Characterizing univariate and multivariate

test-retest reliability of resting state functional connectomes in depressed adolescents

# Christopher C. Camp1, Stephanie Noble2, Dustin Scheinost2, Argyris Stringaris3, and Dylan M. Nielson4

1**Interdepartmental Neuroscience Program, Yale University**

2**Department of Radiology and Biomedical Imaging, Yale University School of Medicine, Yale University**

3**Division of Psychiatry and Division of Psychology and Language Sciences, University College**

**London**

4**Center for Multimodal Neuroimaging, National Institute of Mental Health, NIH**

# ABSTRACT

Keywords:

# INTRODUCTION

One of the foremost goals of neuroimaging work is to develop clinical tools for the diagnosis or prediction of psychiatric illnesses. Resting state functional connectivity is a key measure in this endeavor, as evidence suggests that it may contain useful information on internal neural function (CITE). However, clinical utility may be restricted by the reproducibility of neuroimaging findings. The “replication crisis” has highlighted the difficulty of reproducing the results of underpowered neuroimaging results. This lack of power is inseparable from the reliability of the data: reliability places an upper bound on the observable effect size, in turn thresholding statistical power (Zuo et al., 2019). Thus, quantifying the “test-retest reliability” - that is, the reliability of a test or measurement over repetitions - is critical to interpreting the validity of results. Unfortunately, surveys of neuroimaging studies, in particular functional connectivity, have observed poor test-retest reliability (Noble et al., 2017, 2019).

The test-retest reliability of fMRI data may be influenced by many factors. Longer scans and shorter intervals between scans can increase reliability, while artifact correction can decrease it - underscoring the separation between validity and reliability, as noise such as motion can be highly reliable (Noble et al., 2021). It is therefore crucial to have a nuanced understanding of reliability and its interpretation to optimize our data collection and processing methods.

Adding to this complexity, the measurement of test-retest reliability can be done through different metrics, each of which reflect unique forms of reliability subject to interpretation. These metrics can be univariate – reflecting the reliability of each test item or measurement individually, or multivariate – reflecting the stability of multidimensional data such as an entire scan. The standard reliability metric is intraclass correlation coefficient (ICC), a univariate anda parametric (assuming Gaussian structure) measure of the ratio of within-item variance to between-item variance. Different forms of ICC model variance and stability in different ways. We will focus here on ICC (2,1), or absolute agreement, two-way random effects ICC. For a more detailed explanation of ICC forms and how they relate to fMRI, please see Noble et al. (2021).

Univariate measures of reliability are limited in their application to fMRI data because they can miss higher dimensionality variance structure. Functional connectivity in particular can be reliable at the connectome level despite having low edge- or connection-level reliability (such as ICC). Indeed, high multivariate reliability in functional connectivity has been observed through several different measures (Bridgeford et al., 2021; Horien et al., 2019; Noble et al., 2017). Two metrics of interest in particular are fingerprinting (Finn et al., 2015) and discriminability (Bridgeford et al., 2021). Functional connectome fingerprinting reflects the proportion of subjects whose connectomes are most correlated with their own at a later timepoint. High fingerprinting accuracy has been observed in several datasets, suggesting that functional connectivity data is stable and unique enough to reliably identify subjects (Horien et al., 2019). Discriminability was developed as a multivariate reliability metric that is robust to noise and provides an upper bound on multivariate effect sizes. Initial results suggest that functional connectivity data are highly discriminable (Bridgeford et al., 2021). While we focus on these three metrics, many other means of measuring reliability exist. I2C2 and Kernel methods have been applied to fMRI data but have additional limitations and do not confer significant advantage to the present metrics.

Despite the high multivariate reliability of functional connectivity, fMRI-based prediction of psychiatric illnesses remains a challenge. Among these, major depressive disorder (MDD) is especially difficult to predict (Winter et al., 2022). Depression affects 4.4% of the global population, or 322 million people, and is the most common cause of disability (WHO, 2017). It remains unclear whether depression or other psychiatric illnesses impact reliability. Furthermore, few studies have assessed reliability in adolescence, the most common period of depression onset. This reflects a critical gap in the literature in assessing the reliability of functional connectivity in a target population for classification and prediction.

We aim to determine the test-retest reliability of functional connectivity in a cohort of adolescents with and without major depressive disorder. By characterizing the stability of their connectomes over a one-year period, we can begin to understand how age and psychiatric illness might affect reliability, informing clinical applications of fMRI. We employ a range of reliability metrics – ICC, fingerprinting, and discriminability, to determine how these measures may reflect different facets of reliability and offer unique perspectives on the data.

# METHODS

# Participants

Study participants are adolescent volunteers (age 12–19 years) recruited through mail, online advertisement and direct referrals from clinical sources. Participants provided informed consent to a protocol approved by the NIH Institutional Review Board (clinical trial no. NCT03388606) before completing questionnaires and an in-person evaluation with a medical practitioner at the NIH clinical center to guarantee their suitability to enroll in the study. Both healthy volunteers (not satisfying criteria for any diagnosis according to DSM-5) and patients with a primary diagnosis of major depression (MDD) or sub-threshold depression were included. All participants received the same scripted instructions for their participation in this study. The full list of inclusion and exclusion criteria is outlined in the Supplementary Material.

## fMRI data acquisition

Following in-person screening, participants were scanned in a General Electric (Waukesha, WI) Signa 3-Tesla MR-750s magnet, being randomly assigned to one of two similar scanners. Both scanners were housed in the NMR suite of the NIH clinical center. The fixation stimulus was displayed via back-projection from a head-coil-mounted mirror. Foam padding was used to constrain head movement. 47 oblique axial slices (3.0 mm thickness) per volume were obtained using a T2-weighted echo-planar sequence (echo time, 30 ms; flip angle, 75°; 64 × 64 matrix; field of view, 240 mm; in-plane resolution, 2.5 mm × 2.5 mm; repetition time was 2000 ms). To improve the localization of activations, a high-resolution structural image was also collected from each participant during the same scanning session using a T1-weighted standardized magnetization prepared spoiled gradient recalled echo sequence with the following parameters: 176 1 mm axial slices; repetition time, 8100 ms; echo time, 32 ms; flip angle, 7°; 256 × 256 matrix; field of view, 256 mm; in-plane resolution, 0.86 mm × 0.86 mm; NEX, 1; bandwidth, 25 kHz. During this structural scanning session, all participants watched a short neutral-mood documentary movie about bird migration.

We collected fMRI data from XX volunteers that passed our inclusion criteria. Of this sample, X participants were excluded from all reported analysis, X due to artifacts during data collection. This left 88 participants (57 MDD; 64 females; median age at baseline: 15.71).

## Data preprocessing

Processing of fMRI data was performed using Analysis of Functional and Neural Images (AFNI; Cox, 1996) software (version 19.3.14). Standard preprocessing of EPI data included slice-time correction, motion correction, spatial smoothing with a 6 mm full-width half-maximum Gaussian smoothing kernel, normalization into Talairach space and a 3D nonlinear registration. Each participant’s data were transformed to a percent signal change using the voxel-wise time-series mean blood oxygen-level-dependent (BOLD) activity. Time series were analyzed using multiple regression, where the entire trial was modeled using a gamma-variate basis function. The model also included six nuisance variables modeling the effects of residual translational (motion in the x, y, and z planes), rotational motion (roll, pitch, and yaw), and a regressor for baseline plus slow drift effect, modeled with polynomials (baseline being defined as the non-modeled phases of the task). Echo-planar images (EPIs) were visually inspected to confirm image quality and minimal movement. The code for generating the full processing stream for each participant was created using the afni\_proc.py command. This script creates also a quantitative and qualitative quality control (QC) outputs, which were used to verify the processing in the present study. We then ran a whole-brain, group-level ANOVA (3dMVM in AFNI) with the weights of the Primacy or the Recency model as between-participant covariates of each of these neural activations (each participant’s neural activity was represented by a single whole-brain image of activation across all trials).

## Functional connectivity analysis

Nodes were defined with the 122-cluster Bootstrapped Analysis of Stable Clusters (BASC) atlas. Voxel timeseries were thus averaged within each cluster and the resulting mean timeseries were used to calculate the resting state functional connectivity. Using the Nilearn software package, pairwise correlations were computed be- tween all nodes to generate a 122x122 functional connectivity matrix (the “connectome”) for each individual (Abraham et al., 2014). These matrices form the basis of the subsequent reliability analysis.

## Intraclass correlation coefficient

To assess univariate reliability of the functional connectomes, ICCs were calculated using the *psych* R package. We performed absolute agreement, two-way random effects model reliability assessment, or Shrout and Fleiss ConventionICC(2,1), which models the raters (in this case, scanners) as randomly selected from a larger group (Shrout & Fleiss, 1979). Subjects were bootstrapped 100 times with replacement to ascertain confidence intervals.

## Fingerprinting index

Fingerprinting is described in depth by Finn et. al Finn et al., 2015. For each functional connectivity matrix at timepoint 1, Pearson correlation coefficients are calculated with all connectomes at timepoint 2. The proportion of times a subject is most correlated with their own scan at the second timepoint is the fingerprinting index for that sample. The correlations are repeated between timepoint 2 scans with every timepoint 1 scan and the two values are averaged to get the overall fingerprinting accuracy.

## Discriminability

Calculation of discriminability is described in detail by Bridgeford et al. E. W. Bridgeford et al., 2021. Using a chosen distance function (the default is Euclidean), all between-measurement distances are computed. To calculate the discriminability of functional connectomes, we treat each connection between two ROIs (the correlation values that make up the functional connectome) as a measurement. Thus, for 88 122x122 connectivity matrices, the resulting distance matrix is 10736x10736 distances. Discriminability is then calculated as the proportion of within-subject distances that are smaller than between-subject distances.

## Effect size analysis

To determine if the reliability of functional connections were related to groupwise differences in functional connectivity, we correlated the edge-level effect size of each edge with a continuous edge-level measurement of each reliability metric. To generate the effect sizes, we computed the edgewise Cohen’s *d* effect size of depressed adolescents - healthy volunteers using the session 2 scan subtracted by session 1. We generated continuous edge-level forms of the three reliability metrics as follows: *ICC*: Mean ICC of bootstraps for each edge, *Fingerprinting*: Differential power and group consistency values for each edge, and *Discriminability*: Mean between-edge distance for each edge.

## Correlations with clinical measures and motion

We also compared these continuous measures to the clinical questionnaires administered at baseline and after one year. These included the Mood and Feelings Questionnaire (MFQ), SHAPS, ARI – 1 week, and SCARED. We compared the continuous reliability measures derived from each metric to both the mean value between the two visits and the change in value across visits. Similarly, medication status was categorized as a binary value (taking/not taking) for psychiatric medications and nonpsychiatric medications. To examine the effect of changing medications on reliability, we compared the change in medication status to the continuous measures. As head motion has a significant impact on reliability, we correlated the measures with average motion and average max framewise displacement. We also explored possible associations between participant age and different reliability measures.

# RESULTS

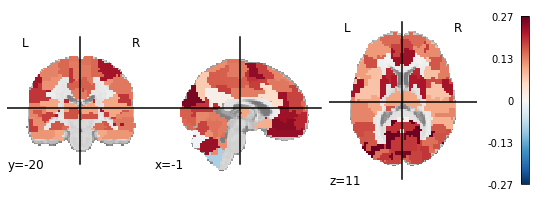
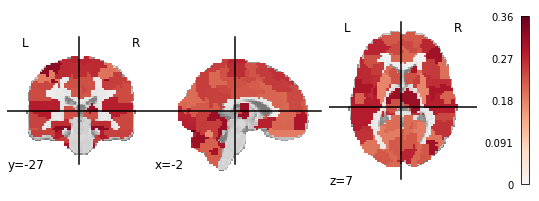
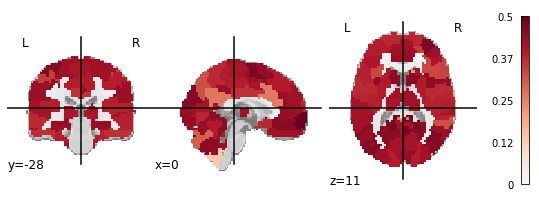
## Univariate Reliability

Edgewise reliability of functional connectivity was poor for both depressed individuals and healthy volunteers [Fig. 1]. However, the depressed participants had higher mean ICC, indicating improved reliability (0.35 > 0.24; Wilcoxon rank-sum: *p* < .001). The largest groupwise differences occurred in these connections XXX.

Chart

Description automatically generated with medium confidence

**Figure 1**: Mean edge-level ICC values for MDD (A) and HV (B), and mean contrast MDD - HV (C).



## Multivariate Reliability

Multivariate metrics reflected improved reliability compared to univariate. Fingerprinting values were moderate (XXX), but significantly greater than chance (Poisson (1) test *p* < , see (Wang et al., 2021). Fingerprinting accuracy did not differ between groups. Both groups were highly discriminable (better than chance XXX) and did not differ.

## Effect size analysis

**Intraclass Correlation Coefficient**

Edge-level ICCs were higher in MDD participants than HV participants (*µHV* = *.*24; *µMDD* = *.*35; Fig. 1). Reliability was poor in both groups, but ICCs in the depressed participants were at the higher range of expected values for resting state fMRI.

*Calculate bootstrapped means and t-test between groups with confidence intervals. Look at effects of medication status and comorbidity.*

## Fingerprinting

Fingerprinting was low (all subjects *µ* = *.*47) and did not differ between groups (*FIHV* = *.*57; *FIMDD* = *.*51). For MDD subjects with 4 month and 8 month scans, fingerprinting did not vary across timepoints (Fig. 2). Edges connected to a node in the left temporal lobe had the highest group consistency (Fig. 3).

*Add continuous measure of fingerprinting. Look at effects of medication status and comorbidity.*

## Discriminability

Participants were discriminable (*p < .*01) and did not differ between groups (*DiscrHV* = *.*75; *DiscrMDD* = *.*77; Fig. 4).

*Statistically compare discriminability values*

**Figure 2. A**: Fingerprinting matrix representing correlations between all subjects at baseline and one year. + is the highest correlation for each session 1 connectome **B**: Distribution of correlation sizes for best correlations (blue) and all correlations (red).**C**: Fingerprinting index between timepoints for first session with second session (red) and vice versa (blue). **D**: Fingerprinting for depressed participants (red) and healthy volunteers (blue).

**Figure 3.** Differential power (red) and group consistency (blue) of edges across all subjects.

**Figure 4.** Discriminability across groups.

# DISCUSSION

* There is a complex relationship between ICC, FI, discriminability, and utility.
* Fingerprinting is comparable to other adolescent datasets without motion exclusions (Fig. **??**A).
* High discriminability suggests multivariate inference may be possible.
* We should explore if the differences in ICC across groups generalize in other datasets.

references