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 in psychiatry is to elucidate the brain correlates of psychiatric illnesses. However, a major barrier to identifying these is the reproducibility of neuroimaging findings. The “replication crisis” has highlighted the difficulty of reproducing underpowered neuroimaging results. This lack of power is inseparable from the reliability of the data. Reliability places an upper bound on any observable effect size, and in turn limits 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.

Nevertheless, quantifying the reliability of fMRI is complex with many factors influencing it. Longer scans and shorter intervals between scans increase reliability, while artifact correction decrease it—underscoring the separation between validity and reliability, as noise such as motion can be highly reliable (Noble et al., 2021). Validity, the accuracy with which a measure represents the “ground truth” of a desired construct, is impossible to fully determine in fMRI due to the thousands of complex interactions that produce the blood oxygenation-level dependent (BOLD) signal. However, measures must be reliable to be valid. 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 whole-brain patterns. Univariate measures, including the widely used intraclass correlation coefficient (ICC), are typically poor in many neuroimaging modalities (Elliott et al., 2020; Noble et al., 2017, 2019). In contrast, functional connectivity has high multivariate reliability (Bridgeford et al., 2021; Horien et al., 2019; Noble et al., 2017), putatively because multivariate approaches incorporate higher dimensionality variance structure. Consequently, there have been several efforts to understand the nuance of reliability and its interpretation to optimize our data collection and processing methods.

Despite these effects and the larger goals of neuroimaging in psychiatric research, few works have investigated test-retest reliability of functional connectivity in a psychiatric population and compared it to a similar healthy population. Those that have were focused on adult populations with either mild cognitive impairment (Blautzik et al., 2013) or schizophrenia (Manoach et al., 2001). Assessing reliability in psychiatric populations is necessary for guiding the search for brain-behavior associations that can predict, diagnose, or explain illnesses (Zuo & Xing, 2014).

We investigated the test-retest reliability of functional connectivity in a cohort of adolescents with and without major depressive disorder. Major depressive disorder is the leading cause of disability worldwide and has proven one of the most challenging illnesses for reproducible biomarker identification (Fried et al., 2022; Nielson et al., 2020; WHO, 2017). Depression typically develops during adolescence, marking this a critical time for investigating neural changes that could indicate depression onset. By characterizing the stability of 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 univariate ICC and two multivariate measures of reliability: 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 is 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). We thus combined ICC, fingerprinting, and discriminability to determine how these measures may reflect different facets of reliability and offer unique perspectives on the data. In conducting these analyses, we expected adolescents with depression to have less reliable connectomes than their healthy peers in line with previous work on other psychiatric illness. We also hypothesized that multivariate reliability (fingerprinting and discriminability) would be higher than univariate (ICC). Through this investigation, we will clarify the test-retest reliability of functional connectivity in a clinically relevant population, guiding the search for biomarkers that can revolutionize psychiatry.

# METHODS

# Participants

Participants were part of the National Institute of Mental Health Characterization and Treatment of Depression (NIMH CAT-D) cohort, a longitudinal case-control study. Adolescent volunteers (age 12–19 years) were 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. After one year, participants returned for a follow-up visit consisting of questionnaires, a clinical interview, and an additional scan. The full list of inclusion and exclusion criteria is outlined in the Supplementary Material.

**Clinical Interview and Questionnaires**

Participants were diagnosed through the Kiddie-Schedule for Affective Disorders and Schizophrenia (K-SADS), a semi-structured interview (Kaufman & Schweder, 2004). Questionnaires administered included the Child Self-Report: Short Version Mood and Feelings Questionnaire (S-MFQ; Angold et al., 1995), Affective Reactivity Index – 1 week (ARI; Stringaris et al., 2012), Snaith-Hamilton Pleasure Scale (SHAPS; Snaith et al., 1995), and Screen for Child Anxiety Related Disorders (SCARED; Birmaher et al., 1997).

## fMRI data acquisition – get details from Dylan

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 – get details from Dylan

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.

## Functional connectivity analysis

Nodes were defined with the 122-cluster Bootstrapped Analysis of Stable Clusters (BASC) atlas. Voxel timeseries were 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 between all nodes to generate a 122x122 functional connectivity matrix – also known as a 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. The criteria for ICC values are typically represented as: poor < 0.4, fair 0.4-0.59, good 0.6-0.74, excellent ≥ 0.75 (Cicchetti & Sparrow, 1981).

**Fingerprinting index**

Fingerprinting is described in depth by 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

Discriminability is a nonparametric multivariate reliability metric conceived by Bridgeford et al. (2021). We calculated discriminability and ran corresponding statistical tests using the *MGC* package in R. 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.

## Edge-level and individual-level reliability associations

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 at baseline and at one year. We generated continuous edge-level forms of the three reliability metrics as follows: *ICC* – Mean bootstrapped ICC for each edge; *Fingerprinting* – Differential power and group consistency values for each edge; and *Discriminability* – Mean between-edge distance for each edge.

We also compared individual-level measures to the clinical questionnaires administered at baseline and after one year. These consisted of the Mood and Feelings Questionnaire (MFQ), SHAPS, ARI (1-week timescale), and SCARED. We derived individual-level continuous measures of reliability from the original metrics as follows: *ICC* – within subject variance for each individual, *Fingerprinting* – Ratio of mean correlation with one’s own connectome to mean correlation with others’ connectomes, and *Discriminability* – Mean between-edge distance for each individual.

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

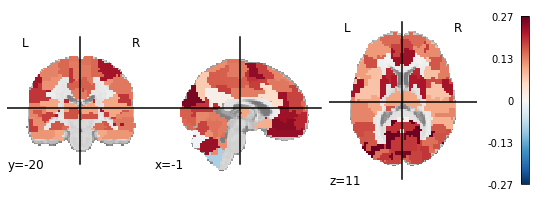
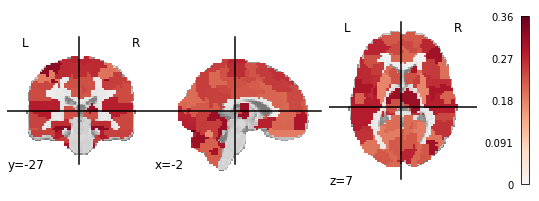
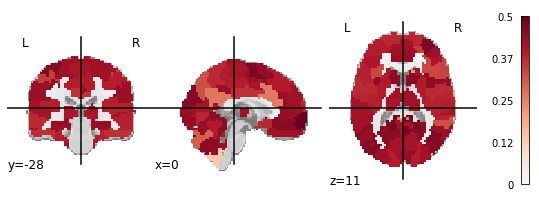
## Univariate Reliability

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).



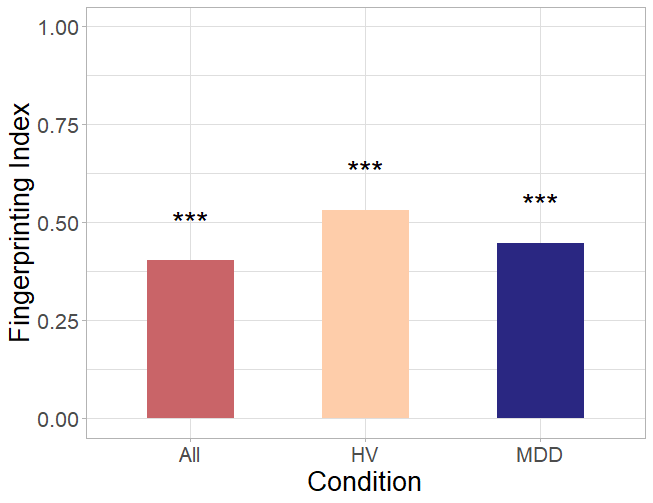
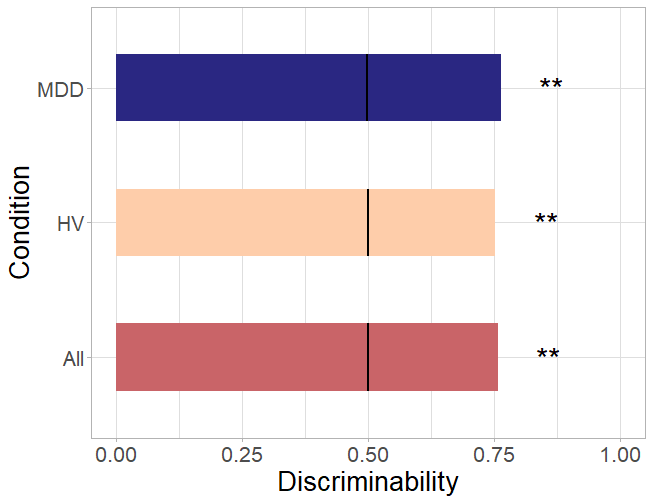
## Multivariate Reliability

Multivariate features of functional connectivity in both groups were reliable [Fig. 2]. Fingerprinting values were moderate (FIHV = 0.53; FIMDD = 0.45) but greater than chance as estimated by a Poisson (1) distribution (*p* < .001) (Wang et al., 2021). Fingerprinting accuracy did not differ between groups, *X2* (1, *N* = 88) = 0.68, *p* > .05. Connections to a region in the prefrontal cortex had the highest group consistency in both healthy and depressed participants, indicating that they reduced identifiability in all subjects [Fig. 3]. By contrast, the differential power analysis indicated that a diverse array of connections across the brain drove identifiability and the most identifiable connections were different between healthy and depressed participants.

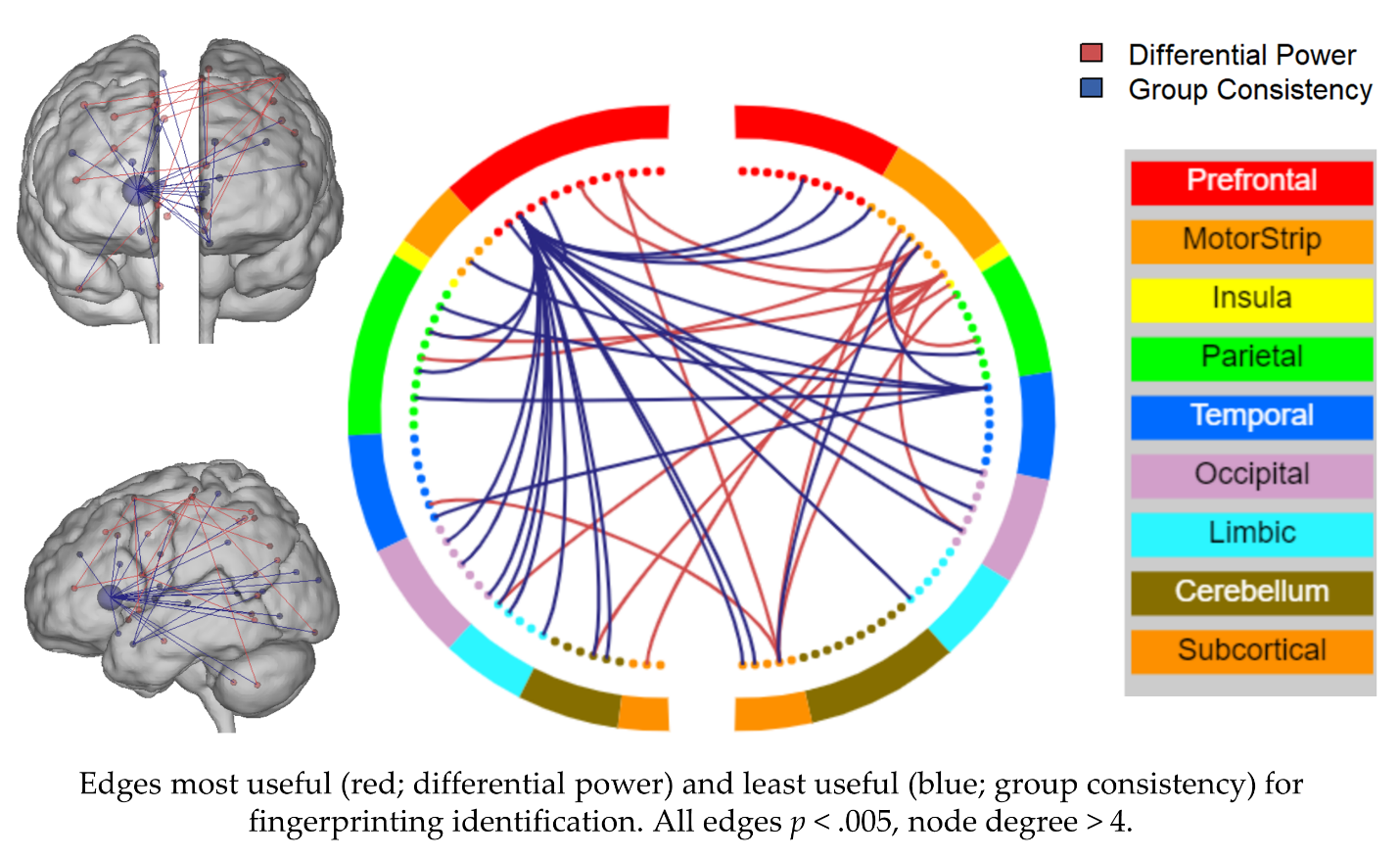
Both groups were discriminable (*Discr*HV = 0.75; *Discr*MDD = 0.76; 500-fold permutation test *p* < .01). Discriminability has a deterministic relationship with ICC, which means that an approximately equivalent ICC(2,1) value can be computed for discriminability values (Wang et al., 2020). The overall *Discr.* of 0.76 corresponds to an ICC of 0.83, suggesting that the reliability of multivariate features is much greater than at the univariate level. To put this further into context, we can calculate equivalent discriminability values for the observed univariate ICCs. The mean ICC of all participants of 0.31 is equivalent to a *Discr.* of 0.56, only just above chance discriminability.

**Power Analysis**

Reliability places a limit on the statistical power of an effect, increasing the number of subjects needed for sufficient power to observe a given effect. We conducted a power analysis to determine the number of subjects necessary to observe a small effect size (Cohen’s *d* = 0.2) with a two-sample t-test at 80% power for a given reliability value [Fig. 4]. The average univariate ICC across groups was 0.31, which would necessitate a large sample size – 1274 in each group. However, the multivariate ICC of 0.83 (ICC equivalent of *Discr.* = 0.76) reduces this figure to 472.



**Figure 2**: Fingerprinting (A) and Discriminability (B) across groups. Black lines indicate chance discriminability. \*\*: p < .01, \*\*\*: p < .001.



**Figure 3**: Edges that most improved (red; differential power) and least improved (blue; group consistency) fingerprinting identification. All edges p < .005, node degree > 4.

## Chart Description automatically generated

2000

1500

1000

500

300

150

50

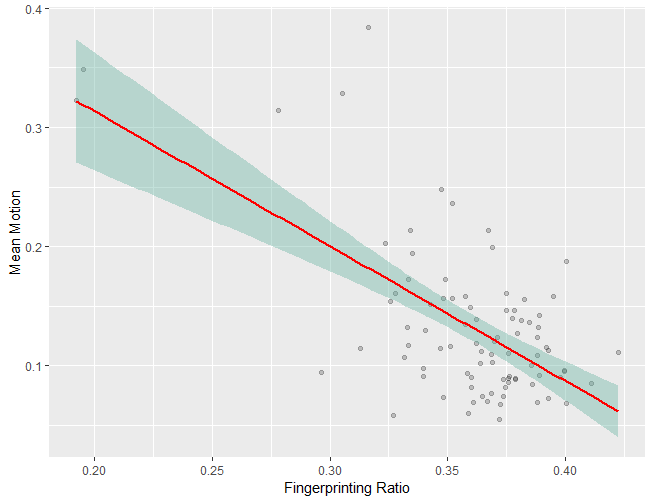
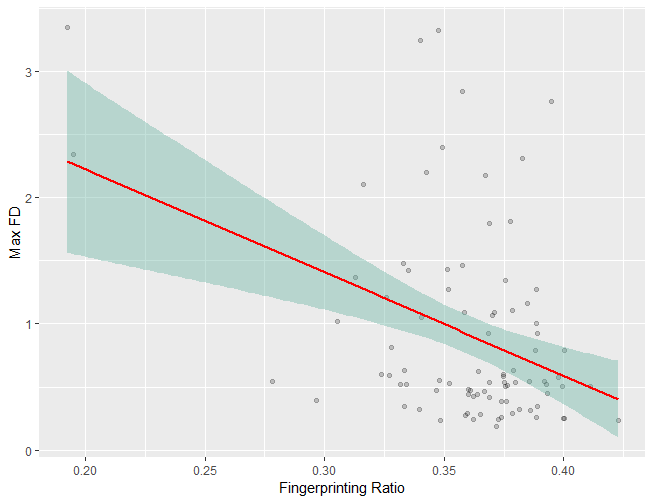
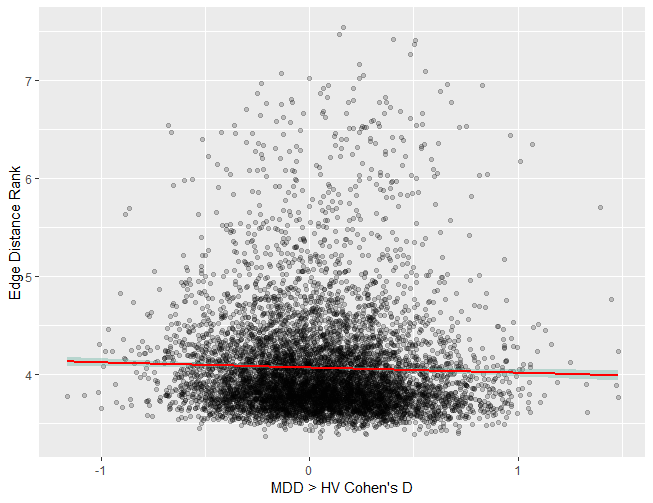
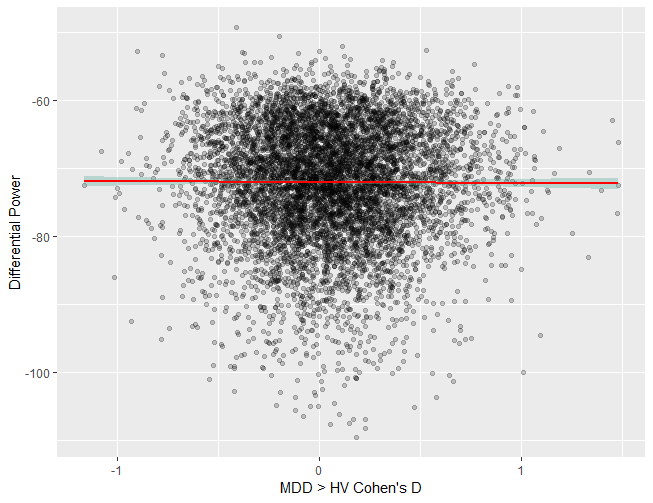
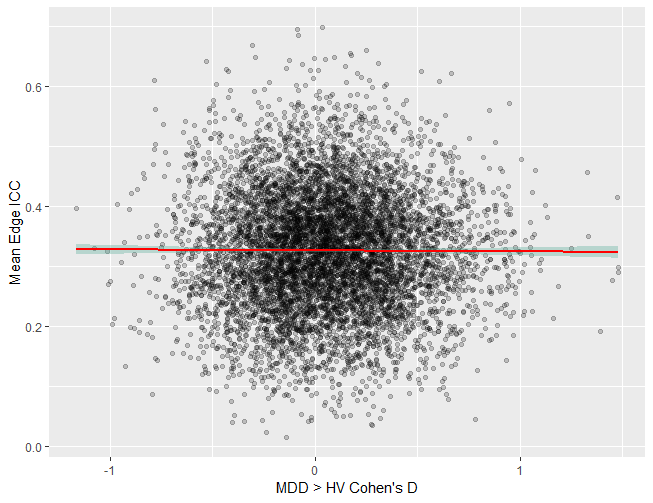
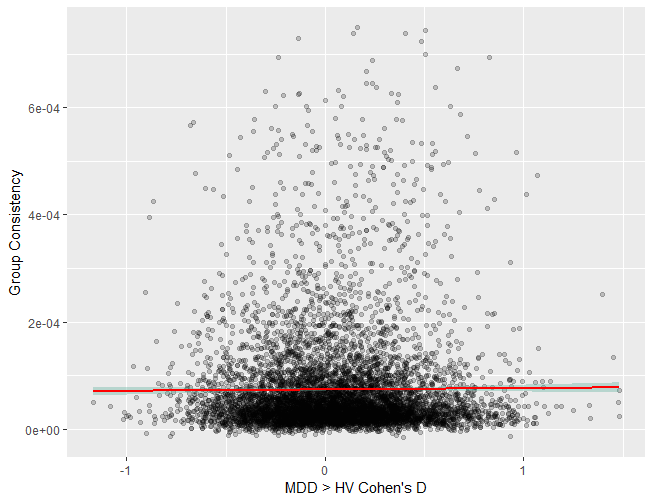
**Figure 4:** Sample size required for 80% power with a 2-sample t-test, assuming equal variance.

## Association of test-retest reliability with behavioral and clinical measures

## The reliability of edges was not associated with the edge-level effect size of group differences between depressed and healthy participants. There was no correlation between ICC, group consistency, differential power, or distance rank [Fig. 4] with MDD-HV effect size at baseline or one year.

## Continuous measures of reliability were not correlated with the change or between-session mean of the MFQ, SHAPS, ARI 1 week, or SCARED, nor did they correlate with age or medication status. The continuous measure of fingerprinting was correlated with maximum framewise displacement (Spearman’s *r* = -.38, *p* < .05) and between-session mean framewise displacement (*r* = -.63, *p* < .05) after correction for multiple comparisons [Fig. 5].

**Figure 5:** Edge-level ICC (A), Differential Power (B), Group Consistency (C), and Distance Rank (D) correlated with MDD-HV Cohen’s d effect size.



**Figure 6:** Spearman correlation between continuous measure of fingerprinting and max framewise displacement (A) and between-session mean motion (B)

**DISCUSSION**

The reliability of resting state functional connectivity and what it means for the clinical utility of fMRI is an open line of inquiry. We sought to explore several critical questions in this domain through a multifaceted reliability analysis of a clinically relevant population, namely, depressed adolescents. Over a one-year period, adolescents with depression were more reliable at the edge level than healthy participants, though both groups had poor reliability. Furthermore, both groups were similarly reliable at the multivariate connectome level. This suggests that depression does not significantly impact the test-retest reliability of functional connectivity.

**Functional connectomes in depressed adolescents are as reliable as in healthy adolescents**

Overall, we did not find evidence for an association between depression and reliability. Although the depressed group had higher ICCs on average, both groups were within the expected range of “poor” univariate reliability (<0.4). There were no significant differences in multivariate reliability, with both groups having high discriminability and moderate fingerprinting.

The observed ICC values provide further evidence for the low univariate reliability of fMRI functional connectivity (Noble et al., 2019). While shorter interscan intervals, task-based paradigms, and analyses restricted to cortical nodes can improve reliability, these options offer limited improvement and may not be feasible (Noble et al., 2017, 2021). While increased reliability does not necessarily indicate increased validity, low reliability places a ceiling on observable effect sizes, increasing type II error and restricting validity.

Contrary to our hypothesis, depressed participants were no less reliable than healthy volunteers. Although previous studies (Blautzik et al., 2013; Manoach et al., 2001) have found evidence that clinical populations are less reliable, these were conducted with task-based designs in adults with illnesses with cognitive impairment (amnesiatic mild cognitive impairment and schizophrenia). Their sample sizes were also less than 15 participants per group, limiting generalizability. Further reliability analyses in larger and more diverse clinical populations will determine if our findings generalize to other psychiatric illnesses and age groups.

**Both groups have poor univariate reliability but high multivariate reliability**

Our findings are consistent with previous work that has found resting state connectivity to be unreliable at the edge level. A review by Noble et al. (2019) found the average ICC of resting state reliability to be 0.29 (95% CI = 0.23-0.36) across 25 studies. Both depressed and healthy adolescents had ICCs within this range, further underscoring the poor reliability of resting state functional connectivity. However, we did find depressed adolescents to be more reliable than their healthy counterparts, with the greatest differences occurring in the orbitofrontal cortex, anterior cingulate cortex, occipital lobe, and sensory areas. This could be the result of increased correlated noise despite corrections for motion and physiological noise in the preprocessing pipeline. However, we did not observe an association between motion and ICC. There is some evidence that the executive function impairments associated with depression begin in adolescence, which could suggest delayed frontal cortex development (Vilgis et al., 2015). This would result in decreased within-subject variance, potentially improving reliability. Finally, this could be due to decreased between-subject variation in the depressed sample. Although depression is very heterogenous, it could still be the case that a group of depressed participants are more similar than a healthy sample. With within-subject variance constant, decreased between-subject variance would result in increased reliability.

Between-group differences in reliability were only present at the univariate level. Both groups had high multivariate reliability. Directly comparing FI across populations is difficult due to substantial impacts from sample size, scan length, and age (Horien et al., 2018; Waller et al., 2017). However, a meta-analysis of four datasets by Horien et al. (2019) found similar ID rates in an adolescent dataset of a similar size. Fingerprinting is noted to be sensitive to motion and other sources of noise (Bridgeford et al., 2021; Horien et al., 2019). We similarly observed a negative association between motion and fingerprinting. Discriminability suggested the highest reliability, with equivalent ICCs in the “excellent” range. This supports the claim that discriminability is robust to noise (Bridgeford et al., 2021). Further study is needed to determine if high discriminability is indeed reflective of increased validity, i.e., the ability to observe a ground truth association. However, we find support for discriminability as a multivariate reliability metric that captures a level of stability within the connectome.

**Reliability is not associated with behavior at the edge or individual level**

We observed no association between edge reliability from any of the metrics and between-group effect size. This supports previous findings that more reliable edges do not have greater utility – that is, they are not better for prediction or observations of brain-behavior associations (Noble et al., 2017). Similarly, we found no association between any of the clinical measures and reliability at the individual level. This underscores the distinction between reliability and validity, especially at the edge level. While edges with low reliability would obscure any true effects, those with high reliability are not necessarily more useful. Noise can be reliable, and some processing methods like motion regression decrease reliability while increasing the chance of observing a true effect. Similarly, an individual’s reliability was not associated with their symptoms, even for ICC, where there was a between-group difference. Optimizing reliability is a balancing act that must be done with careful regard for the goal of validity. This motivates the design and testing of new reliability metrics that are tied to validity.

**Power gains from multivariate methods can substantially decrease sample size**

We were able to directly compare the sample sizes required for adequate power between univariate (ICC) and multivariate (Discriminability) reliability values. Power is determined by reliability via attenuation of effect sizes (Kanyongo et al., 2007). With true effect size constant, lower reliability increases the number of participants necessary for 80% power. Reliable fMRI measures are thus critical for feasibility in terms of cost and dataset sizes. Marek et al. (2022) found that brain-wide association studies (BWAS) may require thousands of participants to observe reproducible univariate and multivariate brain-behavior associations. However, our results indicate that increasing reliability can have a significant impact. The improvement in reliability granted by a multivariate approach reduced the sample size from roughly 3000 total participants to just under 1000 to observe a small true effect size (*d* = 0.2). For a moderate effect size (*d* = 0.5), this drops to ~150 participants, a manageable size for an independent lab. Multivariate models, densely sampled populations, and perturbation studies are among the strategies that have been suggested to improve effect sizes and reliability (Finn & Rosenberg, 2021; Noble et al., 2022). As we both our reliability metrics and the reliability of our measures, we can approach the true effect sizes of brain-behavior associations by accounting for and reducing the attenuation from reliability.

**Limitations**

Our sample was small and predominantly consisted of white, high socioeconomic status (SES) youths from the DC-Maryland-Virginia area. As this was an exploratory investigation, further study is needed to determine if these results replicate in larger and more diverse depressed populations. In investigating the relationship between reliability and validity, we used between-group effect size as a rough proxy for edgewise brain-behavior associations. This was not intended to be a thorough analysis of differences in functional connectivity between depressed and healthy adolescents; rather, we sought to determine if more reliable edges were more likely to have larger effect sizes. Similarly, continuous measures derived from our reliability metrics are not directly comparable to the metrics themselves. These were also intended to reflect relative individual or edgewise reliability. We performed Spearman’s correlations to account for this, which only consider monotonic rank correlations rather than scaled.

**CONCLUSIONS**

We characterized the test-retest reliability of rsfMRI functional connectivity in 88 adolescents with and without major depressive disorder. Both depressed and healthy adolescents had low univariate reliability but high multivariate reliability, supporting the increasing shift towards multivariate models in the search for reproducible brain-behavior associations. Reliability was not associated with any clinical measures, although depressed adolescents had higher univariate reliability than healthy volunteers. Furthermore, we documented a direct relationship between reliability, observed effect size, and sample size, demonstrating that multivariate methods can decrease the requisite sample size to observe a given effect.