Reliability of graph analysis of resting state fMRI using test-retest dataset from the Human Connectome Project

M. Termenon[a,b,1,*], A. Jaillard[c,d,e], C. Delon-Martin[a,b], S. Achard[f,g]

[a] *Univ. Grenoble Alpes, Grenoble Institute des Neurosciences, GIN, F-38000 Grenoble, France*
[b] *INSERM, U1216, F-38000 Grenoble, France*
[c] *Pole Recherche, CHU Grenoble, F-38000 Grenoble, France*
[d] *IRMaGe, Inserm US17 CNRS UMS 3552, F-38000 Grenoble, France*
[e] *AGEIS EA7407, Univ. Grenoble Alpes, F-38000 Grenoble, France*
[f] *Univ. Grenoble Alpes, GIPSA-lab, F-38000, Grenoble, France*
[g] *CNRS, GIPSA-lab, F-38000, Grenoble, France*

Abstract

The exploration of brain networks with resting-state fMRI (rs-fMRI) combined with graph theoretical approaches has become popular, with the perspective of finding network graph metrics as biomarkers in the context of clinical studies. A preliminary requirement for such findings is to assess the reliability of the graph based connectivity metrics. In previous test-retest (TRT) studies, this reliability has been explored using intraclass correlation coefficient (ICC) with heterogeneous results. But the issue of sample size has not been addressed. Using the large TRT rs-fMRI dataset from the Human Connectome Project (HCP), we computed ICCs and their corresponding p-values (applying permutation and bootstrap techniques) and varied the number of subjects (from 20 to 100), the scan duration (from 400 to 1200 time points), the cost and the graph metrics, using the Anatomic-Automatic Labelling (AAL) parcellation scheme. We quantified the reliability of the graph metrics computed both at global and regional level depending, at optimal cost, on two key parameters, the sample size and the number of time points or scan duration. In the cost range between 20% to 35%, most of the global graph metrics are reliable with 40 subjects or more with long scan duration (14 min 24 s). In large samples (for instance, 100 subjects) most global and regional graph metrics are reliable for a minimum scan duration of 7 min 14s. Finally, for 40 subjects and long scan duration (14 min 24 s), the reliable regions are located in the main areas of the default mode network (DMN), the motor and the visual networks.

*Corresponding author
[1] Author funded by Region Rhone-Alpes.

Introduction

Graph theoretical approaches provide a powerful way to analyze complex networks and quantify brain functional systems using resting state functional MRI (rs-fMRI); see a review and references therein (De Vico Fallani et al., 2014). In the last few decades, graph theory has led to understand how different complex systems can share the same key representational systems that can be characterized by different network properties. Such network properties include, among others, global and local efficiency, betweenness centrality, clustering coefficient and small word topology (Achard et al., 2006). One application of graph theory is to use graph metrics to quantify differences between patients and controls or between groups over time, and more specifically to use network properties as diagnostic and recovery biomarkers in the context of clinical trials and longitudinal studies (Bullmore and Sporns, 2009).

However, the validation of sensitive longitudinal imaging biomarkers relying on graphs requires rigorous evaluation of the test-retest (TRT) reliability of graph metric measures (Nakagawa and Schielzeth, 2010). TRT reliability is typically evaluated by acquiring at least two scanning sessions of the same subject at different times. The second session is performed after a time interval varying from a few minutes, when the two acquisitions are performed during the same session (intra-session reliability), to several hours, days or months, for the assessment of intersession reliability. The analysis of TRT data on brain connectivity is necessary to identify network features that are intrinsic to the functioning of brain (called biomarkers in this paper) and not biased by subject variablity or artefacts from acquisition. To our knowledge, graph metrics TRT reliability of the whole brain using a parcellation scheme has been assessed in 6 studies (Braun et al., 2012; Cao et al., 2014; Liang et al., 2012; Wang et al., 2011; Schwarz and McGonigle, 2011; Guo et al., 2012), or at the voxel level (Liao et al., 2013; Du et al., 2015), and in a recent meta-analysis (Andellini et al., 2015).

Intraclass correlation coefficient (ICC) (Shrout and Fleiss, 1979) has been used as a measure of reliability, resulting in a large range of ICC values, presumably due to considerable heterogeneity in the methodological approaches. Despite this heterogeneity, several factors have been shown to influence graph reliability, such as the preprocessing steps (smoothing, global signal regression, movement regression) (Braun et al., 2012; Shirer et al., 2015), the frequency range (Liang et al., 2012; Liao et al., 2013; Shirer et al., 2015), the computation of the edges of the graph (Liang et al., 2012; Fiecas et al., 2013), the type of graph metrics (Wang et al., 2011; Braun et al., 2012; Cao et al., 2014), the cost/sparsity (Braun et al., 2012), the type of network (binary or weighted (Braun et al., 2012; Guo et al., 2012; Liang et al., 2012; Liao et al., 2013; Schwarz and McGonigle, 2011), the brain parcellation scheme (Wang et al., 2011; Cao et al., 2014), the use of voxel-wise metrics (Zuo and Xing, 2014), and most importantly, the scan duration (Braun et al., 2012; Liao et al., 2013; Wang et al., 2011). In a seed-based approach, Birn et al. (2013), exploring TRT reliability of rs-fMRI connectivity for scan duration ranging from 3 to 27 min, found improvement in intersession reliability by increasing scan duration

up to 9 min, suggesting that functional connectivity computed from a 10 min acquisition duration averages slow changes and provides a more stable estimate of the connectivity strength. In a recent meta-analysis based on graph theory TRT studies, the same trend was observed with increased reliability for longer acquisition duration (Andellini et al., 2015).

Another aspect accounting for the large range of ICC and compromising the reliability of primary studies is low statistical power, even when all other factors are ideal. Low power, resulting in low sensitivity, low positive predictive value, and effect inflation, is mainly related to low sample size (Button et al., 2013). The sample size defines the number of degrees of freedom, which is a key element in statistical analyses, so that the ICC at group level is expected to depend on the sample size included in the analysis. Indeed, many TRT studies findings were obtained from small sample sizes, ranging from 11 to 33 subjects. In the literature on fMRI activation, 20 subjects has been found to be the minimum number that permits reliable fMRI results in appropriate acquisition conditions (Thirion et al., 2007). However, as far as we know, the issue of the minimum sample size required to get reliable metrics remains to be addressed in the rs-fMRI literature. Accordingly, we aimed to examine the parameters influencing graph metrics' TRT reliability in a larger sample size. Recent methodological advances and the increasing availability of large datasets gave us the opportunity to analyze a TRT rs-fMRI dataset from the Human Connectome Project (HCP) [2], on a large sample size ($n = 100$ subjects), collected over a long duration (14 min 24 s duration).

Here, we tested for the first time the combined effect of the sample size and duration on TRT reliability. The rs-fMRI dataset recently released by the HCP was acquired using multiband, allowing the combination of large number of volumes (1200) and high spatiotemporal resolution (2 mm isotropic and 720 ms TR or 1.39 Hz sampling frequency). In the literature, ICC is used as a measure of intersession reliability to examine the effect of the sample size, of the scan duration/time points, of the cost for creating the graph and the choice of graph metrics on the intersession reliability. We computed ICCs in sub-datasets corresponding to five subgroups of 20 to 100 subjects, for 400 to 1200 time points corresponding to scan duration from 4min 48s to 14min 24s, and costs ranging from 2.5% to 75%. In line with previous studies, the classical Anatomic-Automatic Labeling (AAL) was used as parcellation scheme (Tzourio-Mazoyer et al., 2002). To further assess the influence of the parcellation scheme, the Harvard-Oxford structural parcellation (Diedrichsen et al., 2009), an AAL based finner parcellation scheme composed of 459 regions (Alexander-Bloch et al., 2012), the Craddock functional parcellation (Craddock et al., 2012) and ICA-based parcellation[3]. Thanks to the large amount of data, we performed an extensive bootstrap study in order to be able to evaluate the standard deviation and p-values of ICC directly from the data.

[2] `http://www.humanconnectome.org/`

[3] Node timeseries available in the HCP website. were also explored

In parallel to global measures of network topology, graph metrics can also be estimated from individual nodes at the regional level (Achard et al., 2006, 2012). In neurological and psychiatric diseases, regional graph metrics allow the quantification of differences between groups of patients and controls (Bullmore and Sporns, 2009). For example, changes in regional betweenness centrality (Wang et al., 2011) and local and global efficiency (Yin et al., 2014) have been reported in subcortical stroke. In addition, these regional metrics were able to discriminate patients from age-matched control groups, and changes in the regional motor network topology correlated with motor outcome (Wang et al., 2011; Yin et al., 2014), suggesting the potential of key regions in the brain for translational research as biomarkers.

However, the TRT reliability of graph metrics needs to be assessed prior to using them in clinical trials and longitudinal studies. Here, we tested the reliability of both the global and the regional values of graph metrics with their respective p-values, at increasing costs. We also explored whether graph metrics would be a relevant approach to classify key regions in networks. Finally, on the basis of this analysis, we aimed to make recommendations to obtain reliable graph metrics with respect to the scan duration, sample size and the cost of the graph.

Methods

Subjects and data acquisition

The dataset used for this experiment was selected from a large sample of rs-fMRI dataset publicly released as part of the Human Connectome Project (HCP), WU-Minn Consortium. Our sample includes 100 subjects: 99 young healthy adults from 20 to 35 years old (54 females) and 1 healthy adult older than 35. Each subject underwent two rs-fMRI acquisitions on different days. Subjects were instructed to keep their eyes open and to let their mind wander while fixating a cross-hair projected on a dark background (Smith et al., 2013).

Data were collected on the 3T Siemens Connectome Skyra MRI scanner with a 32-channel head coil. All functional images were acquired using a multiband gradient-echo EPI imaging sequence with the following parameters: 2 mm isotropic voxels, 72 axial slices, TR = 720 ms, TE = 33.1 ms, flip angle = 52°, field of view = 208x180 mm^2, matrix size = 104x90 and a multiband factor of 8. A total of 1200 images was acquired for a scan duration of 14 min and 24 s. For more detailed parameters, see (Smith et al., 2013). Two high resolution structural images T1-weighted (T1w) and T2-weighted (T2w) were further collected. They were acquired with a 3D MPRAGE sequence and a 3D T2-SPACE sequence, respectively. The main MR parameters for the T1w image were: TR = 2.4 s, TE = 2.14 ms, TI = 1000 ms, flip angle = 8°, field of view = 224 x224 mm^2 and 0.7 mm isotropic voxels and for the T2w: TR = 3.2 s, TE = 565 ms, flip angle = variable, field of view = 224 x224 mm^2 and 0.7 mm isotropic voxels.

Data preprocessing

Structural data were preprocessed according to the pipeline described by Glasser et al. (2013). In brief, it corrects T1w and T2w for bias field and distortions, coregisters them together and registers them to the MNI152 atlas using linear and nonlinear registrations, using FSL's FLIRT and FNIRT functions. After registration to the atlas image, we segmented the individual T1w in six different brain tissues to obtain a grey matter (GM) probability maps that will be later used to extract the time series to compute the graphs.

Functional data were corrected for distortions and subject motion. They were registered to the individual structural image and further to the MNI152 atlas space using the transforms applied to the structural image. All of these preceding transforms were concatenated, together with the structural-to-MNI nonlinear warp field, so that a single resulting warp (per time point) was applied to the original time series to achieve a single resampling into MNI space with a final isotropic voxel size of 2 mm. Finally, the 4D image was normalized to a global mean and the bias field was removed, and non-brain voxels were masked out. No spatial smoothing was applied. For more details of the spatial preprocessing pipeline, see Glasser et al. (2013).

Parcellation scheme

Among the neuroimaging community, there is no consensus about the best parcellation for the investigation of the test-retest reliabilityof brain networks. Mainly, two types of templates exist: those based on anatomical features (either structural T1 or diffusion based) and those based on functional features. Among the structural based templates, the AAL has attracted lots of interest since it is a precisely defined template based on a single subject that includes parcellation of the cerebellum (Tzourio-Mazoyer et al., 2002) and additionally it was mainly used in the previous test-retest studies. However, it may not be representative of the brain populations and thus another atlas based on the structural images of 37 healthy adult subjects was developed, currently known as the Harvard-Oxford atlas (Desikan et al., 2006). More recently, functional connectivity based atlases have been proposed by for example aggregating regions based on their functional similarity using different algorithms such as spatially-constrained spectral clustering algorithm (Craddock et al., 2012) or by using independent component analysis (Filippini et al., 2009). In the present study, we first used a modified version of the classical Anatomic-Automatic Labeling (AAL) (Tzourio-Mazoyer et al., 2002) composed of 89 regions (see Supplementary Material for more information) and a finer one derived from the same parcellation but subdivided into 459 regions (Alexander-Bloch et al., 2012) [4], denoted AAL89 and AAL459 respectively. In order to evaluate the influence of the parcellation scheme on the TRT reliability, we used in a second step additional templates. The structural

[4]Both templates are available on request.

Harvard-Oxford template[5] (HO117) was used together with the cerebellar atlas (Diedrichsen et al., 2009) (note that we merged some parts of the cerebellum to have the same parcellation as in AAL89, see Supplementary Material). As a functional alternative, we used a parcellation with 100 regions provided by Craddock[6] (Crad100), using temporal correlation between voxel-time courses as similarity metrics and a group level clustering based in a two-level scheme in which the data of each participant are clustered separately. Finally, we used the ICA maps available from "node timeseries" in the HCP website[7] for 50, 100 and 200 independent spatial maps (named ICA50, ICA100 and ICA200, respectively), in which the full set of ICA maps was used as spatial regressors against the full data, estimating one time series for each ICA map.

Time series extraction and analysis using wavelets

In each parcel, regional mean time series were estimated by averaging, at each time point, the fMRI voxel values weighted by the GM probability of these voxels. This weighting limits the contamination of the time-series by white matter signals and cerebrospinal fluids. We reduce the influence of the partial volume effect related to voxels that contains both GM and WM or GM and CSF. The problem of regressing out WM and CSF in the functional data is that it may remove also some GM signal. The mean white matter and cerebrospinal fluid signals were thus not regressed. Residual head motion were eventually removed by regressing out motion parameters and their first derivative's time series. Global signal regression was not applied, since it was shown to introduce severe artifacts (Murphy et al., 2009), resulting in correlation pattern distortions (Saad et al., 2012).

The resulting time series were decomposed in 5 scales using discrete dyadic wavelet transformation. Wavelet transforms perform a time-scale decomposition that partitions the total energy of a signal over a set of compactly supported basis functions, or little waves, each of which is uniquely scaled in frequency and located in time (Achard et al., 2006). We applied the maximal overlap discrete wavelet transform (MODWT) to each regional mean time series and estimated the pairwise inter-regional correlations at each of the five wavelet scales. The comparison between wavelets and band pass filtering was already tested in Guo et al. (2012). They found that "ROI matrix reliability improved substantially when ROI time series correlations were computed after wavelet transformation". We performed our analysis at wavelet scale 4. Indeed, resting state signal is currently analyzed in frequencies below 0.1 Hz (Biswal et al., 1995; Fox and Raichle, 2007), thus the relevant information for rs-fMRI data is mainly contained within the scale 4 that represents the frequency interval 0.043–0.087 Hz. Scale 3 is omitted because it belongs to the frequency range

[5] Available in http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/Atlases

[6] http://ccraddock.github.io/cluster_roi/atlases.html

[7] http://www.humanconnectome.org/documentation/S500/HCP500_GroupICA+NodeTS+Netmats_Summary_28aug2014.pdf

between $[0.087 - 0.17]$, thus it contains signal from frequencies higher than 0.1 Hz. The frequency bands extracted using wavelets are reported in the table 3. For a comparison with classical acquisitions using a higher TR, the table reports also the wavelet frequency bands obtained with a TR of 2 s. As the interest in resting-state fMRI study is on low frequencies, the most important parameter is the time duration of the acquisition. The table 3 provides details to link the duration of the scan to the number of points and the corresponding frequency bands of interest.

Graph computation

All pairs of scale 4-specific wavelet correlations between regions are further pooled into a correlation matrix for each of the subjects. To compute the graph, we first extracted the minimum spanning tree based on the absolute correlation matrix (Alexander-Bloch et al., 2012) to keep the graph fully connected, and the remaining absolute values of correlation matrices were thresholded to create an adjacency matrix that defines an unweighted graph for each subject. A threshold R was calculated in order to produce a fixed number of edges M to be able to compare the extracted graphs. As a consequence, the threshold value is subject dependent. The ratio between the number of selected edges and all possible edges is termed "cost", implying that the higher the cost the larger number of edges is considered in the computation of the graph. For example, with a parcellation of 89 regions, the number of edges are 391 at 10% cost and 1564 at 40% cost. Each of these extracted graphs comprised N=89 nodes corresponding to the anatomical regions, and M undirected edges corresponding to the significant correlation values above the threshold R (Achard et al., 2012). There exists no straightforward way to select the appropriate cost (De Vico Fallani et al., 2014). Achard and Bullmore (2007) introduced the small-world regime which defines a range of cost that is a vector of values of cost. The low limit of the range is defined by a sufficiently large number of edges so that the graph is different from regular or random graphs. The upper limit is reached when the graph has too many edges and cannot be differentiated from random or regular graphs.

Computation of Graph metrics

It has been shown that graph metrics have different properties and highlight different topological characteristics of the graphs, see Boccaletti et al. (2006) for a review. Global efficiency, minimum path length or betweenness centrality are interpreted as measures to facilitate functional integration (Rubinov and Sporns, 2010), quantifying how information is propagating in the whole network. Moreover, local efficiency or clustering coefficient are measures associated to segregation functions (Rubinov and Sporns, 2010) and can be regarded as measures of information transfer in the immediate neighborhood of each node. All these measures were used to quantify the graph metrics at the global level with the extraction of one quantity for each graph, subject and session. However, these metrics can also be evaluated at the nodal or regional level, i.e. one value is

Parameter	Character	Description
Regional networks parameters		
Degree	k_i	Number of edges connected to a node i.
Global efficiency	e_{g_i}	Communication efficiency of a node i with all other nodes.
Local efficiency	e_{l_i}	Communication efficiency of a node i with its neighbors.
Betweenness centrality	b_i	Influence of a node i over the information flow between all other nodes.
Clustering coefficient	c_i	Fraction of the neighbors of a node i that are also neighbors of each other.
Global network parameters		
Global efficiency	E_g	Mean of e_{g_i}.
Local efficiency	E_l	Mean of e_{l_i}.
Modularity	Q	Strength of division of a network into clusters of nodes highly connected between them.
Betweenness centrality	B	Mean of b_i.
Small world topology	σ	Extent of a network between randomness and order.
Clustering coefficient	C	Mean of c_i.

Table 1: Description of the network metrics. Detailed information and metrics computation can be found in (Rubinov and Sporns, 2010). We explored both regional metrics computed at the level of the nodes of the graphs and global metrics that correspond to the average of the regional metrics other the whole graph.

computed for each node of the graph or region of the brain. For each subject, session and graph, we computed a vector of parameters quantifying the same characteristics but at the regional level. Table 1 presents a summary of each metric used in the paper. The detailed formulas can be found in (Rubinov and Sporns, 2010). Network parameters computation was preformed in R using *brainwaver* and *igraph* libraries, tools that are freely available on CRAN[8,9].

Test-retest reliability

The assessment of reliability using proper statistical methods needs caution in terms of interpretation. The first studies date back to the last century and the work of Fisher (Fisher, 1925), who proposed to use an ANOVA with a separation of within-subject and between-subject variability. In this study, the adopted statistical model for the observations Y_{ij} for the jth session of the ith subject, is defined as

$$Y_{ij} = \mu + S_i + e_{ij},$$

where μ is the mean of all the observations in the population, the group effects S_i are identically distributed with mean 0 and variance σ_A^2, the residual errors e_{ij} are identically distributed with mean 0 and variance σ_e^2, and the S_i and e_{ij} are independent (Donner, 1986). This model is frequently used in several fields of research, such as, for example, epidemiology, psychology and neuroimaging as shown in a recent review on graph metrics (Welton et al., 2015), and in a

[8]http://cran.r-project.org/web/packages/brainwaver/index.html
[9]http://cran.r-project.org/web/packages/igraph/index.html

meta-analysis of reliability graph metrics of rs-fMRI brain networks (Andellini et al., 2015). The intraclass correlation coefficient is then defined as the following ratio, $\rho = \sigma_A^2/(\sigma_A^2 + \sigma_e^2)$.

In Müller and Büttner (1994), authors highlight the difficulties to choose proper statistical measures of reliability depending on the design of the experiment. In this study, our aim was to test the reliability of inter-session acquisitions. To determine the level of reliability between two acquisitions (McGraw and Wong, 1996), we used intraclass correlation coefficient (ICC), which is based on the comparison of the within-subject and between-subject variability. This coefficient may not be adequate to test the conformity of methods or interchangeability as pointed out by (Bland and Altman, 1986), however it provides a quantitative value to easily build statistical comparisons.

Intraclass correlation coefficient (ICC)

ICC, as defined in the previous section, assesses the reliability of graph connectivity metrics by comparing the variability of these metrics during different sessions of the same subject to the total variation across all sessions and all subjects.

In line with several previous studies (Birn et al., 2013; Wang et al., 2011; Liang et al., 2012), we have applied a one-way random effect model, noted ICC(1,1) following Shrout and Fleiss (1979). This provides an estimation of ρ defined by,

$$ICC = \frac{s_b - s_w}{s_b + (k-1)s_w} \tag{1}$$

where s_b is the variance between subjects, s_w is the variance within subjects and k is the number of sessions per subject. ICC is close to 0 when the reliability is low, and close to 1 when the reliability is high. Note that ICC, as estimation of ρ using equation (1), may take negative values when the variance within subjects is larger than between subjects. This is due to statistical errors given a particular data set and should be considered as non reliable estimation.

A first approach to interpret the ICC is to classify its values into different categories with commonly-cited cutoffs (Cicchetti, 1994; Sampat et al., 2006): less than 0.4 indicates low reliability, 0.4 to 0.6 indicates fair reliability; 0.6 to 0.75 indicates good reliability and greater than 0.75 indicates excellent reliability. However, there are several limitations of ICC approaches, as described by Müller and Büttner (1994). First, ICC estimation may vary according to the estimation method leading to different versions of ICCs, based usually on parametric and non parametric approaches. In parametric approaches, ICCs vary according to the distribution and the equality of variances of the population. In addition, ICCs are dependent on the range of the measuring scale. Consequently, there is no reason to judge an absolute ICC as indicating good consistency, and it has been recommended to calculate confidence intervals (CI) in addition to ICCs (Shrout and Fleiss, 1979).

The ICCs and their CI evidenced a large range for different graph metrics throughout the test-retest literature. CI are computed using F-distribution (*e.g.*

the reviews of Boardman (1974); Donner and Wells (1986)) with degrees of freedom depending on the number of groups and number of subjects. In Cao et al. (2014), the authors computed the CI of ICCs and they reported, for example, that for an ICC of 0.45 for the E_g, the confidence interval was evaluated to be equal to $[0.09-0.71]$, for an ICC of 0.26 in b_i, the confidence interval was ranging between $[0.05-0.55]$, and for an ICC of 0.24 for e_{g_i}, the confidence interval was evaluated to be equal to $[0.04-0.55]$. These values of CI were computed with 26 subjects scanned twice. This example where ICCs were ranging from not reliable to good reliability, highlights that confidence intervals are unstable and difficult to interpret, especially in the context of fMRI studies with small sample size. In order to cut the margin of error in half, it is needed to approximately quadruple our sample size (Shrout and Fleiss, 1979). In a paper exploring several methods for constructing intervals for ICC, small sample size studies and normality assumption violation resulted in wide average interval width (Ionan et al., 2014).

p-values of ICC using permutation tests

In addition to the study of absolute values of ICC, p-values of ICC can be used. The addition of p-values allows a precise statistical analysis to evaluate the accuracy and significance of the extracted ICCs. The difficulty of working with p-values comes from the necessity to have access to the law of the estimators under the null hypothesis. In the case of ICC, it is possible to define an F-test to determine whether the ICC is significantly different from zero for a given level of confidence (McGraw and Wong, 1996). However, this parametric approach can be too restrictive when the sample of the data is too small or far from the Gaussian assumption.

Therefore, we propose to use a recent development of permutation tests to get a data-driven non parametric approach (Boardman, 1974). Each permutation consists in shuffling the acquisition sessions so that for each new subject the two sessions correspond to two different initial subjects, in particular, we shuffle the order of the subjects in the second session. The aim is to model the randomness of the measurements. For each permutation, we computed ICCs which produce a distribution of values where the two sessions correspond to a random choice of subjects, all or some of the paired sessions were disturbed. The true value of ICC obtained with the correct pairs of session of the same subject was then compared to the obtained distribution, hence the p-value is computed. The up-to-date statistical methods, based on Monte Carlo simulation (Metropolis and Ulam, 1949), test the reliability of our sample by randomly permuting the sessions between subjects. Two different tests were constructed. The first one concerns the global network level, where the goal is, for a given cost, to compute the p-value of the ICC for each metric. For that purpose, we use Simctest (Gandy, 2009). It is an open-ended sequential algorithm for computing the p-value of a test using Monte Carlo simulation. It guarantees that the resampling risk, the probability of a different decision than the one based on the theoretical p-value, is uniformly bounded by an arbitrarily small constant. Although the algorithm is open-ended, the expected number of steps is finite, except when

the p-value is on the threshold between rejecting and not rejecting. In the sequel of the paper, the ICC is used with p-values (of ICCs), with the aim of modeling the randomness of the measurements. We consider as reliable ICCs with a p-value≤ 0.05.

A second issue concerns the regional network level, where the tests are applied for each region of the parcellation scheme. In this case, we apply MMCtest (Gandy and Hahn, 2014) which is based on Simctest and includes a correction for multiple comparisons that is crucial when manipulating a large number of regions. Here, we applied the Benjamini-Hochberg procedure that controls the false discovery rate (FDR). These tools are freely available on CRAN [10].

In addition to the permutation, a step of bootstrap was associated to take advantage of the large size of the data set. For example, the results derived for 20 subjects were performed by first choosing at random without replacement a set of 20 subjects among the 100 in the original data set, and the p-values were computed using permutations of the restricted 20 subjects data set. This bootstrap test is repeated N times with a new set of 20 subjects for each repetition. We performed these tests considering 20, 30, 40, 60, 80 and 100 subjects to study how the reliability of the graph metrics depends on the subjects sampling procedure. In the case of $20-80$ subjects, we repeated the Simctest $N = 1000$ times and MMCtest, $N = 100$ times, selecting each time a random subsample of the data. We also repeated these bootstrap tests considering different number of volumes/time points. The original scan duration has 1200 time points at a TR=720ms, corresponding to a total duration of 14 min 24 s. We split it into four: 400 time points (4 min 48 s), 600 time points (7 min 12 s), 800 time points (9 min 36 s) and 1000 time points (12 min 00 s). All the subdivisions were extracted from the beginning of the time series up to each threshold.

Results

We analyzed the reliability of the graphs with respect to different factors that may influence ICCs and p-values: the sample size (number of subjects), the number of time points (duration), the graph metrics (global and regional), and the cost.

Between, within variances, ICC and p-values for E_g with respect to cost

ICC is based on the variance between and within subjects (see Eq. (1) and Fig. 1). Fig. 1 illustrates the computations of ICC, and p-values. The p-values are obtained using permutation techniques, and the error bars are obtained by using bootstrap on the number of subjects (no error bars can be computed using the whole set of 100 subjects). The first column of Fig. 1 displays the values of the between-subject variance s_b, and the within-subject variance s_w. Whatever the cost, for E_g, the between-subject variance s_b was found higher

[10] http://cran.r-project.org/web/packages/simctest/index.html

than the within-subject variance s_w with a maximum difference in the 15%-30% range. At high cost, these values are very small, and very close to each other so that the denominator of the ICC formula (s_b+s_w) is small, and results in high values of ICC (second column of Fig. 1). The p-values are displayed in the third column, so that in addition to the absolute values of ICC, the p-values are given an indication of confidence of these values compared to the randomness of the measurements.

Influence of the number of subjects

On average, s_b, s_w and ICC values are very similar whatever the number of subjects (see Fig. 1) but we can observe a decrease in the standard deviation as the number of subjects increases, resulting in a decreasing p-value with increasing number of subjects. Below 20% cost, the p-values of the ICCs reach significance only with 100 subjects. With 20 subjects the ICCs are only significant by chance. On average, with 30 subjects the ICCs are only significant from around 25% to 45% cost while with 40 subjects, they are significant from around 20% to 72.5% cost. In the case of 100 subjects, ICCs remain significant from 2.5% to 75%. At global level, we plot in Fig. 2 the ICCs and their respective p-values of E_g and E_l for costs between 2.5% to 75.0% at 600 and 1200 time points. The results for 20, 30, 40, 60, 80 and 100 subjects are displayed (computing the mean ICC and p-values of 1000 bootstraps in the first five cases). We observed that for a given ICC value, its significance depends on the number of subjects and on the cost range. The less significant results were observed for costs below 20%. Considering the experimental conditions with 1200 time points, with 20 subjects, ICCs were not significant for the analyzed metrics. With 30 subjects, we can obtain significant ICCs from 25% to 45% cost for E_g and for E_l from 30% to 35% cost. When considering 40 subjects, significant ICCs are observed for E_g in the cost range from 20% to 72.5%, and for E_l, in the cost range between 7.5% to 35%. With 60 subjects, significant ICCs are observed from 12.5% to 75% cost for E_g and for E_l from 7.5% to 40% cost. With 80 subjects, ICCs are significant from 5% to 75% cost for E_g and for E_l from 5% to 45% cost. With 100 subjects, in the case of E_g, ICCs were found significant at any cost, while with E_l, in the range between 2.5% to 52.5%. With 600 time points (Fig. 2 (a)), a similar evolution with the number of subjects is observed but with less significant values corresponding to smaller cost range.

Influence of the number of points in time

At global level, we found that the reliability increases with the number of time points. In Fig. 2 (b), the p-values are plot with respect to ICC for 1200 time points (corresponding to 14 min 24 s) and in Fig. 2 (a) for 600 time points (7 min 12 s), a duration currently used in the rs-fMRI literature (though, usually, with a TR=2s). We can observe that with 600 time points, the ICCs are reliable from 60 subjects, whereas with 1200 time points, reliable E_g and E_l can be achieved with groups of 40 and even 30 subjects at different cost range, as we mentioned in previous section.

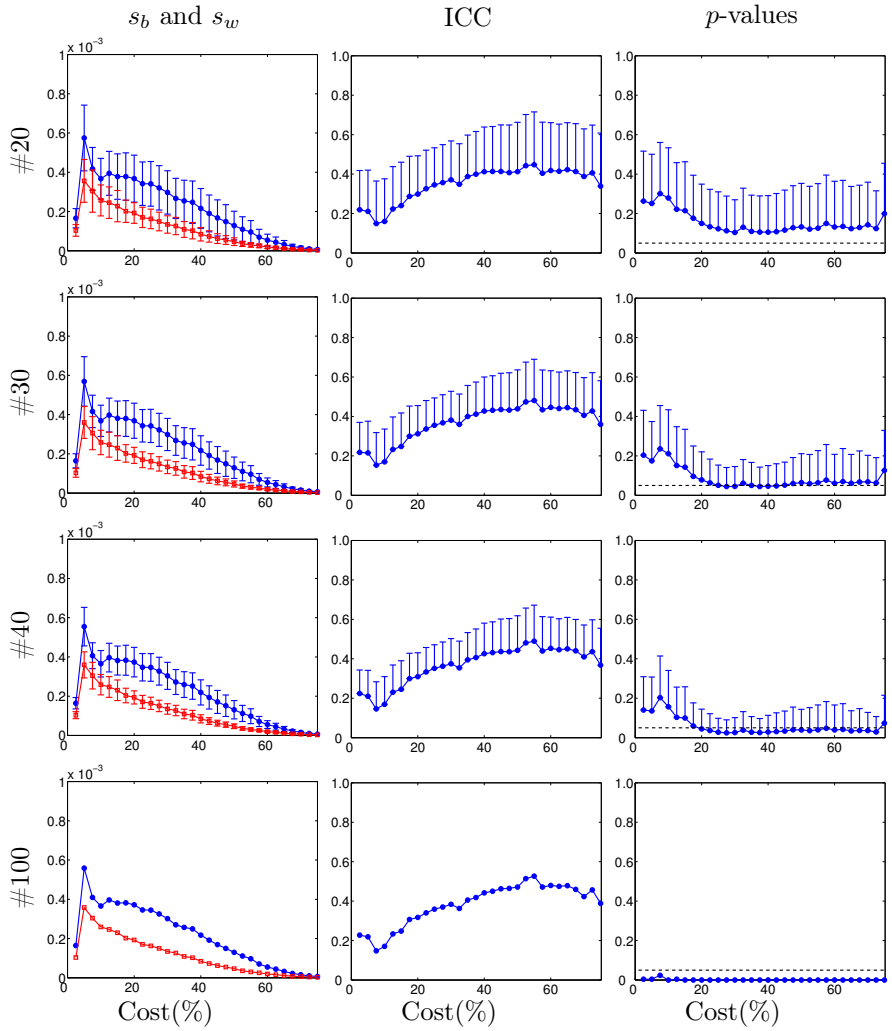

Figure 1: Reliability measures using ICC for global efficiency (E_g) and AAL89 as parcellation scheme. Each curve represents the between and within subjects variance (first column, respectively s_b in blue and s_w in red), values of ICC (second column) and associated p-values (third column) for E_g at 1200 time points as a function of the cost from 2.5% to 75%, in steps of 2.5%. Each row represents a different number of subjects (20, 30, 40 and 100 subjects). Error bars indicate one standard deviation of the bootstrap procedure. 1000 bootstraps were computed to select different subsamples of 20, 30 and 40 subjects. As the number of subjects is increasing, the p-values are decreasing, and the reliability is increasing. For 20 subjects, no p-values are significant, showing a poor reliability. However, for 40 subjects, p-values are significant for a large range of cost and reliable results are expected.

In Fig. 3, we display the p-values of ICC for E_g at 20% cost with respect to the number of time points for different groups of subjects. This result shows that it is not possible to achieve reliable results with 400 time points at 20%

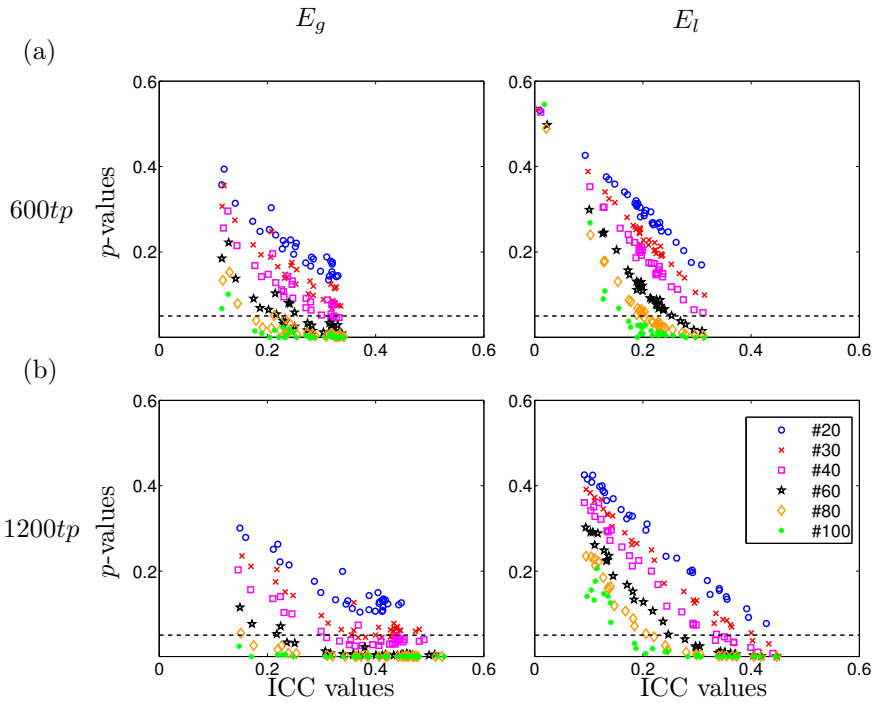

Figure 2: Reliability results in terms of number of subjects and scan duration. p-values of ICC (y-axis) as a function of ICC values (x-axis) for different number of subjects and different scan duration. (a) Using 600 time points, which corresponds to a scan duration of 7 min 12 s and (b) using 1200 time points (14 min 24 s). Mean result after 1000 bootstraps for cost values ranging from 2.5% to 75% are plot for 20, 30, 40, 60 and 80 subjects and for two global network parameters: E_g (left panels) and E_l (right panels). Note that increasing the scan duration and the number of subjects resulted in decreased p-values and that ICCs increase as the scan duration increases. Results correspond to the AAL89 parcellation scheme.

cost, even with 100 subjects. At 20% cost, significant p-values of the ICC are found to be achieved with 1200 time points and 40 subjects or for length above 600 time points and 60 subjects. At 400 time points, ICCs for the E_g are only significant for the range $37.5 - 55.0$% cost in 100 subjects (not shown).

Graph metrics reliability

At the global network level, in Fig. 4, we plot the p-values of ICC at different costs and for 6 different graph metrics. Plots on the left six panels correspond to 40 subjects randomly chosen 1000 times, and on the right panels computed with 100 subjects. In the former, with 40 subjects, E_g and B are significant from 20–60% cost, E_l and C from 7.5–35% cost, σ from 7.5–20% cost and Q from 10–25% cost, approximately. In the latter, with 100 subjects, all metrics are significant from 10% to around 40% cost.

At the nodal network level, Fig. 5 represents Pearson's correlation matrix between the p-values of the ICC of the different graph metrics at 1200 time

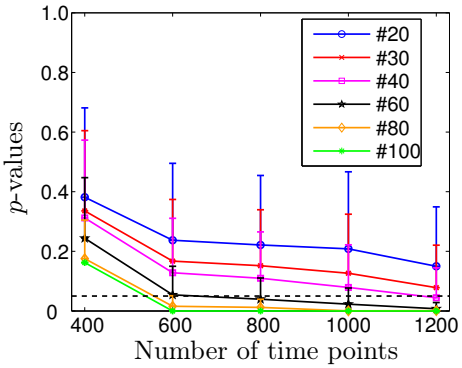

Figure 3: Reliability trade-off between number of subject and number of points in time using AAL89 as parcellation scheme. Evolution of the significance of global efficiency at global network level when increasing the number of points in time and the number of subjects. E_g mean p-values at different points in time applying 1000 bootstraps of 20, 30, 40, 60 and 80 subjects and for 100 subjects. All the results shown are computed at 20% cost. The number of subjects to achieve reliable results depends on the number of time points: a larger number of subjects is needed for a short scan duration. The correspondence between scan duration and time points is as follows: 400 time points (4 min 48 s), 600 time points (7 min 12 s), 800 time points (9 min 36 s), 1000 time points (12 min 00 s) and 1200 time points (14 min 24 s). All the subdivisions were extracted from the beginning of the time series up to each threshold.

points, with 30 subjects and 20% cost. It is possible to see that p-values of the ICC of e_{l_i} and c_i are highly correlated (89.00%) and p-values of the ICC for e_{g_i} and k_i are correlated (72.49%), while the rest of the metrics are not. This high correlation between p-values of the ICC of metrics means that same regions in the brain have similar significance reliability between those metrics.

Regional metrics reliability

Fig. 6 illustrates our first finding in terms of the number of regions that reach significant ICCs; the number of significant regions is dependent on the number of subjects and scan duration. When increasing the number of subjects from 20 to 100, the number of regions with significant reliability is 21 for 20 subjects, 42 for 30 subjects, 57 for 40 subjects, 77 for 60 subjects, 85 for 80 subjects and up to 87 for 100 subjects. The e_{g_i}, e_{l_i} and b_i with their p-values of all the AAL89 ROIs for 1200 tp at 20% cost can be found in Supplementary Material in Table 4 for 40 subjects and Table 6 for 100 subjects.

The locations of these regions are displayed in Fig. 7. Only significant regions are shown for 20, 30 and 40 subjects (100 permutations) and for 100 subjects at 20% cost for e_{g_i} and e_{l_i}.

We finally analyze the reliability of the regional values of global efficiency metrics with the p-values. As high values of regional e_{g_i} and b_i graph metrics are potential indicators of brain key regions (Bullmore and Sporns, 2009), we plot these metrics with their respective p-values at two different costs (20%

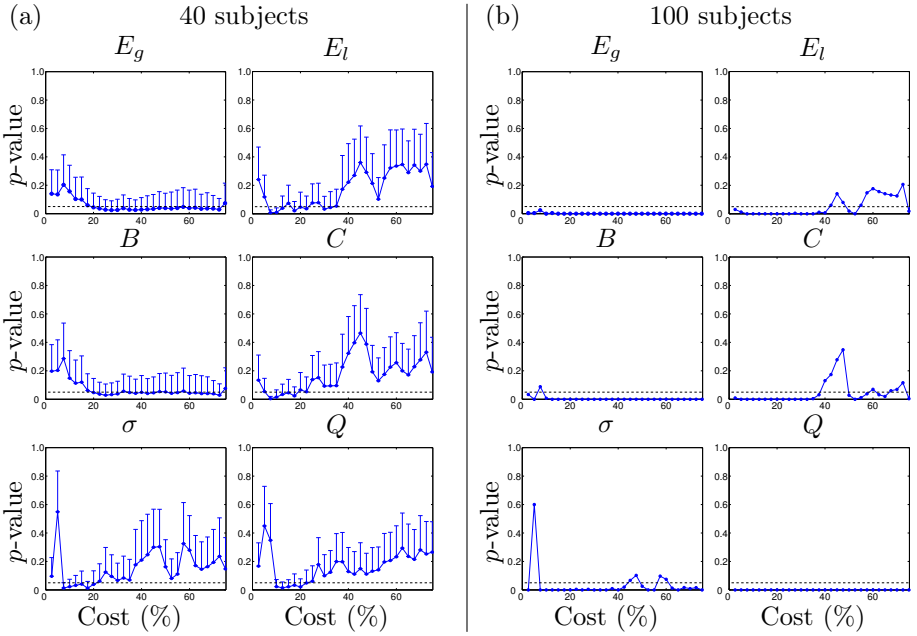

Figure 4: Reliability evaluation of different metrics using AAL89 as parcellation scheme. Mean p-values of ICC and standard deviation of 6 different global network parameters: global efficiency (E_g), local efficiency (E_l), betweenness centrality (B), clustering (C), small worldness (σ) and modularity (Q). Cost ranges from 2.5 to 75%. (a) 1000 bootstraps with 40 subjects randomly selected are shown using error bars with one standard deviation; (b) with 100 subjects.

in blue and 40% in red) for 40 subjects and 1200 tp in Fig. 8. On average, higher e_{g_i} values are associated with smaller p-values (but not always) at both costs. At 20% cost, with an e_{g_i} of 0.35–0.45, we found a 53% of nodes that are significantly reproducible, while from 0.55–0.65, there are 69%. Contrary, in the case of b_i, there are few significant nodes at both costs, not necessarily the nodes with highest b_i value. In terms of brain networks, this suggests that reliable key regions are better determined using e_{g_i} than b_i. Accordingly, we propose a classification of regions (Table 2) based on high e_{g_i} and on their p-values higher or lower than 0.05 to define: regions with high e_{g_i} and low p-value as 'reliable key regions', regions with high e_{g_i} and high p-value as 'non-reliable key regions', regions with low e_{g_i} and low p-value as 'reliable non-key regions', regions with low e_{g_i} and high p-value as 'non-reliable non-key regions'. The threshold for the proposed classification of e_{g_i} was set at the 65th percentile corresponding to values higher than 0.58.

Reliability versus parcellation

In Fig. 9, we show the p-values for E_g and E_l using the parcellation AAL459 and we compare the results for 40 and 100 subjects. With 40 subjects, both E_g

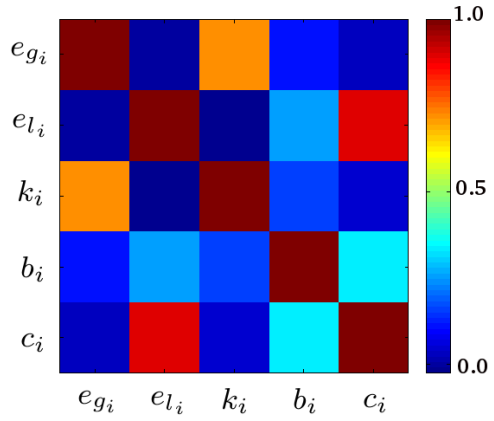

Figure 5: Correlation of reliability of graph metrics. Correlation matrix between the p-values of the ICCs of the AAL89 using 5 different regional network parameters. Results are computed at 1200 time points, 20% cost and 100 bootstraps of 30 subjects randomly selected: global efficiency (e_{g_i}), local efficiency (e_{l_i}), node degree (k_i), betweenness centrality (b_i) and clustering (c_i). A high correlation value between two metrics implies that the regions in the brain present similar reliability between those metrics. The chosen metrics do not show high correlation except between node degree and global efficiency and between local efficiency and clustering as it can be inferred from the definition of these metrics.

	$p \leq 0.05$	$p > 0.05$
$\geq \mid e_{g_i} \mid 0.58$	PrecGy (L/R), FrontMid (R), SMA (L/R), CingMid (L), Calcarine (L/R), Cuneus (L/R), Lingual (R), Occipital (L/R), Fusiform (R), ParietalSup (L/R), Precuneus (L/R), TempSup (L/R), TempMid (L), TempInf(L), RolandOperc (L/R), Cereb (VII, VIII, IX, X) (R)	FrontSup (L/R), FrontMid (L), CingMid (R), Lingual (L), Fusiform (L), PoscGy (L/R), TempMid (R), TempInf (R), Insula (R), Cuneus(R), FrontMid (L)
< 0.58	FrontSupOrb (R), FrontMidOrb (L/R), FrontInfOperc (L/R) FrontInfTri (L/R), RolandOperc (L), FrontSupMed (L/R), Insula (L), CingAnt(R), CingPost(R), Hippocampus (L), ParaHippoc (L), ParietalInf (L/R), SupraMarginal (L/R), Angular (L), ParacentralLob (L/R), Caudate (L/R), Putamen (L), Thalamus (L), Heschl (L), TempPole (L/R), TempInf (L), Cereb (I, II) (L/R), Cereb (III, IV, V, VI) (L/R) Cereb (VII, VIII, IX, X) (L), Vermis	FrontSupOrb (L), FrontInfOrb (L/R), Olfactory (L/R), FrontMedOrb (L/R), CingAnt(L), CingPost(L), Hippocampus (R) ParaHippoc (R), Amygdala (L/R), Angular (R) Putamen (R), Pallidum (L/R), Thalamus (R), Heschl (R)

Table 2: Regions with strong global efficiency (e_{g_i}) for AAL89 parcellation scheme. Classification of regions according to their e_{g_i} value and their p-value. We consider as key regions the nodes with the 33% of the highest e_{g_i} values (in this case the threshold is $e_{g_i} \geq 0.58$) and p-value $p \leq 0.05$, corrected for multiple comparisons. Some regions not classified as key regions are also found to be reliable. Results are computed using 100 bootstraps of 40 subjects at 1200 time points and a 20% cost.

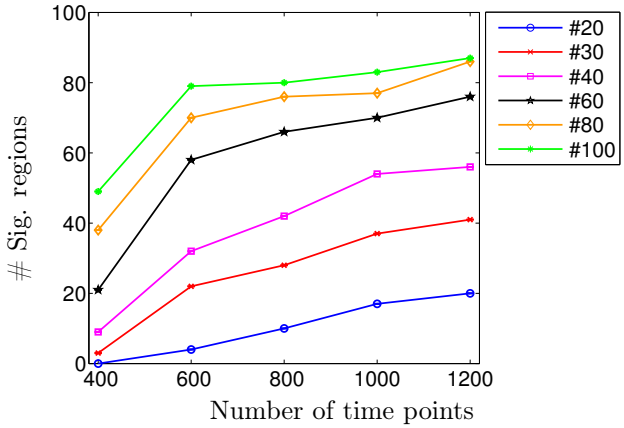

Figure 6: Reliability at the regional level using global efficiency and AAL89 parcellation scheme. Number of significant regions, computed using e_{g_i}, as function of the points in time for different number of subjects (corrected for multiple comparisons using a false discovery rate procedures at 0.05%). All the results shown are computed at 20% cost. The correspondence between scan duration and time points is as follows: 400 time points (4 min 48 s), 600 time points (7 min 12 s), 800 time points (9 min 36 s), 1000 time points (12 min 00 s) and 1200 time points (14 min 24 s). All the subdivisions were extracted from the beginning of the time series up to each threshold.

and E_l becomes significant; from $5-40\%$ cost in the case of E_l and $10-37.5\%$ cost for E_g. When considering 100 subjects both metrics are significant at almost every cost. This figure can be compared to the first row of Fig. 4 that displays the same plot with the AAL89 parcellation scheme. Both E_g and E_l are showing also reliable measures with a finer parcellation scheme. The two schemes were designed using different methods and the regions present different characteristics. The AAL89 parcellation is based on anatomical considerations with regions of different sizes, while the finer parcellation was designed using an algorithm to optimize the size of the regions and the covering of the brain.

In Fig. 10, we show the ICCs and their respective p-values for E_g and E_l using several anatomical and functional parcellation schemes at 20% cost and 1200 tp. In the case of E_g, we observe that the p-values are very close for the different parcellations except for HO117. A number of subjects between 30 and 40 is sufficient to achieve reliability on ICC except for HO117, where 60 subjects are required. In the case of E_l, p-values present bigger differences between parcellation schemes with lowest p-values for ICA200, followed by AAL459, then ICA100, Crad100 and AAL89, then HO117 and finally ICA50. Accordingly, for E_l, the number of subjects required to achieve reliability on ICC depends on the parcellation (lower row). We also show that ICC values on E_g and E_l are dependent on the parcellation scheme (upper row). With E_g, we found the highest ICC values for the finer parcellations: ICA200, ICA100 and AAL459; then for ICA50, AAL89 and Crad100 and finally, for HO117 parcellation. With E_l, highest ICC values were found with AAL459, followed by ICA200, ICA100

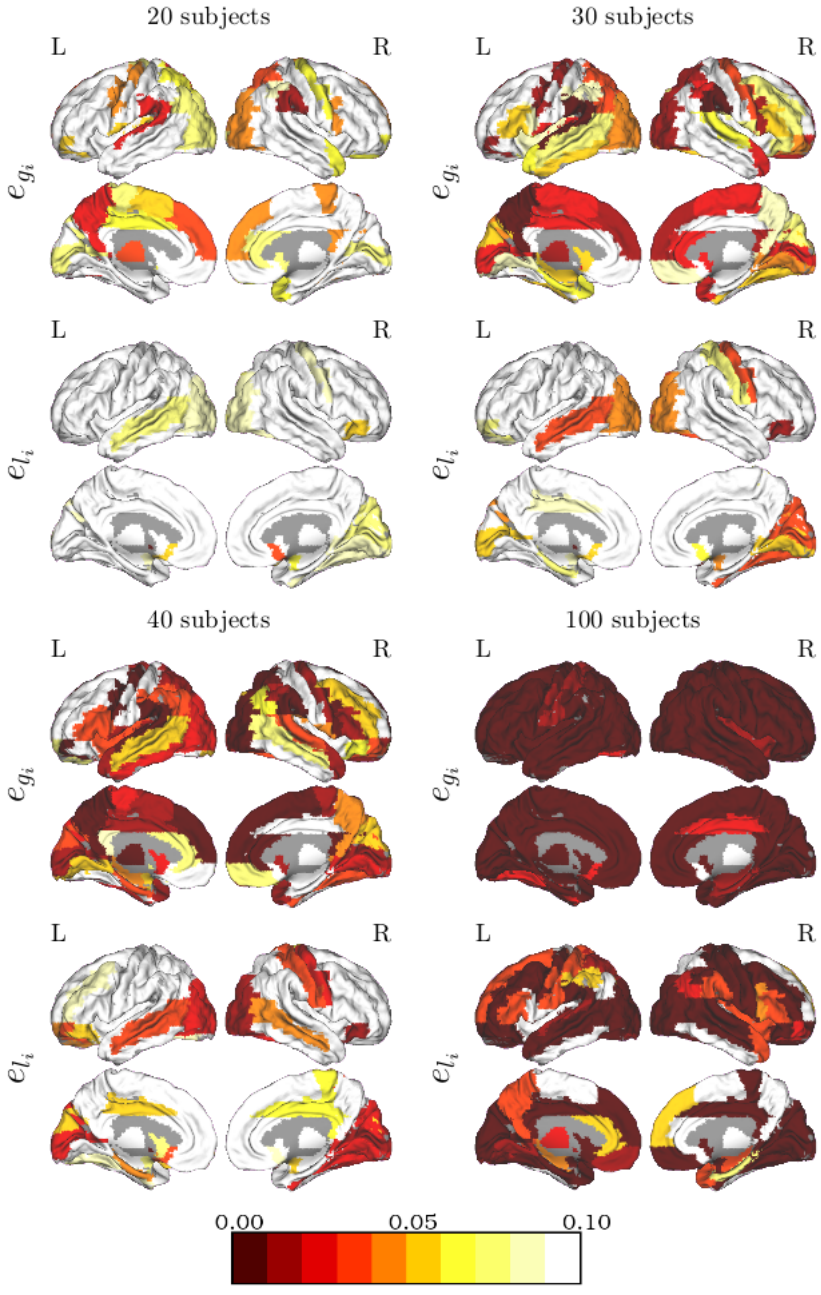

Figure 7: Brain maps of reliable regions for AAL89 parcellation scheme. Cortical surface representation of nodes that demonstrated significant regions on the brain using two regional network parameters: global efficiency (e_{g_i}) and local efficiency (e_{l_i}). The displayed p-values are the ones corrected for multiple comparisons using a false discovery rate at 0.05%. First two rows, e_{g_i} for 20, 30, 40 (100 bootstraps) and 100 subjects. Last two rows, e_{l_i} for 20, 30, 40 (100 bootstraps) and 100 subjects. Results were obtained at 1200 time points and 20% cost. The tables of ROI's names with their p-values from 20 to 100 subjects can be found in Appendix together with the e_{g_i} and e_{l_i} of the ROI for 40 subjects.

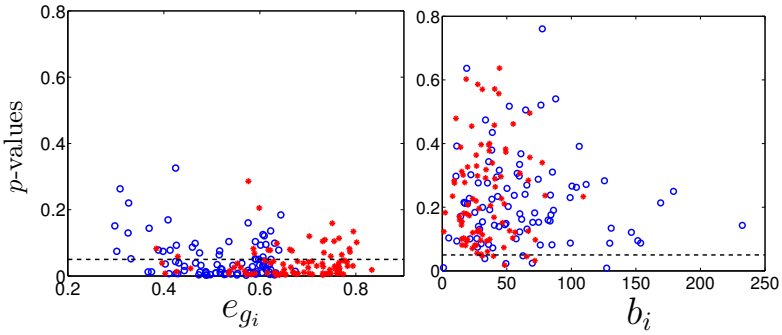

Figure 8: Reliability of brain regions in terms of cost using AAL89 as parcellation scheme. Mean e_{g_i} (left) and b_i (right) with their mean p-values. Computed for 40 subjects, 20% cost (in blue), 40% cost (in red) with 100 bootstraps (error bars are not shown) at 1200 points in time. Interestingly, the number of significant reliable regions obtained with betweenness is less than the one obtained with global efficiency. This may show that global efficiency is better at characterizing reliable hubs.

and AAL89, then Crad100 and HO117 and finally ICA50. In the case of B, the results show dependence of the p-values and on the ICC with the template used to compute the graph (Supplementary Material, Fig. 11).

These results argue for more thorough studies on this topic which is out of the scope of this paper.

Discussion

The present study has investigated the test-retest reliability of brain network properties/metrics derived from graph theory methods using rs-fMRI acquisition. Original extensive statistical analyses have been conducted by using a large number of subjects (100 healthy participants) and a high number of time points (1200 time points/volumes; 14 min 24 s duration) provided by the test-retest data set of the Human Connectome Project.

In this paper, we first showed that ICC can measure reliability only when combined with confidence intervals or p-values. For each tested parameter, we found a smaller within subjects variance in comparison to the between subjects variance. This suggests that the extraction of graph metrics from rs-fMRI brings valuable information that are consistent with a test-retest analysis. However, having a positive difference does not mean that it is statistically significant, and the use of p-values is crucial to quantify the reliability of the rs-fMRI brain connectivity using graph metrics. These p-values were found to be significant for a whole set of parameters showing that the rs-fMRI brain connectivity networks present common characteristics that are shared by a large number of subjects and also individual features that make each subject unique.

Although ICCs are widely used to quantify test-retest reliability, a large discrepancy in ICC values was found in a systematic review of the literature (Welton et al., 2015). As ICC values depend on both cost and scan length (Fig. 2),

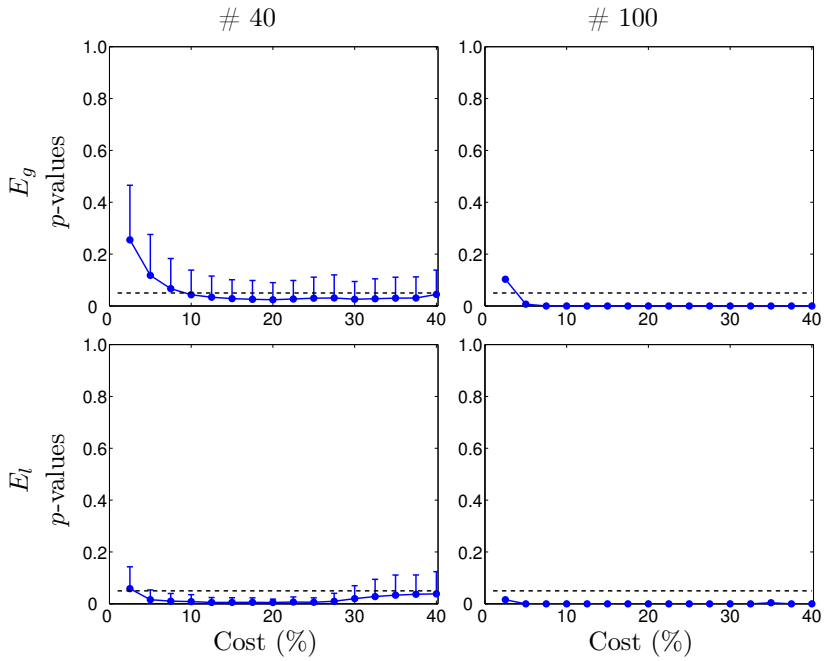

Figure 9: Comparison between the reliability of 40 and 100 subjects using the parcellation of 459 regions. Two global network parameters (Global and local efficiency) are compared. E_g (first row) and E_l (second row) p-values; on the left, 40 subjects (1000 bootstraps) and 100 subjects on the right. Results computed for cost from 2.5 to 40%. This figure can be compared to the first row of Fig. 4.

high ICC scores do not necessarily indicate reliable results. Therefore using confidence intervals (Braun et al., 2012; Cao et al., 2014; Liang et al., 2012; Liao et al., 2013; Birn et al., 2013) or p-values in addition to ICCs offers an efficient procedure for dealing with sometimes inconsistent results among studies jointly characterized by small sample sizes, short scan duration and different fMRI acquisition and analysis techniques. As shown in this work, the advantage of p-values is to provide direct information on the significance of the ICCs, and permutations tests are a very efficient way to compute p-values, when sufficient data are available. Accordingly, we recommend that ICC values should be accompanied by p-values to assess TRT reliability.

The main parameters analyzed in this study are the sample size i.e. number of subjects, scan length, i.e. the duration, choice of metrics and optimal cost. The increase of sample size and scan length was characterized by a decrease on the p-values of ICC illustrating the major role of these two parameters in reaching statistical significance to obtain reliable metrics at the global level. The ability of graph methods to quantify the role of each node of the graph (or region of the brain) allows us to study regional reliability. The p-values of ICC scores were extracted for each region separately and after a correction for multiple comparisons, the decrease of the p-values was confirmed when increasing the

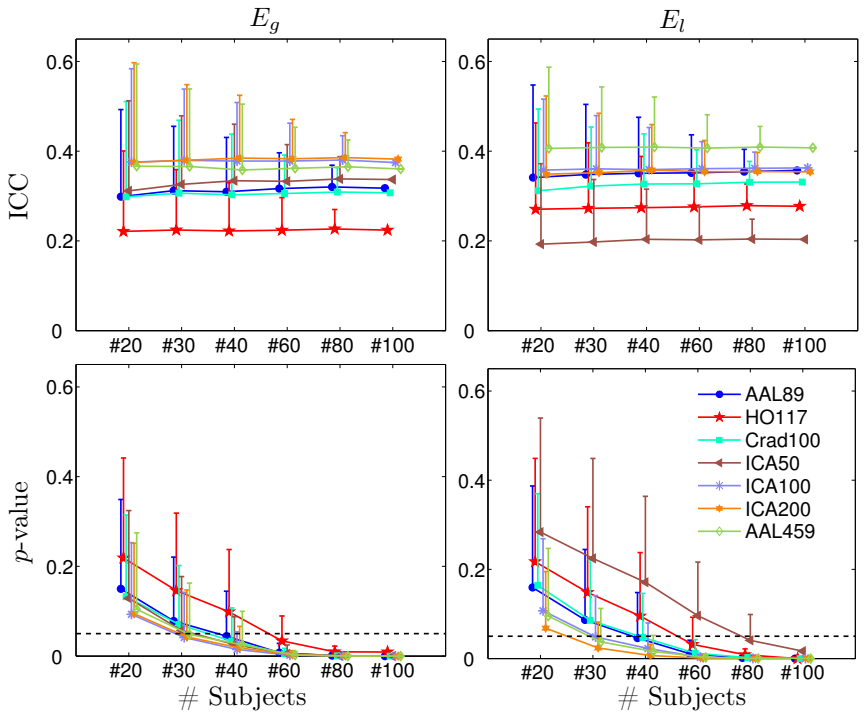

Figure 10: ICCs and their p-values for different parcellation templates and different number of subjects. Results are shown for global efficiency (E_g) and local efficiency (E_l) in first and second column, respectively, at 1200tp and 20% cost.

sample size or scan length. Major reliable regions, for study with 40 subjects, were found in the default mode network, the motor and the visual networks.

Brain connectivity graphs to find potential biomarkers

Up to now, brain connectivity is mainly studied for the discrimination of groups of patients and used to characterize the disruption in the connectivity affected by a certain disease (De Vico Fallani et al., 2014). Moreover, graph representation of brain connectivity has the potential to extract a unique representation for each patient and to provide a unique tool to quantify the brain connectivity networks at the individual level. However, from a translational perspective, graph metrics could serve as biomarkers for diagnosis, follow-up and treatment efficacy only if it is proved to be reliable across acquisitions and subjects. We provide in this study the combination of parameters that allow the graph representation of brain connectivity networks to be used as potential biomarkers. As we have observed along the results, the reliability of the graph metrics depends on the number of subjects, the number of time points and the cost of the graph.

Influence of the number of subjects. Reliability of global metrics is strongly related to the number of participants that are included in the study, as shown in Fig. 1, where the p-values progressively decrease when adding participants. As can be observed in Fig. 2 and 3, there is a scale in effect for both E_g and E_l, with a threshold number of 40 subjects. Indeed, a minimum number of 80 to 100 participants is recommended to carry out reliable graph analysis at both global and regional network level. This is, to our knowledge, the first report showing the influence of the number of participants in the resting state literature, since previous studies were all performed with a limited number of subjects (between 11 and 33). Indeed, the poor reproducibility of scientific works reported by statisticians (Button et al., 2013; Ioannidis, 2014) would be, at least for a part, due to the low statistical power because of low sample size. This is a strong push for running multisite studies that are powered to obtain reliable results (Button et al., 2013). This kind of approach is strongly encouraged both by institutions such as the Meta Research Innovation Center at Stanford [11] and by the neuroimaging community (Consortium for Reliability and Reproducibility (CoRR) [12], with the need to quantify site-related variance. Our findings are derived using a population of healthy volunteers and we focus on reliability as established by TRT studies. The goal of identifying difference between groups of subjects is different and our recommendations may be over evaluated. Indeed, the potentiality to find differences depends on the amplitude of the unknown differences between the two groups. However, if very few assumptions are known when comparing two groups of subjects using our recommendations would minimize the risk of not being able to identify differences between groups.

Influence of scan duration on TRT reliability. The amount of time points or the duration of the acquisition required to obtain reliable results is also another open issue in brain connectivity fMRI analysis with graphs. Thanks to the long scan acquisition (1200 time points) and short repetition time (720 ms) of the HCP data set, we could assess test-retest reliability with different numbers of time points, i.e. different scan duration. We considered 5 different number of time points (from 400 to 1200) corresponding to scan duration ranging from 4 min 48 s to 14 min 24 s. As scan duration increases, reliability increases in parallel with a minimum of 600 time points or 7 min duration that is required to achieve reliable results for global metrics (Fig. 3). In the literature on whole brain graph analysis of rs-fMRI, the issue of the influence of scan duration on test-retest (TRT) reliability has been previously addressed (Liao et al., 2013; Cao et al., 2014). Liao et al. (2013) found increased ICC values as the scan duration increased during the first 5-6 min and confidence intervals above 0 for a minimum duration of 5 min. No gain was observed for the ICC values after 6 min (Liao et al., 2013), but the confidence intervals are not shown after 5 min, limiting the validity of their findings for longer duration. Similarly,

[11] http://metrics.stanford.edu/
[12] http://fcon_1000.projects.nitrc.org/indi/CoRR/html/

Braun et al. (2012) observed that shortening in a group of 33 subjects the scan duration from 5 to 3 min led to decrease significantly the reliability, leading the authors to recommended the acquisition of longer time series. In a recent meta-analysis based on TRT reliability of graph metrics studies, Andellini et al. found a significant increase in the TRT reliability for time series longer than 5 min compared to times series shorter than 5 min (Andellini et al., 2015). As an effect of scan duration was questioned for longer duration ($5 - 15$ min) by studies outside TRT (Van Dijk et al., 2010; Whitlow et al., 2011), Birn et al. (2013) explored the influence of scan duration using 9 different scan duration from 3 to 27 min, in the context of a seed based approach TRT study (25 subjects). Increase duration had a significant effect on TRT reliability, until a plateau reached around 13 min for intra-session reliability and 9 min for inter-session reliability. Authors concluded that TRT reliability can be improved by increasing the scan duration to 12 min. Although the methodology of analysis applying graphs is different from theirs, our findings are consistent with those of Birn et al. (2013).

Influence of the cost. At low cost, typically 10.0% or below, global metrics such as global efficiency E_g and betweenness centrality, B, are not reliable up to 40 subjects (Fig. 4). This might be explained by the low sparsity of the graph, thus preventing robust calculation of these metrics. Conversely, above 30.0% cost, more edges are added to the graph that becomes more uniform so that potential variability within the network is reduced. For such costs, the metrics related to local features in the graph such as clustering coefficient C, modularity Q or local efficiency E_l appear poorly reliable. This might be related to an insufficient sparsity in the graph. In the literature, costs are currently in a range from 1.0% to 40.0%, or computed using values integrated on all the cost range. The influence of cost ranging from 10.0% to 40.0% on the reliability of global and local metrics has been previously investigated by Braun et al. (2012) and Wang et al. (2011) without clear results, given the large uncertainty on ICC values. Here, aiming at finding reliability for all global metrics, we considered a cost around 20.0% or 25.0% to be a good trade off. Accordingly, most of the figures presented in this study were performed at 20.0% cost.

Reliability of the different graph metrics. The sample size appears to be a key parameter to achieve reliable global graph metrics (Fig. 4). At the global network level, Fig. 4 also shows that in small samples global metrics are cost dependent. Global efficiency, and in a lesser extent betweenness centrality, appear to be the most robust metrics in small samples (40 subjects) in a large cost range, while all metrics are reliable at intermediate costs in larger samples, here 100 subjects (Fig. 4). At regional level, for 20.0% cost, we found that degree and global efficiency present similar reliability, in line with Du et al. (2015) and Wang et al. (2011), as well as clustering coefficient and local efficiency (Fig. 5).

Trade off between number of subjects, number of time points and cost. Our findings showed that reliability of the graph metrics computed at both global

and regional level depends on two main factors, the number of subjects and the number of time points. This suggests that good reliability of graph metrics cannot be easily achieved with a sample size of 10 to 30 subjects and relative short duration (5 to 10 min.) that are sometimes used in studies of human cognition and clinical research. Fig. 4 shows how reliability is also influenced by the cost and the type of graph metric. Therefore, reaching the appropriate sample size may become an issue given the particular study design (recruitment of patients, rare inclusion criteria, complex cognitive study) and hypothesized effects. An alternative way to obtain reliable metrics consists of increasing the scan duration and to select the cost according to the metric of interest. For instance, as shown in Fig. 3, at the global network level and at 20.0% cost, global efficiency is significantly reliable at 1200 time points for 40 subjects or above, or at 600 time points for 60 subjects. Taking together the influence of the number of subjects, the number of time points, the cost and the metrics, it appears that reliability at the global level can be achieved through a trade off between these parameters.

Influence of the parcellation scheme

The main results given in this study were all obtained with the AAL parcellation scheme, slightly modified to merge the individual cerebellar lobules based on their anatomical and functional properties: vermis (median anatomical structure), lobules I-II (vestibulo-cerebellum); lobules III-VI (quadrangular lobule engaged in motor control) and lobules VII-X of the posterior cerebellar lobe engaged in cognition) (Duvernoy, 1995; Schmahmann et al., 1999; Stoodley and Schmahmann, 2009; O'Reilly et al., 2010) (see Supplementary Material for details). This classification, resulting in larger parcels, could have modified the reliability and the values of the graph metrics in the cerebellum.

Another potential issue is the parcellation scheme size. Using a number of nodes in the graphs ranging from 84 up to 4320, Fornito and collaborators (Zalesky et al., 2010; Fornito et al., 2010) found that the graph metrics depend on the number of nodes in the networks and on the cost of the graph. Noticeably, they observed larger differences in global metrics at 10% cost than above 20% cost, indicating that low cost might not be a good choice for graph analysis, a result coherent with our findings even if their studies were not TRT. Here, using a finer parcellation with 459 regions, we found an enlarged cost range from 5% to 40% for computing reliable local efficiency, E_l, as compared to the cost range of 10% to 20% using the AAL. This is confirmed with the use of other parcellation schemes. However, the results depend on the metric used. For example local efficiency or clustering are more reliable when using the finer parcellations. While, when computing global efficiency, the parcellations with a low number of regions are still providing reliable results. These findings are in line with a recent review of TRT literature with graphs, in which it was found that the reliability of graph metrics was influenced by the parcellation scheme but which of those yields to the highest reliability remains unclear (Andellini et al., 2015).

Exploring reliability at the regional network level

Another option to design clinical and cognitive studies for graph analysis is to focus at the regional network level to target reliable regions of interest. Fortunately, when a large number of subjects is available (80 or 100), almost all the brain parcellated regions are reliable (Table 6), and the values of the graph metrics can be used to classify nodes in terms of specific roles in networks. However, it is important to keep in mind that with a small sample size (below 40), the characterization of key regions has first to be referred for their reliable value, as only a part of the brain regions are reliable. Fig. 7 shows that, even with a small number of subjects (20 to 40 participants), e_{g_i} and e_{l_i} can be explored in a set of significantly reliable regions. The number of reliable regions depends on the time points and, for a given number of subjects, it increases from 400 to 800 time points to reach a plateau at 1000-1200 time points (Fig. 6). This plateau corresponds to 12 min., a duration that is also recommended by Birn et al. (2013) using a different methodological approach. Reliable regions as a function of sample size subgroups (20, 30, 40 and 100 participants) can be found in Table 5 for global efficiency at a cost of 20.0% with the AAL89 parcellation. As indicated in this table, the number of subjects can be adapted to the set of regions of interest in the study. For example, a sample size of 20 subjects is nearly sufficient to study the DMN, whereas, the exploration of SMA connectivity requires at least 30 subjects. To determine the regions which are most reliable depending on the type of metric in small samples, we report the values of e_{g_i}, e_{l_i} and b_i with their p-values for each AAL89 regions in Table 4 for graphs computed at 20.0% cost, for 40 subjects and 1200 time points. In these conditions, e_{g_i} presents the highest number of significant regions (58 regions among 89), e_{l_i} (19 regions) and b_i the lowest number of significant regions (6 among 89). It should be noted that the low reliability of betweenness centrality has been previously reported by Guo et al. (2012) when studying the default mode and the salience networks and by Du et al. (2015). Therefore, we recommend to explore e_{g_i} and e_{l_i} rather than b_i for exploring graphs in small sample studies. These figures and table are thought to be useful for designing future rs-fMRI studies. In order to determine reliable key region in small samples, we used e_{g_i} values to rank the regions with their p-value of ICC at regional level for 1200 time points (corresponding to 14 min and 24 sec) in 40 subjects, at 20.0% cost (Table 2). Considering as reliable key regions the ones above $65th$ percentile for e_{g_i} values (higher or equal 0.58) and p-value$\leq$ 0.05, we found a set of 25 reliable key regions, which are listed in bold in Table 4.

Defining reliable key regions with graph metrics in small samples. In parallel, the measure of the metrics reliability can also be applied to identify significant or robust key regions in small samples. The present study showed that global efficiency (or similarly minimum path length) and node degree were the most reliable metrics. As commented above, considering 40 subjects, we found 25 regions that remain highly connected and reliable over time, printed in bold in Table 4 and in the upper left part of Table 2. All these regions belong to three main groups of resting state networks: the default mode networks (bilateral

precuneus, left middle part of the cingulum and left inferior, left middle and bilateral superior temporal lobe), the sensorimotor networks (bilateral precentral gyri, bilateral SMA, right rolandic operculum) and the visual networks (bilateral calcarine cortex, bilateral occipital areas, bilateral cuneus, right lingual and right fusiform gyri), in agreement with previous studies (Liao et al., 2013; Wang et al., 2011; Agcaoglu et al., 2015).

Defining reliable key regions in subcortical regions. At subcortical level, all cerebellar regions show significant reliability, even for 20 subjects (Table 5). Tomasi and Volkow (2012) have found the cerebellum to be a reliable and global key region using high functional connectivity density. The lobules VII, VIII, IX and X that are comprised in the posterior lobe of the cerebellum, are incorporated into a distributed neural circuits subserving complex movement, language, working memory and emotion (Stoodley and Schmahmann, 2009; O'Reilly et al., 2010). Indeed, impairment in motor control, behavioral and cognitive functions can be observed after cerebellar lesions (Stoodley et al., 2012). Of note, the left cerebellar lobe did not show high graph metrics, which could be related to acquisition limitations (Table 6). In this study, when examining graph metrics in the basal ganglia and the thalamus, we found the caudate nucleus to be reliable in the small sample of 40 subjects, although graph metric values were average (Table 4). In the 100 subject sample, the striatum (putamen and caudate), pallidum and left thalamus regions that are part of the basal ganglia network (Agcaoglu et al., 2015; Malherbe et al., 2014) are all reliable (Table 6). The engagement of the basal ganglia and more particularly the implication of the dorsal caudate in executive functioning, working memory and learning has been documented in a body of neuroimaging, anatomical and behavioral studies (Middleton and Strick, 2000; Grahn et al., 2009; Malherbe et al., 2014). Although previous TRT graph studies have not reported the basal ganglia as key regions (Liao et al., 2013), the fact that striatal regions show relatively high b_i values, combined with a possible loss of signal related to acquisition and segmentation limitations (see limitations subsection) suggest that their role as a multimodal key region needs to be more specifically addressed.

Methodological considerations

The normal resting adult human heart rate ranges from $60 - 80$ bpm, which belongs to a frequency band between $[1.00 - 1.33]$ Hz. Given the short TR acquisition value, the maximal frequency is $1/(2\ \text{TR}) = 0.69$Hz. This band is aliased within the frequency range of $[0.05 - 0.38]$ Hz mainly located within wavelet scales 2, 3 and 4. Thus, there could be some contamination of our results by cardiac signal. The respiratory signal at rest is in the frequency band between $[0.20 - 0.25]$ Hz, typically located in the wavelet scale 2. It is, thus, not contaminating our results.

While in most rs-fMRI studies the sampling rate used is about 0.5 Hz (corresponding to TR = 2 s), the multiband acquisition technique applied to fMRI (Feinberg et al., 2010) offers the advantage of increasing the sampling rate

up to 1.67 Hz (corresponding to TR = 0.6 s) for a whole brain coverage. Previous studies have used multiband rs-fMRI data to test the reliability of brain network regions (Liao et al., 2013; Tomasi et al., 2015). A main limitation of low sampling rate (or long TR) is that the cardiac and respiratory contributions are aliased in the data leading to artifacts. An advantage of multiband sequence is thus to prevent artifacts related to these physiological factors.

The equivalence between time points and scan duration may be questionable, as number of time points used in the wavelet approach depends on the duration, the repetition time, and the frequency band. In Table 3 (see Supplementary Material), we show the equivalences between these parameters. For example, in order to compute the graphs with at least 50 independent points in time, we need 800 time points for a TR = 720 ms (scan duration of 9 min 36 s) or 400 time points for a TR = 2 s (scan duration of 13 min 20 s).

Finally, the dyadic wavelet transform applied in this study splits the frequency band in five sub-bands. This permits to compute correlations that relate to each particular sub-band, a strategy which was found to improve reliability (Guo et al., 2012). The results obtained here are thus not fully comparable with those found with larger frequency bands, such as the currently used 0.01 to 0.1 Hz band. Because of these differences as compared to other graph preprocessing, the results found here might not be strictly equivalent to other preprocessing strategies.

Limitations

The dataset used in this study was acquired in healthy young adults aged from 20 to 36 years old and included 54 females and 46 males. The analysis performed in this study did not attempt to discriminate between gender and age. Age effect was reported in structural and functional connectivity among resting-state networks across the human lifespan (Betzel et al., 2014) but in the small age range of this study, this effect was considered as negligible. Moreover, during adulthood, the age effect might be moderate (Song et al., 2012) and the results found here could probably be extended to subjects up to 60 years old. Gender effect has also been reported in youth (between 9 and 22 years of age) (Satterthwaite et al., 2015) and in the lateralization of the resting-state networks (Agcaoglu et al., 2015). Agcaoglu et al. (2015) found that differences related to gender are consistent in samples above 250 subjects. Given the relatively lower number of subjects studied here, the effect of gender was not considered.

In order to compute the time series in the gray matter alone, the T1w structural images were segmented to obtain a gray matter mask. This permits to eliminate most of the white matter and cerebrospinal fluid contribution in the time series. The T1w structural image was acquired with a MPRAGE sequence, which is efficient at presenting a good gray matter/white matter contrast in the cortical surface. However, the high iron content of the subcortical nuclei (especially caudate, pallidum and putamen) (Haacke et al., 2005) further shortens the T1 in these deep structures. As a consequence, they exhibit reduced contrast from white matter in T1w images and are often misclassified by segmentation

algorithms (Helms et al., 2009). The time series may thus not be optimally extracted in these deep structures, reflecting the relative low reliability of the graph metrics in the basal ganglia (putamen, pallidum, caudate) and in the thalamus.

The dataset provided by the Human Connectome Project comprised two different acquisitions, one with the phase encoding direction from left to right and the other one from right to left. As the mental state of the subjects may differ between them, rather than pooling them, we chose to use only the first dataset, the left to right one. Despite the distortion corrections schemes that were applied in the preprocessing stage, the inhomogeneity of the static field may have led to some residual distortions which can be seen in the inferior part of the temporal lobe, and to some loss of signal in regions such as the olfactory, orbito-frontal and amygdala regions.

Acknowledgments

M.T. is supported by a grant from the Rhône-Alpes Région, France. This work was partly funded by France Life Imaging, grant ANR-11-INBS-0006. S.A. was partly funded by a grant from la Région Rhône-Alpes and a grant from AGIR-PEPS, Université Grenoble Alpes–CNRS.

Data were provided by the Human Connectome Project, WU-Minn Consortium (Principal Investigators: David Van Essen and Kamil Ugurbil; 1U54MH091657) funded by the 16 NIH Institutes and Centers that support the NIH Blueprint for Neuroscience Research; and by the McDonnell Center for Systems Neuroscience at Washington University.

We want to thanks Prof. Axel Gandy and Ph.D. Georg Hahn for their help to run the MMCtest tool and Ph.D. Petra E. Vértes for fruitful discussions.

1. Supplementary Material

1.1. Modified version of classical AAL parcellation scheme

The classical AAL parcellation scheme is composed by 116 regions including the cerebellum. We have merged some of the regions, reducing the parcellation to 89 regions. Merged regions are: frontal medial orbital and rectus (one region for left and one for right hemisphere); occipital superior, middle and inferior (one region for left and one for right hemisphere); temporal pole superior and medial (one region for left and one for right hemisphere); the cerebral crus (one region for left and one for right hemisphere); areas III, IV, V and VI of cerebellum (one region for left and one for right hemisphere); areas VII, VIII, IX, X of cerebellum (one region for left and one for right hemisphere) and finally, the vermis (one single region for both hemispheres).

	TR=2s; Fmax=0.25Hz				TR=720ms; Fmax= 0.695Hz			
	Scan duration Initial Nb Time points	$13min20s$ 400 tp	$6min30s$ 200 tp			$14min24s$ 1200 tp	$9min36s$ 800 tp	$4min45s$ 400 tp
				1	0.35-0.69 Hz	600 tp	400 tp	200 tp
1	0.12-0.25 Hz	200 tp	100 tp	2	0.17-0.35 Hz	300 tp	200 tp	100 tp
2	0.06-0.12 Hz	100 tp	50 tp	3	0.085-0.17 Hz	150 tp	100 tp	50 tp
3	0.03-0.06 Hz	50 tp	**25 tp**	4	0.042-0.085 Hz	75 tp	50 tp	**25 tp**
4	0.015-0.03 Hz	**25 tp**	**12 tp**	5	0.024-0.042 Hz	37 tp	**25 tp**	**12 tp**

Table 3: Illustration of reliability for two different datasets using two different TR; numbers in bold indicate insufficient number of points in time to get reliability using graph analysis according to p-values obtained on ICC.

1.2. TR, scan duration and frequency band

The problem of choosing the scan duration is linked to the sampling rate of the acquisition and the interested frequency bands. The following tables present two examples, one with a TR equal to 720 ms (the one used in this study), and a TR equal to 2 s used classically with non multiband acquisitions.

1.3. Standard ICC computation

Considering a graph metric (e_j^i) per subject ($i \in [1, n]$) per session ($j \in [1, k]$):

$$\mathbf{X} = [\mathbf{x_0}\, \mathbf{x_1} \ldots \mathbf{x_k}] = \begin{bmatrix} e_1^1 & e_2^1 & \ldots & e_k^1 \\ e_1^2 & e_2^2 & \ldots & e_k^2 \\ \ldots & \ldots & \ldots & \ldots \\ e_1^n & e_2^n & \ldots & e_k^n \end{bmatrix},$$

where the mean graph metric value per subject across sessions is:

$$\mathbf{t}(\mathbf{X}) = \begin{bmatrix} t^1 \\ t^2 \\ \ldots \\ t^n \end{bmatrix};$$

where $t^i = \frac{1}{k}\sum_{j=1}^{k} e_j^i$ and the total mean $\bar{x} = \frac{1}{nk}\sum_{i=1}^{n}\sum_{j=1}^{k} e_j^i$.

The within subjects mean squares (s_w) difference is then obtained by computing the mean of the square difference between each session value and the mean value among sessions:

$$s_w = \frac{1}{n(k-1)}\sum_{i=1}^{n}\sum_{j=1}^{k} p_j^i, \tag{2}$$

where $p_j^i = (e_j^i - t^i)^2$.

and the between subjects mean square (s_b) difference:

$$s_b = \frac{k}{n-1}\sum_{i=1}^{n} b^i, \tag{3}$$

where $b^i = (t^i - \bar{x})^2$.

Finally, the ICC value is computed as follows:

$$ICC = \frac{s_b - s_w}{s_b + (k-1)s_w} \tag{4}$$

1.4. Reliability of betweenness centrality for different parcellation schemes

See Fig. 11

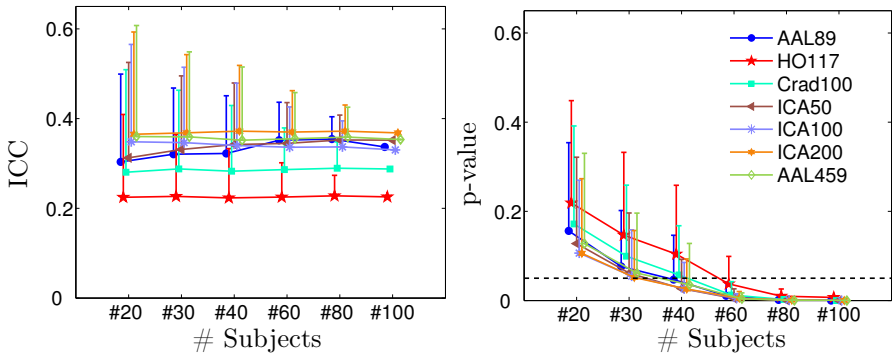

Figure 11: ICCs and their p-values for different parcellation templates and different number of subjects. Results are shown for betweenness centrality (B) at 1200tp and 20% cost.

1.5. Nodal network metrics

See Table 4

References

Achard, S., Bullmore, E., 2007. Efficiency and Cost of Economical Brain Functional Networks. PLoS Comput Biol 3, e17. URL: `http://dx.plos.org/10.1371/journal.pcbi.0030017`, doi:`10.1371/journal.pcbi.0030017`.

Achard, S., Delon-Martin, C., Vértes, P.E., Renard, F., Schenck, M., Schneider, F., Heinrich, C., Kremer, S., Bullmore, E.T., 2012. Hubs of brain functional networks are radically reorganized in comatose patients. Proceedings of the National Academy of Sciences , 201208933URL: `http://www.pnas.org/content/early/2012/11/21/1208933109`, doi:`10.1073/pnas.1208933109`. PMID: 23185007.

Achard, S., Salvador, R., Whitcher, B., Suckling, J., Bullmore, E., 2006. A resilient, low-frequency, small-world human brain functional network with highly connected association cortical hubs. The Journal of Neuroscience 26, 63–72. URL: `http://www.jneurosci.org/content/26/1/63`, doi:`10.1523/JNEUROSCI.3874-05.2006`. 00831.

Region	e_{g_i} (p-value)	e_{l_i} (p-value)	bc_i (p-value)	Region	e_{g_i} (p-value)	e_{l_i} (p-value)	bc_i (p-value)
CingMid L	**0.67(0.014)**	**0.74**(0.054)	**232.41**(0.143)	FrontInfTri R	0.52(0.020)	0.74(0.122)	59.89(0.334)
CingMid R	0.64(0.184)	0.76(0.065)	179.12(0.250)	ParacentralLob L	0.52(0.022)	0.86(0.399)	28.01(0.140)
TempMid R	0.64(0.078)	0.75(0.048)	169.19(0.214)	ParacentralLob R	0.52(0.011)	0.83(0.063)	25.22(0.184)
PrecGy L	**0.63(0.009)**	**0.80**(0.131)	99.23(0.087)	CingAnt L	0.52(0.076)	0.75(0.284)	64.58(0.505)
TempInf R	0.63(0.099)	0.79(0.255)	130.84(0.134)	Angular R	0.52(0.064)	0.77(0.127)	38.77(0.435)
Precuneus L	**0.63(0.003)**	0.76(0.191)	**153.70**(0.087)	FrontInfTri L	0.51(0.034)	0.72(0.129)	49.42(0.149)
TempMid L	**0.62(0.050)**	0.75(0.031)	**151.43**(0.095)	Heschl L	0.51(0.009)	0.85(0.155)	14.31(0.172)
Fusiform L	0.62(0.136)	0.82(0.089)	99.07(0.230)	FrontSupMed R	0.51(0.009)	0.75(0.368)	74.32(0.288)
ParietalSup L	**0.62(0.019)**	**0.83**(0.150)	60.01(0.122)	FrontInfOperc R	0.50(0.004)	0.80(0.224)	29.02(0.131)
TempSup L	**0.62(0.005)**	**0.80**(0.142)	83.81(0.238)	Cereb I_II R	0.49(0.003)	0.83(0.158)	64.14(0.240)
PrecGy R	**0.62(0.012)**	**0.82**(0.025)	74.67(0.153)	CingAnt R	0.49(0.010)	0.79(0.516)	27.89(0.276)
TempSup R	**0.62(0.034)**	**0.80**(0.117)	**103.94**(0.262)	TempPole R	0.49(0.011)	0.74(0.239)	55.78(0.237)
Occipital L	**0.61(0.024)**	**0.84**(0.021)	62.49(0.047)	Cereb I_II L	0.49(0.003)	0.81(0.068)	61.48(0.162)
Calcarine L	**0.61(0.019)**	**0.84**(0.023)	**75.93**(0.081)	ParietalInf R	0.49(0.013)	0.80(0.125)	18.93(0.636)
FronMid L	0.61(0.095)	0.73(0.089)	146.52(0.121)	Angular L	0.48(0.032)	0.75(0.358)	31.14(0.154)
FronSup R	0.61(0.121)	0.75(0.327)	125.63(0.283)	TempPole L	0.48(0.012)	0.65(0.111)	60.95(0.368)
Lingual R	**0.61(0.028)**	**0.86**(0.027)	50.15(0.259)	FrontInfOrb R	0.47(0.070)	0.68(0.012)	51.94(0.517)
Precuneus R	**0.61(0.045)**	0.79(0.694)	**100.46**(0.266)	Vermis	0.47(0.004)	0.74(0.092)	36.74(0.192)
Lingual L	0.61(0.057)	0.86(0.123)	44.19(0.316)	FrontInfOperc L	0.47(0.039)	0.80(0.458)	22.13(0.302)
PostcGy L	0.61(0.125)	0.85(0.262)	48.53(0.145)	FrontInfOrb L	0.47(0.099)	0.69(0.057)	42.94(0.234)
Fusiform R	**0.61(0.031)**	**0.82**(0.025)	65.66(0.131)	CingPost L	0.46(0.087)	0.73(0.104)	35.04(0.276)
SMA L	**0.61(0.010)**	**0.82**(0.300)	71.60(0.182)	Caudate L	0.44(0.029)	0.59(0.073)	77.52(0.760)
PostcGy R	0.60(0.093)	0.84(0.039)	57.24(0.101)	Caudate R	0.44(0.016)	0.59(0.093)	85.00(0.082)
Cuneus L	**0.60(0.033)**	**0.87**(0.054)	39.14(0.072)	ParaHippoc L	0.43(0.039)	0.49(0.047)	60.10(0.179)
FrontMid R	**0.60(0.050)**	0.75(0.138)	**129.74**(0.087)	CingPost R	0.43(0.007)	0.65(0.068)	16.84(0.215)
Calcarine R	**0.59(0.013)**	**0.86**(0.015)	43.95(0.139)	FrontMedOrb L	0.43(0.093)	0.55(0.115)	65.49(0.270)
Occipital R	**0.59(0.007)**	**0.87**(0.016)	35.87(0.083)	FrontMidOrb L	0.43(0.008)	0.54(0.037)	42.26(0.295)
RolandOperc R	**0.59(0.050)**	**0.83**(0.132)	51.07(0.202)	Thalamus R	0.42(0.326)	0.55(0.428)	20.92(0.228)
FronSup L	0.59(0.104)	0.75(0.180)	111.42(0.272)	Hippocampus L	0.42(0.044)	0.46(0.194)	86.15(0.190)
ParietalSup R	**0.59(0.006)**	**0.85**(0.323)	33.00(0.038)	FrontMedOrb R	0.41(0.078)	0.57(0.107)	59.63(0.298)
Cuneus R	**0.59(0.052)**	**0.88**(0.024)	26.84(0.101)	Thalamus L	0.41(0.007)	0.55(0.182)	27.27(0.074)
TempInf L	**0.59(0.028)**	0.78(0.644)	82.00(0.160)	ParaHippoc R	0.41(0.169)	0.47(0.251)	35.96(0.343)
SMA R	**0.58(0.009)**	**0.82**(0.705)	76.42(0.521)	FrontalMidOrb R	0.40(0.036)	0.53(0.165)	33.49(0.474)
Cereb VII - X R	**0.58(0.011)**	0.78(0.050)	**127.37**(0.009)	FrontSupOrb L	0.40(0.074)	0.53(0.042)	21.50(0.101)
Insula R	0.58(0.160)	0.79(0.118)	87.87(0.540)	Putamen R	0.39(0.078)	0.37(0.113)	106.10(0.391)
RolandOperc L	**0.58(0.017)**	**0.85**(0.245)	38.44(0.225)	FrontSupOrb R	0.37(0.013)	0.44(0.090)	17.55(0.212)
SupraMarginal R	0.57(0.002)	0.80(0.138)	49.30(0.023)	Hippocampus R	0.37(0.144)	0.34(0.186)	30.14(0.085)
ParietalInf L	0.56(0.038)	0.77(0.169)	58.04(0.306)	Putamen L	0.37(0.012)	0.38(0.046)	70.07(0.152)
Cereb VII - X L	0.55(0.008)	0.80(0.097)	83.54(0.157)	Amygdala L	0.33(0.052)	0.07(0.072)	19.59(0.212)
SupraMarginal L	0.55(0.003)	0.84(0.092)	40.41(0.143)	Olfactory L	0.33(0.220)	0.09(0.036)	10.61(0.297)
Insula L	0.54(0.032)	0.83(0.532)	38.32(0.379)	Amygdala R	0.33(0.130)	0.08(0.057)	11.23(0.392)
FrontSupMed L	0.54(0.006)	0.74(0.139)	84.19(0.176)	Olfactory R	0.31(0.263)	0.04(0.081)	5.01(0.103)
Heschl R	0.54(0.104)	0.87(0.254)	31.25(0.199)	Pallidum L	0.30(0.075)	0.01(0.005)	0.96(0.010)
Cereb III - VI R	0.53(0.007)	0.78(0.144)	69.77(0.024)	Pallidum R	0.30(0.151)	0.11(0.178)	11.25(0.094)
Cereb III - VI L	0.53(0.004)	0.77(0.074)	85.33(0.310)				

Table 4: Regional network parameters and their p-values for each region in AAL89 parcellation at 20% cost, considering 40 subjects at 1200 trials. Key regions are defined as regions with high e_{g_i} and significant p-value. The threshold to consider a high e_{g_i} is fix to the percentile 65%, in this case, $e_{g_i} \geq 0.58$. Other regional network parameters and their respective p-values are also shown: local efficiency e_{l_i} and betweenness centrality b_i.

Agcaoglu, O., Miller, R., Mayer, A., Hugdahl, K., Calhoun, V., 2015. Lateralization of resting state networks and relationship to age and gender. NeuroImage 104, 310 – 325. URL: http://www.sciencedirect.com/science/article/pii/S1053811914007435, doi:http://dx.doi.org/10.1016/j.neuroimage.2014.09.001.

Alexander-Bloch, A., Lambiotte, R., Roberts, B., Giedd, J., Gogtay, N., Bullmore, E., 2012. The discovery of population differences in network community structure: New methods and applications to brain functional networks in schizophrenia. NeuroImage 59, 3889–3900. URL: http://www.sciencedirect.com/science/article/pii/S1053811911013164, doi:http://dx.doi.org/10.1016/j.neuroimage.2011.11.035.

Andellini, M., Cannatà, V., Gazzellini, S., Bernardi, B., Napolitano, A., 2015. Test-retest reliability of graph metrics of resting state {MRI} functional brain networks: A review. Journal of Neuroscience Methods 253, 183 – 192. URL: http://www.sciencedirect.com/science/article/pii/S0165027015002113, doi:http://dx.doi.org/10.1016/j.jneumeth.2015.05.020.

Betzel, R.F., Byrge, L., He, Y., Goñi, J., Zuo, X.N., Sporns, O., 2014. Changes in structural and functional connectivity among resting-state networks across the human lifespan. NeuroImage 102, Part 2, 345 – 357. URL: http://www.sciencedirect.com/science/article/pii/S1053811914006508, doi:http://dx.doi.org/10.1016/j.neuroimage.2014.07.067.

Birn, R.M., Molloy, E.K., Patriat, R., Parker, T., Meier, T.B., Kirk, G.R., Nair, V.A., Meyerand, M.E., Prabhakaran, V., 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. NeuroImage 83, 550–558. URL: http://www.sciencedirect.com/science/article/pii/S1053811913006010, doi:10.1016/j.neuroimage.2013.05.099.

Biswal, B., Zerrin Yetkin, F., Haughton, V.M., Hyde, J.S., 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar mri. Magnetic Resonance in Medicine 34, 537–541. URL: http://dx.doi.org/10.1002/mrm.1910340409, doi:10.1002/mrm.1910340409.

Bland, J.M., Altman, D.G., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310.

Boardman, T.J., 1974. Confidence intervals for variance components–a comparative monte carlo study. Biometrics 30, 251–262.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D., 2006. Complex networks: Structure and dynamics. Physics Reports 424, 175–308. URL: http://dx.doi.org/10.1016/j.physrep.2005.10.009, doi:10.1016/j.physrep.2005.10.009.

Braun, U., Plichta, M.M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., Mier, D., Mohnke, S., Heinz, A., Erk, S., Walter, H., Seiferth, N., Kirsch, P., Meyer-Lindenberg, A., 2012. Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. NeuroImage 59, 1404–1412. URL: http://www.sciencedirect.com/science/article/pii/S105381191100961X, doi:10.1016/j.neuroimage.2011.08.044.

Bullmore, E., Sporns, O., 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nature Reviews Neuroscience 10, 186–198. URL: http://www.nature.com/nrn/journal/v10/n3/full/nrn2575.html, doi:10.1038/nrn2575. 01680.

Button, K.S., Ioannidis, J.P.A., Mokrysz, C., Nosek, B.A., Flint, J., Robinson, E.S.J., Munafò, M.R., 2013. Power failure: why small sample size undermines the reliability of neuroscience. Nat Rev Neurosci 14, 365–376. URL: http://dx.doi.org/10.1038/nrn3475, doi:10.1038/nrn3475.

Cao, H., Plichta, M.M., Schäfer, A., Haddad, L., Grimm, O., Schneider, M., Esslinger, C., Kirsch, P., Meyer-Lindenberg, A., Tost, H., 2014. Test-retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. NeuroImage 84, 888–900. doi:10.1016/j.neuroimage.2013.09.013. PMID: 24055506.

Cicchetti, D.V., 1994. Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological assessment 6, 284.

Craddock, R.C., James, G.A., Holtzheimer, P.E., Hu, X.P., Mayberg, H.S., 2012. A whole brain fmri atlas generated via spatially constrained spectral clustering. Human brain mapping 33, 1914–1928. URL: http://europepmc.org/articles/PMC3838923, doi:10.1002/hbm.21333.

De Vico Fallani, F., Richiardi, J., Chavez, M., Achard, S., 2014. Graph analysis of functional brain networks: practical issues in translational neuroscience. Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences 369. doi:10.1098/rstb.2013.0521.

Desikan, R.S., Ségonne, F., Fischl, B., Quinn, B.T., Dickerson, B.C., Blacker, D., Buckner, R.L., Dale, A.M., Maguire, R.P., Hyman, B.T., Albert, M.S., Killiany, R.J., 2006. An automated labeling system for subdividing the human cerebral cortex on {MRI} scans into gyral based regions of interest. NeuroImage 31, 968 – 980. URL: http://www.sciencedirect.com/science/article/pii/S1053811906000437, doi:http://dx.doi.org/10.1016/j.neuroimage.2006.01.021.

Diedrichsen, J., Balsters, J.H., Flavell, J., Cussans, E., Ramnani, N., 2009. A probabilistic {MR} atlas of the human cerebellum. NeuroImage 46, 39 – 46. URL: http://www.sciencedirect.

com/science/article/pii/S1053811909000809, doi:http://dx.doi.org/10.1016/j.neuroimage.2009.01.045.

Donner, A., 1986. A review of inference procedures for the intraclass correlation coefficient in the one-way random effects model. International Statistical Review/Revue Internationale de Statistique 54, 67–82.

Donner, A., Wells, G., 1986. A comparison of confidence interval methods for the intraclass correlation coefficient. Biometrics 42, 401–412.

Du, H.X., Liao, X.H., Lin, Q.X., Li, G.S., Chi, Y.Z., Liu, X., Yang, H.Z., Wang, Y., Xia, M.R., 2015. Test-retest reliability of graph metrics in high-resolution functional connectomics: A resting-state functional mri study. CNS Neurosci Ther URL: http://dx.doi.org/10.1111/cns.12431, doi:10.1111/cns.12431.

Duvernoy, H., 1995. The Human Brain Stem and Cerebellum: Surface, Structure, Vascularization, and Three-dimensional Sectional Anatomy with MRI. Spring-Verlag. URL: https://books.google.fr/books?id=cRlkQgAACAAJ.

Feinberg, D.A., Moeller, S., Smith, S.M., Auerbach, E., Ramanna, S., Glasser, M.F., Miller, K.L., Ugurbil, K., Yacoub, E., 2010. Multiplexed echo planar imaging for sub-second whole brain fmri and fast diffusion imaging. PLoS ONE 5, e15710. URL: http://dx.doi.org/10.1371%2Fjournal.pone.0015710, doi:10.1371/journal.pone.0015710.

Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A., Feng, D., 2013. Quantifying temporal correlations: A test-retest evaluation of functional connectivity in resting-state fMRI. NeuroImage 65, 231–241. URL: http://www.sciencedirect.com/science/article/pii/S1053811912009627, doi:10.1016/j.neuroimage.2012.09.052. 00008.

Filippini, N., MacIntosh, B.J., Hough, M.G., Goodwin, G.M., Frisoni, G.B., Smith, S.M., Matthews, P.M., Beckmann, C.F., Mackay, C.E., 2009. Distinct patterns of brain activity in young carriers of the apoe-epsilon4 allele. Proceedings of the National Academy of Sciences of the United States of America 106, 7209–7214. URL: http://europepmc.org/articles/PMC2678478, doi:10.1073/pnas.0811879106.

Fisher, R.A., 1925. Theory of Statistical Estimation. Mathematical Proceedings of the Cambridge Philosophical Society 22, 700–725. doi:10.1017/S0305004100009580.

Fornito, A., Zalesky, A., Bullmore, E.T., 2010. Network scaling effects in graph analytic studies of human resting-state fmri data. Frontiers in Systems Neuroscience 4. URL: http://www.frontiersin.org/systems_neuroscience/10.3389/fnsys.2010.00022/abstract, doi:10.3389/fnsys.2010.00022.

Fox, M.D., Raichle, M.E., 2007. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. Nat Rev Neurosci 8, 700–711. URL: http://dx.doi.org/10.1038/nrn2201, doi:10.1038/nrn2201.

Gandy, A., 2009. Sequential implementation of monte carlo tests with uniformly bounded resampling risk. Journal of the American Statistical Association 104, 1504–1511. URL: http://dx.doi.org/10.1198/jasa.2009.tm08368, doi:10.1198/jasa.2009.tm08368.

Gandy, A., Hahn, G., 2014. A Safe Algorithm for Implementing Multiple Monte Carlo Tests. Scandinavian Journal of Statistics 41, 1083–1101. URL: http://onlinelibrary.wiley.com.gate2.inist.fr/doi/10.1111/sjos.12085/abstract, doi:10.1111/sjos.12085.

Glasser, M.F., Sotiropoulos, S.N., Wilson, J.A., Coalson, T.S., Fischl, B., Andersson, J.L., Xu, J., Jbabdi, S., Webster, M., Polimeni, J.R., Van Essen, D.C., Jenkinson, M., 2013. The minimal preprocessing pipelines for the human connectome project. NeuroImage 80, 105–124. URL: http://www.sciencedirect.com/science/article/pii/S1053811913005053, doi:10.1016/j.neuroimage.2013.04.127.

Grahn, J.A., Parkinson, J.A., Owen, A.M., 2009. The role of the basal ganglia in learning and memory: Neuropsychological studies. Behavioural Brain Research 199, 53 – 60. URL: http://www.sciencedirect.com/science/article/pii/S016643280800630X, doi:http://dx.doi.org/10.1016/j.bbr.2008.11.020. special issue on the role of the basal ganglia in learning and memory.

Guo, C.C., Kurth, F., Zhou, J., Mayer, E.A., Eickhoff, S.B., Kramer, J.H., Seeley, W.W., 2012. One-year test-retest reliability of intrinsic connectivity network fMRI in older adults. NeuroImage 61, 1471–1483. URL: http://www.sciencedirect.com/science/article/pii/S105381191200300X, doi:10.1016/j.neuroimage.2012.03.027. 00034.

Haacke, E.M., Cheng, N.Y., House, M.J., Liu, Q., Neelavalli, J., Ogg, R.J., Khan, A., Ayaz, M., Kirsch, W., Obenaus, A., 2005. Imaging iron stores in the brain using magnetic resonance imaging. Magnetic Resonance Imaging 23, 1 – 25. URL: http://www.sciencedirect.com/science/article/pii/S0730725X04002905, doi:http://dx.doi.org/10.1016/j.mri.2004.10.001.

Helms, G., Draganski, B., Frackowiak, R., Ashburner, J., Weiskopf, N., 2009. Improved segmentation of deep brain grey matter structures using magnetization transfer (mt) parameter maps. NeuroImage 47, 194 – 198. URL: http://www.sciencedirect.com/science/article/pii/S105381190900295X, doi:http://dx.doi.org/10.1016/j.neuroimage.2009.03.053.

Ioannidis, J.P.A., 2014. How to make more published research true. PLoS Med 11, e1001747. URL: http://dx.doi.org/10.1371/journal.pmed.1001747, doi:10.1371/journal.pmed.1001747.

Ionan, A.C., Polley, M.Y.C., McShane, L.M., Dobbin, K.K., 2014. Comparison of confidence interval methods for an intra-class correlation coefficient (icc). BMC medical research methodology 14, 121. URL: http://europepmc.org/articles/PMC4258044, doi:10.1186/1471-2288-14-121.

Liang, X., Wang, J., Yan, C., Shu, N., Xu, K., Gong, G., He, Y., 2012. Effects of different correlation metrics and preprocessing factors on small-world brain functional networks: A resting-state functional MRI study. PLoS ONE 7, e32766. URL: http://dx.doi.org/10.1371/journal.pone.0032766, doi:10.1371/journal.pone.0032766. 00032.

Liao, X.H., Xia, M.R., Xu, T., Dai, Z.J., Cao, X.Y., Niu, H.J., Zuo, X.N., Zang, Y.F., He, Y., 2013. Functional brain hubs and their test-retest reliability: a multiband resting-state functional MRI study. NeuroImage 83, 969–982. URL: http://www.sciencedirect.com/science/article/pii/S1053811913008276, doi:10.1016/j.neuroimage.2013.07.058. 00007.

Malherbe, C., Messe, A., Bardinet, E., Pelegrini-Issac, M., Perlbarg, V., Marrelec, G., Worbe, Y., Yelnik, J., Lehericy, S., Benali, H., 2014. Combining spatial independent component analysis with regression to identify the subcortical components of resting-state FMRI functional networks. Brain Connect 4, 181–192.

McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychological Methods 1, 30–46. doi:10.1037/1082-989X.1.1.30.

Metropolis, N., Ulam, S., 1949. The Monte Carlo Method. J. Amer. Stat. Assoc. 44, 335–341.

Middleton, F.A., Strick, P.L., 2000. Basal ganglia output and cognition: Evidence from anatomical, behavioral, and clinical studies. Brain and Cognition 42, 183 – 200. URL: http://www.sciencedirect.com/science/article/pii/S0278262699910990, doi:http://dx.doi.org/10.1006/brcg.1999.1099.

Müller, R., Büttner, P., 1994. A critical discussion of intraclass correlation coefficients. Statistics in Medicine 13, 2465–2476. URL: http://onlinelibrary.wiley.com.gate2.inist.fr/doi/10.1002/sim.4780132310/abstract, doi:10.1002/sim.4780132310. 00360.

Murphy, K., Birn, R.M., Handwerker, D.A., Jones, T.B., Bandettini, P.A., 2009. The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? NeuroImage 44, 893–905. URL: http://dx.doi.org/10.1016/j.neuroimage.2008.09.036, doi:10.1016/j.neuroimage.2008.09.036.

Nakagawa, S., Schielzeth, H., 2010. Repeatability for gaussian and non-gaussian data: a practical guide for biologists. Biol Rev Camb Philos Soc 85, 935–956. URL: http://dx.doi.org/10.1111/j.1469-185X.2010.00141.x, doi:10.1111/j.1469-185X.2010.00141.x.

O'Reilly, J.X., Beckmann, C.F., Tomassini, V., Ramnani, N., Johansen-Berg, H., 2010. Distinct and overlapping functional zones in the cerebellum defined by resting state functional connectivity. Cerebral cortex (New York, N.Y. : 1991) 20, 953–965. URL: http://europepmc.org/articles/PMC2837094, doi:10.1093/cercor/bhp157.

Rubinov, M., Sporns, O., 2010. Complex network measures of brain connectivity: uses and interpretations. NeuroImage 52, 1059–1069. doi:10.1016/j.neuroimage.2009.10.003.

Saad, Z.S., Gotts, S.J., Murphy, K., Chen, G., Jo, H.J., Martin, A., Cox, R.W., 2012. Trouble at rest: how correlation patterns and group differences become distorted after global signal regression. Brain connectivity 2, 25–32. URL: http://europepmc.org/articles/PMC3484684, doi:10.1089/brain.2012.0080.

Sampat, M.P., Whitman, G.J., Stephens, T.W., Broemeling, L.D., Heger, N.A., Bovik, A.C., Markey, M.K., 2006. The reliability of measuring physical characteristics of spiculated masses on mammography. The British Journal of Radiology 79 Spec No 2, S134–140. doi:10.1259/bjr/96723280.

Satterthwaite, T.D., Wolf, D.H., Roalf, D.R., Ruparel, K., Erus, G., Vandekar, S., Gennatas, E.D., Elliott, M.A., Smith, A., Hakonarson, H., Verma, R., Davatzikos, C., Gur, R.E., Gur, R.C., 2015. Linked sex differences in cognition and functional connectivity in youth. Cereb Cortex 25, 2383–2394. URL: http://dx.doi.org/10.1093/cercor/bhu036, doi:10.1093/cercor/bhu036.

Schmahmann, J.D., Doyon, J., McDonald, D., Holmes, C., Lavoie, K., Hurwitz, A.S., Kabani, N., Toga, A., Evans, A., Petrides, M., 1999. Three-dimensional {MRI} atlas of the human cerebellum in proportional stereotaxic space. NeuroImage 10, 233 – 260. URL: http://www.sciencedirect.com/science/article/pii/S1053811999904597, doi:http://dx.doi.org/10.1006/nimg.1999.0459.

Schwarz, A.J., McGonigle, J., 2011. Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. NeuroImage 55, 1132–1146. URL: http://www.sciencedirect.com/science/article/pii/S1053811910016332, doi:10.1016/j.neuroimage.2010.12.047. 00047.

Shirer, W.R., Jiang, H., Price, C.M., Ng, B., Greicius, M.D., 2015. Optimization of rs-fmri pre-processing for enhanced signal-noise separation, test-retest reliability, and group discrimination. NeuroImage 117, 67 – 79. URL: http://www.sciencedirect.com/

science/article/pii/S105381191500395X, doi:http://dx.doi.org/10.1016/j.neuroimage.2015.05.015.

Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychological Bulletin 86, 420–428.

Smith, S.M., Beckmann, C.F., Andersson, J., Auerbach, E.J., Bijsterbosch, J., Douaud, G., Duff, E., Feinberg, D.A., Griffanti, L., Harms, M.P., Kelly, M., Laumann, T., Miller, K.L., Moeller, S., Petersen, S., Power, J., Salimi-Khorshidi, G., Snyder, A.Z., Vu, A.T., Woolrich, M.W., Xu, J., Yacoub, E., Uğurbil, K., Van Essen, D.C., Glasser, M.F., 2013. Resting-state fMRI in the human connectome project. NeuroImage 80, 144–168. URL: http://www.sciencedirect.com/science/article/pii/S1053811913005338, doi:10.1016/j.neuroimage.2013.05.039.

Song, J., Desphande, A.S., Meier, T.B., Tudorascu, D.L., Vergun, S., Nair, V.A., Biswal, B.B., Meyerand, M.E., Birn, R.M., Bellec, P., Prabhakaran, V., 2012. Age-related differences in test-retest reliability in resting-state brain functional connectivity. PLoS ONE 7, e49847. URL: http://dx.doi.org/10.1371/journal.pone.0049847, doi:10.1371/journal.pone.0049847. 00014.

Stoodley, C.J., Schmahmann, J.D., 2009. Functional topography in the human cerebellum: A meta-analysis of neuroimaging studies. NeuroImage 44, 489 – 501. URL: http://www.sciencedirect.com/science/article/pii/S1053811908009725, doi:http://dx.doi.org/10.1016/j.neuroimage.2008.08.039.

Stoodley, C.J., Valera, E.M., Schmahmann, J.D., 2012. Functional topography of the cerebellum for motor and cognitive tasks: An fmri study. NeuroImage 59, 1560 – 1570. URL: http://www.sciencedirect.com/science/article/pii/S1053811911009827, doi:http://dx.doi.org/10.1016/j.neuroimage.2011.08.065.

Thirion, B., Pinel, P., Mériaux, S., Roche, A., Dehaene, S., Poline, J.B., 2007. Analysis of a large fMRI cohort: Statistical and methodological issues for group analyses. NeuroImage 35, 105–120. doi:10.1016/j.neuroimage.2006.11.054.

Tomasi, D., Shokri-Kojori, E., Volkow, N.D., 2015. High-resolution functional connectivity density: Hub locations, sensitivity, specificity, reproducibility, and reliability. Cerebral Cortex URL: http://cercor.oxfordjournals.org/content/early/2015/07/28/cercor.bhv171.abstract, doi:10.1093/cercor/bhv171.

Tomasi, D., Volkow, N.D., 2012. Laterality patterns of brain functional connectivity: gender effects. Cereb Cortex 22, 1455–1462. URL: http://dx.doi.org/10.1093/cercor/bhr230, doi:10.1093/cercor/bhr230.

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., Mazoyer, B., Joliot, M., 2002. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage 15, 273–289. doi:10.1006/nimg.2001.0978.

Van Dijk, K.R.A., Hedden, T., Venkataraman, A., Evans, K.C., Lazar, S.W., Buckner, R.L., 2010. Intrinsic functional connectivity as a tool for human connectomics: theory, properties, and optimization. J Neurophysiol 103, 297–321. URL: http://dx.doi.org/10.1152/jn.00783.2009, doi:10.1152/jn.00783.2009.

Wang, J.H., Zuo, X.N., Gohel, S., Milham, M.P., Biswal, B.B., He, Y., 2011. Graph theoretical analysis of functional brain networks: Test-retest evaluation on short- and long-term resting-state functional MRI data. PLoS ONE 6, e21976. URL: http://dx.doi.org/10.1371/journal.pone.0021976, doi:10.1371/journal.pone.0021976. 00069.

Welton, T., Kent, D., Auer, D., Dineen, R., 2015. Reproducibility of graph-theoretic brain network metrics: A systematic review. Brain Connectivity 5, 193–202. doi:10.1089/brain.2014.0313.

Whitlow, C.T., Casanova, R., Maldjian, J.A., 2011. Effect of resting-state functional MR imaging duration on stability of graph theory metrics of brain network connectivity. Radiology 259, 516–524. doi:10.1148/radiol.11101708. 00020 PMID: 21406628.

Yin, D., Song, F., Xu, D., Sun, L., Men, W., Zang, L., Yan, X., Fan, M., 2014. Altered topological properties of the cortical motor-related network in patients with subcortical stroke revealed by graph theoretical analysis. Human Brain Mapping 35, 3343–3359. URL: http://dx.doi.org/10.1002/hbm.22406, doi:10.1002/hbm.22406.

Zalesky, A., Fornito, A., Harding, I.H., Cocchi, L., Yücel, M., Pantelis, C., Bullmore, E.T., 2010. Whole-brain anatomical networks: Does the choice of nodes matter? NeuroImage 50, 970 – 983. URL: http://www.sciencedirect.com/science/article/pii/S1053811909013159, doi:http://dx.doi.org/10.1016/j.neuroimage.2009.12.027.

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 and biobehavioral reviews 45, 100–118. doi:10.1016/j.neubiorev.2014.05.009, pmid:24875392.

Region	e_{g_i} (#40)	#20	#30	#40	#100	Region	e_{g_i} (#40)	#20	#30	#40	#100
PrecGy L	0.63	0.044	0.018	0.009	0.004	Lingual L	0.61	0.152	0.080	0.057	0.004
PrecGy R	0.62	0.070	0.039	0.012	0.004	Lingual R	0.61	0.099	0.046	0.028	0.004
FontSup L	0.59	0.162	0.122	0.104	0.004	Occipital L	0.61	0.078	0.043	0.024	0.004
FontSup R	0.61	0.218	0.144	0.121	0.009	Occipital R	0.59	0.046	0.015	0.007	0.004
FrontSupOrb L	0.40	0.182	0.122	0.074	0.004	Fusiform L	0.62	0.277	0.223	0.136	0.021
FrontSupOrb R	0.37	0.063	0.029	0.013	0.004	Fusiform R	0.61	0.114	0.059	0.031	0.004
FrontMid L	0.61	0.222	0.146	0.095	0.009	Postcentral L	0.61	0.195	0.142	0.125	0.018
FrontMid R	0.60	0.148	0.073	0.050	0.004	Postcentral R	0.60	0.180	0.113	0.093	0.004
FrontMidOrb L	0.43	0.057	0.017	0.008	0.004	ParietalSup L	0.62	0.066	0.035	0.019	0.004
FrontMidOrb R	0.40	0.099	0.049	0.036	0.004	ParietalSup R	0.59	0.039	0.015	0.006	0.004
FrontInfOperc L	0.47	0.153	0.083	0.039	0.004	ParietalInf L	0.56	0.151	0.090	0.038	0.004
FrontInfOperc R	0.50	0.048	0.012	0.004	0.004	ParietalInf R	0.49	0.087	0.029	0.013	0.004
FrontInfTri L	0.51	0.113	0.054	0.034	0.004	SupraMarginal L	0.55	0.020	0.005	0.003	0.004
FrontInfTri R	0.52	0.105	0.056	0.020	0.004	SupraMarginal R	0.57	0.011	0.003	0.002	0.004
FrontInfOrb L	0.46	0.203	0.141	0.099	0.009	Angular L	0.48	0.072	0.036	0.032	0.004
FrontInfOrb R	0.47	0.126	0.079	0.070	0.004	Angular R	0.51	0.158	0.100	0.064	0.004
RolandOperc L	0.57	0.057	0.022	0.017	0.004	Precuneus L	0.63	0.021	0.006	0.003	0.004
RolandOperc R	0.59	0.131	0.071	0.050	0.004	Precuneus R	0.61	0.134	0.083	0.045	0.004
SMA L	0.60	0.058	0.024	0.010	0.004	ParacentralLob L	0.52	0.079	0.028	0.022	0.004
SMA R	0.58	0.098	0.030	0.009	0.004	ParacentralLob R	0.52	0.047	0.023	0.011	0.004
Olfactory L	0.33	0.290	0.265	0.220	0.028	Caudate L	0.44	0.110	0.057	0.029	0.004
Olfactory R	0.31	0.325	0.291	0.263	0.118	Caudate R	0.44	0.074	0.029	0.016	0.004
FrontSupMed L	0.54	0.035	0.010	0.006	0.004	Putamen L	0.37	0.068	0.027	0.012	0.004
FrontSupMed R	0.51	0.044	0.014	0.009	0.004	Putamen R	0.39	0.156	0.116	0.078	0.004
FrontMedOrb L	0.43	0.186	0.149	0.093	0.004	Pallidum L	0.30	0.192	0.120	0.074	0.004
FrontMedOrb R	0.41	0.127	0.088	0.078	0.004	Pallidum R	0.30	0.245	0.202	0.151	0.011
Insula L	0.54	0.119	0.084	0.032	0.004	Thalamus L	0.41	0.038	0.014	0.007	0.004
Insula R	0.58	0.220	0.166	0.160	0.014	Thalamus R	0.42	0.347	0.341	0.326	0.160
CingAnt L	0.52	0.183	0.130	0.076	0.004	Heschl L	0.51	0.050	0.016	0.009	0.004
CingAnt R	0.49	0.073	0.026	0.010	0.004	Heschl R	0.53	0.158	0.092	0.104	0.004
CingMid L	0.67	0.067	0.024	0.014	0.004	TempSup L	0.62	0.029	0.009	0.005	0.004
CingMid R	0.64	0.241	0.188	0.184	0.028	TempSup R	0.62	0.098	0.060	0.034	0.004
CingPost L	0.46	0.136	0.101	0.087	0.004	TempPole L	0.47	0.118	0.067	0.012	0.004
CingPost R	0.43	0.047	0.017	0.007	0.004	TempPole R	0.49	0.068	0.023	0.011	0.004
Hippocampus L	0.42	0.113	0.059	0.044	0.004	TempMid L	0.62	0.145	0.072	0.050	0.004
Hippocampus R	0.37	0.243	0.171	0.144	0.014	TempMid R	0.64	0.222	0.130	0.078	0.009
ParaHippoc L	0.43	0.117	0.060	0.039	0.004	TempInf L	0.58	0.112	0.057	0.028	0.004
ParaHippoc R	0.41	0.283	0.213	0.169	0.018	TempInf R	0.63	0.184	0.124	0.099	0.004
Amygdala L	0.33	0.126	0.059	0.052	0.004	Cereb I II L	0.49	0.009	0.003	0.003	0.004
Amygdala R	0.32	0.196	0.168	0.130	0.018	Cereb I II R	0.49	0.014	0.003	0.003	0.004
Calcarine L	0.61	0.078	0.026	0.019	0.004	Cereb III - VI L	0.53	0.021	0.005	0.004	0.004
Calcarine R	0.59	0.071	0.028	0.013	0.004	Cereb III - VI R	0.53	0.047	0.013	0.007	0.004
Cuneus L	0.60	0.105	0.050	0.033	0.004	Cereb VII - X L	0.55	0.035	0.008	0.008	0.004
Cuneus R	0.59	0.138	0.079	0.052	0.004	Cereb VII - X R	0.58	0.038	0.015	0.011	0.004
						Vermis	0.47	0.036	0.010	0.004	0.004

Table 5: *P*-values for each brain region of the AAL89 for 20, 30, 40 and 100 subjects. Graph metric: regional global efficiency (e_{g_i}). Cost of 20% at 1200 time points.

Region	e_{g_i} (p-value)	e_{l_i} (p-value)	b_i (p-value)	Region	e_{g_i} (p-value)	e_{l_i} (p-value)	b_i (p-value)
CingMid L	**0.67(0.005)**	0.74(0.002)	**232.75**(0.005)	ParacentralLob L	0.52(0.005)	0.86(0.140)	28.06(0.050)
CingMid R	**0.64(0.028)**	0.76(0.002)	**180.31**(0.026)	FrontalInfTri R	0.52(0.005)	0.74(0.023)	61.55(0.273)
TempMid R	**0.64(0.009)**	0.75(0.002)	**168.89**(0.064)	ParacentralLob R	0.52(0.005)	0.83(0.002)	25.84(0.026)
Precentral L	**0.63(0.005)**	0.79(0.004)	99.96(0.003)	CingAnt L	0.52(0.005)	0.75(0.062)	63.55(0.607)
TempInf R	**0.63(0.005)**	0.79(0.124)	**129.30**(0.017)	Angular R	0.51(0.005)	0.77(0.034)	38.34(0.413)
Precuneus L	**0.63(0.005)**	0.76(0.032)	**154.14**(0.004)	FrontInfTri L	0.51(0.005)	0.72(0.023)	49.18(0.038)
Fusiform L	**0.62(0.021)**	**0.82**(0.007)	98.74(0.036)	Heschl L	0.51(0.005)	0.85(0.011)	14.26(0.084)
TempSup L	**0.62(0.005)**	**0.80**(0.013)	83.03(0.051)	FrontSupMed R	0.51(0.005)	0.74(0.052)	76.78(0.033)
ParietalSup L	**0.62(0.005)**	**0.83**(0.017)	58.00(0.006)	FrontInfOperc R	0.50(0.005)	0.80(0.043)	29.97(0.028)
TempMid L	**0.62(0.005)**	0.75(0.004)	**149.03**(0.002)	Cereb I II R	0.49(0.005)	0.83(0.011)	63.00(0.114)
TempSup R	**0.62(0.005)**	**0.80**(0.002)	**104.30**(0.102)	Cereb I II L	0.49(0.005)	0.81(0.002)	63.90(0.136)
Precentral R	**0.62(0.005)**	**0.82**(0.002)	76.88(0.026)	CingAnt R	0.49(0.005)	0.79(0.457)	28.05(0.219)
FrontMid L	**0.61(0.009)**	0.73(0.002)	**150.74**(0.005)	ParietalInf R	0.49(0.005)	0.80(0.007)	19.00(0.679)
Occipital L	**0.61(0.005)**	**0.84**(0.002)	60.29(0.002)	TempPoleSupMid R	0.49(0.005)	0.73(0.029)	57.68(0.147)
Calcarine L	**0.61(0.005)**	**0.84**(0.002)	74.92(0.001)	Angular L	0.48(0.005)	0.75(0.140)	31.19(0.021)
FrontSup R	**0.61(0.009)**	0.75(0.284)	**127.58**(0.142)	TempPoleSupMid L	0.48(0.005)	0.65(0.007)	63.09(0.353)
Precuneus R	**0.61(0.005)**	0.78(0.846)	**103.00**(0.037)	Vermis	0.47(0.005)	0.74(0.004)	38.09(0.060)
Lingual R	**0.61(0.005)**	**0.86**(0.002)	49.49(0.112)	FrontInfOrb R	0.47(0.005)	0.68(0.002)	52.25(0.412)
Postcentral L	**0.61(0.018)**	**0.85**(0.047)	47.79(0.021)	FrontInfOper L	0.47(0.005)	0.79(0.413)	22.07(0.284)
Fusiform R	**0.61(0.005)**	**0.82**(0.002)	66.48(0.028)	FrontInfOrb L	0.46(0.009)	0.68(0.002)	43.92(0.110)
SMA L	**0.61(0.005)**	**0.82**(0.295)	72.68(0.044)	CingulumPost L	0.46(0.005)	0.73(0.004)	35.81(0.081)
Lingual L	**0.61(0.005)**	**0.86**(0.011)	44.14(0.258)	Caudate R	0.44(0.005)	0.59(0.007)	87.00(0.002)
Postcentral R	**0.60(0.005)**	**0.84**(0.004)	56.75(0.001)	Caudate L	0.44(0.005)	0.59(0.004)	76.01(0.857)
Cuneus L	**0.60(0.005)**	**0.87**(0.002)	38.57(0.002)	CingulumPost R	0.43(0.005)	0.66(0.002)	16.93(0.069)
FrontMid R	**0.60(0.005)**	0.75(0.002)	**128.87**(0.005)	ParaHippocampal L	0.43(0.005)	0.50(0.002)	60.74(0.008)
Occipital R	**0.59(0.005)**	**0.87**(0.002)	35.48(0.003)	FrontalMidOrb L	0.43(0.005)	0.54(0.004)	44.60(0.080)
Calcarine R	**0.59(0.005)**	**0.86**(0.002)	42.86(0.011)	Thalamus R	0.42(0.160)	0.54(0.500)	20.98(0.104)
RolandicOper R	**0.59(0.005)**	**0.83**(0.007)	50.88(0.084)	FrontMedOrb Rectus L	0.42(0.005)	0.55(0.013)	64.40(0.148)
ParietalSup R	**0.59(0.005)**	**0.85**(0.368)	32.56(0.003)	Hippocampus L	0.42(0.005)	0.46(0.042)	87.67(0.037)
Cuneus R	**0.59(0.005)**	**0.88**(0.002)	26.68(0.013)	FrontMedOrb Rectus R	0.41(0.005)	0.57(0.004)	58.91(0.184)
FrontSup L	**0.59(0.005)**	0.75(0.029)	**110.37**(0.029)	Thalamus L	0.41(0.005)	0.54(0.033)	27.34(0.013)
TempInf L	**0.58(0.005)**	0.78(0.923)	81.56(0.024)	Parahippocampal R	0.41(0.018)	0.46(0.080)	36.05(0.219)
SMA R	**0.58(0.005)**	**0.82**(0.923)	75.50(0.571)	FrontMidOrb R	0.40(0.005)	0.52(0.027)	33.21(0.413)
Cereb VII - X R	**0.58(0.005)**	0.78(0.017)	**125.59**(0.001)	FrontSupOrb L	0.40(0.005)	0.53(0.002)	20.68(0.038)
RolandicOper L	**0.58(0.005)**	**0.85**(0.108)	38.65(0.097)	Putamen R	0.39(0.005)	0.37(0.007)	107.80(0.326)
Insula R	**0.58(0.014)**	0.79(0.032)	89.42(0.607)	FrontSupOrb R	0.37(0.005)	0.43(0.013)	17.18(0.068)
SupraMarginal R	0.57(0.005)	0.80(0.035)	48.93(0.001)	Hippocampus R	0.37(0.014)	0.33(0.025)	31.04(0.012)
ParietalInf L	0.56(0.005)	0.78(0.056)	58.13(0.103)	Putamen L	0.37(0.005)	0.38(0.002)	69.49(0.008)
SupraMarginal L	0.55(0.005)	0.84(0.011)	40.08(0.052)	Amygdala L	0.33(0.005)	0.07(0.002)	21.03(0.062)
Cereb VII - X L	0.55(0.005)	0.80(0.021)	84.82(0.053)	Olfactory L	0.33(0.028)	0.09(0.002)	11.09(0.441)
Insula L	0.54(0.005)	0.83(0.607)	38.01(0.337)	Amygdala R	0.32(0.018)	0.08(0.002)	11.35(0.536)
FrontSupMed L	0.54(0.005)	0.73(0.004)	83.51(0.024)	Olfactory R	0.31(0.118)	0.04(0.194)	4.38(0.153)
Heschl R	0.54(0.005)	0.87(0.127)	31.73(0.045)	Pallidum L	0.30(0.005)	0.01(0.004)	0.87(0.017)
Cereb III - VI R	0.53(0.005)	0.78(0.004)	70.97(0.001)	Pallidum R	0.30(0.011)	0.11(0.154)	12.45(0.084)
Cereb III - VI L	0.53(0.005)	0.77(0.002)	88.47(0.157)				

Table 6: P-values of each brain region (AAL89) for 100 subjects using different graph metrics: regional global efficiency (e_{g_i}), regional local efficiency (e_{l_i}) and regional betweenness centrality (b_i). Cost of 20% at 1200 time points.