# Test-retest reliability of functional connectivity in depressed adolescents

# Chris C. Camp1, Stephanie Noble2, Dustin Scheinost2, Argyris Stringaris3,4, and Dylan M. Nielson5

**1** **Interdepartmental Neuroscience Program, Yale University School of Medicine, Yale University, New Haven, CT**

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

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

**41st Department of Psychiatry, National and Kapodistrian University of Athens, Aiginition Hospital**

**5Center for Multimodal Neuroimaging, National Institute of Mental Health, Bethesda, MD**

# ABSTRACT

**Introduction:** The test-retest reliability of fMRI functional connectivity is critical to identifying reproducible biomarkers for psychiatric illness. Low reliability limits the observable effect size of brain-behavior associations. Despite this, few studies have explored reliability in populations with psychiatric illnesses or across age groups. Several reliability metrics exist that can offer complementary perspectives for exploring this issue; univariate metrics capture reliability of individual connections, while multivariate metrics reflect stability of the entire connectome. **Methods:** Here, we investigate functional connectivity reliability in a longitudinal cohort of 88 adolescents with and without major depressive disorder (MDD). We compare a univariate metric, intraclass correlation coefficient (ICC), and two multivariate metrics, fingerprinting and discriminability. **Results:** Depressed adolescents had marginally higher mean ICC (μMDD=0.34, 95% CI=0.27–0.42; μHV=0.24, 95% CI=0.17–0.31), but both groups had poor average ICCs (<0.4). Fingerprinting was greater than chance and did not differ between groups (FIHV = 0.53; FIMDD = 0.45; Poisson tests *p* < .001). Discriminability indicated high multivariate reliability in both groups (*Discr*HV = 0.75; *Discr*MDD = 0.76; permutation tests *p* < .01), comparable to a univariate ICC of 0.83. Neither univariate nor multivariate reliability was associated with symptom severity or edge-level effect size of group differences. **Conclusions:** Overall, we find little evidence for a relationship between depression and reliability of functional connectivity in adolescence. These findings suggest that biomarker identification in depression is not limited due to reliability relative to healthy samples and support the shift towards multivariate analysis for improved power and reliability.

Keywords:

fMRI, functional connectivity, test-retest reliability, psychiatry, major depressive disorder, development, adolescents

# INTRODUCTION

One of the foremost goals of neuroimaging work in psychiatry is to identify the brain correlates of psychiatric illnesses. However, a major barrier to this is the reproducibility of neuroimaging findings. The “replication crisis” has highlighted the difficulty of reproducing underpowered neuroimaging results, especially from functional magnetic resonance imaging (fMRI). This lack of power is inseparable from the reliability of the data. Reliability places an upper bound on any observable effect size, limiting statistical power (Nielson et al., 2020; Zuo et al., 2019). Thus, quantifying “test-retest reliability”—the stability of a measurement over repeated tests—is critical to interpreting the validity of results.

Nevertheless, quantifying the reliability of fMRI is complex. Longer scans and shorter intervals between scans increase reliability, while artifact correction can decrease it—underscoring the separation between validity and reliability, as noise such as motion can be highly reliable but not a valid measurement of cognition (Noble et al., 2021). However, measures must be reliable to be valid and useful. Consequently, there have been several efforts to understand the nuance of reliability and its interpretation to optimize our data collection and processing methods. 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 fMRI functional connectivity (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. In other words, individual differences and similarities may appear as patterns across the connectome, rather than in individual connections or edges.

Despite its importance in psychiatric research, few works have investigated the 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 (Nielson et al., 2020; Pilmeyer et al., 2022; Zuo & Xing, 2014).

We investigated the test-retest reliability of resting state functional connectivity in a cohort of adolescents with and without major depressive disorder. By characterizing the stability of connectomes in adolescents over a four-month and one-year period, we can begin to understand how age and psychiatric illness might affect reliability, informing clinical applications of fMRI. For this reason, we focus on resting state functional connectivity, which is frequently used for biomarker identification due to its accessibility for an array of clinical populations (Nour et al., 2022). 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. Fingerprinting accuracy well above chance has been observed in several datasets, including in adolescents, 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 classification accuracy (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 addition to assessing group differences, we explored potential associations between reliability and individual differences in symptoms. Finally, we expanded previous work finding that more reliable edges are not associated with greater predictive utility into a psychiatric context (Noble et al., 2017). In conducting these analyses, we expected adolescents with depression to have less reliable connectomes than their healthy peers, as the episodic nature of the illness may result in increased within-subject variability across timepoints. This relationship would follow the limited previous work with other psychiatric illness (Blautzik et al., 2013; Manoach et al., 2001). We also predicted high multivariate reliability (fingerprinting and discriminability) and poor univariate reliability (ICC), in line with previous results (Bridgeford et al., 2021; Horien et al., 2019; Noble et al., 2019). 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 AND MATERIALS

# 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 (Gorham et al., 2022; Sadeghi et al., 2022). 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 (protocol no. 18-M-0037, clinical-trials.gov 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. Participants completed an fMRI scan after evaluation. After one year, participants returned for a follow-up visit consisting of questionnaires, a clinical interview, and an additional scan. Depressed participants also completed scans and questionnaires at four-month intervals between evaluation and the one-year timepoint. See Supplemental Table 1 for participant information.

**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

Following in-person screening, participants were scanned in a General Electric (Waukesha, WI) Discovery MR-750s 3 tesla scanner with 32 channel head coils, 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. Resting-state data was collected with a multi-echo T2\*-weighted echo-planar sequence with 32 oblique axial slices (4.0 mm thickness) (4 echoes: 15.2 ms, 28.8 ms, 42.4 ms, 56 ms; flip angle, 80°; 64 × 64 matrix; field of view, 240 mm; in-plane resolution, 3.75 mm × 3.75 mm; repetition time, 2500 ms; pixel bandwidth, 7812.5 Hz). 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 sagittal slices; inversion time 900 ms; time between inversion pulses, 2530 ms; repetition time, 7.7 ms; echo time, 3.436 ms; flip angle, 7°; 256 × 256 matrix; field of view, 256 mm; in-plane resolution, 1.0 mm × 1.0 mm; pixel bandwidth, 195.312 Hz. During this structural scanning session, all participants watched a short neutral-mood documentary movie about bird migration.

We collected fMRI data from 202 volunteers that passed our inclusion criteria. Of this sample, 10 participants were excluded from all reported analyses due to issues with data collection, quality, or processing. To have a consistent time interval between participants, we only included baseline and one-year sessions and only included participants who had usable data from both sessions, excluding 104 participants. This left 88 participants (57 MDD; 64 females; median age at baseline: 15.71). We also explored the impact of inter-scan interval on reliability in 80 depressed participants who had usable data at baseline and at a four-month scan (63 females; median age at baseline: 15.73).

## Data preprocessing

Preprocessing was performed using fMRIPrep 21.0.0 (Esteban et al., 2019; RRID:SCR\_016216), which is based on Nipype 1.6.1 (Esteban et al., 2022; Gorgolewski et al., 2011; RRID:SCR\_002502). Details of anatomical and functional data preprocessing are available in the Supplemental Materials.

## Functional connectivity analysis

Nodes were defined with the 122-cluster Bootstrapped Analysis of Stable Clusters (BASC) atlas (Bellec et al., 2010) with a manually added node for the dorsolateral prefrontal cortex. Voxel timeseries were averaged within each cluster and the resulting mean timeseries were used to calculate the resting state functional connectivity. We regressed out motion and physiological confounds and included cosine terms to control for temporal drifts. Regressors for heart rate, respiration, and respiration volume per time were created from cardiac and respiratory data with AFNI’s RetroTS.py (Cox, 1996). We included 24 motion terms, translations and rotations in each axis, those values squared, their derivatives, and derivatives squared. 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 with the healthy and depressed populations (Revelle, 2022). 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; see Noble et al. 2021 for a discussion on ICC models in neuroimaging). Subjects were bootstrapped 1000 times with replacement to ascertain confidence intervals. ICCs were then averaged across all edges to obtain group means. 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’s baseline scan has the greatest correlation with their own scan at the second timepoint (compared to all other scans at that 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. We also calculated group consistency and differential power using scripts from Horien et al. (2019). Group consistency identifies connections that were the least useful in fingerprinting, whereas differential power identifies those that were most useful (for derivations, see Finn et al., 2015). Fingerprinting confidence intervals were obtained using bootstrapping with replacement at 80% of the sample size (Horien et al., 2019).

## 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 (Bridgeford et al., 2020). To calculate the discriminability of functional connectomes, we generate a matrix of the Euclidean distances between all scans. Thus, for 88 subjects with two functional connectivity matrices each, the resulting distance matrix is 176x176 distances. Discriminability is 1 minus the proportion of between-subject measurement distances that are smaller than within-subject measurement distances (i.e. the distance between scan 1 and 2 of a subject). Wang et al. (2020) derive a formula converting discriminability to a comparable ICC value of a normally distributed univariate measurement:

We determined confidence intervals of discriminability by bootstrapping 1000 times without replacement using 80% of the sample size. This is necessary to avoid inflating values by including more scans from a single subject.

## Edge-level and individual-level reliability associations

To determine if the reliability of functional connections was related to predictive utility in psychiatric illnesses, we correlated groupwise differences in connectome edges with an edge-level measurement of each reliability metric. We computed the edgewise Cohen’s d effect size of depressed adolescents - healthy volunteers averaged between scans. We then generated edge-level forms of the three reliability metrics from all participants as follows: *ICC* – Mean bootstrapped ICC for each edge; *Fingerprinting* – Differential power and group consistency values for each edge; and *Discriminability* – z-scored discriminability of samples with edge omitted.

We also compared individual-level measures to the clinical questionnaires administered at baseline and after one year. 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 rank 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. We also explored possible confounding associations with age, medication status, and head motion. We calculated Pearson and Spearman correlations of measures with the *sjPlot* package in R (Lüdecke et al., 2022) to account for the approximation of individual-level metrics.

# RESULTS

## Univariate reliability

Edgewise reliability of functional connectivity was poor for both depressed individuals and healthy volunteers [Fig. 1]. Depressed participants had higher mean ICC, indicating greater reliability, although bootstrapped confidence intervals overlapped (μMDD=0.34, 95% CI=0.27–0.42; μHV=0.24, 95% CI=0.17–0.31) [Supplemental Figure 1]. 1000-fold bootstrapped values were nearly identical to full-sample estimates. We also measured the edgewise reliability of 80 depressed participants who completed scans four months after baseline. Reliability of depressed adolescents at four months was nearly equivalent to one year (μ=0.31; 95% CI=0.24–0.36) [Supplemental Figure 1], suggesting that the low ICC was not a function of a longer period between scans. All ICC matrices are in Supplemental Figure 2.

Chart

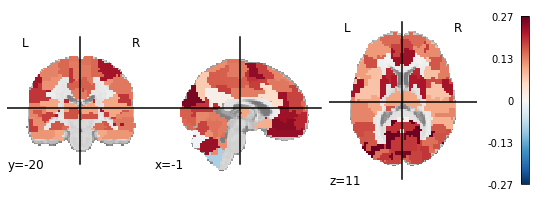
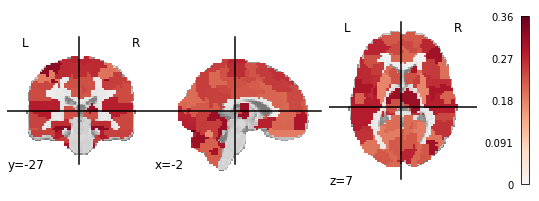
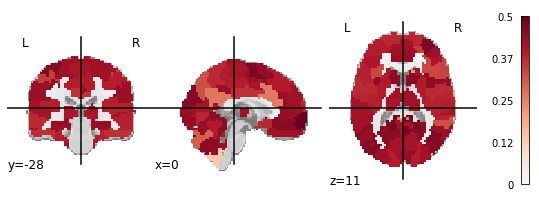
Description automatically generated with medium confidence

**B**

**A**

**C**

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



## Multivariate reliability

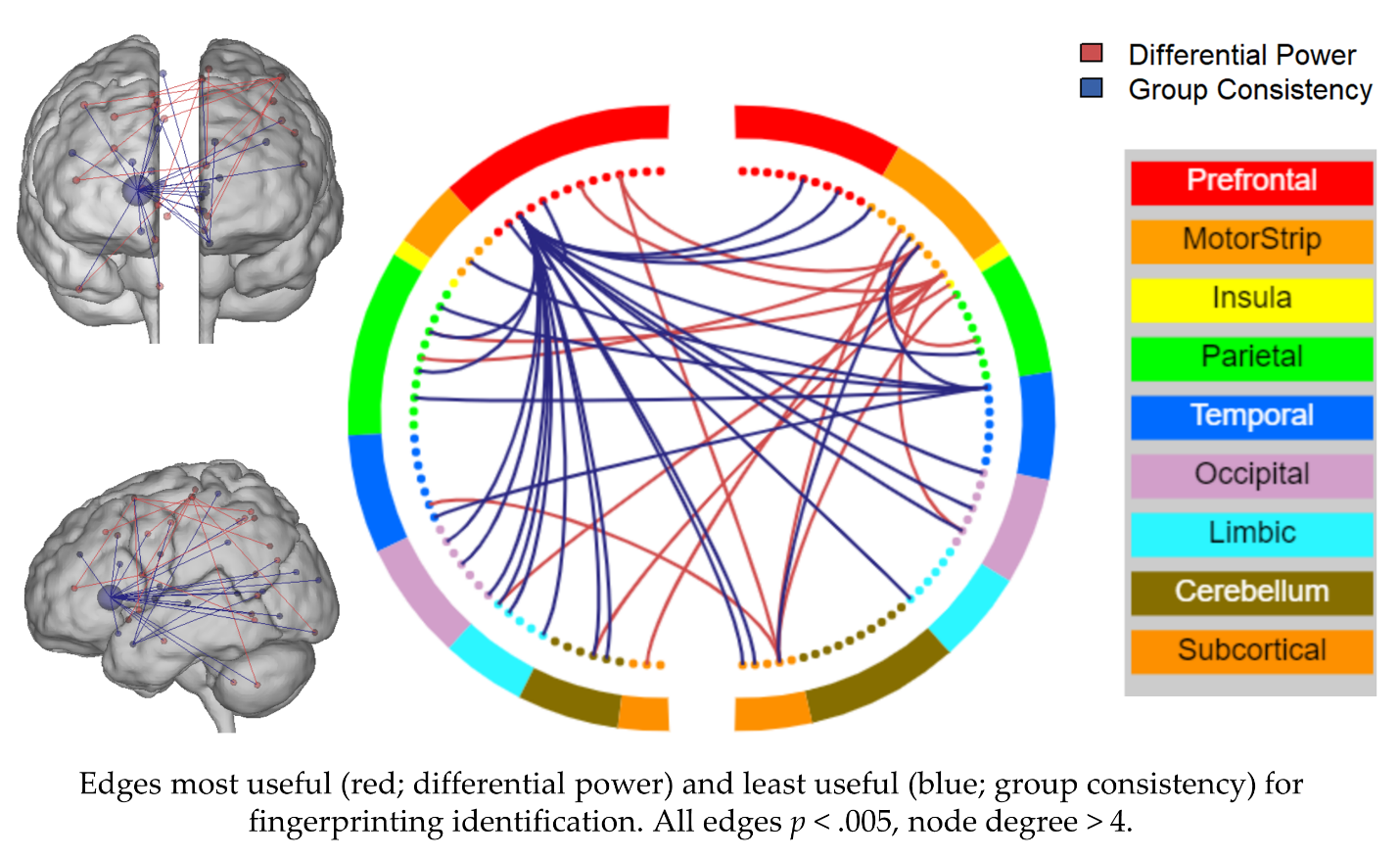
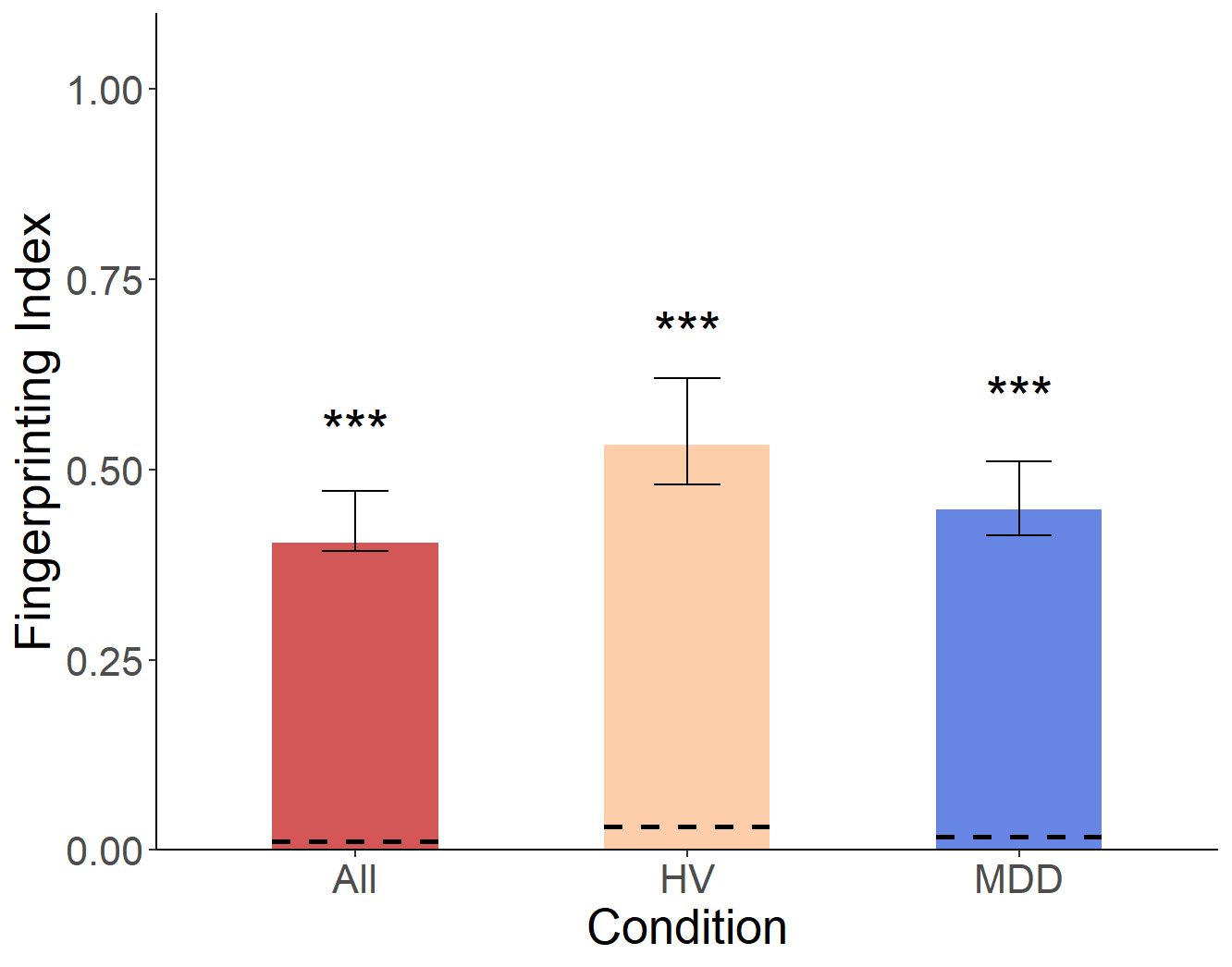
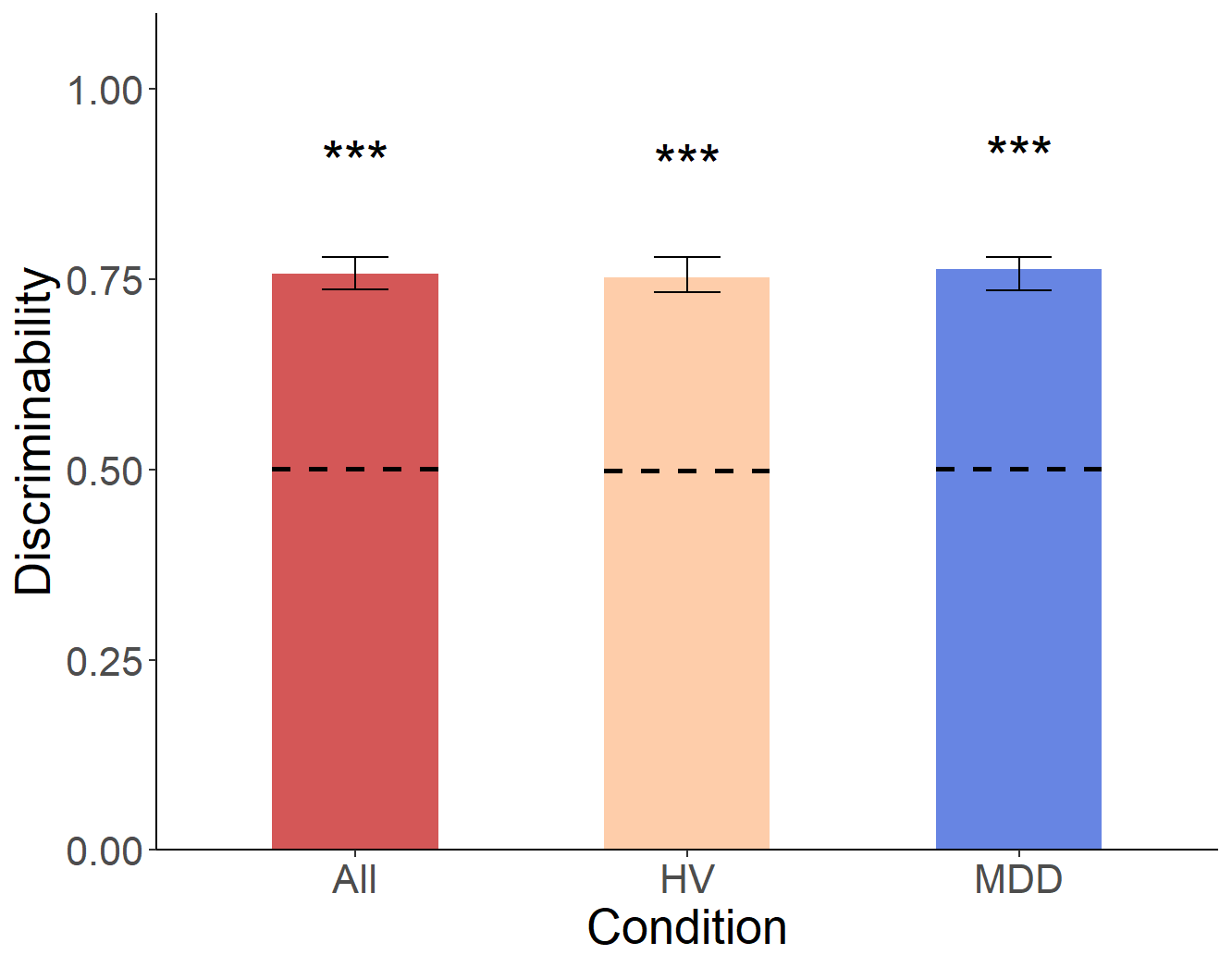
Multivariate features of functional connectivity in both groups were reliable [Fig. 2]. Fingerprinting values were greater than chance as estimated by a Poisson (1) distribution (FIMDD = 0.45, 95% CI=0.41–0.51; FIHV = 0.53, 95% CI=0.48–0.62; *p* < .001) (Wang et al., 2021). Fingerprinting accuracy was not found to differ between groups (*X2* (1, *N* = 88) = 0.49, *p* = .49), although the difference in sample sizes limits direct comparison. 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. 2C, in blue]. Notably, these edges had roughly average univariate reliability (μ = 0.27), and group consistency was not correlated with ICC (Pearson’s *r* = 0.11, *p* < .001). The differential power analysis indicated that a diverse array of connections across the brain drove identifiability [Fig. 2C, in red]. We did not calculate fingerprinting for the depressed adolescents who scanned at four months due to the differences in sample sizes.

Both groups were more discriminable than chance (*Discr*MDD = 0.76, 95% CI=0.74–0.78; *Discr*HV = 0.75, 95% CI=0.74–0.78; 500-fold permutation test *p* < .01) [Fig. 2B]. Depressed adolescents at four months had identical discriminability to those at one year. Discriminability has a deterministic relationship with ICC, which means that a comparable ICC(2,1) value can be computed for discriminability values (Wang et al., 2020). The overall *Discr.* of 0.76 corresponds to a normally distributed univariate measure with an ICC of 0.83, suggesting that the reliability of multivariate features is much greater than at the univariate level. The maximum observed ICC value across all subjects was 0.79. To put this further into context, we can calculate equivalent discriminability values for the observed univariate ICCs. The all-subjects mean ICC of 0.31 is equivalent to a *Discr.* of 0.56, only just above chance discriminability.

**B**

**A**

**C**



**Figure 2**: Fingerprinting (A) and discriminability (B) across groups. Error bars represent 95% confidence intervals derived from bootstrapping (Fingerprinting bootstraps are sampled at 80% of sample size, increasing the confidence intervals). Significance values reflect tests against chance (dashed black lines). \*\*: p < .01, \*\*\*: p < .001. C) Edges that most improved (red; differential power) and least improved (blue; group consistency) fingerprinting identification across all participants. All edges p < .005, node degree > 4.

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

## The reliability of edges was not found to be associated with the edge-level Cohen’s d effect size of between-group differences between depressed and healthy participants. Pearson correlations between ICC, group consistency, differential power, or distance rank with MDD-HV effect size had a max r = .18 (group consistency) [Fig. 3, Supplemental Table 2]. Spearman rank correlations were also minimal (max ρ = .12, group consistency).

## Continuous measures of reliability were not found to be correlated with the change or between-session mean of the MFQ, SHAPS, ARI 1 week, or SCARED, nor were correlations observed with age or medication status (max Pearson’s r = .22, max Spearman’s ρ = .22) [Supplemental Figure 3, Supplemental Table 3].

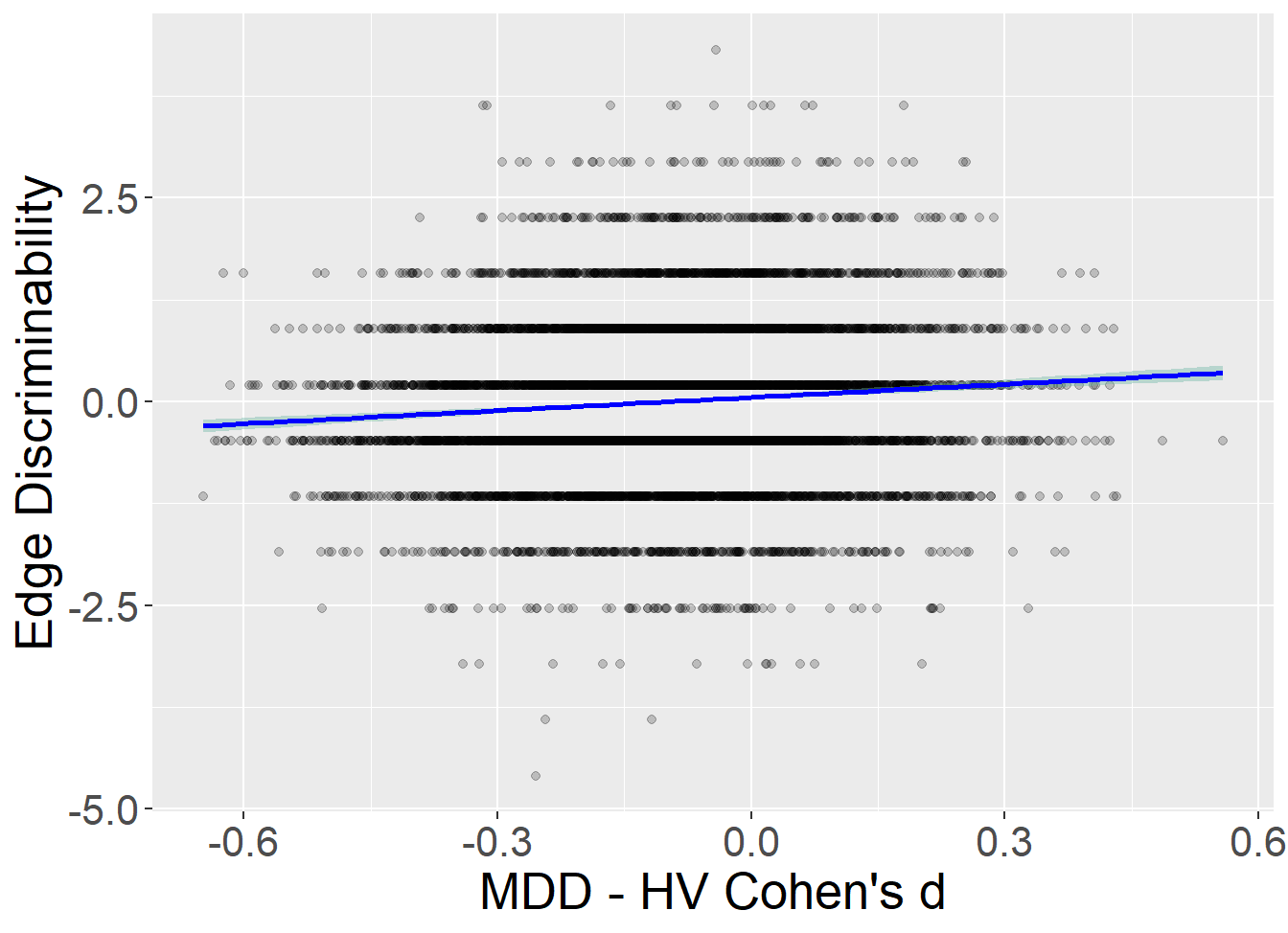
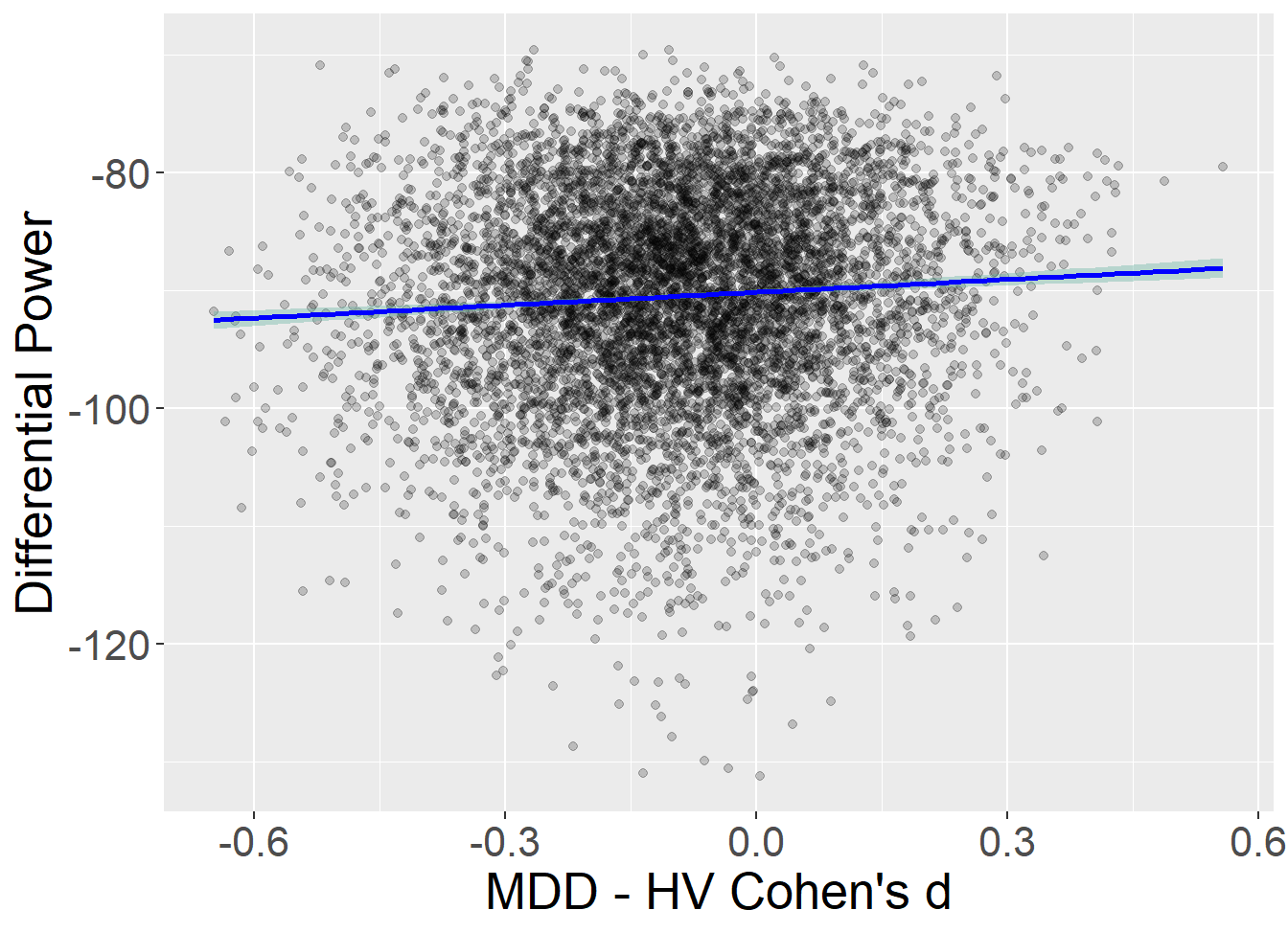
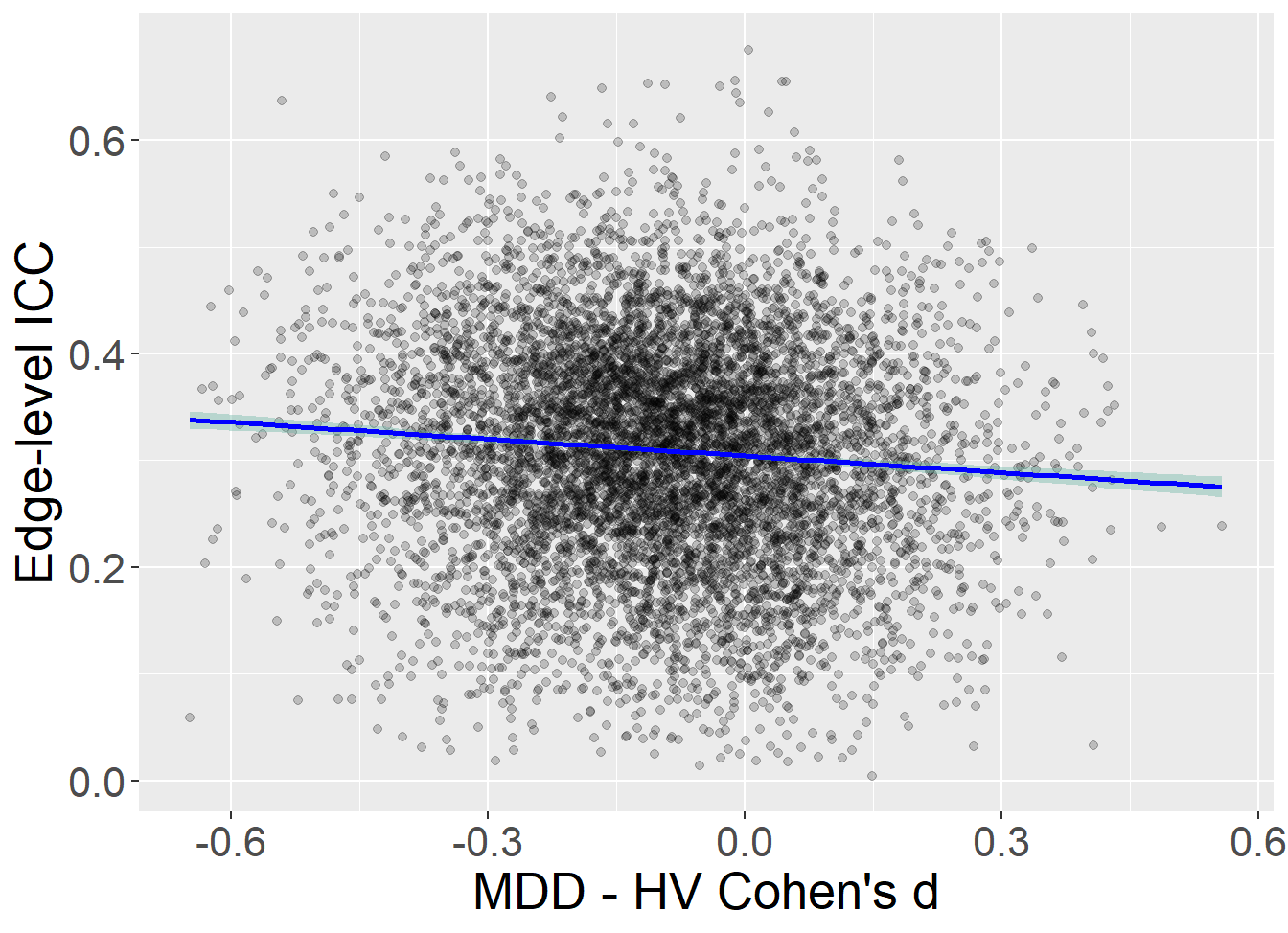
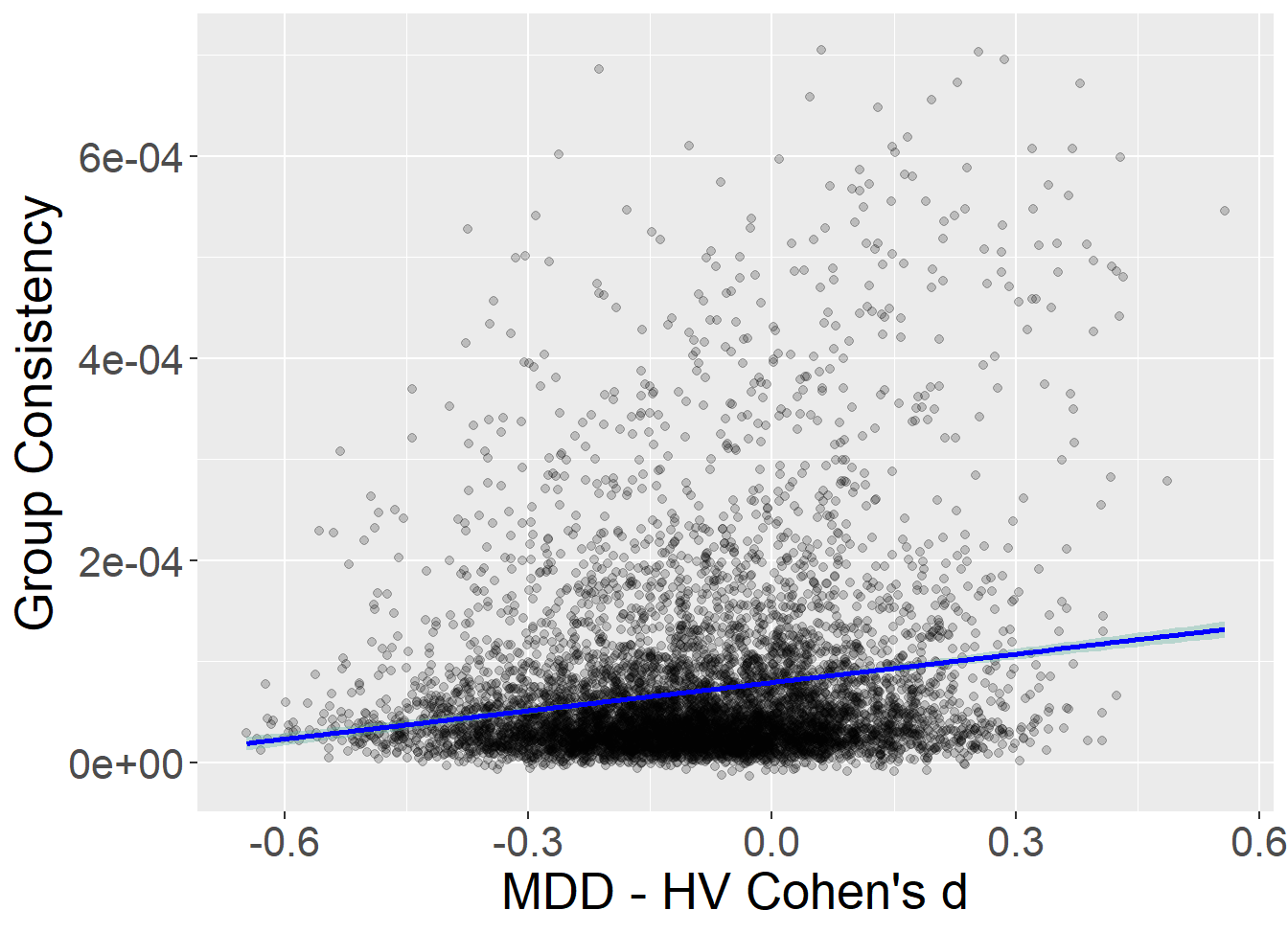
**A**

**B**

**C**

**D**

**Figure 3:** Edge-level ICC (A), differential power (B), group consistency (C), and edge discriminability (D) correlated with MDD-HV Cohen’s d effect size.



**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 four-month and one-year period. Contrary to our hypothesis, depressed adolescents were overall not less reliable than their healthy peers. Both groups exhibited similarly poor reliability at the univariate (or edge) level but high reliability at the multivariate (or connectome) level. Reliability at four months and one year were similar, suggesting that the interscan interval had limited impact on the observed reliability. Importantly, reliability was not associated with any clinical symptoms. As reliability bounds statistical power and replication, our results can inform future investigations of biomarkers of mood disorders. Depression has proven one of the most challenging illnesses for reproducible biomarker identification: Symptom presentation is highly heterogenous across patients, and recurring episodes are difficult to predict (Fried, 2022; Nielson et al., 2021; WHO, 2017). Our results suggest that despite these sources of variance, the reliability of fMRI functional connectivity is unaffected. As a result, functional connectivity may hold potential for identifying reproducible biomarkers of depression and similar illnesses, especially with multivariate methods that can improve reliability.

**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.

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). Increased reliability does not necessarily indicate increased validity, but low reliability places a ceiling on observable effect sizes, increasing type II error and restricting validity. As a result, certain univariate analyses can provide misleading results that fail to replicate without thousands of participants (Marek et al., 2022).

Discriminability suggested high reliability comparable to ICCs in the “excellent” range. While the relationship between power and reliability is dependent on the specific analysis performed, higher discriminability is associated with improved power (Bridgeford et al., 2021). We find support for discriminability as a multivariate reliability metric that captures a level of stability within the connectome and is robust to noise. The high multivariate reliability observed here supports previous recommendations that shifting towards multivariate methods can improve reproducibility (Finn & Rosenberg, 2021; Kragel et al., 2021; Marek et al., 2022; Tetereva et al., 2022). Multivariate reliability in functional connectivity may be higher than univariate due to high-dimensionality variance structures in the connectome. Whereas a single edge may be unreliable, zooming out and looking at the larger picture can reveal stable patterns. And while these approaches may reduce interpretability or effect localization, this consequence is typically modest compared to the increase in power (Noble et al., 2022).

Directly comparing fingerprinting 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 fingerprinting index values in an adolescent dataset of a similar size. Fingerprinting is sensitive to motion and other sources of noise, which tend to be amplified in younger populations. However, we did not observe an association between motion and fingerprinting after preprocessing.

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

While both groups had similar univariate and multivariate reliability, we observed greater univariate ICC in depressed adolescents. The greatest differences occurred in the frontoparietal and visual association networks. This observation 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. Delayed frontal cortex development has been observed in depressed adolescents (Straub et al., 2019), which could decrease within-subject variance and potentially increase ICC. Finally, this observation 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 to each other than a healthy sample. With within-subject variance constant, decreased between-subject variance would result in increased ICC. Nonetheless, the small effect size of this difference in combination with the multivariate and individual-level results indicate that this is unlikely to be a reproducible effect of depression on ICC.

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.

**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 necessarily 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 carry no guarantee of strong effects. Noise from sources like head motion or respiration rates can be reliable, and some processing methods like motion regression decrease reliability while increasing the chance of observing a true effect (Noble et al., 2017). 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.

**Limitations**

Psychiatric disorders are highly heterogenous. Thus, further study is needed to determine if these results replicate in other populations with depression or different psychiatric illnesses. Comparing reliability metrics is difficult – this study is a first step in interpreting the perspectives given by a range of approaches. In investigating the relationship between reliability and validity, we used between-group effect size as a rough proxy for edgewise brain-behavior associations. This analysis 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 rank correlations to account for this, which only consider monotonic correlations rather than scaled. Comparisons of multivariate and univariate reliability are approximate, as they represent different dimensions of variance. Multivariate and univariate analyses have unique conditions and requirements, and reliability can affect the results of these analyses quite differently. We converted discriminability values to ICC values to contextualize the values from a new metric, but these conversions should be viewed as experimental.

**CONCLUSIONS**

We characterized the test-retest reliability of resting-state fMRI 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 for improved reproducibility. Overall, we found that individuals with major depressive disorder were no less reliable than healthy participants, suggesting that approaches to optimize reliability and improve biomarker identification may generalize well to this population.

**REFERENCES**

Angold, A., Costello, E. J., Messer, S. C., & Pickles, A. (1995). Development of a short questionnaire for use in epidemiological studies of depression in children and adolescents. *International Journal of Methods in Psychiatric Research*, *5*(4), 237–249.

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

Birmaher, B., Khetarpal, S., Brent, D., Cully, M., Balach, L., Kaufman, J., & Neer, S. M. (1997). The Screen for Child Anxiety Related Emotional Disorders (SCARED): Scale construction and psychometric characteristics. *Journal of the American Academy of Child and Adolescent Psychiatry*, *36*(4), 545–553. https://doi.org/10.1097/00004583-199704000-00018

Blautzik, J., Keeser, D., Berman, A., Paolini, M., Kirsch, V., Mueller, S., Coates, U., Reiser, M., Teipel, S. J., & Meindl, T. (2013). Long-Term Test-Retest Reliability of Resting-State Networks in Healthy Elderly Subjects and Patients with Amnestic Mild Cognitive Impairment. *Journal of Alzheimer’s Disease*, *34*(3), 741–754. https://doi.org/10.3233/JAD-111970

Bridgeford, E. W., Shen, C., Wang, S., & Vogelstein, J. (2020). *mgc: Multiscale Graph Correlation* (2.0.2). https://CRAN.R-project.org/package=mgc

Bridgeford, E. W., Wang, S., Wang, Z., Xu, T., Craddock, C., Dey, J., Kiar, G., Gray-Roncal, W., Colantuoni, C., Douville, C., Noble, S., Priebe, C. E., Caffo, B., Milham, M., Zuo, X.-N., & Vogelstein, J. T. (2021). Eliminating accidental deviations to minimize generalization error and maximize replicability: Applications in connectomics and genomics. *PLoS Computational Biology*, *17*(9), e1009279. https://doi.org/10.1371/journal.pcbi.1009279

Cicchetti, D. V., & Sparrow, S. A. (1981). Developing criteria for establishing interrater reliability of specific items: Applications to assessment of adaptive behavior. *American Journal of Mental Deficiency*, *86*(2), 127–137.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research, an International Journal*, *29*(3), 162–173. https://doi.org/10.1006/cbmr.1996.0014

Elliott, M. L., Knodt, A. R., Ireland, D., Morris, M. L., Poulton, R., Ramrakha, S., Sison, M. L., Moffitt, T. E., Caspi, A., & Hariri, A. R. (2020). What Is the Test-Retest Reliability of Common Task-Functional MRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science*, 0956797620916786. https://doi.org/10.1177/0956797620916786

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., Kent, J. D., Goncalves, M., DuPre, E., Snyder, M., Oya, H., Ghosh, S. S., Wright, J., Durnez, J., Poldrack, R. A., & Gorgolewski, K. J. (2019). fMRIPrep: A robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*(1), Article 1. https://doi.org/10.1038/s41592-018-0235-4

Esteban, Oscar, Markiewicz, Christopher J., Burns, Christopher, Goncalves, Mathias, Jarecka, Dorota, Ziegler, Erik, Berleant, Shoshana, Ellis, David Gage, Pinsard, Basile, Madison, Cindee, Waskom, Michael, Notter, Michael Philipp, Clark, Daniel, Manhães-Savio, Alexandre, Clark, Dav, Jordan, Kesshi, Dayan, Michael, Halchenko, Yaroslav O., Loney, Fred, … Ghosh, Satrajit. (2022). *nipy/nipype: 1.8.3* (1.8.3). Zenodo. https://doi.org/10.5281/ZENODO.596855

Finn, E. S., & Rosenberg, M. D. (2021). Beyond fingerprinting: Choosing predictive connectomes over reliable connectomes. *NeuroImage*, *239*, 118254. https://doi.org/10.1016/j.neuroimage.2021.118254

Finn, E. S., Shen, X., Scheinost, D., Rosenberg, M. D., Huang, J., Chun, M. M., Papademetris, X., & Constable, R. T. (2015). Functional connectome fingerprinting: Identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, *18*(11), Article 11. https://doi.org/10.1038/nn.4135

Fried, E. I. (2022). Studying Mental Health Problems as Systems, Not Syndromes. *Current Directions in Psychological Science*, 09637214221114089. https://doi.org/10.1177/09637214221114089

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, *5*. https://doi.org/10.3389/fninf.2011.00013

Gorham, L. S., Sadeghi, N., Eisner, L., Taigman, J., Haynes, K., Qi, K., Camp, C. C., Fors, P., Rodriguez, D., McGuire, J., Garth, E., Engel, C., Davis, M., Towbin, K., Stringaris, A., & Nielson, D. M. (2022). Clinical utility of family history of depression for prognosis of adolescent depression severity and duration assessed with predictive modeling. *Journal of Child Psychology and Psychiatry*, *63*(8), 939–947. https://doi.org/10.1111/jcpp.13547

Horien, C., Noble, S., Finn, E. S., Shen, X., Scheinost, D., & Constable, R. T. (2018). Considering factors affecting the connectome-based identification process: Comment on Waller et al. *NeuroImage*, *169*, 172–175. https://doi.org/10.1016/j.neuroimage.2017.12.045

Horien, C., Shen, X., Scheinost, D., & Constable, R. T. (2019). The individual functional connectome is unique and stable over months to years. *NeuroImage*, *189*, 676–687. https://doi.org/10.1016/j.neuroimage.2019.02.002

Kaufman, J., & Schweder, A. E. (2004). The Schedule for Affective Disorders and Schizophrenia for School-Age Children: Present and Lifetime version (K-SADS-PL). In *Comprehensive handbook of psychological assessment, Vol. 2: Personality assessment* (pp. 247–255). John Wiley & Sons, Inc.

Kragel, P. A., Han, X., Kraynak, T. E., Gianaros, P. J., & Wager, T. D. (2021). Functional MRI Can Be Highly Reliable, but It Depends on What You Measure: A Commentary on Elliott et al. (2020). *Psychological Science*, *32*(4), 622–626. https://doi.org/10.1177/0956797621989730

Lüdecke, D., Bartel, A., Schwemmer, C., Powell, C., Djalovski, A., & Titz, J. (2022). *sjPlot: Data Visualization for Statistics in Social Science* (2.8.12). https://CRAN.R-project.org/package=sjPlot

Manoach, D. S., Halpern, E. F., Kramer, T. S., Chang, Y., Goff, D. C., Rauch, S. L., Kennedy, D. N., & Gollub, R. L. (2001). Test-Retest Reliability of a Functional MRI Working Memory Paradigm in Normal and Schizophrenic Subjects. *American Journal of Psychiatry*, *158*(6), 955–958. https://doi.org/10.1176/appi.ajp.158.6.955

Marek, S., Tervo-Clemmens, B., Calabro, F. J., Montez, D. F., Kay, B. P., Hatoum, A. S., Donohue, M. R., Foran, W., Miller, R. L., Hendrickson, T. J., Malone, S. M., Kandala, S., Feczko, E., Miranda-Dominguez, O., Graham, A. M., Earl, E. A., Perrone, A. J., Cordova, M., Doyle, O., … Dosenbach, N. U. F. (2022). Reproducible brain-wide association studies require thousands of individuals. *Nature*, *603*(7902), Article 7902. https://doi.org/10.1038/s41586-022-04492-9

Nielson, D. M., Keren, H., O’Callaghan, G., Jackson, S. M., Douka, I., Vidal-Ribas, P., Pornpattananangkul, N., Camp, C. C., Gorham, L. S., Wei, C., Kirwan, S., Zheng, C. Y., & Stringaris, A. (2020). Great Expectations: A Critical Review of and Suggestions for the Study of Reward Processing as a Cause and Predictor of Depression. *Biological Psychiatry*. https://doi.org/10.1016/j.biopsych.2020.06.012

Nielson, D. M., Keren, H., O’Callaghan, G., Jackson, S. M., Douka, I., Vidal-Ribas, P., Pornpattananangkul, N., Camp, C. C., Gorham, L. S., Wei, C., Kirwan, S., Zheng, C. Y., & Stringaris, A. (2021). Great Expectations: A Critical Review of and Suggestions for the Study of Reward Processing as a Cause and Predictor of Depression. *Biological Psychiatry*, *89*(2), 134–143. https://doi.org/10.1016/j.biopsych.2020.06.012

Noble, S., Mejia, A. F., Zalesky, A., & Scheinost, D. (2022). Improving power in functional magnetic resonance imaging by moving beyond cluster-level inference. *Proceedings of the National Academy of Sciences*, *119*(32), e2203020119. https://doi.org/10.1073/pnas.2203020119

Noble, S., Scheinost, D., & Constable, R. T. (2019). A decade of test-retest reliability of functional connectivity: A systematic review and meta-analysis. *NeuroImage*, *203*, 116157. https://doi.org/10.1016/j.neuroimage.2019.116157

Noble, S., Scheinost, D., & Constable, R. T. (2021). A guide to the measurement and interpretation of fMRI test-retest reliability. *Current Opinion in Behavioral Sciences*, *40*, 27–32. https://doi.org/10.1016/j.cobeha.2020.12.012

Noble, S., Spann, M. N., Tokoglu, F., Shen, X., Constable, R. T., & Scheinost, D. (2017). Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility. *Cerebral Cortex (New York, NY)*, *27*(11), 5415–5429. https://doi.org/10.1093/cercor/bhx230

Nour, M. M., Liu, Y., & Dolan, R. J. (2022). Functional neuroimaging in psychiatry and the case for failing better. *Neuron*, *110*(16), 2524–2544. https://doi.org/10.1016/j.neuron.2022.07.005

Pilmeyer, J., Huijbers, W., Lamerichs, R., Jansen, J. F. A., Breeuwer, M., & Zinger, S. (2022). Functional MRI in major depressive disorder: A review of findings, limitations, and future prospects. *Journal of Neuroimaging*, *32*(4), 582–595. https://doi.org/10.1111/jon.13011

Revelle, W. (2022). *psych: Procedures for Psychological, Psychometric, and Personality Research* (2.2.9). https://CRAN.R-project.org/package=psych

Sadeghi, N., Fors, P. Q., Eisner, L., Taigman, J., Qi, K., Gorham, L. S., Camp, C. C., O’Callaghan, G., Rodriguez, D., McGuire, J., Garth, E. M., Engel, C., Davis, M., Towbin, K. E., Stringaris, A., & Nielson, D. M. (2022). Mood and Behaviors of Adolescents With Depression in a Longitudinal Study Before and During the COVID-19 Pandemic. *Journal of the American Academy of Child & Adolescent Psychiatry*. https://doi.org/10.1016/j.jaac.2022.04.004

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, *86*(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

Snaith, R. P., Hamilton, M., Morley, S., Humayan, A., Hargreaves, D., & Trigwell, P. (1995). A scale for the assessment of hedonic tone the Snaith-Hamilton Pleasure Scale. *The British Journal of Psychiatry: The Journal of Mental Science*, *167*(1), 99–103. https://doi.org/10.1192/bjp.167.1.99

Stringaris, A., Goodman, R., Ferdinando, S., Razdan, V., Muhrer, E., Leibenluft, E., & Brotman, M. A. (2012). The Affective Reactivity Index: A concise irritability scale for clinical and research settings. *Journal of Child Psychology and Psychiatry*, *53*, 1109–1117. https://doi.org/10.1111/j.1469-7610.2012.02561.x

Tetereva, A., Li, J., Deng, J. D., Stringaris, A., & Pat, N. (2022). Capturing brain-cognition relationship: Integrating task-based fMRI across tasks markedly boosts prediction and test-retest reliability. *NeuroImage*, *263*, 119588. https://doi.org/10.1016/j.neuroimage.2022.119588

Waller, L., Walter, H., Kruschwitz, J. D., Reuter, L., Müller, S., Erk, S., & Veer, I. M. (2017). Evaluating the replicability, specificity, and generalizability of connectome fingerprints. *NeuroImage*, *158*, 371–377. https://doi.org/10.1016/j.neuroimage.2017.07.016

Wang, Z., Bridgeford, E., Wang, S., Vogelstein, J. T., & Caffo, B. (2020). Statistical Analysis of Data Repeatability Measures. *ArXiv:2005.11911 [Math, Stat]*. http://arxiv.org/abs/2005.11911

Wang, Z., Sair, H. I., Crainiceanu, C., Lindquist, M., Landman, B. A., Resnick, S., Vogelstein, J. T., & Caffo, B. (2021). On statistical tests of functional connectome fingerprinting. *Canadian Journal of Statistics*, *49*(1), 63–88. https://doi.org/10.1002/cjs.11591

WHO. (2017). *Depression and other common mental disorders: Global health estimates* (p. 24 p.) [Technical documents]. World Health Organization.

Zuo, X.-N., & Xing, X.-X. (2014). Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: A systems neuroscience perspective. *Neuroscience & Biobehavioral Reviews*, *45*, 100–118. https://doi.org/10.1016/j.neubiorev.2014.05.009

Zuo, X.-N., Xu, T., & Milham, M. P. (2019). Harnessing reliability for neuroscience research. *Nature Human Behaviour*, *3*(8), Article 8. https://doi.org/10.1038/s41562-019-0655-x