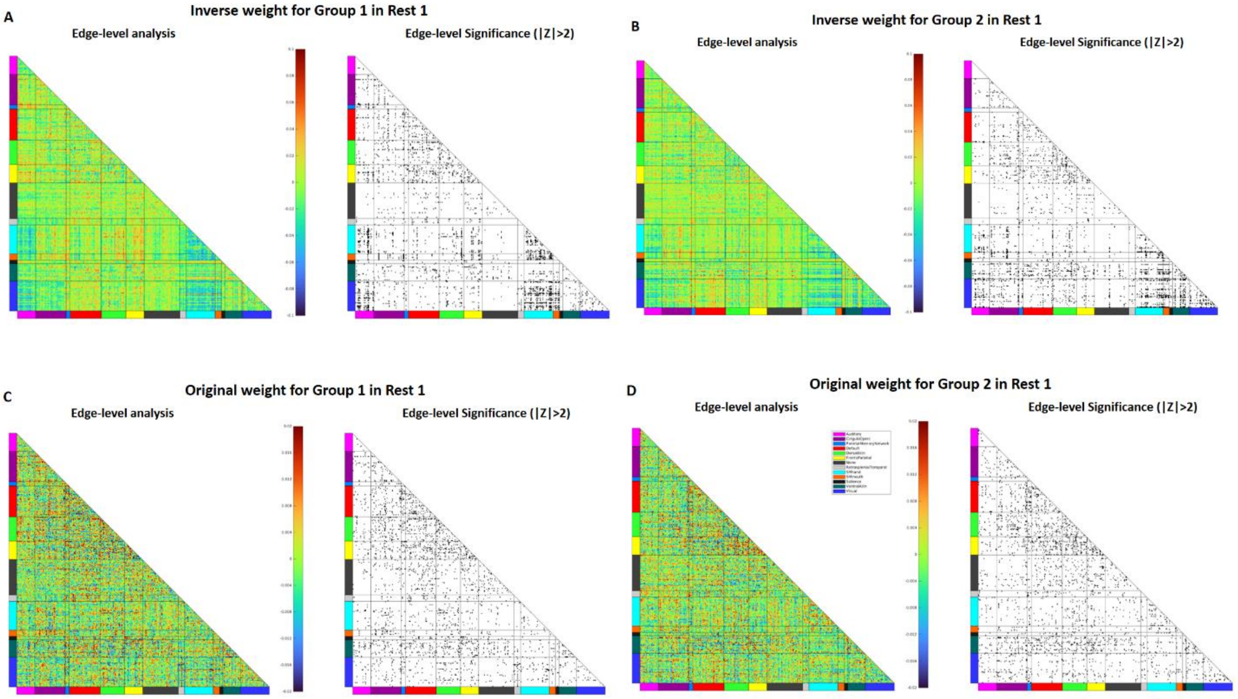

Figure 4 Nominally significant ROI pairs selected by two weights for two groups on day 1.

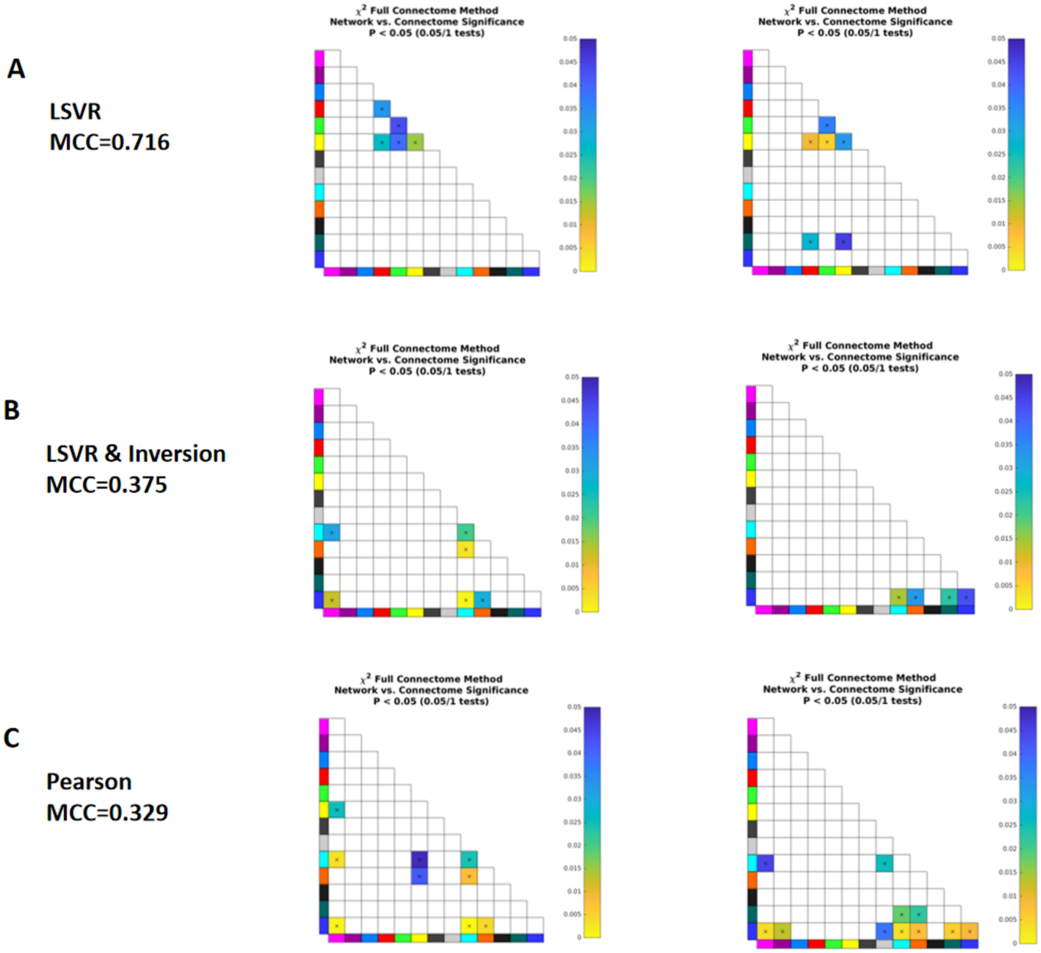

Figure 5 Significant network blocks selected by χ^2 test of two groups on day 1 with three different sets of inputs

Specifically, the inputs were obtained from **(A)** LSVR model, **(B)** LSVR model with inversion, **(C)** Pearson r-correlation. (Left) split1, (Right) split2

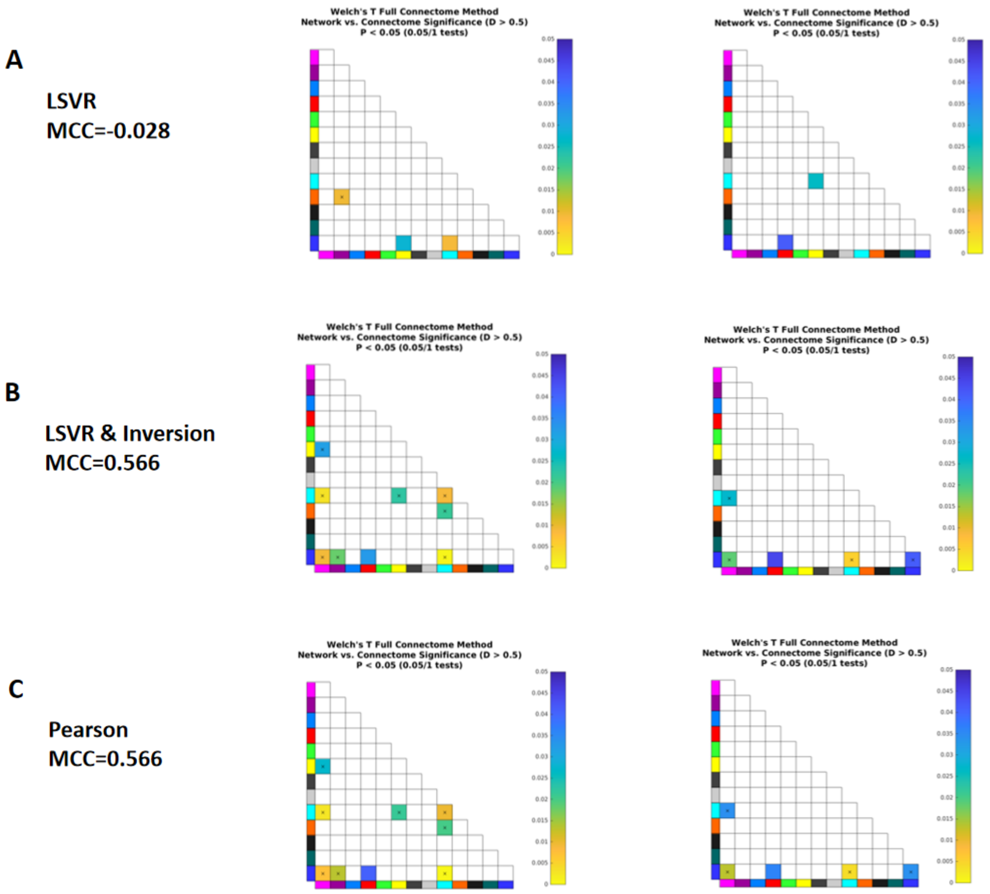

Figure 6 Significant network blocks selected by Welch's T test of two groups on day 1 with three different sets of inputs

Specifically, the inputs were obtained from **(A)** LSVR model, **(B)** LSVR model with inversion, **(C)** Pearson r-correlation. (Left) split1, (Right) split2

Chapter 4: Discussion

The primary objective of this study was to ascertain whether the application of the bagging method enhances the reliability of brain-behavior associations, focusing on two key perspectives: time series and subjects. To address the former, the CBB method was employed to resample the functional magnetic resonance imaging (fMRI) data of each subject, followed by the averaging of all resampled data to derive averaged resting-state functional connectivity (FC) data for each subject. Upon analysis, it was observed that the reliability of univariate FC-behavior correlations across two groups on first day did not exhibit significant improvement. In response, we considered the utilization of the LSVR model to explore the relationship between functional connectome data and behavioral scores. This analysis incorporated various factors including the disparities in reliability levels between different samples and sessions, as well as the examination of beta weights, weight inversions, and univariate Pearson correlations. Ultimately, our findings indicate that the bagging method indeed enhances the reliability of HCP data, particularly demonstrating higher reliability in weight inversions compared to univariate Pearson correlations.

4.1 Bagging on Individual Time Series Data

Some study have indicated that irrespective of the particular parcellation strategy utilized, employing bagging could serve as a pivotal technique for enhancing functional parcellation and bridging the gap between functional neuroimaging-based measurements and clinical applicability (Nikolaidis et al., 2020). However, our investigation into the impact of bagging on individual time series data yielded different conclusions. Despite employing the CBB method to resample and average daily FMRI data, the application of bagging did not significantly enhance the reliability of the data, as evidenced by the similar Pearson correlation coefficients between the functional connectivity and behaviour scores across two groups on first scan day. Additionally, network-level analyses did not reveal persistent significant differences in results, suggesting that the bagging method's efficacy might not be contingent upon the quantity of resting-state data utilized. These findings underscore the importance of considering alternative approaches or refining existing methodologies to address the challenge of improving data reliability in time series analyses.

4.2 Bagging on Subject-Level Data

Conversely, our exploration of employing bagging methods on subject-level data yielded more promising outcomes. By three different models to fit the relationship between the functional connectivity and behaviour scores across two groups on first scan day, we observed notable improvements in model weights between two groups when utilizing the bagging method. Leveraging various measures including univariate marginal Pearson correlations, LSVR models, and inversion of LSVR beta weights, we identified significant network blocks predictive of sleep-related behavioral scores. Notably, the LSVR model demonstrated enhanced network-level reliability compared to the Pearson correlations, as evidenced by higher Matthews Correlation Coefficients (MCC) scores. However, it is crucial to note that although beta weights demonstrated the highest MCC score in the chi-square test, caution must be exercised in interpreting their reliability. This discrepancy could be attributed to the utilization of thresholded and binarized data in the Chi-square test, leading to MCC evaluation solely based on the location of binarized high magnitude weights, regardless of sign. In contrast, Welch's T test incorporates weighted and signed data, reflecting mean shifts in the total network block. Additionally, while edge-level raw ML beta weights are interspersed within network blocks with both negative and positive weights, Pearson and Haufe Inversion weights tend to exhibit uniformity within a network block, either all negative or all positive. Consequently, raw ML beta weights may exhibit poorer MCC scores when assessed using Welch's T test due to this mixed representation within network blocks. While the beta weights derived directly from the predictive LSVR model offer insights into the influence of resting-state functional connectivity on label prediction, they inherently lack direct biological interpretation. In contrast, the inversion model introduces a crucial layer of biological context by preserving the individual effects of each feature while encompassing the holistic model framework. This aspect is pivotal for advancing our understanding beyond mere prediction accuracy toward a more nuanced exploration of the underlying neurobiological mechanisms.

4.3 Limitations and Future Work

Our study was confined to the analysis of Human Connectome Project (HCP) data, thus limiting the extrapolation of our findings to other datasets. The applicability and efficacy of the bagging method on diverse datasets remain uncertain and warrant exploration in future research endeavors. While we integrated inversion models to enhance biological interpretability, we did not assess the impact of bagging methods on model accuracy, which is a crucial aspect often intertwined with reliability. Furthermore, in our study, each result has been computed only once. To thoroughly analyze the variance and ensure the robustness of our findings, it is essential to conduct multiple computations. Lastly, it's imperative to acknowledge the variability in network-level analyses across different datasets, necessitating tailored approaches and potentially disparate testing methodologies. The choice of analysis techniques can significantly influence the efficacy of the bagging method, highlighting the importance of carefully considering experimental design and analytical strategies in future investigations.

Chapter 5: Conclusion

Our study investigated the reliability of Human Connectome Project (HCP) data and the efficacy of the bagging method in improving data consistency. Despite efforts to mitigate individual variability through bagging on individual time series data, the reliability of HCP data remained insufficient. However, employing the bagging method on subject-level data demonstrated promising results, particularly in enhancing the reliability of model weights. Notably, the Linear Support Vector Regression (LSVR) model exhibited enhanced network-level reliability compared to univariate Pearson correlations. Moving forward, further research is warranted to explore the generalizability of these findings across diverse datasets and to refine methodologies for optimizing data reliability and interpretability in neuroimaging studies.

References/Bibliography/Works Cited

Bellec, P., Marrelec, G., & Benali, H. (2008). A bootstrap test to investigate changes in brain connectivity for functional MRI. *Statistica Sinica*, *18*, 1253–1268.

Bellec, P., Rosa-Neto, P., Lyttelton, O. C., Benali, H., & Evans, A. C. (2010). Multi-level bootstrap analysis of stable clusters in resting-state fMRI. *NeuroImage*, *51*(3), 1126–1139. https://doi.org/10.1016/j.neuroimage.2010.02.082

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. https://doi.org/10.1007/BF00058655

Buysse, D. J., Reynolds, C. F., Monk, T. H., Berman, S. R., & Kupfer, D. J. (1989). The Pittsburgh Sleep Quality Index: A new instrument for psychiatric practice and research. *Psychiatry Research*, *28*(2), 193–213. https://doi.org/10.1016/0165-1781(89)90047-4

Chen, J., Ooi, L. Q. R., Tan, T. W. K., Zhang, S., Li, J., Asplund, C. L., Eickhoff, S. B., Bzdok, D., Holmes, A. J., & Yeo, B. T. T. (2023). Relationship between prediction accuracy and feature importance reliability: An empirical and theoretical study. *NeuroImage*, *274*, 120115. https://doi.org/10.1016/j.neuroimage.2023.120115

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., & 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. https://doi.org/10.1016/j.neuroimage.2017.03.020

Crum, R. M., Anthony, J. C., Bassett, S. S., & Folstein, M. F. (1993). Population-based norms for the Mini-Mental State Examination by age and educational level. *JAMA*, *269*(18), 2386–2391.

D’Orazio, M. (2021). *Distances with mixed type variables some modified Gower’s coefficients* (arXiv:2101.02481). arXiv. https://doi.org/10.48550/arXiv.2101.02481

Dudoit, S., & Fridlyand, J. (2003). Bagging to improve the accuracy of a clustering procedure. *Bioinformatics (Oxford, England)*, *19*(9), 1090–1099. https://doi.org/10.1093/bioinformatics/btg038

Dworetsky, A., Seitzman, B. A., Adeyemo, B., Neta, M., Coalson, R. S., Petersen, S. E., & Gratton, C. (2021). Probabilistic mapping of human functional brain networks identifies regions of high group consensus. *NeuroImage*, *237*, 118164. https://doi.org/10.1016/j.neuroimage.2021.118164

Efron, B., & Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. Chapman and Hall/CRC. https://doi.org/10.1201/9780429246593

Fair, D. A., Miranda-Dominguez, O., Snyder, A. Z., Perrone, A., Earl, E. A., Van, A. N., Koller, J. M., Feczko, E., Tisdall, M. D., van der Kouwe, A., Klein, R. L., Mirro, A. E., Hampton, J. M., Adeyemo, B., Laumann, T. O., Gratton, C., Greene, D. J., Schlaggar, B. L., Hagler, D. J., … Dosenbach, N. U. F. (2020). Correction of respiratory artifacts in MRI head motion estimates. *NeuroImage*, *208*, 116400. https://doi.org/10.1016/j.neuroimage.2019.116400

Fischer, B., & Buhmann, J. (2003). Bagging for path-based clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, *25*, 1411–1415. https://doi.org/10.1109/TPAMI.2003.1240115

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N., Rosen, B., & Dale, A. M. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron, 33*(3), 341–355. https://doi.org/10.1016/s0896-6273(02)00569-x

Folstein, M. F., Folstein, S. E., & McHugh, P. R. (1975). “Mini-mental state”: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research, 12*(3), 189–198. https://doi.org/10.1016/0022-3956(75)90026-6

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., & WU-Minn HCP Consortium. (2013). The minimal preprocessing pipelines for the Human Connectome Project. *NeuroImage, 80*, 105–124. https://doi.org/10.1016/j.neuroimage.2013.04.127

Gordon, E. M., Laumann, T. O., Adeyemo, B., Huckins, J. F., Kelley, W. M., & Petersen, S. E. (2016). Generation and Evaluation of a Cortical Area Parcellation from Resting-State Correlations. *Cerebral Cortex (New York, N.Y.: 1991), 26*(1), 288–303. https://doi.org/10.1093/cercor/bhu239

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., Ortega, M., Hoyt-Drazen, C., Gratton, C., Sun, H., Hampton, J. M., Coalson, R. S., Nguyen, A. L., McDermott, K. B., Shimony, J. S., Snyder, A. Z., Schlaggar, B. L., Petersen, S. E., Nelson, S. M., & Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron, 95*(4), 791-807.e7. https://doi.org/10.1016/j.neuron.2017.07.011

Haufe, S., Meinecke, F., Görgen, K., Dähne, S., Haynes, J.-D., Blankertz, B., & Bießmann, F. (2014). On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, *87*, 96–110. https://doi.org/10.1016/j.neuroimage.2013.10.067

Hocke, K., & Kämpfer, N. (2009). Gap filling and noise reduction of unevenly sampled data by means of the Lomb-Scargle periodogram. *Atmospheric Chemistry and Physics*, *9*(12), 4197–4206. https://doi.org/10.5194/acp-9-4197-2009

Insel, T., Cuthbert, B., Garvey, M., Heinssen, R., Pine, D. S., Quinn, K., Sanislow, C., & Wang, P. (2010). Research Domain Criteria (RDoC): Toward a New Classification Framework for Research on Mental Disorders. *American Journal of Psychiatry*, *167*(7), 748–751. https://doi.org/10.1176/appi.ajp.2010.09091379

Laumann, T. O., Gordon, E. M., Adeyemo, B., Snyder, A. Z., Joo, S. J., Chen, M.-Y., Gilmore, A. W., McDermott, K. B., Nelson, S. M., Dosenbach, N. U. F., Schlaggar, B. L., Mumford, J. A., Poldrack, R. A., & Petersen, S. E. (2015). Functional System and Areal Organization of a Highly Sampled Individual Human Brain. *Neuron*, *87*(3), 657–670. https://doi.org/10.1016/j.neuron.2015.06.037

Li, J., Segel, A., Feng, X., Tu, J. C., Eck, A., King, K., Adeyemo, B., Karcher, N. R., Chen, L., Eggebrecht, A. T., & Wheelock, M. D. (2023). *Network level enrichment provides a framework for biological interpretation of machine learning results* (p. 2023.10.14.562358). bioRxiv. https://doi.org/10.1101/2023.10.14.562358

Miezin, F. M., Maccotta, L., Ollinger, J. M., Petersen, S. E., & Buckner, R. L. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, *11*(6 Pt 1), 735–759. https://doi.org/10.1006/nimg.2000.0568

Miranda-Dominguez, O., Feczko, E., Grayson, D. S., Walum, H., Nigg, J. T., & Fair, D. A. (2018). Heritability of the human connectome: A connectotyping study. *Network Neuroscience*, *2*(2), 175–199. https://doi.org/10.1162/netn_a_00029

Nikolaidis, A., Solon Heinsfeld, A., Xu, T., Bellec, P., Vogelstein, J., & Milham, M. (2020). Bagging improves reproducibility of functional parcellation of the human brain. *NeuroImage*, *214*, 116678. https://doi.org/10.1016/j.neuroimage.2020.116678

O'Connor, D., Lake, E. M. R., Scheinost, D., & Constable, R. T. (2021). Resample aggregating improves the generalizability of connectome predictive modeling. *NeuroImage*, *236*, 118044. https://doi.org/10.1016/j.neuroimage.2021.118044

Politis, D. N., & Romano, J. P. (1994). The Stationary Bootstrap. *Journal of the American Statistical Association*, *89*(428), 1303–1313. https://doi.org/10.1080/01621459.1994.10476870

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *Neuroimage*, *59*(3), 2142–2154. https://doi.org/10.1016/j.neuroimage.2011.10.018

Power, J. D., Mitra, A., Laumann, T. O., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2014). Methods to detect, characterize, and remove motion artifact in resting state fMRI. *NeuroImage*, *84*, 10.1016/j.neuroimage.2013.08.048. https://doi.org/10.1016/j.neuroimage.2013.08.048

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., & Greicius, M. D. (2009). Neurodegenerative diseases target large-scale human brain networks. *Neuron*, *62*(1), 42–52. https://doi.org/10.1016/j.neuron.2009.03.024

Seitzman, B. A., Gratton, C., Marek, S., Raut, R. V., Dosenbach, N. U. F., Schlaggar, B. L., Petersen, S. E., & Greene, D. J. (2020). A set of functionally-defined brain regions with improved representation of the subcortex and cerebellum. *NeuroImage*, *206*, 116290. https://doi.org/10.1016/j.neuroimage.2019.116290

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E. J., Johansen-Berg, H., Bannister, P. R., De Luca, M., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., De Stefano, N., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, *23 Suppl 1*, S208-219. https://doi.org/10.1016/j.neuroimage.2004.07.051

Tian, Y., & Zalesky, A. (2021). Machine learning prediction of cognition from functional connectivity: Are feature weights reliable? *NeuroImage*, *245*, 118648. https://doi.org/10.1016/j.neuroimage.2021.118648

Van Essen, D. C., Ugurbil, K., Auerbach, E., Barch, D., Behrens, T. E. J., Bucholz, R., Chang, A., Chen, L., Corbetta, M., Curtiss, S. W., Della Penna, S., Feinberg, D., Glasser, M. F., Harel, N., Heath, A. C., Larson-Prior, L., Marcus, D., Michalareas, G., Moeller, S., … WU-Minn HCP Consortium. (2012). The Human Connectome Project: A data acquisition perspective. *NeuroImage*, *62*(4), 2222–2231. https://doi.org/10.1016/j.neuroimage.2012.02.018

Wheelock, M. D., Lean, R. E., Bora, S., Melzer, T. R., Eggebrecht, A. T., Smyser, C. D., & Woodward, L. J. (2021). Functional Connectivity Network Disruption Underlies Domain-Specific Impairments in Attention for Children Born Very Preterm. *Cerebral*

Cortex (New York, N.Y.: 1991), *31*(2), 1383–1394. https://doi.org/10.1093/cercor/bhaa303

Woo, B. F. Y., Lee, J. X. Y., & Tam, W. W. S. (2017). The impact of the advanced practice nursing role on quality of care, clinical outcomes, patient satisfaction, and cost in the emergency and critical care settings: A systematic review. *Human Resources for Health*, *15*(1), 63. https://doi.org/10.1186/s12960-017-0237-9

Yan, C.-G., Cheung, B., Kelly, C., Colcombe, S., Craddock, R. C., Di Martino, A., Li, Q., Zuo, X.-N., Castellanos, F. X., & Milham, M. P. (2013). A comprehensive assessment of regional variation in the impact of head micromovements on functional connectomics. *NeuroImage*, *76*, 183–201. https://doi.org/10.1016/j.neuroimage.2013.03.004