Cerebral Cortex, November 2017;27: 5415–5429

doi: 10.1093/cercor/bhx230
Advance Access Publication Date: 12 September 2017
Original Article

ORIGINAL ARTICLE

Influences on the Test–Retest Reliability of Functional Connectivity MRI and its Relationship with Behavioral Utility

Stephanie Noble[1], Marisa N. Spann[2], Fuyuze Tokoglu[3], Xilin Shen[3], R. Todd Constable[1,3,4] and Dustin Scheinost[3]

[1]Interdepartmental Neuroscience Program, Yale University, New Haven, CT 06520, USA, [2]Department of Psychiatry, College of Physicians and Surgeons, Columbia University, New York, NY 10032, USA, [3]Department of Radiology and Biomedical Imaging, Yale School of Medicine, New Haven, CT 06520, USA and [4]Department of Neurosurgery, Yale School of Medicine, New Haven, CT 06520, USA

Address correspondence to Stephanie Noble, Department of Radiology and Biomedical Imaging, Yale School of Medicine, 300 Cedar Street, PO Box 208043, New Haven, CT 06520-8043, USA. Email address: stephanie.noble@yale.edu

Abstract

Best practices are currently being developed for the acquisition and processing of resting-state magnetic resonance imaging data used to estimate brain functional organization—or "functional connectivity." Standards have been proposed based on test–retest reliability, but open questions remain. These include how amount of data per subject influences whole-brain reliability, the influence of increasing runs versus sessions, the spatial distribution of reliability, the reliability of multivariate methods, and, crucially, how reliability maps onto prediction of behavior. We collected a dataset of 12 extensively sampled individuals (144 min data each across 2 identically configured scanners) to assess test–retest reliability of whole-brain connectivity within the generalizability theory framework. We used Human Connectome Project data to replicate these analyses and relate reliability to behavioral prediction. Overall, the historical 5-min scan produced poor reliability averaged across connections. Increasing the number of sessions was more beneficial than increasing runs. Reliability was lowest for subcortical connections and highest for within-network cortical connections. Multivariate reliability was greater than univariate. Finally, reliability could not be used to improve prediction; these findings are among the first to underscore this distinction for functional connectivity. A comprehensive understanding of test–retest reliability, including its limitations, supports the development of best practices in the field.

Key words: behavioral prediction, multivariate, resting state functional connectivity, test–retest reliability, whole brain

Introduction

The cornerstones of scientific progress are reliability, reproducibility, and validity. These concepts provide complementary facets for understanding the accuracy of a measure: reliability and reproducibility refer to the consistency of a measure across repeated tests and/or varying conditions, and validity refers to the association of a measure with a predefined measure of "ground truth" (Cronbach 1988; Open Science Collaboration 2015). Unfortunately, many scientists argue that biomedical research is in the midst of a sort of "crisis of reproducibility" (Baker 2016; cf.

Begley and Ellis 2012; Pashler and Wagenmakers 2012; Open Science Collaboration 2015). Accordingly, the field of neuroimaging has begun to compile best practices to increase reliability, reproducibility, and validity (Nichols et al. 2017; Poldrack et al. 2017). A fundamental measure increasingly used to inform these best practices is "test–retest reliability," intended to reflect intrinsic stability over repeated tests.

Given the promise of resting-state functional connectivity as a research and clinical tool (Smith 2012), characterizing its test–retest reliability is an active line of research and has resulted in estimates of test–retest reliability ranging from poor to good (Shehzad et al. 2009; Van Dijk et al. 2010; Anderson et al. 2011; Birn et al. 2013; Shou et al. 2013; Laumann et al. 2015; Noble et al. 2016; Shah et al. 2016; Tomasi et al. 2016; O'Connor et al. 2017; Pannunzi et al. 2017). Some of these conflicting results are due to differences in measures of test–retest reliability (Birn et al. 2013; Laumann et al. 2015) and the selection of regions used for connectivity analyses (Anderson et al. 2011; Shah et al. 2016). This variability across studies has given rise to widely varying recommendations regarding the amount of data needed for the reliable measurement of functional connectivity. As such, understanding how much data is needed to achieve different levels of test–retest reliability for different connections across the whole brain remains an open question. A whole-brain scope is relevant to exploratory or data-driven research where anatomical constraints are not established a priori. Furthermore, mounting evidence suggests that there is rich anatomical variability in test–retest reliability across the brain (Shehzad et al. 2009; Birn et al. 2013; Laumann et al. 2015; Noble et al. 2016; Shah et al. 2016; Tomasi et al. 2016; O'Connor et al. 2017). Additionally, novel methods are being developed that treat connectivity as a multivariate object (as opposed to univariate analyses operating at the level of single connections, or "edges"; Shirer et al. 2012; Smith et al. 2015; La Rosa et al. 2016; Shen et al. 2017) or incorporate test–retest reliability into various analyses (Strother et al. 2004; Shou et al. 2014; Mueller et al. 2015; Shirer et al. 2015). However, there is a lack of work investigating test–retest reliability using multivariate frameworks or associating test–retest reliability of any single connection with measures of its behavior utility (i.e., its association with or ability to predict behavior).

Here, we used the generalizability theory framework (Webb and Shavelson 2005) to investigate the test–retest reliability of whole-brain resting-state functional connectivity. The matrix connectivity approach employed here represents a generalization of classical seed connectivity (Finn et al. 2014), allowing us to examine the connectivity amongst 268 seed regions spanning the whole brain. To assess test–retest reliability, we collected a state-of-the-art high spatial and temporal resolution dataset from 12 subjects scanned over four 36-min sessions, each session about a week apart, on 2 separate but harmonized scanners. Finally, we replicated a portion of the test–retest reliability results and related each connection's test–retest reliability to its utility in predicting behavior using the Human Connectome Project (HCP) 900 Subjects Data Release (Van Essen et al. 2013). In the following, we investigate the influence on test–retest reliability of increasing runs and sessions, the spatial distribution of reliable edges, multivariate test–retest reliability and discriminability, and, crucially, how test–retest reliability relates to the prediction of behavior.

Methods

Test–Retest Cohort

In total, 12 healthy subjects between the ages of 27 and 56 (mean = 40, standard deviation [SD] = 11) (6 males and 6 females) were recruited. Exclusion criteria included history of psychiatric illness or any magnetic resonance imaging (MRI) contraindications. All subjects gave informed consent and were compensated for their participation.

Data were acquired on 2 identically configured Siemens 3 T Tim Trio scanners at Yale University using a 32-channel head coil. After the localizer, high-resolution T1-weighted 3D anatomical scans were acquired using a magnetization prepared rapid gradient echo (MPRAGE) sequence (208 contiguous slices acquired in the sagittal plane, TR = 2400 ms, TE = 1.18 ms, flip angle=8°, thickness=1 mm, in-plane resolution=1 mm × 1 mm, matrix size = 256 × 256). Next, T1-weighted 3D anatomical scans were acquired using a fast low angle shot (FLASH) sequence (75 contiguous slices acquired in the axial plane, TR = 440 ms, TE = 2.61 ms, flip angle=70°, thickness=2 mm, in-plane resolution=2 mm × 2 mm, matrix size = 256 × 256). Next, functional images were acquired in the same slice locations as the axial T1-weighted data using a multiband echo-planar imaging (EPI) pulse sequence (75 contiguous slices acquired in the axial plane, TR = 1000 ms, TE = 30 ms, flip angle=55°, thickness=2 mm, in-plane resolution = 2 mm × 2 mm, matrix size = 110 × 110).

Each subject underwent scans over 4 sessions approximately 1 week apart (mean inter-scan interval=9.4 days, SD = 5.3 days), with all 4 scans completed within a maximum period of 1.5 months. Each subject was scanned using the 2 scanners; 2 sessions used "Scanner A", and the other 2 sessions used "Scanner B." The order of visits to scanners was counterbalanced across subjects, and visits to a single scanner were not necessarily consecutive (e.g., all visit orders were allowed). Six functional runs were collected at each session; a single 6-min run of resting-state functional data consisted of 360 continuous EPI functional volumes. Altogether, 144 min of data was collected for each subject (4 sessions/subject × 6 runs/session × 6 min/run), as shown in Figure 1. Subjects were instructed to remain still in the scanner with their eyes open, keeping their gaze on a fixation cross.

Human Connectome Project Cohort

The Human Connectome Project (HCP) 900 Subjects Data Release (Van Essen et al. 2013) was used to assess behavioral utility in a much larger set of subjects. The full dataset contains 877 individuals; this dataset was narrowed by restricting analysis only to individuals who had resting-state data from all 4 scans (n = 823/877), low motion (mean frame-to-frame displacement [mFFD] <0.1 mm; n = 610/823), and behavioral records of fluid intelligence (gF; n = 606/610). Thus, behavioral predictions were made from a final cohort of 606 subjects.

Image Analysis

Test–Retest Preprocessing

All analyses were performed using BioImage Suite (Joshi et al. 2011) unless otherwise noted. Functional images were motion-corrected using SPM5 (http://www.fil.ion.ucl.ac.uk/spm/software/spm5/). The data were then iteratively smoothed to an equivalent smoothness of a 2.5 mm Gaussian kernel in order to ensure uniform smoothness across the dataset (Friedman et al. 2006; Scheinost et al. 2014). White matter and CSF were defined on a MNI-space template brain and eroded in order to minimize inclusion of grey matter in the mask. The template was then warped to subject space using a series of transformations described in the next section. This ensured that mainly grey matter voxels were used in subsequent analyses. The following noise covariates were regressed from the data: linear, quadratic,

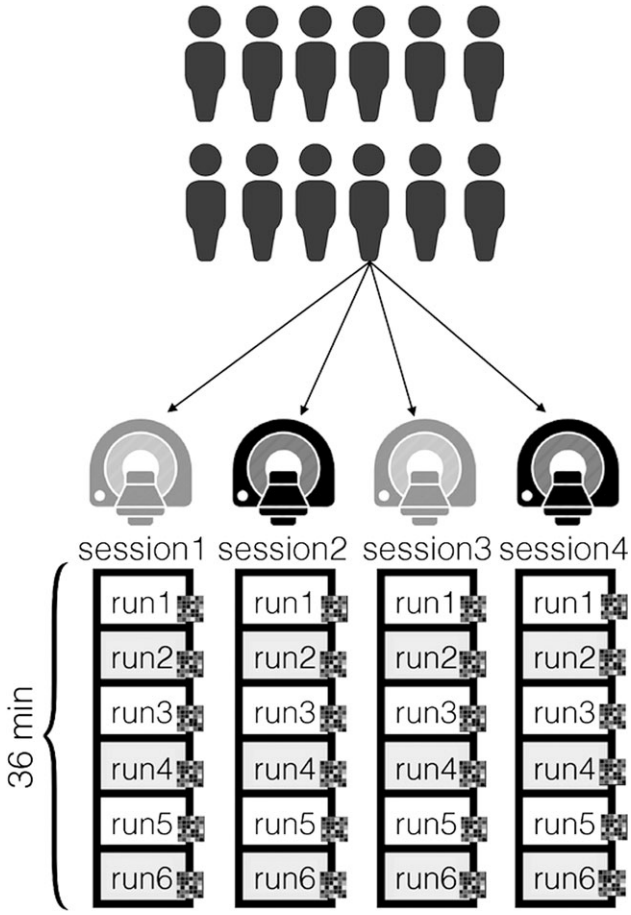

Figure 1. Study design. A total of 12 subjects were each scanned at 4 sessions (2 scanners × 2 days), with each session comprising six 6-min runs for a total of 36 min of data per session. All scans were acquired with subjects at rest with eyes open. Connectivity matrices were obtained independently for each run.

and cubic drift, a 24-parameter model of motion (Satterthwaite et al. 2013), mean cerebrospinal fluid signal, mean white matter signal, and mean global signal. Finally, data were temporally smoothed with a zero mean unit variance Gaussian filter (cutoff frequency=0.19 Hz).

Test–Retest Common Space Registration

In order to warp single subject results into MNI space, a series of linear and nonlinear transformations were estimated using BioImage Suite. Anatomical data were first skull-stripped using FSL (Smith 2002). Functional data for each subject, scanner, and session were linearly registered to the corresponding FLASH images. FLASH images were then linearly registered to MPRAGE images. Next, an average MPRAGE image for each subject was created by linearly registering and averaging all 4 anatomical images (from 2 scanner × 2 sessions) for each subject. These average MPRAGE images were used for an iterative nonlinear registration to MNI space. The use of the average anatomical images and a single nonlinear registration for each subject ensures that any potential anatomical distortions due to the different scanners does not introduce a systematic bias into the registration. The average MPRAGE images were nonlinearly registered to an evolving group average template in MNI space as described previously (Scheinost et al. 2017). All transformation pairs were calculated independently and then combined into a single transform that warps single participant results into common space. From this, all subjects' images can be transformed into common space using a single transformation, which reduces interpolation error.

HCP Preprocessing

Starting with the minimally preprocessed HCP data (Glasser et al. 2013), further preprocessing steps were performed using BioImage Suite. These included regressing 24 motion parameters, regressing the mean time courses of the white matter and CSF as well as the global signal, removing the linear trend, and low-pass filtering as described for the test–retest cohort.

Matrix Connectivity

Regions were delineated according to a 268-node whole-brain grey matter atlas (Shen et al. 2013). This atlas, defined in an independent dataset, provides a parcellation of the whole gray matter (including subcortex) into 268 contiguous, functionally coherent regions. These 268 nodes have been also grouped into 10 functionally coherent "networks" (cf. Finn et al. 2015) using the same parcellation procedure, and by anatomy into 10 "anatomical regions." Note that the subcortical-cerebellar network (previously Network 4 in Finn et al. 2015) is further divided here into 3 networks: cerebellar, subcortical, and limbic.

For each scan, the average timecourse within each region was obtained, and the Pearson's correlation between the mean time courses of each pair of regions was calculated. These correlation values provided the edge strengths for a 268 × 268 symmetric correlation matrix for each combination of subject, session, and run. These correlations were converted to z-scores using a Fisher transformation, providing a "connectivity matrix" for each combination of subject, session, and run. This connectivity matrix is analogous to a conventional seed connectivity analysis wherein a volumetric region of interest is defined and correlations are calculated between that region and all whole-brain grey matter voxels, except that other regions are used instead of voxels. Therefore, the connectivity matrix represents a set of regions ("nodes") and connections between each pair of regions ("edges").

Test–Retest Reliability Analyses in the Test–Retest Cohort

Univariate Test–Retest Reliability

The generalizability theory (G theory) framework was used to assess test–retest reliability. G theory is a generalization of classical test theory that allows the inclusion of more than one facet of measurement, or source of error (Webb and Shavelson 2005; Webb et al. 2006). G theory has been previously used to assess the test–retest reliability of multisite functional connectivity (Forsyth et al. 2014; Gee et al. 2015; Noble et al. 2016). The first step is a generalizability study (G-study), which involves estimating variance components for all factors: the "object" of measurement (here, people), "facets" of measurement (here, sessions and runs), and their interactions. The residual contains variance due to both the 3-way interaction and residual error.

A 3-way ANOVA model was used to estimate variance due to all factors modeled as random—to maximize generalizability—using the Matlab OLS-based function "anovan." Negative variance components were small in magnitude and therefore set to 0, as previously described (Shavelson et al. 1993). The model of variance is as follows, with subscripts representing factors p = person, s = session, r = run, and e = residual:

$$\sigma^2(X_{psr}) = \sigma_p^2 + \sigma_s^2 + \sigma_r^2 + \sigma_{ps}^2 + \sigma_{pr}^2 + \sigma_{sr}^2 + \sigma_{psr,e}^2.$$

These variance components can then be used to estimate test–retest reliability. For this study, we calculated "absolute reliability," which is measured by the dependability coefficient (*D*-coefficient, Φ) and reflects the absolute agreement of measurements. The *D*-coefficient is a form of the intraclass correlation coefficient (ICC; Shrout and Fleiss 1979), which is based on the ratio of variance due to the object of measurement versus sources of error (akin to an *F*-statistic). For the univariate case, *D*-coefficients (Φ_{UV}) were calculated for each edge *x* as follows:

$$\Phi_{UV}(x) = \frac{\sigma_p^2(x)}{\sigma_p^2(x) + \frac{\sigma_s^2(x)}{n_s'} + \frac{\sigma_r^2(x)}{n_r'} + \frac{\sigma_{ps}^2(x)}{n_s'} + \frac{\sigma_{pr}^2(x)}{n_r'} + \frac{\sigma_{sr}^2(x)}{n_s' \bullet n_r'} + \frac{\sigma_{psr,e}^2(x)}{n_s' \cdot n_r'}}$$

where, $\sigma_{i\ldots}^2$ represents a variance component associated with factor *i* or an interaction between facets, n_i' represents the number of conditions of facet *i* used for an average measurement (discussed in the next paragraph), and the factors are, again, *p* (person), *s* (session), and *r* (run). Note that the treatment of facets as possibly similar across subjects yet randomly selected from a larger population makes this *D*-coefficient of the form ICC(2,1) (cf. Webb et al. 2006). To summarize all edges, the mean and SD of the *D*-coefficient across all edges is presented. *D*-coefficients (like most ICC coefficients) range from 0 to 1, and can be interpreted as follows: <0.4 poor; 0.4–0.59 fair; 0.60–0.74 good; > 0.74 excellent (Cicchetti and Sparrow 1981).

This formulation enables the construction of a decision study (D-study), which provides information about what combination of measurements from each facet of measurement yields the desired level of test–retest reliability. To build a D-study matrix, *D*-coefficients are re-calculated with n_i' allowed to vary as free parameters. Increasing n_i' in the calculation of the *D*-coefficients is akin to assessing the test–retest reliability of connectivity matrices averaged over multiple sessions or runs, for example, how test–retest reliability increases with the addition of more data from either sessions or runs. Therefore, test–retest reliability obtained from $n_s' = 1$ and $n_r' = 1$ is the projected test–retest reliability of a connectivity matrix obtained from a single 6-min run (6 min total), whereas test–retest reliability obtained from $n_s' = 2$ and $n_r' = 4$ is the projected test–retest reliability of a connectivity matrix obtained from the average of matrices over 2 sessions of four 6-min runs of data (48 min total). After calculating test–retest reliability across all edges, test–retest reliability was then summarized by taking the mean within (1) nodes, (2) networks, and (3) anatomical regions. Using the univariate D-study results, the minimum scan duration required to obtain different mean levels of test–retest reliability was calculated.

To facilitate comparisons with previous studies, test–retest reliability was also assessed within the context of classical test theory. In brief, 2 versions of the ICC were calculated: (1) a coefficient describing short-term (intrasession) test–retest reliability, and (2) a coefficient describing long-term (intersession) test–retest reliability. Details can be found in the Supplementary Methods.

Region and Motion Influences on Univariate Test–Retest Reliability

The extent to which test–retest reliability was associated with node size (voxel-wise volume) and location was explored via Matlab's "fitlm." Three models were constructed: (1) test–retest reliability ~ node size, (2) test–retest reliability ~ location, and (3) test–retest reliability ~ node size + location. Location was coded as a binary variable (1 = cortex, 0 = subcortex). The fits of both single-predictor models (1 or 2) were compared against the dual-predictor model (3) via *F*-test.

The influence of motion on test–retest reliability was investigated by comparing test–retest reliability estimated in lower and higher motion groups. To estimate motion, mFFD was calculated for each subject as the average FFD across all runs and sessions. A threshold of mFFD = 0.1 was chosen to separate subjects into higher and lower motion groups, based on the mean FFD across medium motion adults in Power et al. (2014; see Fig. S1). Test–retest reliability was then calculated separately for both groups.

Multivariate Analyses

Two multivariate analyses were performed: (1) whole-brain multivariate test–retest reliability, using a method derived from the image intraclass correlation coefficient (I2C2; Shou et al. 2013), and (2) whole-brain discriminability, using 2 discrimination procedures related to the "fingerprinting" approach (Finn et al. 2015). All 35 778 unique edges in the 268-node atlas were used for these analyses.

Multivariate Test–Retest Reliability

For multivariate test–retest reliability, a single test–retest reliability coefficient was calculated by summing variance components over all edges, as described by Shou et al. (2013):

$$\Phi_{MV} = \frac{\sum_{x=1}^{X} \sigma_p^2(x)}{\sum_{x=1}^{X} \left(\sigma_p^2(x) + \frac{\sigma_s^2(x)}{n_s'} + \frac{\sigma_r^2(x)}{n_r'} + \frac{\sigma_{ps}^2(x)}{n_s'} + \frac{\sigma_{pr}^2(x)}{n_r'} + \frac{\sigma_{sr}^2(x)}{n_s' \bullet n_r'} + \frac{\sigma_{psr,e}^2(x)}{n_s' \cdot n_r'} \right)}$$

where, as above, $\sigma_{i\ldots}^2$ represents the variance component associated with factor *i* or an interaction between factors, n_i' represents the number of levels in facet i, and factors are *p* (person), *s* (session), and *r* (run). In accordance with Shou et al. (2013), the I2C2 approach represents a multivariate image measurement error model because these combined variance components reflect the true overall image variance, in contrast with the univariate ICC which reflects a univariate (marginal) measurement error model. Scan durations required for each level of test–retest reliability were then calculated as in the univariate analysis.

Multivariate Discriminability

The fingerprinting approach (Finn et al. 2015) is a complementary method of assessing the reproducibility of functional connectivity. We used 2 related discrimination approaches to determine whether subjects could be identified (using the fingerprinting procedure) or separated (using a new, related procedure). Accordingly, 2 summary statistics were calculated: the identification success rate (ISR) and the perfect separability rate (PSR). Summary statistics were calculated 6 times to assess the effect of different amounts of data, each time adding a 6-min run before re-calculating the connectivity matrix (e.g., for 6, 12, 18, 24, 30, and 36 min). Runs were added sequentially (i.e., in the order acquired) to reflect real-world conditions.

The ISR measures whether scans from the same subject can be accurately selected from a set of all scans. Each of the 48 scans (12 subjects × 4 sessions = 48 scans) served as a reference once. For each reference scan, the maximum correlation between the reference scan matrix and all other 47 scan matrices was selected. If the selected matrix belonged to the same subject as the reference matrix, the identification for that scan

was coded as a success. After repeating this procedure for all scans, the ISR was then calculated as follows:

$$\mathrm{ISR} = \frac{100}{N_p * N_s} * \sum_{p=1}^{N_p} \sum_{s=1}^{N_s} (identification_{ps} == 1)$$

where p = person and s = session. Each successful identification can be compared with the probability of choosing one of 3 correct scans out of 47 total scans at random.

The PSR is more challenging than the ISR; it describes whether all within-subject correlations exceeded all between-subject correlations for each scan. Again, each of the 48 scans were used as reference. For each reference scan, if all 3 correlations between the reference and other within-subject scan matrices exceeded all 44 correlations between the reference and between-subject scan matrices—that is, the reference subject did not once look more like another subject than him/herself—then the separation for the reference scan was recorded as a success. The PSR was calculated as follows:

$$\mathrm{PSR} = \frac{100}{N_p * N_s} * \sum_{p=1}^{N_p} \sum_{s=1}^{N_s} (separation_{ps} == 1)$$

where, p = person and s = session. Each perfect separation for a scan can be compared with the probability of the minimum of 3 randomly chosen integers exceeding the maximum of 47 randomly chosen integers.

Mean correlations within and between subjects for the above procedures are also presented. Note that correlations will always exceed ICC (Müller and Büttner 1994); therefore, a perfect Pearson's correlation can occur in the absence of identical measurements, but identical measurements must always be accompanied by a high correlation.

Estimating Scanner and Day Effects in the Test–Retest Cohort

We investigated the effect of acquiring test–retest data using identical scanners at a single site. This step is needed to determine whether it is appropriate to collapse scanner and day factors into a single session factor, which was done in the above analyses in order to estimate the presence of a session effect. A model including person, scanner, day, run, and all interactions was created to assess the presence of scanner and day effects, with subscripts representing factors p = person, s = scanner, d = day, r = run, and e = residual:

$$\sigma^2(X_{psd}) = \sigma_p^2 + \sigma_s^2 + \sigma_d^2 + \sigma_r^2 + \sigma_{ps}^2 + \sigma_{pd}^2 + \sigma_{pr}^2 + \sigma_{sd}^2 + \sigma_{sr}^2 + \sigma_{rd}^2 + \sigma_{psd}^2 + \sigma_{psr}^2 + \sigma_{pdr}^2 + \sigma_{sdr}^2 + \sigma_{prsd,e}^2.$$

These variance components were then compared with the variance components obtained with the model including person, session, run, and all interactions.

Next, we estimated whether person, scanner, and day factors were associated with significant variance across different numbers of runs, as described previously (Noble et al. 2016). We first assessed for the effect of each factor (via ANOVA), then assessed for the effect of each individual level within each factor (via GLM). For each number of runs, a connectivity matrix was re-calculated over that number of runs. Then the regression estimation procedure was repeated using that number of runs. Run was not explicitly included in either model due to the potential for inflation of estimates of significance resulting from the within-factor repeated-measures nature of the data.

Effects due to each factor were assessed as follows. The contribution of all factors to the variability in connectivity was estimated using a 3-way ANOVA with all factors modeled as random effects, as above, except without the interactions. This ensures more accurate estimates for the denominator of the F-test. The model is as follows, with subscripts representing p = person, s = session, d = day, and e = residual:

$$\sigma^2(X_{psd}) = \sigma_p^2 + \sigma_s^2 + \sigma_d^2 + \sigma_e^2.$$

Next, the F-test statistic was used to assess whether each factor was associated with significant variability in connectivity. Finally, correction via estimation of the false discovery rate (FDR) was performed separately for each factor using "mafdr" in Matlab (Storey 2002).

Next, effects due to each individual person, scanner, and day were assessed using a general linear model (GLM). Each of the 3 factors (person, scanner, and day) was modeled separately, so that 3 GLMs—1 per factor—were constructed for each edge or voxel and fit using the Matlab function "glmfit." Since the aims of this analysis are exploratory, GLMs were independently estimated to facilitate interpretation of the direct effects (Hayes 2013). The results of this analysis show whether any of the measures are significantly different from the group mean as a function of these factors. While the inclusion of the level of interest in the grand mean can decrease the power of this test, this setup is useful for comparing the effects of each level to one another because each is compared with a common reference. As above, FDR correction was performed separately for each level of each factor.

Test–Retest Reliability Analysis in the HCP Cohort

For the HCP cohort, we performed univariate and multivariate test–retest reliability analyses similar to those used in the test–retest cohort. The following model was used to estimate variance components, with subscripts representing p = person, s = session, and e = residual:

$$\sigma^2(X_{psd}) = \sigma_p^2 + \sigma_s^2 + \sigma_{ps,e}^2.$$

The 2 runs acquired in the same day for the HCP cohort were acquired with different phase encodings, which could artificially increase within-session variance and confound test–retest reliability. As such, connectivity matrices for the right-left and left-right phase encoding acquired on the same day (i.e., REST2_LR and REST2_RL) were averaged together and no run factor was included in the test–retest reliability analysis. Thus, test–retest reliability estimated with $n_s' = 1$ is the projected test–retest reliability from a 30-min session.

Behavioral Analysis Using Connectome-Based Predictive Modeling in the HCP Cohort

We then explored the association between an edge's test–retest reliability and its behavioral utility in the HCP cohort, using Connectome-based Predictive Modeling (CPM; Shen et al. 2017) to define useful edges. CPM is a data-driven protocol for developing predictive models of brain-behavior relationships from connectomes using leave-one-subject-out cross-validation. Full details of the CPM protocols are provided elsewhere (Shen et al. 2017). Briefly, CPM is composed of (1) selecting edges that are significantly correlated with behavior ($P < 0.05$), (2) summing those edge strengths into a single-subject summary score, (3) building a linear model to predict behavior based on the single-subject

summary scores, and (4) testing this predictive model in novel subjects. Fluid intelligence scores (gF) obtained via the Raven Progressive Matrices ("pmat" variable in HCP dataset) were used as the behavioral measure, and connectivity matrices were averaged over both days (60 min of data in total). Our final predictive model was only composed of edges that appeared in every round of cross-validation (cf. Rosenberg et al. 2015).

Results

Test–Retest Reliability of Univariate Functional Connectivity

Overall Test–Retest Reliability of Individual Edges

For a single, 6-min session, test–retest reliability across all edges was found to be poor ($\Phi_{UV} = 0.18 \pm 0.13$). This is mainly due to the large contribution of the residual (65.8%) relative to the other variance components, including subject (18.3%; Supplementary Table 1). This large residual indicates that all scans in the dataset are unique in a way that is not explained by the specified factors. All other variance components were relatively small (<2%) except the person by session interaction (12.0%), suggesting that subjects differ in how they progress through sessions.

To assess the influence of quantity of data on test–retest reliability, a Decision Study map was created by estimating test–retest reliability for different combinations of numbers of sessions and runs (Fig. 2). The Decision Study map is asymmetric across the diagonal, indicating that increasing the number of sessions boosts test–retest reliability more than increasing the number of runs within a session. As such, fair test–retest reliability could not be obtained within a single session, even using the maximum amount of data acquired, 36 min ($\Phi_{UV} = 0.39 \pm 0.21$). In contrast, using the minimal amount of data (6 min), it is possible to achieve fair test–retest reliability in 4 sessions ($\Phi_{UV} = 40 \pm 21$; 24 min in total). Good test–retest reliability can be obtained with a minimum of 3 sessions, requiring 35 min of data per session ($\Phi_{UV} = 0.60 \pm 0.23$; 105 min in total). However, excellent test–retest reliability cannot be obtained even using all data acquired (144 min total; $\Phi_{UV} = 0.65 \pm 0.23$).

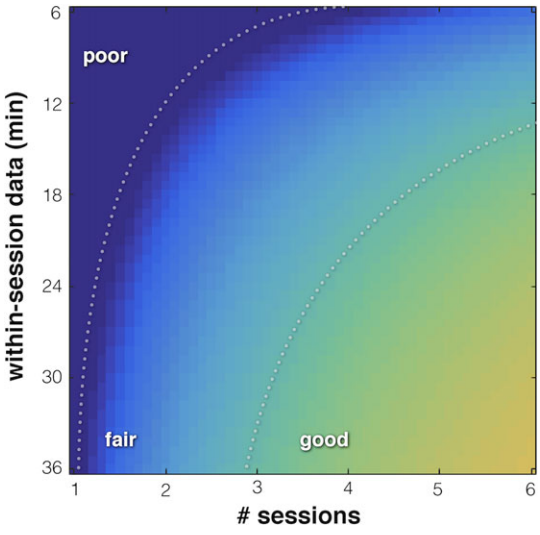

Figure 2. Effect of number of sessions and scan duration on mean test–retest reliability over all edges. A Decision Study was performed to estimate absolute reliability (Φ) as a function of scan duration (min) and number of sessions. Brighter colors correspond to higher levels of test–retest reliability, and results are categorized as follows: poor < 0.4, fair = 0.4–0.59, good = 0.6–0.74, excellent≥0.74 (Cicchetti and Sparrow 1981). The asymmetry across the diagonal indicates that test–retest reliability improves more quickly with increasing number of sessions than with increasing scan durations. Accordingly, for a single session, even with 36 min of data, only poor test–retest reliability is obtained. However, fair test–retest reliability can be obtained with only 24 min of data collected over 4 sessions of 6 min each. Good test–retest reliability requires 96–108 min of data divided over 3–4 sessions. Excellent test–retest reliability cannot be achieved using the maximum amount of data collected (4 sessions × 36 min, or 2.4 h in total).

Test–Retest Reliability of Edges Summarized by Nodes and Networks

Although test–retest reliability over all edges was poor at a single session even using the full scan duration (36 min), edges associated with certain nodes showed fair or good mean test–retest reliability (Φ = 0.4–0.6; Fig. 3*a*). Edges associated with cortical nodes ($\Phi_{ctx} = 0.20 \pm 0.05$) were more reliable than those associated with subcortical nodes ($\Phi_{subctx} = 0.12 \pm 0.04$). More reliable cortical nodes attained fair test–retest reliability with less data per subject (Fig. 3*b*). These differences are attributed to a combination of node size ($size_{ctx} = 5211 \pm 1854$ voxels; $size_{subctx} = 3408 \pm 1116$ voxels) and location (cortex or subcortex), since both were found to uniquely contribute to test–retest reliability when both were simultaneously included as predictors ($\beta_{size} = 0.67$, $P < 0.0001$; $\beta_{location} = 0.15$, $P < 0.001$). This model explained significantly more variance than either single-predictor model (Supplementary Fig. 1).

Edges associated with different large-scale functional networks exhibited different levels of test–retest reliability (Fig. 4*a*, Supplementary Fig. 2*a*). Networks with more reliable edges required less data per subject to obtain a fair level of test–retest reliability (Fig. 4*b*, Supplementary Fig. 2*b*). For cortical networks, within-network test–retest reliability (values on the diagonal in Fig. 4*a*) was greater than between-network test–retest reliability (values off the diagonal in Fig. 4*a*). For all networks, test–retest reliability increased with more data per subject. Results organized by anatomical region are also included (Supplementary Fig. 3).

Test–retest reliability estimated via classical test theory was also found to be poor with a similar spatial distribution; test–retest reliability and associated variance components for classical test theory can be found in Supplementary Figure 4.

Influence of Motion on Univariate Test–Retest Reliability

Average motion for each subject ranged between 0.0374 and 0.1818 mm mFFD (mean = 0.0993 mm, SD = 0.0506 mm; Supplementary Table 2). Overall, both higher and lower motion groups showed similar mean test–retest reliability. Test–retest reliability was found to be $\Phi_{UV} = 0.15 \pm 0.12$ for the higher motion group (mFFD > 0.1 mm; $n = 5$) and $\Phi_{UV} = 0.15 \pm 0.13$ for the lower motion group (mFFD < 0.1 mm; $n = 7$). In both cases, test–retest reliability was estimated to be lower than that obtained using the entire cohort. The spatial distribution of test–retest reliability was also found to be similar across higher and lower motion groups, with higher test–retest reliability of within-network compared with between-network edges (Supplementary Fig. 5). However, this spatial pattern was more pronounced in the lower motion group compared with the higher motion group.

Multivariate Test–Retest Reliability and Discriminability of Functional Connectivity

Multivariate Test–Retest Reliability

The D-study for multivariate test–retest reliability in shown in Figure 5. For a single session with 6 min of data, multivariate

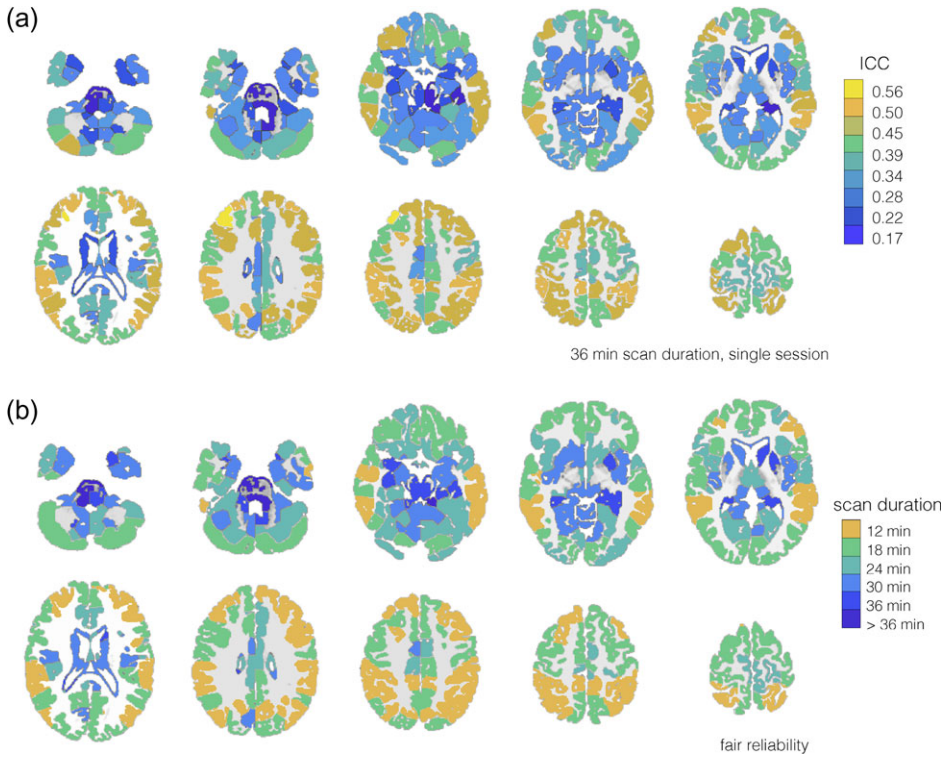

Figure 3. Spatial distribution of test–retest reliability, organized by node. (*a*) Mean test–retest reliability (Φ) of connectivity at each node. For each node, the mean test–retest reliability of all edges associated with that node was calculated for a single session with a 36-min scan duration. Brighter colors correspond to higher levels of test–retest reliability, and results are categorized as follows: poor < 0.4, fair = 0.4–0.59, good = 0.6–0.74, excellent≥0.74 (Cicchetti and Sparrow 1981). Cortical nodes exhibited greater test–retest reliability than noncortical nodes. (*b*) Minimum scan duration needed to achieve mean fair test–retest reliability at each node for a single session. Brighter colors correspond to shorter scan durations, and are scaled differently than for (*a*). Cortical nodes became reliable at shorter scan durations than noncortical nodes.

test–retest reliability was greater than univariate test–retest reliability ($\Phi_{MV} = 0.23$, $\Phi_{UV} = 0.19 \pm 0.13$). As for univariate test–retest reliability, the D-study map is asymmetric across the diagonal. Fair test–retest reliability ($\Phi_{MV} \geq 0.4$) was obtained within a single session using 18 or more min of data. Using the minimum amount of data per session (6 min), 3 sessions were required to achieve fair test–retest reliability (18 min in total). Good test–retest reliability was obtained with a minimum of 2 sessions, requiring 23 min of data per session (46 min in total). The minimum amount of data per session required to attain good test–retest reliability is 9 min, requiring 4 sessions (36 min in total). Excellent test–retest reliability was obtained with 4 sessions of 22 min of data (88 min in total). Not only is less data required to achieve higher test–retest reliability for multivariate, compared with univariate, measures, but there is a greater rate of change in test–retest reliability with increasing amounts of data.

Multivariate Discriminability

The correlations between all sessions of all subjects were calculated (Supplementary Fig. 6) and used in 2 measures of between-subject discrimination. Overall, subjects were found to be unique, but not perfectly so (Table 1, Supplementary Fig. 7*a*). With 6 min of data, the ISR was 98% (only a single session was misidentified) and reached the maximum of 100% with ≥12 min of data, which corresponded with poor univariate and multivariate test–retest reliability. The PSR, which requires all within-subject correlations to exceed all between-subject correlations, was lower. PSR was 71% when using 6 min of data and 90% when using 36 min of data. Correlations between the reference session and the session with the maximum similarity increased from $r = 0.59$ using 6 min of data to $r = 0.82$ using 36 min of data. As the amount of data per subject increased, the worst within-subject correlation increased at a greater rate than the best between-subject correlation (Supplementary Fig. 7*b*), which underlies the improvement in discrimination with increasing data.

Day and Scanner Effects

Variance was estimated for a model including scanner and day (Supplementary Fig. 8*a*). No main effects of scanner or day were found for any amount of data (except at 36 min, where a single edge showed a scanner effect). In other words, subject connectivity does not appear to systematically increase or decrease from the first scanner to the second or the first day to the second. In contrast, many edges were influenced by person with 6 min of data (~80% edges by factor, ~5% edges by level), and the number of affected edges increased with larger amounts of within-session data (at 36 min, 100% edges by factor, 25% edges by level; Supplementary Fig. 8*b*). For correspondence with the session effect, see Supplementary Results. Finally, mean matrix correlations were high within-person across scanners and days ($r_{min} = 0.75 \pm 0.07$ $r_{max} = 0.80 \pm 0.06$; Supplementary Fig. 8*c*). Altogether, these results suggest minimal systematic session or scanner effects and support pooling data across sessions or scanners, in the case of harmonized scanners.

Replication of Test–Retest Reliability in the HCP Cohort

With 30 min of data, univariate test–retest reliability was largely similar in the HCP cohort compared with the TRT cohort

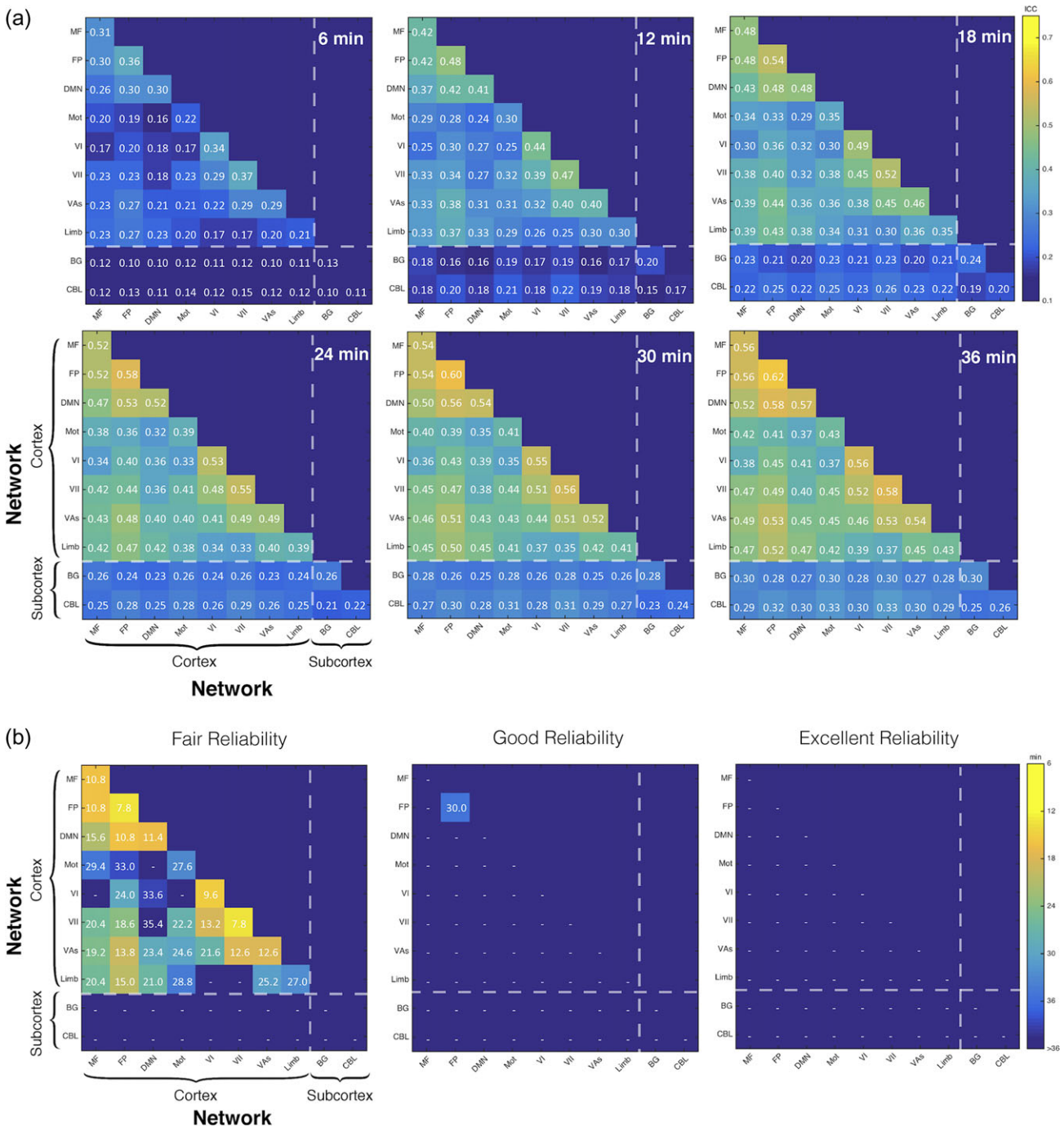

Figure 4. (*a*) Spatial distribution of test–retest reliability, summarized by network. For each pair of networks, the mean test–retest reliability (Φ) of all edges between those networks was calculated for a single session with variable scan duration (as noted in figures, scan duration=6, 12, 18, 24, 30, and 36 min). The frontoparietal, medial frontal, and secondary visual networks showed the highest test–retest reliability. Brighter colors correspond to higher levels of test–retest reliability, and results are categorized as follows: poor < 0.4, fair = 0.4–0.59, good = 0.6–0.74, excellent≥0.74 (Cicchetti and Sparrow 1981). (*b*) Minimum scan duration needed to achieve mean fair, good, and excellent test–retest reliability, summarized by network. For each pair of networks, the mean test–retest reliability of all edges between those networks was calculated (Φ) for a single session with variable scan duration (scan duration=6, 12, 18, 24, 30, and 36 min), as above. The minimum scan duration resulting in mean fair, good, and excellent test–retest reliability for that network pair was then determined. Only within-network frontoparietal connectivity reached good test–retest reliability in a single session. No network pair reached excellent test–retest reliability. Subcortical regions never reached fair test–retest reliability. Brighter colors correspond with shorter scan durations. MF, medial frontal; FP, frontoparietal; DMN, default mode; Mot, Motor; VI, visual I; VII, visual II; VAs, visual association; Lim, limbic; BG, basal ganglia (including thalamus and striatum); CBL, cerebellum.

($\Phi_{UV\text{-}HCP,30\,min}$ = 0.28 ± 0.15; $\Phi_{UV\text{-}TRT,30\,min}$ = 0.37 ± 0.20). Test–retest reliability (Φ_{UV}) was correlated between the 2 cohorts at the single edge level (r_{UV} = 0.57, P < 0.0001; Supplementary Fig. 9a). Between cohorts, mean reliabilities were almost perfectly correlated at the network level ($r_{network}$ = 0.94, P < 0.0001; Supplementary Fig. 9b) and were highly correlated at the node- and anatomical region-levels (r_{node} = 0.79, P < 0.0001; $r_{anat.region}$ = 0.91, P < 0.0001). Multivariate test–retest reliability

was slightly lower for the HCP compared with the TRT cohort ($\Phi_{MV\text{-}HCP,30\,min} = 0.33$; $\Phi_{MV\text{-}TRT,30\,min} = 0.48$).

Association Between Test–Retest Reliability and Behavioral Utility

The model obtained from the CPM analysis significantly predicted gF (correlation between predicted and observed value, $r = 0.22$, $P < 0.0001$; Supplementary Fig. 10*a*). Edges used to predict gF were associated with both cortical and subcortical regions. The difference between test–retest reliability (Φ_{UV}) of predictive and nonpredictive edges exhibited a small effect size (Cohen's $d = 0.14$; Fig. 6). Similarly, using all edges, the association between the test–retest reliability of an edge and its behavioral relevance (i.e., magnitude of its correlation with gF) was found to be small ($r = 0.05$), with test–retest reliability accounting for less than 1% of the variance in behavioral relevance of edges (Fig. 7). Finally, we repeated our CPM analysis after removing the least reliable edges (or the most reliable edges) from the analysis in an iterative fashion, increasing the number of edges removed at each iteration by 3200 (9% of all edges; Fig. 8; see Supplementary Fig. 10*b* for additional results from positive and negative predictive networks). Prediction performance remained mainly stable when removing edges in either direction (reliable or unreliable). The influence of removing edges in either direction was inconsistent, with performance never increasing by more than +0.025 from performance with all edges included. Altogether, these results illustrate a dissociation between an edge's test–retest reliability and its utility in predicting behavior.

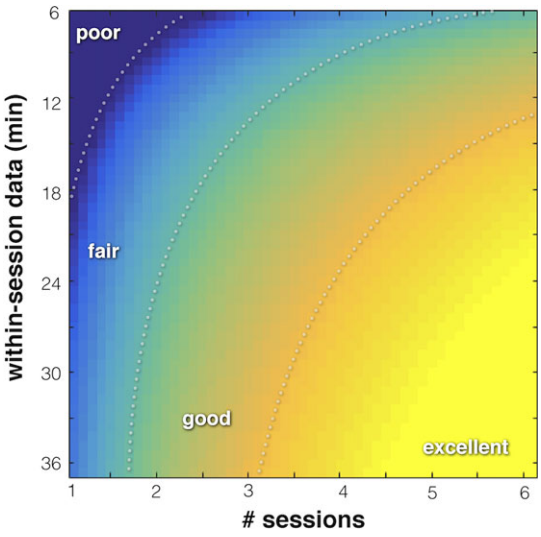

Figure 5. Effect of number of sessions and scan duration on multivariate test–retest reliability of the connectivity matrix. A Decision Study was performed to estimate multivariate absolute test–retest reliability (Φ) as a function of scan duration (min) and number of sessions. Brighter colors correspond to higher levels of test–retest reliability, and results are categorized as follows: poor < 0.4, fair = 0.4–0.59, good = 0.6–0.74, excellent≥0.74 (Cicchetti and Sparrow 1981). The asymmetry across the diagonal indicates that test–retest reliability improves more quickly with increasing number of sessions than with increasing scan durations. Accordingly, fair test–retest reliability can be obtained with less total data if data is acquired with multiple sessions than if only a single session is used. Unlike in the univariate case, mean fair test–retest reliability can be achieved in a single session and mean excellent test–retest reliability can be achieved using the maximum amount of data collected (4 sessions × 36 min).

Discussion

As studies continue to suggest the promise of resting-state functional connectivity, full characterization of its reliability is necessary. To add to previous studies of the test–retest reliability of resting-state functional connectivity, we investigated the influence of the amount of data per subject on test–retest reliability, characterized the whole-brain spatial distribution of test–retest reliability, compared univariate and multivariate measures of test–retest reliability, and explored the association between test–retest reliability and behavioral prediction. Overall, our results suggest that the historical 5-min resting-state scan is associated with poor average test–retest reliability of whole-brain connectivity, and support ongoing efforts to collect more data per subject (Laumann et al. 2015). While test–retest reliability averaged over all edges was poor, this was mainly due to the lower test–retest reliability of noncortical edges. Within-network connectivity was the most reliable, consistent with previous studies (Shehzad et al. 2009; Birn et al. 2013; O'Connor et al. 2017). Multivariate test–retest reliability was substantially greater than univariate test–retest reliability, indicating that the connectivity matrix, or connectome, as a whole contains more stable information than any particular edge. Finally, we found that an edge's test–retest reliability was not meaningfully correlated with that edge's contribution to behavioral prediction and that removing the least and most reliable edges did not substantially change prediction performance. These findings are among the first to underscore the distinction between reliability and utility,

Table 1 Multivariate discriminability increases with scan duration

	6 min	12 min	18 min	24 min	30 min	36 min
Discriminability						
ISR	98%	100%	100%	100%	100%	100%
PSR	71%	73%	79%	91%	90%	90%
Mean correlations						
Match	0.5938	0.7101	0.7615	0.792	0.8105	0.8191
Within-sub	0.5512	0.6643	0.7152	0.7477	0.7683	0.7791
Between-sub	0.3828	0.4644	0.502	0.5265	0.5424	0.551
Worst within	0.5036	0.617	0.6679	0.7041	0.723	0.7348
Best between	0.4686	0.5501	0.594	0.6133	0.6272	0.6353

Two discriminability measures were used here: the identification success rate (ISR) and the perfect separability rate (PSR). The ISR measures whether matrices from the same subject can be accurately chosen from a set of all other 47 scan matrices. The PSR measures whether, for each reference scan, all 3 within-subject correlations exceed all 44 between-subject correlations. The following metrics are reported for each scan duration from 6 to 36 min: ISR, PSR, mean match correlation, mean within-subject correlation, mean worst (lowest) within-subject correlation, and mean best (highest) between-subject correlation.

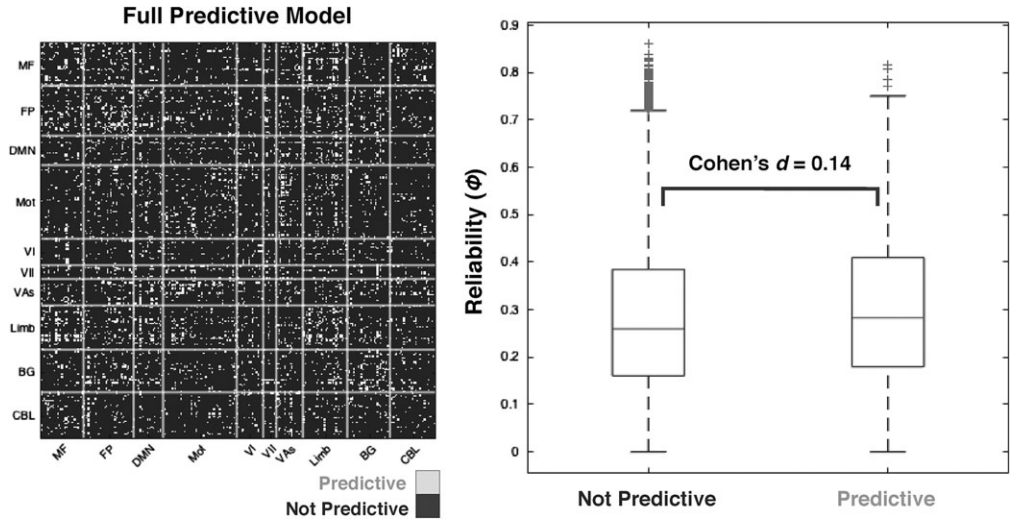

Figure 6. Difference in test–retest reliability between edges predictive and not predictive of gF. Edges categorized as predictive of gF are those included in predictive networks for every fold of the cross-validation; the distribution of predictive edges is shown at left. Tukey boxplots show median (red line), data between first and third quartile (edges of box), and suspected outliers (whiskers and red crosses; beyond 1.5 inter-quartile range [IQR]).

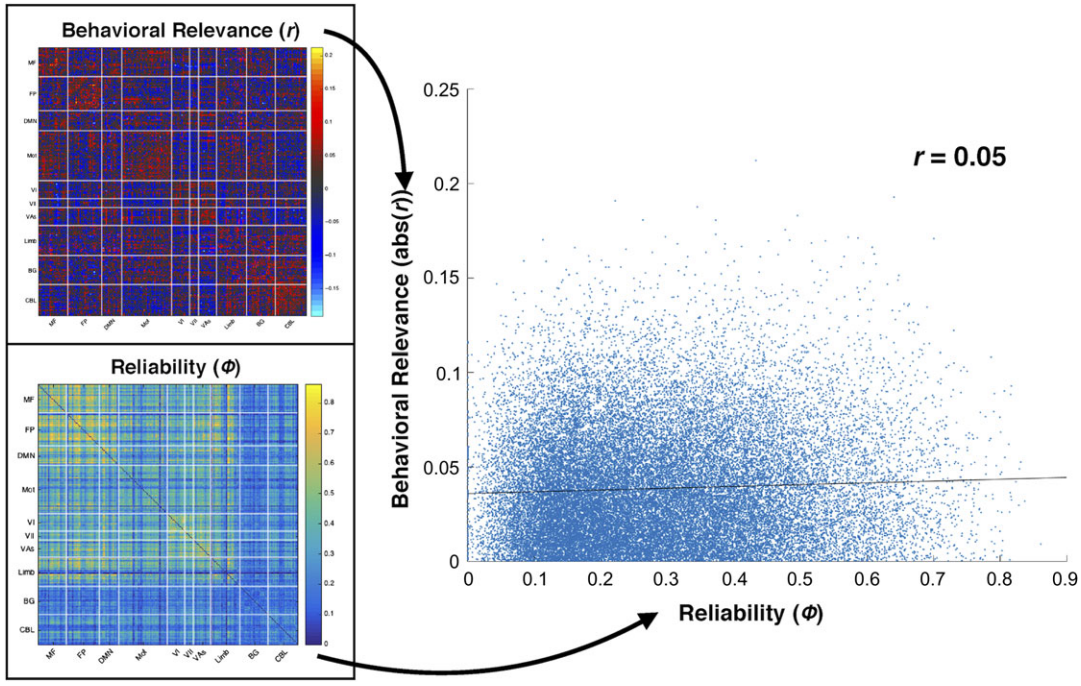

Figure 7. Edge-wise relationship between test–retest reliability and behavioral relevance. Behavioral relevance refers to the correlation between that edge's strength and fluid intelligence (gF) across all subjects (abs(*r*)). For the plot of the spatial distribution of behavioral relevance (top left), warmer colors are more positively correlated with behavior, and cooler colors are more negatively correlated with behavior. For the plot of test–retest reliability (bottom left), brighter colors are more reliable. For the scatterplot (right), each point represents a single edge. Here, behavioral relevance is the absolute value of behavioral relevance to facilitate finding effects related to magnitude of relevance.

suggesting that the most reliable edges are not necessarily the most informative edges, and vice versa.

Support for More Resting-State Functional Connectivity Data per Subject

Our results suggest that a relatively large amount of data per subject (>36 min) is needed to reach good test–retest reliability of functional connectivity using both univariate and multivariate test–retest reliability metrics. This concurs with prior evidence that 5–10 min of data may not result in reliable functional connectivity (Shou et al. 2013) and that reproducibility improves with more data (Anderson et al. 2011; Birn et al. 2013), even up to 90 min (Laumann et al. 2015)—but may seem in conflict with other recent studies suggesting that about 10 min of data is sufficient to generate a stable measurement of

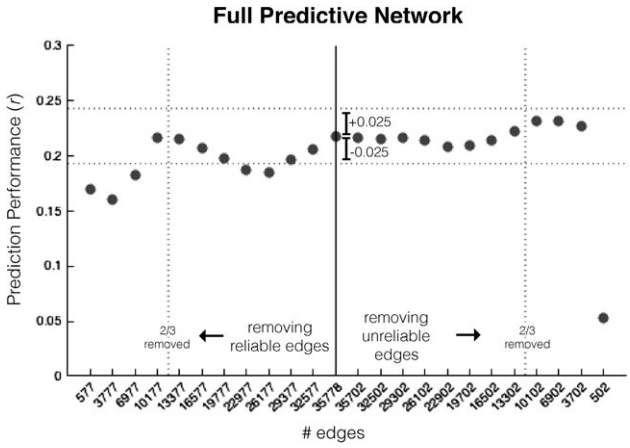

Figure 8. Influence of reliable and unreliable edges on prediction of fluid intelligence. An increasing number of unreliable edges are removed toward the right of the x-axis, and an increasing number of reliable edges are removed toward the left. Removal of unreliable edges starts with all edges showing $\Phi = 0$, then removing in intervals of 3200; removal of reliable edges also occurs in increments of 3200 (+1 for the first interval). The removal of 2/3 of all edges (23 852 edges removed) are marked with vertical lines; changes in performance greater than 0.025 from performance with all edges are marked with horizontal lines.

connectivity (Shehzad et al. 2009; Van Dijk et al. 2010; Birn et al. 2013; Choe et al. 2015; Tomasi et al. 2016). These contradictions arise primarily from differences in data summarization strategies and measures of similarity. For example, while Shehzad et al. (2009) report moderate to high test–retest reliability with less than 10 min of data, this result is specific to a subset of within-network data, and they too detail poor test–retest reliability over all edges (ICC < 0.4 from Table 1 in Shehzad et al. 2009). These results are consistent with the overall poor test–retest reliability we report in Figure 1 and the fair to good within-network test–retest reliability we report in Figure 4*a*. Crucially, several reports have determined the amount of needed connectivity data (ranging from 6 to 12 min) by showing limited gains in reproducibility with increasing the amount of data (Van Dijk et al. 2010; Birn et al. 2013; Tomasi et al. 2016). While a plateau in test–retest reliability serves as a practical measure for maximizing the tradeoff between test–retest reliability and the amount of data collected, it is not an endorsement of good test–retest reliability. The latter studies still show poor between-session test–retest reliability across all scan durations (ICC < 0.4 in Fig. 7 from Tomasi et al. 2016, and in Fig. 3*a* from Birn et al. 2013) and do not in themselves support high test–retest reliability of average whole-brain connectivity using approximately 10 min of data.

Within-Network Cortical Edges are Most Reliable; Subcortical Edges are Least Reliable

While subsets of nodes have been used to demonstrate the spatial distribution of test–retest reliability (Shehzad et al. 2009; Shah et al. 2016) or to summarize the average influence of scan duration on connectivity (Birn et al. 2013; Shah et al. 2016), to our knowledge, the spatial distribution of edge-wise test–retest reliability from a whole-brain atlas and its association with scan duration have not been previously shown.

Connectivity was most reliable within networks defined a priori, particularly the frontoparietal and default mode networks. The frontoparietal network has been shown to underlie individual differences (Finn et al. 2015; Gordon et al. 2016), the default mode network has served as the bread and butter of much functional connectivity research (Broyd et al. 2009; Whitfield-Gabrieli and Ford 2012; Raichle 2015), and altered connectivity in these networks has been linked to a variety of neuropsychiatric disorders (Uddin and Menon 2009; Öngür et al. 2010; Whitfield-Gabrieli and Ford 2012; Baker et al. 2014; Di Martino et al. 2014; Northoff 2014; Abbott et al. 2015; Alonso-Solís et al. 2015; Kaiser et al. 2015). Therefore, that connectivity within frontoparietal, default mode, and the other canonical networks was more reliable is encouraging, suggesting that these networks may be reliable targets in studies of normal cognition and disorders.

Still, a few caveats bear mentioning. Test–retest reliability results were variable at the level of individual edges, meaning that unreliable edges will be found within reliable networks and vice versa. Furthermore, anatomy may play a role in these findings, for example, frontoparietal anatomy in particular has been found to vary across subjects (Hill et al. 2010).

In contrast, connectivity between noncortical regions was the least reliable, as previously suggested (Shah et al. 2016). This is likely due in part to the smaller size of subcortical nodes, since node size was found to be correlated with test–retest reliability. However, subcortical location was found to independently contribute to test–retest reliability in addition to node size; this may be due to other factors such as unique activity, proximity to non-neuronal sources (e.g., susceptibility variations associated with breathing [Raj et al. 2001], cerebrospinal fluid), or lower central SNR with highly parallel array coils (Wiggins et al. 2009).

The Connectome as a Whole is More Reliable than Individual Edges

Our multivariate results imply that the connectivity matrix in its entirety holds more reliable information than simply the sum of its parts. Similarly, recent work has demonstrated the relative unreliability of edges compared with whole-brain connectivity fingerprinting (Pannunzi et al. 2017). We interpret these findings as highlighting the potential of multivariate methods—such as the CPM approach used here (Shen et al. 2017), or related approaches (Shirer et al. 2012; Smith et al. 2015; La Rosa et al. 2016)—when analyzing connectivity data, instead of univariate comparison of single edges. This is loosely analogous to the classic comparison between task-based activation and multivoxel pattern analysis (MVPA) where each variable contributes unequally and uniquely to the final measure (Norman et al. 2006).

Data can be Pooled Across Well-Harmonized Scanners and Sessions

We found no main effect of scanner or day when scanners are identically configured and harmonized, suggesting that it is acceptable to pool data across different scanning sessions and harmonized scanners. Pooling data across different scanning sessions may enable the acquisition of large amounts of data per subject in populations that cannot tolerate longer scanning sessions. For example, children and elderly subjects could be scanned in multiple shorter sessions, and then data can be pooled across sessions.

People are Highly Variable Across Runs and Sessions

In contrast, the non-negligible person-by-session interaction suggests that each subject varied substantially across sessions.

Other work has shown similar indications of greater between- than within-session variance, often in the form of smaller intersession compared with intrasession test–retest reliability (Shehzad et al. 2009; Birn et al. 2013; Pannunzi et al. 2017). Similarly, data from different tasks completed within the same session were found to be more similar than different sessions (O'Connor et al. 2017). The present work, coupled with the work of others, suggests a possibility that repeated measurements across sessions rather than runs may result in a more stable trait measurement of an individual. Although this high intersession variability may reflect meaningful individual variation, it may also increase the likelihood of finding an effect by chance, especially for studies with repeated measures (e.g., pre-/post-treatment studies). Additionally, the large residual (50–60%)—which has also been found in previous task-based and resting-state fMRI (Forsyth et al. 2014; Gee et al. 2015; Noble et al. 2016)—suggests the presence of substantial individual variability across all runs and sessions. This may be from either random or structured sources unaccounted for here. Parsing the contribution of real, brain-derived dynamics underlying variability at shorter timescales remains a challenge (Hutchison et al. 2013; Tagliazucchi and Laufs 2014; Hindriks et al. 2016; Laumann et al. 2016). It is possible that the mental states of individuals change from measurement to measurement—that, according to that ancient adage attributed to Heraclitus, "You can't stand in the same river twice."

The Most Reliable Data is not Necessarily the Most Useful Data

Potentially our most important result is that higher test–retest reliability is not meaningfully associated with higher data utility. We observed high discriminability in the absence of good test–retest reliability, suggesting that meaningful information unique to each individual can be captured by data with relatively low test–retest reliability. Similarly, we showed that an edge's test–retest reliability does not account for its usefulness in predicting behavior. Since the test–retest reliability of a measure establishes the upper limit of its predictive validity (Carmines and Zeller 1979), it is tempting to incorporate test–retest reliability information into analytical procedures to improve utility (Strother et al. 2004; Shou et al. 2014; Mueller et al. 2015; Shirer et al. 2015). However, this may not ultimately facilitate the development of predictive biomarkers of behavior. While it is possible that such procedures can support utility for specific behavioral traits or in under specific conditions, benefits must be determined on a case-by-case basis. Others have also shown a disconnect between the processing procedures that optimize for test–retest reliability versus those that optimize for prediction of behavioral traits (Strother et al. 2004; Shirer et al. 2015). Altogether, these results suggest that optimizing data acquisition and analysis only with respect to test–retest reliability, while ignoring other metrics closer to utility, may not provide the most meaningful results.

Limitations

The statistical models used here are designed to maximize generalizability to a healthy, post-adolescent population. Generalizability to different populations is less certain, as others have demonstrated (Somandepalli et al. 2015). Additionally, differences in acquisition and processing procedures also impact test–retest reliability (Zuo et al. 2013; Aurich et al. 2015; Shirer et al. 2015; Varikuti et al. 2017). For example, increasing the temporal resolution of the data has also been shown to increase test–retest reliability (Birn et al. 2013; Zuo and Xing 2014; Shah et al. 2016). Notably, global signal regression has heterogeneous effects on test–retest reliability of functional connectivity (Shirer et al. 2015; Varikuti et al. 2017), complicated by the high test–retest reliability of motion itself (Zuo and Xing 2014). This study is therefore most relevant to others using this common denoising technique (cf. Power et al. 2015 for a more complete discussion of motion denoising methods). Finally, test–retest reliability may be related in different ways to other models and measures of behavior. While these considerations are not expected to impact the generalizability of the central conclusions presented here—namely, that test–retest reliability improves with more data and can be distinct from the utility of the data—it is important to try to parse the effects of population, acquisition, processing, and analysis on characterizing test–retest reliability (cf. Poldrack et al. 2017).

Conclusion

In conclusion, this work helps to address some problems in characterizing the test–retest reliability of resting-state functional connectivity. Our results support the need for more data per subject to improve test–retest reliability and shows the spatial distribution of test–retest reliability of connectivity spanning the whole brain. Significantly, our results are among the first to highlight the increase in test–retest reliability when treating the connectivity matrix as a multivariate object and the dissociation between test–retest reliability and behavioral utility. As standards for the acquisition and analysis of functional connectivity data continue to be developed (Nichols et al. 2017), these results provide important considerations for establishing best practices in the field.

Supplementary Material

Supplementary data is available at *Cerebral Cortex* online.

Funding

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship Program under Grant Number DGE-1122492 (S.M.N.).

Notes

Data were provided in part by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University. We also thank Abigail Greene and Corey Horien for their comments during the preparation of this article. *Conflict of Interest*: None declared.

References

Abbott AE, Nair A, Keown CL, Datko M, Jahedi A, Fishman I, Müller R-A. 2015. Patterns of atypical functional connectivity and behavioral links in autism differ between default, salience, and executive networks. Cereb Cortex. 26(10): 4034–4045.

Alonso-Solís A, Vives-Gilabert Y, Grasa E, Portella MJ, Rabella M, Sauras RB, Roldán A, Núñez-Marín F, Gómez-Ansón B, Pérez V. 2015. Resting-state functional connectivity alterations in the default network of schizophrenia patients with persistent auditory verbal hallucinations. Schizophrenia Res. 161:261–268.

Anderson JS, Ferguson MA, Lopez-Larson M, Yurgelun-Todd D. 2011. Reproducibility of single-subject functional connectivity measurements. AJNR Am J Neuroradiol. 32:548–555.

Aurich NK, Alves Filho JO, Marques da Silva AM, Franco AR. 2015. Evaluating the reliability of different preprocessing steps to estimate graph theoretical measures in resting state fMRI data. Front Neurosci. 9:48.

Baker JT, Holmes AJ, Masters GA, Yeo BT, Krienen F, Buckner RL, Öngür D. 2014. Disruption of cortical association networks in schizophrenia and psychotic bipolar disorder. JAMA Psychiatry. 71:109–118.

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. Nature. 533:452–454.

Begley CG, Ellis LM. 2012. Drug development: raise standards for preclinical cancer research. Nature. 483:531–533.

Birn RM, Molloy EK, Patriat R, Parker T, Meier TB, Kirk GR, Nair VA, Meyerand ME, Prabhakaran V. 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. Neuroimage. 83:550–558.

Broyd SJ, Demanuele C, Debener S, Helps SK, James CJ, Sonuga-Barke EJ. 2009. Default-mode brain dysfunction in mental disorders: a systematic review. Neurosci Biobehavioral Rev. 33:279–296.

Carmines EG, Zeller RA. 1979. Reliability and validity assessment. Sage publications. p. 17.

Choe AS, Jones CK, Joel SE, Muschelli J, Belegu V, Caffo BS, Lindquist MA, van Zijl PC, Pekar JJ. 2015. Reproducibility and temporal structure in weekly resting-state fMRI over a period of 3.5 years. PloS One. 10:e0140134.

Cicchetti DV, Sparrow SA. 1981. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. Am J Mental Def. 86(2): 127–137.

Cronbach LJ. 1988. Five perspectives on validity argument. In: Wainer H, Braun H, editors. Test validity. Hillsdale, NJ: Lawrence Erlbaum. p. 3–17.

Di Martino A, Fair DA, Kelly C, Satterthwaite TD, Castellanos FX, Thomason ME, Craddock RC, Luna B, Leventhal BL, Zuo X-N. 2014. Unraveling the miswired connectome: a developmental perspective. Neuron. 83:1335–1353.

Finn ES, Shen X, Holahan JM, Scheinost D, Lacadie C, Papademetris X, Shaywitz SE, Shaywitz BA, Constable RT. 2014. Disruption of functional networks in dyslexia: a whole-brain, data-driven analysis of connectivity. Biol Psychiatry. 76:397–404.

Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT. 2015. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. Nat Neurosci. 18:1664–1671.

Forsyth JK, McEwen SC, Gee DG, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA, Olvet DM. 2014. Reliability of functional magnetic resonance imaging activation during working memory in a multi-site study: analysis from the North American Prodrome Longitudinal Study. Neuroimage. 97:41–52.

Friedman L, Glover GH, Krenz D, Magnotta V, BIRN F. 2006. Reducing inter-scanner variability of activation in a multicenter fMRI study: role of smoothness equalization. Neuroimage. 32:1656–1668.

Gee DG, McEwen SC, Forsyth JK, Haut KM, Bearden CE, Addington J, Goodyear B, Cadenhead KS, Mirzakhanian H, Cornblatt BA. 2015. Reliability of an fMRI paradigm for emotional processing in a multisite longitudinal study. Hum Brain Map. 36(7):2558–2579.

Glasser MF, Sotiropoulos SN, Wilson JA, Coalson TS, Fischl B, Andersson JL, Xu J, Jbabdi S, Webster M, Polimeni JR. 2013. The minimal preprocessing pipelines for the Human Connectome Project. Neuroimage. 80:105–124.

Gordon EM, Laumann TO, Adeyemo B, Gilmore AW, Nelson SM, Dosenbach NU, Petersen SE. 2016. Individual-specific features of brain systems identified with resting state functional correlations. NeuroImage. 146:918–939.

Hayes AF. 2013. Introduction to mediation, moderation, and conditional process analysis: a regression-based approach. New York, NY: Guilford Press.

Hill J, Dierker D, Neil J, Inder T, Knutsen A, Harwell J, Coalson T, Van Essen D. 2010. A surface-based analysis of hemispheric asymmetries and folding of cerebral cortex in term-born human infants. J Neurosci. 30:2268–2276.

Hindriks R, Adhikari M, Murayama Y, Ganzetti M, Mantini D, Logothetis N, Deco G. 2016. Can sliding-window correlations reveal dynamic functional connectivity in resting-state fMRI? Neuroimage. 127:242–256.

Hutchison RM, Womelsdorf T, Allen EA, Bandettini PA, Calhoun VD, Corbetta M, Della Penna S, Duyn JH, Glover GH, Gonzalez-Castillo J, et al. 2013. Dynamic functional connectivity: promise, issues, and interpretations. Neuroimage. 80: 360–378.

Joshi A, Scheinost D, Okuda H, Belhachemi D, Murphy I, Staib LH, Papademetris X. 2011. Unified framework for development, deployment and robust testing of neuroimaging algorithms. Neuroinformatics. 9:69–84.

Kaiser RH, Andrews-Hanna JR, Wager TD, Pizzagalli DA. 2015. Large-scale network dysfunction in major depressive disorder: a meta-analysis of resting-state functional connectivity. JAMA Psychiatry. 72:603–611.

La Rosa PS, Brooks TL, Deych E, Shands B, Prior F, Larson-Prior LJ, Shannon WD. 2016. Gibbs distribution for statistical analysis of graphical data with a sample application to fcMRI brain images. Stat Med. 35:566–580.

Laumann TO, Gordon EM, Adeyemo B, Snyder AZ, Joo SJ, Chen M-Y, Gilmore AW, McDermott KB, Nelson SM, Dosenbach NU. 2015. Functional system and areal organization of a highly sampled individual human brain. Neuron. 87(3):657–670.

Laumann TO, Snyder AZ, Mitra A, Gordon EM, Gratton C, Adeyemo B, Gilmore AW, Nelson SM, Berg JJ, Greene DJ. 2016. On the stability of bold fMRI correlations. Cereb Cortex.

Mueller S, Wang D, Fox MD, Pan R, Lu J, Li K, Sun W, Buckner RL, Liu H. 2015. Reliability correction for functional connectivity: theory and implementation. Hum Brain Map. 36: 4664–4680.

Müller R, Büttner P. 1994. A critical discussion of intraclass correlation coefficients. Stat Med. 13:2465–2476.

Nichols TE, Das S, Eickhoff SB, Evans AC, Glatard T, Hanke M, Kriegeskorte N, Milham MP, Poldrack RA, Poline JB, et al. 2017. Best practices in data analysis and sharing in neuroimaging using MRI. Nat Neurosci. 20:299–303.

Noble S, Scheinost D, Finn ES, Shen X, Papademetris X, McEwen SC, Bearden CE, Addington J, Goodyear B, Cadenhead KS. 2016. Multisite reliability of MR-based functional connectivity. NeuroImage. 146:959–970.

Norman KA, Polyn SM, Detre GJ, Haxby JV. 2006. Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn Sci. 10:424–430.

Northoff G. 2014. Are auditory hallucinations related to the brain's resting state activity? A 'neurophenomenal

resting state hypothesis'. Clin Psychopharmacol Neurosci. 12:189–195.

O'Connor D, Potler NV, Kovacs M, Xu T, Ai L, Pellman J, Vanderwal T, Parra LC, Cohen S, Ghosh S, et al. 2017. The Healthy Brain Network Serial Scanning Initiative: a resource for evaluating inter-individual differences and their reliabilities across scan conditions and sessions. GigaScience. 6:1–4.

Öngür D, Lundy M, Greenhouse I, Shinn AK, Menon V, Cohen BM, Renshaw PF. 2010. Default mode network abnormalities in bipolar disorder and schizophrenia. Psychiatry Res. 183:59–68.

Open Science Collaboration. 2015. Estimating the reproducibility of psychological science. Science. 349:aac4716.

Pannunzi M, Hindriks R, Bettinardi RG, Wenger E, Lisofsky N, Martensson J, Butler O, Filevich E, Becker M, Lochstet M, et al. 2017. Resting-state fMRI correlations: from link-wise unreliability to whole brain stability. Neuroimage. 157:250–262.

PashlerHWagenmakersEJ2012Editors' introduction to the special section on replicability in psychological science: a crisis of confidence?Perspect Psychol Sci7528530

Poldrack RA, Baker CI, Durnez J, Gorgolewski KJ, Matthews PM, Munafò MR, Nichols TE, Poline J-B, Vul E, Yarkoni T. 2017. Scanning the horizon: towards transparent and reproducible neuroimaging research. Nat Rev Neurosci. 18:115–126.

Power JD, Mitra A, Laumann TO, Snyder AZ, Schlaggar BL, Petersen SE. 2014. Methods to detect, characterize, and remove motion artifact in resting state fMRI. Neuroimage. 84:320–341.

Power JD, Schlaggar BL, Petersen SE. 2015. Recent progress and outstanding issues in motion correction in resting state fMRI. Neuroimage. 105:536–551.

Raichle ME. 2015. The brain's default mode network. Ann Rev Neurosci. 38:433–447.

Raj D, Anderson AW, Gore JC. 2001. Respiratory effects in human functional magnetic resonance imaging due to bulk susceptibility changes. Phys Med Biol. 46:3331.

Rosenberg MD, Finn ES, Scheinost D, Papademetris X, Shen X, Constable RT, Chun MM. 2015. A neuromarker of sustained attention from whole-brain functional connectivity. Nat Neurosci. 19:165–171.

Satterthwaite TD, Elliott MA, Gerraty RT, Ruparel K, Loughead J, Calkins ME, Eickhoff SB, Hakonarson H, Gur RC, Gur RE. 2013. An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. Neuroimage. 64:240–256.

Scheinost D, Kwon SH, Lacadie C, Vohr BR, Schneider KC, Papademetris X, Constable RT, Ment LR. 2017. Alterations in anatomical covariance in the prematurely born. Cereb Cortex. 27:534–543.

Scheinost D, Papademetris X, Constable RT. 2014. The impact of image smoothness on intrinsic functional connectivity and head motion confounds. Neuroimage. 95:13–21.

Shah LM, Cramer JA, Ferguson MA, Birn RM, Anderson JS. 2016. Reliability and reproducibility of individual differences in functional connectivity acquired during task and resting state. Brain Behav. 6(5):e00456.

Shavelson RJ, Baxter GP, Gao X. 1993. Sampling variability of performance assessments. J Educ Measure. 30:215–232.

Shehzad Z, Kelly AM, Reiss PT, Gee DG, Gotimer K, Uddin LQ, Lee SH, Margulies DS, Roy AK, Biswal BB, et al. 2009. The resting brain: unconstrained yet reliable. Cereb Cortex. 19: 2209–2229.

Shen X, Finn ES, Scheinost D, Rosenberg MD, Chun MM, Papademetris X, Constable RT. 2017. Using connectome-based predictive modeling to predict individual behavior from brain connectivity. Nat Protoc. 12:506–518.

Shen X, Tokoglu F, Papademetris X, Constable RT. 2013. Groupwise whole-brain parcellation from resting-state fMRI data for network node identification. Neuroimage. 82:403–415.

Shirer WR, Jiang H, Price CM, Ng B, Greicius MD. 2015. Optimization of rs-fMRI pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. Neuroimage. 117:67–79.

Shirer WR, Ryali S, Rykhlevskaia E, Menon V, Greicius MD. 2012. Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb Cortex. 22:158–165.

Shou H, Eloyan A, Lee S, Zipunnikov V, Crainiceanu A, Nebel M, Caffo B, Lindquist M, Crainiceanu C. 2013. Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). Cogn Affect Behav Neurosci. 13:714–724.

Shou H, Eloyan A, Nebel MB, Mejia A, Pekar JJ, Mostofsky S, Caffo B, Lindquist MA, Crainiceanu CM. 2014. Shrinkage prediction of seed-voxel brain connectivity using resting state fMRI. Neuroimage. 102:938–944.

Shrout PE, Fleiss JL. 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 86:420.

Smith SM. 2002. Fast robust automated brain extraction. Hum Brain Map. 17:143–155.

Smith SM. 2012. The future of FMRI connectivity. Neuroimage. 62:1257–1266.

Smith SM, Nichols TE, Vidaurre D, Winkler AM, Behrens TE, Glasser MF, Ugurbil K, Barch DM, Van Essen DC, Miller KL. 2015. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. Nat Neurosci. 18:1565–1567.

Somandepalli K, Kelly C, Reiss PT, Zuo X-N, Craddock RC, Yan C-G, Petkova E, Castellanos FX, Milham MP, Di Martino A. 2015. Short-term test–retest reliability of resting state fMRI metrics in children with and without attention-deficit/hyperactivity disorder. Dev Cogn Neurosci. 15: 83–93.

Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc. 64:479–498.

Strother S, La Conte S, Hansen LK, Anderson J, Zhang J, Pulapura S, Rottenberg D. 2004. Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. Neuroimage. 23:S196–S207.

Tagliazucchi E, Laufs H. 2014. Decoding wakefulness levels from typical fMRI resting-state data reveals reliable drifts between wakefulness and sleep. Neuron. 82:695–708.

Tomasi DG, Shokri-Kojori E, Volkow ND. 2016. Temporal evolution of brain functional connectivity metrics: could 7 min of rest be enough? Cereb Cortex. 27(8):4153–4165.

Uddin LQ, Menon V. 2009. The anterior insula in autism: under-connected and under-examined. Neurosci Biobehav Rev. 33: 1198–1203.

Van Dijk KR, Hedden T, Venkataraman A, Evans KC, Lazar SW, Buckner RL. 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. J Neurophysiol. 103:297–321.

Van Essen DC, Smith SM, Barch DM, Behrens TE, Yacoub E, Ugurbil K, Consortium W-MH. 2013. The WU-Minn human connectome project: an overview. Neuroimage. 80:62–79.

Varikuti DP, Hoffstaedter F, Genon S, Schwender H, Reid AT, Eickhoff SB. 2017. Resting-state test–retest reliability of a priori defined canonical networks over different preprocessing steps. Brain Struct Funct. 222:1447–1468.

Webb NM, Shavelson RJ 2005. Generalizability theory: overview. Wiley StatsRef: Statistics Reference Online.

Webb NM, Shavelson RJ, Haertel EH. 2006. Reliability coefficients and generalizability theory. Handbook Stat. 26:81–124.

Whitfield-Gabrieli S, Ford JM. 2012. Default mode network activity and connectivity in psychopathology. Ann Rev Clin Psychol. 8:49–76.

Wiggins GC, Polimeni JR, Potthast A, Schmitt M, Alagappan V, Wald LL. 2009. 96-Channel receive-only head coil for 3 Tesla: design optimization and evaluation. Magn Reson Med. 62: 754–762.

Zuo X-N, Xu T, Jiang L, Yang Z, Cao X-Y, He Y, Zang Y-F, Castellanos FX, Milham MP. 2013. Toward reliable characterization of functional homogeneity in the human brain: preprocessing, scan duration, imaging resolution and computational space. Neuroimage. 65:374–386.

Zuo XN, Xing XX. 2014. Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. Neurosci Biobehav Rev. 45:100–118.