BRAIN CONNECTIVITY
Volume 5, Number 4, 2015

DOI: 10.1089/brain.2014.0313

REVIEW ARTICLE

Reproducibility of Graph-Theoretic Brain Network Metrics: A Systematic Review

Thomas Welton,* Daniel A. Kent,* Dorothee P. Auer, and Robert A. Dineen

Abstract

This systematic review aimed to assess the reproducibility of graph-theoretic brain network metrics. Primary research studies of test-retest reliability conducted on healthy human subjects were included that quantified test-retest reliability using either the intraclass correlation coefficient (ICC) or the coefficient of variance. The MEDLINE, Web of Knowledge, Google Scholar, and OpenGrey databases were searched up to February 2014. Risk of bias was assessed with 10 criteria weighted toward methodological quality. Twenty-three studies were included in the review ($n = 499$ subjects) and evaluated for various characteristics, including sample size (5–45), retest interval (< 1 h to > 1 year), acquisition method, and test-retest reliability scores. For at least one metric, ICCs reached the fair range (ICC 0.40–0.59) in one study, the good range (ICC 0.60–0.74) in five studies, and the excellent range (ICC > 0.74) in 16 studies. Heterogeneity of methods prevented further quantitative analysis. Reproducibility was good overall. For the metrics having three or more ICCs reported for both functional and structural networks, six of seven were higher in structural networks, indicating that structural networks may be more reliable over time. The authors were also able to highlight and discuss a number of methodological factors affecting reproducibility.

Key words: connectome; graph theory; network; reliability; reproducibility; systematic review; test-retest

Introduction

GRAPH THEORY HAS recently been applied to brain imaging data and shows promise as an interpretable and generalizable way to model brain networks (Bullmore and Sporns, 2009; Park and Friston, 2013). In graph theory, a graph is a mathematical construct used to model the relationships between objects, in which the objects are called *vertices* and their interconnecting links are called *edges*. In terms of brain networks, regions of interest (ROIs) can be represented by the vertices in a graph, and some measure of connectivity between those ROIs can be represented by the edges. One of the main advantages to this model is that simple, numerical summary descriptors of graph organization can be derived, which describe the graph structure or topology in terms of the whole network (Rubinov and Sporns, 2010). The most common descriptors are *characteristic path length* (a measure of how easy it is to traverse the whole graph), *clustering coefficient* (a measure of local connectivity), and *small-worldness* (the state of being highly clustered, yet having a short average path (Watts and Strogatz, 1998)) considered to be a highly efficient structure (Latora and Marchiori, 2001). These metrics provide a way to characterize the underlying functional and structural brain networks and allow comparisons across time, subjects, or groups of subjects.

There has been a trend toward applying these techniques in studies of patient populations to investigate how, on the level of whole-brain networks, symptoms may emerge from the underlying neurological injury or psychopathology. Studies have demonstrated significant differences in metrics derived from graphs of brain networks between diseased and healthy groups as well as in normal development (Supekar et al., 2009), for example, in multiple sclerosis (He et al., 2009), Alzheimer's (Buckner et al., 2009; Stam et al., 2009), Parkinson's (Göttlich et al., 2013), epilepsy (Quraan et al., 2013), and body dysmorphic disorder (Arienzo et al., 2013) [for reviews see Bassett and Bullmore (2009); Menon (2011); Wang et al. (2010)], and have offered various interpretations of these findings. With this wave of positive results, some authors have suggested the use of graph metrics as surrogate markers in clinical trials (Petrella, 2011) and even suggested that they have potential as diagnostic tools (Quraan et al., 2013; Schoonheim et al., 2013). However, such applications

Sir Peter Mansfield Imaging Centre, University of Nottingham, Nottingham, United Kingdom.
*Joint first authors.

are dependent, alongside validity, on evidence of reliability and responsiveness to intervention.

Several recent studies have aimed to meet this need by measuring the test-retest reliability of graph metrics (Andreotti et al., 2014; Bassett et al., 2011; Braun et al., 2012; Buchanan et al., 2014; Cao et al., 2014; Cheng et al., 2012; Dennis et al., 2012; Deuker et al., 2009; Duda et al., 2014; Fan et al., 2012; Faria et al., 2012; Guo et al., 2012; Jin et al., 2011; Liang et al., 2012; Liao et al., 2013; Niu et al., 2013; Owen et al., 2013; Park et al., 2012; Parker et al., 2012; Schwarz and McGonigle, 2011; Telesford et al., 2010; Vaessen et al., 2010; Wang et al., 2011; Weber et al., 2013). To achieve this, graphs of brain networks derived from healthy volunteers at two or more time points were analyzed to determine their organizational properties, and the level of agreement between the measurements quantified using an intraclass correlation coefficient (ICC). Most studies employed a variation on this design; for example, a common secondary aim was to identify the data preprocessing and graph construction strategies, which resulted in the most reproducible graph metrics. While many of these studies concluded that graph metrics were reliable enough for wider application in future translational research, heterogeneity in their methods and quality and the occurrence of some conflicting results mean that no consensus view is apparent.

In this study, the authors aimed to systematically review and summarize the published literature describing the test-retest reliability of graph-theoretic brain network metrics. Specifically, the authors ask the following: (1) What is the test-retest reliability of graph metrics in brain networks? (2) Based on reliability data, which graph metrics show the greatest promise for translation into clinical neuroscience research? And (3) how do methodological factors in data analysis impact the test-retest reliability of graph metrics?

Materials and Methods

Search strategy

A systematic literature search was performed independently by two researchers (T.W. and D.A.K.) on the 9th of February, 2014, in the *MEDLINE* (www.ncbi.nlm.nih.gov/pubmed/), *Web of Knowledge* (http://wok.mimas.ac.uk/), *Google Scholar* (http://scholar.google.co.uk/), and *OpenGrey* (www.opengrey.eu/) databases. Based on keywords identified from the known literature, the authors used the following search string: (''graph theory'' OR ''graph theoretical'') AND (''TRT'' OR ''test-retest'' OR ''reproducibility''). The authors included all languages and dates in the search. For Google Scholar, results were sorted by relevance and only the top 100 were checked. The authors searched the reference lists of the included articles to identify any additional relevant articles.

In the first phase of screening, articles were identified that attempted to measure the test-retest reliability of summary graph metrics in human brain networks based on the title and abstract. In the second phase of screening, the authors excluded any articles that did not meet all of the following criteria: (1) to avoid the confounding effect of any disease process, the study must use data only from healthy human subjects; (2) to make simple comparisons between studies, the study must measure reproducibility using either the ICC (Shrout and Fleiss, 1979) or coefficient of variance (CV); (3) the article must not be a review or meta-analysis; and (4) the full text of the article must be available.

Qualification of researchers

The literature search was performed by T.W. (who has expertise in computer science and radiological science) and D.A.K. (who has expertise in medicine and radiological science). Both were supervised and trained in the conduct of systematic reviews by the authors D.P.A. and R.A.D., each holding PhDs and experienced in neuroimaging research and systematic review.

Data extraction and synthesis

From each article, the authors recorded and tabulated the number of subjects, the type of scan, the interscan interval, and the conclusions drawn about the reliability of graph metrics. From each article reporting reliability measurements derived from structural networks, the authors also recorded the software tools used for parcellation, registration, diffusion modeling, fiber tracking, and the edge weight definition used. Data were extracted independently by two researchers (T.W. and D.A.K.) and then merged to reduce the chance of data being missed or reported incorrectly. In the case of a conflict, the article in question was reviewed and discussed by both researchers together until an agreement was reached. Because the acquisition protocol used may be a factor translating into graph retest performance, results for structural and functional data were tabulated separately. A qualitative synthesis based on the included articles' findings and about how they relate to reliability was written for each of the following recurrent or important themes identified in the literature: choice of density threshold, type of ICC used, ROI size, retest interval, preprocessing strategy, type of graph metric, and fiber tracking algorithm. The software used for graph thresholding and calculating metrics was assumed to be equivalent; most studies used custom software with the Brain Connectivity Toolbox (Rubinov and Sporns, 2010) and the algorithms for the different graph metrics are well defined.

Risk of bias assessment

To assess the quality of each included study, the authors rated each article using a set of 10 criteria based on previous quality checklists (Downs and Black, 1998; West et al., 2002; Von Elm et al., 2007). Each criterion was assigned a weight of 1, 2, or 3 such that the emphasis was placed on quality of methodology rather than reporting. The highest possible score was 20 and the lowest, 0. The quality of each article was assessed independently by two researchers (T.W. and D.A.K.) and then finally determined by consensus. Low-scoring articles were not omitted, but their conclusions carried less influence within the review.

Results

Literature search

The database search returned 202 results, of which 73 were excluded for being duplicates. In the first phase of screening, 105 of the remaining 129 articles were excluded for not measuring the reliability of graph metrics in brain networks. In the second phase of screening, 1 of the remaining

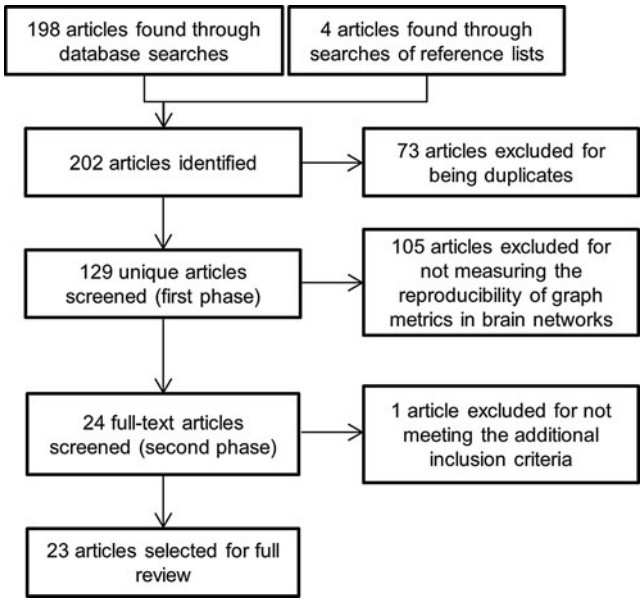

FIG. 1. Flowchart describing the number of results at each stage of the literature search.

24 articles (Faria et al., 2012) was excluded for analyzing the test-retest reliability of individual edge weights instead of summary graph metrics, leaving 23 articles to be included in the review. Figure 1 shows the results of the literature search process at each stage.

Risk of bias

The criteria used for quality assessment and the quality scores for each study are shown in Supplementary Table S1 (Supplementary Data are available online at www.liebertpub.com/brain). The reviewers agreed on quality criteria in 196 (89.1%) of the 230 total checks (10 criteria for each of 23 studies). In the cases where scores conflicted, a consensus was reached by discussion. Quality scores ranged from 16 to 20 with a median score of 20.

Frequent limiting factors in the methodological quality of the reviewed studies were not using a scanner with field strength of greater than 1.5 T (lower signal-to-noise ratio than higher strength magnets) and using small sample sizes. Some of the penalties incurred were due to inadequate reporting, such as failure to describe the type of ICC used or inadequate characterization of the sample. Particular strengths of the included studies were their appropriate choosing of acquisition, processing and graph construction methods, and clarity when reporting them.

Study characteristics

Table 1 gives a summary of each study's design and conclusions. The number of subjects in the studies ranged from 5 to 45 and numbered 499 in total. The most frequent image acquisition methods were functional magnetic resonance imaging (fMRI; 11 instances; 48% of 23) and diffusion tensor imaging (DTI; 10 instances; 43% of 23), but there were also two studies using magnetoencephalography data (10% of 23), one using functional near-infrared spectroscopy (fNIRS) and one using arterial spin labeling (each 5% of 23). The test-retest interval ranged from being shorter than 1 h to being longer than 1 year.

Within each study using functional data, at least one metric reached the excellent range in nine studies (Braun et al., 2012; Deuker et al., 2009; Guo et al., 2012; Liao et al., 2013; Niu et al., 2013; Park and Friston, 2013; Telesford et al., 2010; Wang et al., 2011; Weber et al., 2013) (ICC > 0.74; 64% of 14), the good range in three studies (Cao et al., 2014; Jin et al., 2011; Schwarz and McGonigle, 2011) (ICC 0.60–0.74; 21% of 14), the fair range in one study (Liang et al., 2012) (ICC 0.40–0.59; 7% of 14), the poor range in none of the studies (ICC < 0.40; 0% of 14), and one study did not fully report ICC data (Fan et al., 2012).

Within each study using structural data, at least one metric reached the excellent range in seven studies (Andreotti et al., 2014; Buchanan et al., 2014; Cheng et al., 2012; Duda et al., 2014; Owen et al., 2013; Parker et al., 2012; Vaessen et al., 2010) (78% of 9), the good range in two studies (Bassett et al., 2011; Dennis et al., 2012) (22% of 9), the fair range in none of the studies (0% of 9), and the poor range in 0 studies (0% of 9).

Tables 2 and 3 list the highest ICC measurements from the studies that reported the exact ICC values for those metrics. Table 4 draws a comparison between the methods employed in studies of the test-retest reliability of graph metrics in structural brain networks.

Synthesis of results

Acquisition method. The method used to acquire the test-retest data is one factor influencing reproducibility due to the differences in sensitivity to different physical properties of the brain between methods. Of the metrics for which three or more ICCs were reported for both functional and structural groups, six of seven were higher in the metrics based on structural data (Tables 2 and 3). This difference may have been expected, given the brain's dynamic and rapidly fluctuating hemodynamic state, even at rest, compared with its relatively static structure (Biswal et al., 1995). None of the included studies made a comparison between reliabilities of graph metrics derived from different acquisition methods.

Graph thresholds. In graphs of functional networks, edges are weighted by the correlation coefficient between the time series of two ROIs. In graphs of structural networks, edges are weighted by the number of streamlines connecting two ROIs. Typically, before calculating graph metrics, an arbitrary threshold is chosen below which edge weights are set to zero. Several different approaches were taken when thresholding weighted graphs. The most common was to threshold the graph at a range of densities (the density of a graph is given by the ratio of existing edges in the graph to the number of possible edges). Other approaches were fixed thresholding (Fan et al., 2012), mean degree thresholding (Owen et al., 2013), average path length thresholding (Telesford et al., 2010), and calculating weighted variants of graph metrics (Jin et al., 2011). While no study attempted to isolate the range of density threshold used to determine its effect on reproducibility, one study (Guo et al., 2012) compared the use of a fixed threshold (based on the edge weight alone) with soft and proportional thresholding techniques, but found neither to be significantly more reliable.

Table 1. Summary of Included Studies

Study	n	*Interval*	*Acquisition*	*Conclusions*
Andreotti	19	< 1 hour	DTI	Smaller ROIs produced less reliable graph metrics.
Bassett	6	5.5 days	DTI, DSI	Metrics were more reliable based on DTI data compared with DSI data: DSI is too susceptible to noise. Metrics were more reliable based on structural atlases compared with functional atlases. Larger ROIs produced more reliable graph metrics. Correcting edge weights for ROI size did not improve reliability.
Braun	33	14 days	RS-fMRI	Reducing the length of the time series and global signal regression did not improve reliability. Using the first eigenvariate and the median time series instead of the mean did not improve reliability. Using a broader frequency range significantly improved reliability.
Buchanan	10	2–3 days	DTI	Networks based on white matter ROIs proved more reliable than gray matter ROIs. Using a deterministic or probabilistic tractography algorithm or varying the edge weight definition did not significantly affect reliability.
Cao	26	14 ± 2.1 days	RS-fMRI, n-back fMRI	Global metrics were more reliable than local metrics. Coarse structural atlases were less reliable than finer parcellation schemes.
Cheng	44	1 week	DTI	Reliability was sensitive to the two different edge weighting schemes tested.
Dennis	17	101 ± 18 days	DTI	Most metrics were reliable above densities of 0.3.
Deuker	16	4–6 weeks	RS-MEG, n-back MEG	Reliability was higher during performance of a task. Task practice was associated with reliability. Metrics were more reliable in low-frequency bands.
Duda	21	< 1 hour	DTI	Choice of deterministic tracking algorithm did not significantly affect intrasubject reliability. Between subjects, some metrics are more sensitive than others to the choice of deterministic tracking algorithm. Choice of the anatomical label set affects some metric scores and reliabilities.
Fan	16	1 year	RS-fMRI	Having eyes open or closed during rest affected reliability of graph metrics. Hubs were less sensitive to changes in rest condition.
Guo	24	13 ± 3 months	RS-fMRI	Wavelet transformation of ROI time series improved reliability. Reliability of betweenness centrality was poor compared to degree and clustering coefficient. Soft and proportional thresholding did not improve reliability.
Jin	10	15 ± 8.4 days	RS-MEG	Reliability varies depending on eyes open and closed states, frequency band, metric type, and MEG sensor position. High frequencies were less reliable.
Liang	22, 25	< 1 hour, 11 ± 4 months	RS-fMRI	Demonstration of differences in reliability with different image processing strategies (Pearson vs. partial correlation, global signal regression, and frequency band).
Liao	11	1 week	RS-fMRI	Degree centrality was more reliable at longer scan durations. Degree centrality was more reliable in some brain regions with a shorter TR. Global signal regression reduced reliabilities.
Niu	21	20 minutes	RS-fNIRS	Different hemoglobin concentration types produced different graph metric reliabilities. Denoising the data with ICA did not improve reliability.
Owen	10, 5	60.8 ± 33.6 days	DTI	Metrics were more reliable using individual segmentation than when using a high-resolution atlas. Number of streamlines per voxel and weighting edges by connection strength did not affect metrics' reliability.
Park	12	3 hours (×8)	RS-fMRI	Local efficiency is more reliable than global efficiency.
Parker	28	< 1 hour	DTI	Demonstration of differences in reliability with different image processing strategies. Reliability of the small-worldness index was low at low densities compared with other metrics.
Schwarz	25	< 1 hour, 11 ± 4 months	RS-fMRI	Reliability was higher over a short between-scan interval compared with a long interval. Deconvolution of white matter, CSF, and motion time series increased reliability.
Telesford	45	< 1 hour	Executive task fMRI	Reliability was increased after spatial smoothing. Some metrics are more reliable in network hubs compared with nonhub nodes. Only the most stringent density thresholds significantly affected reproducibility. Mean degree was the least reliable global metric.
Vaessen	6	14 ± 8 days	DTI	Number of diffusion directions and gradient amplitude had no effect on reliability.
Wang	25	< 1 hour, 11 ± 4 months	RS-fMRI	Reliability was higher over a long between-scan interval compared with a short interval. Nodal metrics were more reliable than global metrics. Recommendations were given for graph construction depending on the type of atlas used.
Weber	22	30 minutes	ASL, risk/reward task fMRI	ASL perfusion fMRI produced more reliable graph metrics than BOLD fMRI. Reliability was lowest at low frequencies. Reliability was improved during task performance.

ASL, arterial spin-labeling perfusion functional magnetic resonance imaging; DSI, diffusion spectrum imaging; DTI, diffusion-tensor imaging; fMRI, blood oxygenation level-dependent functional magnetic resonance imaging; fNIRS, functional near-infrared spectroscopy; MEG, magnetoencephalography; ROI, regions of interest; RS, resting state.

Table 2. Reported Intraclass Correlation Coefficient Values for Various Metrics in Each Study That Reported Exact Intraclass Correlation Coefficients from Functional Data

Study	Sample size	ICC type	Acquisition	Clustering coefficient	Characteristic path length	Degree	Small-worldness	Global efficiency	Assortativity	Strength	Modularity	Local efficiency	Betweenness cent	Hierarchy	Synchronizability	Diversity	Transitivity	Mutual information	Cost efficiency	Mean of all metrics
Cao	26	2,1	RS-fMRI, nback	0.27	0.54	0.38	0.57	0.67	0.71	0.59	0.58	0.25	0.28	—	—	0.43	0.69	—	—	—
Deuker	16	—	RS-MEG, nback	0.72	**0.90**	—	**0.77**	**0.87**	0.69	—	—	—	—	**0.83**	**0.76**	—	—	**0.87**	**0.81**	0.62
Braun	33	3,1	RS-fMRI	0.59	0.61	—	**0.76**	0.60	**0.75**	—	0.70	0.49	—	**0.75**	—	—	—	—	—	0.48
Niu	21	1,1	RS-fNIRS	**0.78**	—	**0.84**	—	**0.78**	—	—	—	**0.84**	0.68	—	—	—	—	—	—	0.71
Telesford	45	1,1	Task fMRI	**0.86**	**0.79**	0.29	—	**0.83**	—	—	—	**0.75**	—	—	—	—	—	—	—	**0.86**
Jin	10	3,1	RS-MEG	—	—	0.66	—	—	—	—	—	0.60	0.43	—	—	—	—	—	—	—
Guo	24	3,1	RS-fMRI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	0.63
Liao	11	1,1	RS-fMRI	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—	**0.78**
Median (Range)				0.72 (0.27–0.86)	0.70 (0.54–0.90)	0.52 (0.29–0.84)	**0.77 (0.57–0.76)**	**0.78 (0.60–0.87)**	0.71 (0.69–0.75)	0.59 (—)	0.64 (0.58–0.70)	0.60 (0.25–0.84)	0.43 (0.28–0.68)	**0.79 (0.75–0.83)**	**0.76** (—)	0.43 (—)	0.69 (—)	**0.87** (—)	**0.81** (—)	0.67 (0.48–0.86)

The studies not listed did not report exact ICC values. The highest values were chosen so that the values reported below represent the most reliable method tested within each study. These values could act as a guide for approximate expected values of ICC given different methodological choices. As a rule of thumb, ICC scores are interpreted as follows: <0.40, poor; 0.40–0.59, fair; 0.60–0.74, good; >0.74, excellent (Fleiss et al., 2013) (shown in bold).

ICC, intraclass correlation coefficient.

Table 3. Reported Intraclass Correlation Coefficient Values for Various Metrics in Each Study That Reported Exact Intraclass Correlation Coefficients from Structural Data

Study	Sample size	Gradient directions	ICC type	Acquisition	Clustering Coefficient	Characteristic path length	Degree	Small-worldness	Global efficiency	Assortativity	Strength	Modularity	Local efficiency	Nodal distance	Betweenness cent	Rent's coef	Robustness rand	Robustness targ	Hierarchy	Synchronizability	Diversity	Mean of all metrics
Bassett	6	30	3,1	DTI, DSI	0.48	—	—	—	0.37	0.63	—	0.46	0.43	0.64	0.28	0.70	0.53	0.44	0.43	0.36	—	0.72
Andreotti	19	42	3,1	DTI	**0.93**	**0.91**	**0.90**	**0.91**	**0.92**	**0.87**	**0.87**	**0.91**	**0.93**	**0.76**	0.67	—	—	—	—	—	—	**0.86**
Cheng	44	48	3,1	DTI	0.54	0.28	—	0.55	0.64	—	—	0.67	—	—	—	—	—	—	—	—	0.70	—
Owen	10, 5	30	—	DTI	**0.88**	**0.94**	**0.90**	—	**0.94**	—	—	—	**0.87**	—	**0.94**	—	—	—	—	—	—	—
Dennis	17	94	1,1	DTI	0.58	0.59	—	0.52	0.60	—	—	0.61	—	—	—	—	—	—	—	—	—	—
Buchanan	10	64	3,1	DTI	**0.76**	0.64	0.66	—	—	—	**0.76**	—	—	—	—	—	—	—	—	—	—	—
Vaessen	6	32, 15, 6	3,1	DTI	**0.88**	**0.94**	**0.77**	—	—	—	—	—	—	—	—	—	—	—	—	—	—	—
Median (range)					**0.76 (0.48–0.93)**	**0.78 (0.28–0.94)**	**0.84 (0.66–0.90)**	0.55 (0.52–0.91)	0.64 (0.37–0.94)	**0.75 (0.63–0.87)**	**0.82 (0.76–0.87)**	0.64 (0.46–0.91)	**0.87 (0.43–0.93)**	0.70 (0.64–0.76)	0.67 (0.28–0.94)	0.70 (—)	0.53 (—)	0.44 (—)	0.43 (—)	0.36 (—)	0.70 (—)	**0.79 (0.72–0.86)**

The studies not listed did not report exact ICC values. The highest values were chosen so that the values reported below represent the most reliable method tested within each study. These values could act as a guide for approximate expected values of ICC given different methodological choices. As a rule of thumb, ICC scores are interpreted as follows: <0.40, poor; 0.40–0.59, fair; 0.60–0.74, good; >0.74, excellent (Fleiss et al., 2013) (shown in bold).

Table 4. Comparison of the Approaches Taken in Studies of Graph Metrics' Reliability in Structural Networks

Study	Acquisition	Number of diffusion gradient directions	Parcellation scheme	Registration	Diffusion modeling	Fiber tracking algorithm	Edge weight definition
Andreotti	DTI	42	FreeSurfer	FSL	FSL	FSL	[f]
Bassett	DTI, DSI	30	AAL, HOA, LPBA40, upsampling	FSL	TrackVis	TrackVis	[b, a]
Buchanan	DTI	64	FreeSurfer	FSL	FSL	FSL, FACT	[b, c, e]
Cheng	DTI	48	FreeSurfer	FSL	TrackVis	TrackVis	[c, g]
Dennis	DTI	94	FreeSurfer	FSL	FSL	FSL	[d]
Duda	DTI	34	DKT31	ANTs	CAMINO	FACT, Euler, RK4, TEND	[a]
Owen	DTI	30	FreeSurfer	FSL	FSL	FSL	[d]
Parker	DTI DTI	60	FreeSurfer, NiftySeg	FSL, NiftyReg	FSL, MRTrix	FSL, MRTrix	[a]
Vaessen	DTI	32, 15, 6	WFUpick	CATNAP	CAMINO	CAMINO	[a]

[a]The edge weight is given by the number of connecting streamlines between two ROIs.
[b]The edge weight is given by the sum of the connecting streamlines divided by the mean of the two ROIs' volumes.
[c]As (a), but correcting for streamline length.
[d]The number of fibers connecting the two ROIs normalized to the volume of the selected ROI.
[e]The mean FA value along interconnecting streamlines.
[f]As (c), divided by the sum number of streamlines started from the ROIs, multiplied by the sum size of the two ROIs.
[g]Twice the sum of the connecting streamlines between the two ROIs, divided by the sum volume of the 2 ROIs.

AAL, Automated Anatomical Labeling; ANTs, Advanced Normalization Tools; CATNAP, Coregistration Adjustment and Tensor Solving, A Nicely Automated Program; DKT31, Desikan–Killiany–Tourville; FACT, Fiber Assignment by Continuous Tracking; FSL, FMRIB Software Library; HOA, Harvard–Oxford Atlas; LPBA40, LONI Probabilistic Brain Atlas; RK4, Fourth-order Runge-Kutta; TEND, Tensor Deflection; WFUpick, Wake Forest University Pick.

ROI size. Three studies looked at the effect of the ROI size on reproducibility. One tested the relationship between the ROI size (from within a single structural parcellation) and test-retest reliability of local DTI-based graph metrics over time (Andreotti et al., 2014). Another used three different structural atlases (based on anatomical, as opposed to functional regions) and upsampled them by dividing each region into two, thereby doubling the resolution, and tested the reproducibility of the resulting global DTI-based graph metrics over time (Bassett et al., 2011). These two studies found that graphs based on larger structural ROIs and derived from DTI data produced metrics that were more reliable than those based on upsampled or more finely grained parcellation schemes. By contrast, the third study, which used fMRI data, found metrics derived from graphs based on a high-resolution functional atlas to be more reproducible than those of a lower resolution structural atlas (Cao et al., 2014); however, these results were based on ICCs averaged over three types of tasks, within which there were significant differences in reproducibility. It is unclear whether the reported difference in reliabilities associated with atlas resolution would have retained significance when compared within each task condition or at rest.

Preprocessing strategy. All included studies used different strategies and tools for data preprocessing. One study (Parker et al., 2012) tested two entirely different DTI pipelines, finding differences in CV and ICC values between them; however, interpretation of the results is limited by not being able to identify which of the steps were responsible for the greatest differences in reliability. Another study (Braun et al., 2012) tested seven fMRI pipelines, varying one step of a standard pipeline at a time, and found that including a broader frequency band from the fMRI time series and using global signal regression yielded the most reliable graph metrics. A third study (Cao et al., 2014) tested five different task regression methods and two atlases on fMRI data, identifying two approaches to regression as being the most effective and finding that neither the functional nor the structural atlas produced significantly more reproducible metrics than the other.

Type of graph metric. Of the many summary measures of graph organization, several classifications can be made; for example, global and local metrics or weighted and binary metrics. All of the included articles gathered reliability measurements for different metrics, and many of them drew a direct comparison between the test-retest reliability of different metrics or types of metrics. Two studies (Andreotti et al., 2014; Cao et al., 2014) distinguished between local and global metrics, each finding global metrics to be more reproducible, with local metrics being more variable. One (Braun et al., 2012) noted that first-order metrics (those derived directly from the graph) were less reproducible than second-order metrics (those derived from the first-order metrics). Four studies, each with different acquisition types, focused on the relative reproducibility of individual metrics. The first (Dennis et al., 2012), which acquired DTI data, found that modularity was the most reproducible metric. The second (Niu et al., 2013), which acquired resting-state fNIRS (RS-fNIRS), and the third (Telesford et al., 2010), which acquired fMRI during performance of an executive task, found that the clustering coefficient and global efficiency were both the most reproducible metrics, with the third noting that degree was the least reproducible. In contrast, the fourth study (Wang et al., 2011), which acquired RS-fMRI data, found that degree was the most reproducible metric.

Fiber tracking algorithm. Eight different fiber tracking algorithms were used by the included studies (Table 4). There were two instances where different algorithms were compared within-study to test the reproducibility of graph metrics derived from each. In one (Buchanan et al., 2014), the authors found that for gray matter seeds, neither the FMRIB's diffusion toolbox (FDT) nor fiber assignment by continuous tracking (FACT) algorithms produced significantly more reproducible graph metrics than the other when run with any weighting or waypoint length threshold. The second study (Duda et al., 2014) compared four different algorithms and found that none was consistently more reproducible than the others for any graph metric.

Retest interval. The two studies looking at the effect of the length of the interscan interval on graph metrics' test-retest reliability had divergent conclusions. The first found that the reproducibility of graph metrics measured over a short interval was greater compared with those measured over a long interval (Schwarz and McGonigle, 2011). Despite both using the same publicly available RS-fMRI dataset, the second study found the opposite—that reproducibility was greatest when measured between the scans separated by a long retest interval (Wang et al., 2011). The most overt methodological difference between these studies was that the first measured its long interval between scans >5 months apart, whereas the second measured it between the first scan (>5 months from the second) and the average of the second and third scans (<1 h apart). The two studies also used different atlases for parcellation and removed different sets of confound signals, which could have impacted the result.

ICC type. There are six main types of ICCs, each one of which has a subtly different interpretation (Müller and Büttner, 1994); therefore, choosing the most appropriate version of ICC is an important yet difficult task, which must take into account the aim of the study. In this review, nine studies (Dennis et al., 2012; Fan et al., 2012; Liang et al., 2012; Liao et al., 2013; Niu et al., 2013; Schwarz and McGonigle, 2011; Telesford et al., 2010; Wang et al., 2011; Weber et al., 2013) used the ICC(1,1) version, which is a measure of absolute agreement and is sensitive to differences in means between raters. One study (Cao et al., 2014) used the ICC(2,1) version, which treats raters as random effects and emphasizes interchangeability between raters. Eight studies (Andreotti et al., 2014; Bassett et al., 2011; Braun et al., 2012; Buchanan et al., 2014; Duda et al., 2014; Guo et al., 2012; Jin et al., 2011; Park et al., 2012) used the ICC(3,1) version, which treats raters as a fixed effect and emphasizes inter-rater consistency, that is, association between a finite set of scanners, but is not generalizable beyond those scanners. Other methods used to quantify the test-retest reliability were the CV and Bland–Altman plots. No study compared ICC types or discussed the effect of their choice of statistical test on the interpretation of their results.

Discussion

The authors have reviewed and summarized the published literature that investigates the test-retest reliability of graph-theoretic brain network metrics. The primary aim was to establish the reproducibility of graph metrics of brain networks. The authors find that reported ICC scores were often in the good and excellent ranges, indicating that the test-retest reliability can be adequate under certain conditions. These scores varied between functional and structural networks. For example, across the studies of functional networks, six metrics (the clustering coefficient, characteristic path length, small-worldness, global efficiency, assortativity, and local efficiency) had median ICCs across three or more studies in the good or excellent ranges (Table 2). In the studies of structural networks, seven metrics (the clustering coefficient, characteristic path length, degree, global efficiency, modularity, local efficiency, and betweenness centrality) met the same criteria (Table 3). These were the most reproducible metrics and therefore may be the most promising for future use in clinical neuroscience research. For the metrics having three or more ICCs reported for both functional and structural networks, six of seven were higher in structural networks, indicating that structural networks may be more reliable over time.

Another aim was to understand how different methodological factors affect the reproducibility of graph summary measures. There was limited evidence that, when using structural data, larger ROIs may be preferable, and that when using functional data, smaller ROIs may be preferable. The authors also find that global metrics are more reproducible than local metrics and second-order metrics are more reproducible than first-order metrics. Different metrics are more or less reproducible depending on both the acquisition type and the state of the test subject; for example, Wang and coworkers (2011) show that for resting-state fMRI data, degree was the most reproducible metric, whereas for Telesford and associates (2010), under an executive task fMRI scan, degree was the least reproducible. There was some evidence that the specific fiber tracking algorithm used with DTI data had little effect on graph metrics' reproducibility, and that the preprocessing steps taken can significantly alter metrics' reproducibility. The optimal graph threshold type, retest interval, and ICC type were not clear from the existing literature due to conflicting results, and the sample size and number of gradient directions had no clear correspondence to ICC scores (Table 3).

However, this analysis of methodological factors identifies some important issues to be addressed. A major issue is that the breadth of approaches and the range of reported ICC types in the included articles prevented meta-analysis and complicated the identification of any consensus view. For example, even studies using the same dataset and performing relatively similar analyses report drastically different results (Schwarz and McGonigle, 2011; Wang et al., 2011). There are still many unknowns in the methods being applied, such as the ideal density threshold or range, necessary fMRI scan length (known to affect reliability (Birn et al., 2013; Whitlow et al., 2010), type of atlas, and ROI size; furthermore, the most reproducible of these is not necessarily the most biologically plausible. Variability in the research designs of the individual studies prevented any clear analysis strategy from standing out as superior, so when testing multiple preprocessing pipelines or analysis strategies, the authors recommend that researchers isolate one variable at a time and study its effect on reproducibility rather than varying multiple aspects of the

method at once. In this respect, two articles stand out as good examples of research upon which future studies could be modeled (Braun et al., 2012; Cao et al., 2014). Individual processing steps can have a large impact on results; for example, the use of global signal regression has been shown to obscure the findings of increased cortical power and variance in schizophrenia (Yang et al., 2014). The authors also suggest that replication studies are performed to establish further the generalizability of the ICC measurements across cohorts and across more than one repeat scan (four of the datasets used in the included studies are freely available to download (Buchanan et al., 2014; Duda et al., 2014; Liang et al., 2012; Schwarz and McGonigle, 2011; Wang et al., 2011)).

Previous studies have suggested the use of graph metrics in clinical trials (Petrella, 2011) and as diagnostic tools (Quraan et al., 2013; Schoonheim et al., 2013). There is clear appeal to this approach. Metrics are well defined in terms of the graph itself, and studies in disease populations have reported changes in the direction of metric score that are consistent in relation to the disease status; for example, in schizophrenia where clustering is consistently lower than in healthy people (Anderson and Cohen, 2013; He et al., 2012; Liu et al., 2008; Lynall et al., 2010; Rubinov et al., 2009). However, in addition to the uncertainties regarding the contextual validity of graph summary measures (i.e., why they correlate with some disease processes, and whether some metrics hold any biological significance at all), the issue of reproducibility is critical to address before graph metrics are used in clinical trials or for clinical diagnosis. Several studies have concluded that the reproducibility of this approach is sufficient to allow application in clinical research populations (Bassett et al., 2011; Braun et al., 2012; Niu et al., 2013; Owen et al., 2013; Tomasi and Volkow, 2011), but others have suggested the opposite (Andreotti et al., 2014; Deuker et al., 2009). On the findings of this review, in which the authors have collated the evidence of graph metric reproducibility as identified by systematic review, the authors cannot draw conclusions about clinical relevance. While reproducibility studies have often demonstrated good ICC measurements, reproducibility is not the only criterion for suitability for use in clinical trials; to the authors knowledge there have been no studies examining the responsiveness of brain network properties to intervention. This review of test-retest reproducibility studies of GT metrics has also identified a lack of studies assessing multicenter or multiplatform reproducibility, which will be important to establish if GT metrics are to be adopted in future multicenter treatment trials. Although one of the included studies used data from two different scanners, no comparison was made between them (Braun et al., 2012).

This review is also the first to systematically review data processing strategies used in graph-theoretic analysis of brain networks in the context of test-retest studies. Andreotti and colleagues (2014) performed a short qualitative review, in which they tabulate several parameters of the graph analysis, but only included six studies and did not compare or discuss the table in depth. Zuo and Xing (2014) conducted a qualitative review of the test-retest reliability of resting-state fMRI measurements in human brain networks, but did not focus on graph metrics.

This study was limited primarily by incomplete reporting. The original aim was to meta-analyze the published literature to provide summarized test-retest reliability data for the various graph theory metrics, but it became apparent that meta-analysis was not possible without full reporting of the variances. Additionally, meta-analysis would be severely limited by the heterogeneity of the methods employed in individual studies. To allow future meta-analysis, the authors recommend that studies report the data fully in terms of variances (standard deviation or range depending on normal or nonnormal distribution) as well as the type of ICC calculated. The authors would also warn other authors to take care when interpreting results based on a mixture of imaging modalities.

Conclusion

The authors have identified the graph metrics, which show the most promise for future research use. Reproducibility for these metrics was frequently good and excellent. Methodological factors impact upon reproducibility, and researchers need to take these into account when planning their analyses.

Acknowledgment

This work was supported by a PhD studentship grant from the UK Multiple Sclerosis Society (registered charity 1139257).

Author Disclosure Statement

No competing financial interests exist.

References

Anderson A, Cohen MS. 2013. Decreased small-world functional network connectivity and clustering across resting state networks in schizophrenia: an fMRI classification tutorial. Front Hum Neurosci 7:520.

Andreotti J, Jann K, Melie-Garcia L, Giezendanner S, Dierks T, Federspiel A. 2014. Repeatability analysis of global and local metrics of brain structural networks. Brain Connect 4:203–220.

Arienzo D, et al. 2013. Abnormal brain network organization in body dysmorphic disorder. Neuropsychopharmacology 38: 1130–1139.

Bassett DS, Brown JA, Deshpande V, Carlson JM, Grafton ST. 2011. Conserved and variable architecture of human white matter connectivity. Neuroimage 54:1262–1279.

Bassett DS, Bullmore ET. 2009. Human brain networks in health and disease. Curr Opin Neurol 22:340.

Birn RM, et al. 2013. The effect of scan length on the reliability of resting-state fMRI connectivity estimates. Neuroimage 83:550–558.

Biswal B, Zerrin Yetkin F, Haughton VM, Hyde JS. 1995. Functional connectivity in the motor cortex of resting human brain using echo-planar MRI. Magn Reson Med 34:537–541.

Braun U, et al. 2012. Test–retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. Neuroimage 59:1404–1412.

Buchanan CR, Pernet CR, Gorgolewski KJ, Storkey AJ, Bastin ME. 2014. Test–retest reliability of structural brain networks from diffusion MRI. Neuroimage 86:231–243.

Buckner RL, et al. 2009. Cortical hubs revealed by intrinsic functional connectivity: mapping, assessment of stability,

and relation to Alzheimer's disease. The J Neurosci 29: 1860–1873.
Bullmore E, Sporns O. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. Nat Rev Neurosci 10:186–198.
Cao H, et al. 2014. Test–retest reliability of fMRI-based graph theoretical properties during working memory, emotion processing, and resting state. Neuroimage 84:888–900.
Cheng H, et al. 2012. Characteristics and variability of structural networks derived from diffusion tensor imaging. Neuroimage 61:1153–1164.
Dennis EL, et al. 2012. Test-retest reliability of graph theory measures of structural brain connectivity. Med Image Comput Comput Assist Interv 15(Pt3):305–312.
Deuker L, et al. 2009. Reproducibility of graph metrics of human brain functional networks. Neuroimage 47:1460–1468.
Downs SH, Black N. 1998. The feasibility of creating a checklist for the assessment of the methodological quality both of randomised and non-randomised studies of health care interventions. J Epidemiol Commun Health 52:377–384.
Duda JT, Cook PA, Gee JC. 2014. Reproducibility of graph metrics of human brain structural networks. Front Neuroinform 8:46.
Fan Y, et al. 2012. Test-retest reliability of the graph metrics in different rest conditions and sampling rates. Poster presented at the 18th Annual Meeting of the Organization for Human Brain Mapping, Beijing, China.
Faria AV, et al. 2012. Atlas-based analysis of resting-state functional connectivity: evaluation for reproducibility and multimodal anatomy–function correlation studies. Neuroimage 61:613–621.
Fleiss JL, Levin B, Paik MC. 2013. *Statistical Methods for Rates and Proportions*. New York, NY: John Wiley & Sons.
Göttlich M, Münte TF, Heldmann M, Kasten M, Hagenah J, Krämer UM. 2013. Altered resting state brain networks in Parkinson's disease. PLoS One 8:e77336.
Guo CC, et al. 2012. One-year test–retest reliability of intrinsic connectivity network fMRI in older adults. Neuroimage 61: 1471–1483.
He H, et al. 2012. Altered small-world brain networks in schizophrenia patients during working memory performance. PLoS One 7:e38195.
He Y, et al. 2009. Impaired small-world efficiency in structural cortical networks in multiple sclerosis associated with white matter lesion load. Brain 132:3366–3379.
Jin S-H, Seol J, Kim JS, Chung CK. 2011. How reliable are the functional connectivity networks of MEG in resting states? J Neurophysiol 106:2888–2895.
Latora V, Marchiori M. 2001. Efficient behavior of small-world networks. Phys Rev Lett 87:198701.
Liang X, et al. 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.
Liao X-H, et al. 2013. Functional brain hubs and their test-retest reliability: a multiband resting-state functional MRI study. Neuroimage 83:969–982.
Liu Y, et al. 2008. Disrupted small-world networks in schizophrenia. Brain 131:945–961.
Lynall M-E, et al. 2010. Functional connectivity and brain networks in schizophrenia. J Neurosci 30:9477–9487.
Menon V. 2011. Large-scale brain networks and psychopathology: a unifying triple network model. Trends Cogn Sci 15:483–506.
Müller R, Büttner P. 1994. A critical discussion of intraclass correlation coefficients. Stat Med 13:2465–2476.
Niu H, et al. 2013. Test-retest reliability of graph metrics in functional brain networks: a resting-state fNIRS study. PLoS One 8:e72425.
Owen JP, et al. 2013. Test–retest reliability of computational network measurements derived from the structural connectome of the human brain. Brain Connect 3:160–176.
Park B, Kim JI, Lee D, Jeong S-O, Lee JD, Park H-J. 2012. Are brain networks stable during a 24-hour period? Neuroimage 59:456–466.
Park H-J, Friston K. 2013. Structural and functional brain networks: from connections to cognition. Science 342:1238411.
Parker C, Clark C, Clayden J. 2012. Reproducibility of whole-brain structural connectivity networks. Neuroimage 61:1153–1164.
Petrella JR. 2011. Use of graph theory to evaluate brain networks: a clinical tool for a small world? Radiology 259: 317–320.
Quraan MA, McCormick C, Cohn M, Valiante TA, McAndrews MP. 2013. Altered resting state brain dynamics in temporal lobe epilepsy can be observed in spectral power, functional connectivity and graph theory metrics. PLoS One 8:e68609.
Rubinov M, et al. 2009. Small-world properties of nonlinear brain activity in schizophrenia. Hum Brain Mapp 30:403–416.
Rubinov M, Sporns O. 2010. Complex network measures of brain connectivity: uses and interpretations. Neuroimage 52: 1059–1069.
Schoonheim MM, et al. 2013. Functional connectivity changes in multiple sclerosis patients: a graph analytical study of MEG resting state data. Hum Brain Mapp 34:52–61.
Schwarz AJ, McGonigle J. 2011. Negative edges and soft thresholding in complex network analysis of resting state functional connectivity data. Neuroimage 55:1132–1146.
Shrout PE, Fleiss JL. 1979. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86:420.
Stam C, et al. 2009. Graph theoretical analysis of magnetoencephalographic functional connectivity in Alzheimer's disease. Brain 132:213–224.
Supekar K, Musen M, Menon V. 2009. Development of large-scale functional brain networks in children. PLoS Biol 7:e1000157.
Telesford QK, et al. 2010. Reproducibility of graph metrics in fMRI networks. Front Neuroinform 4:117.
Tomasi D, Volkow ND. 2011. Functional connectivity hubs in the human brain. Neuroimage 57:908–917.
Vaessen M, Hofman P, Tijssen H, Aldenkamp A, Jansen J, Backes WH. 2010. The effect and reproducibility of different clinical DTI gradient sets on small world brain connectivity measures. Neuroimage 51:1106–1116.
Von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. 2007. The Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) statement: guidelines for reporting observational studies. Prev Med 45:247–251.
Wang J-H, Zuo X-N, Gohel S, Milham MP, Biswal BB, 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.
Wang J, Zuo X, He Y. 2010. Graph-based network analysis of resting-state functional MRI. Front Syst Neurosci 4:16.
Watts DJ, Strogatz SH. 1998. Collective dynamics of 'small-world'networks. Nature 393:440–442.
Weber MJ, Detre JA, Thompson-Schill SL, Avants BB. 2013. Reproducibility of functional network metrics and network structure: a comparison of task-related BOLD, resting ASL

with BOLD contrast, and resting cerebral blood flow. Cogn Affect Behav Neurosci 13:627–640.
West S, et al. 2002. Systems to rate the strength of scientific evidence: Summary. In: AHRQ evidence report summaries. Rockville (MD): Agency for Healthcare Research and Quality (US); 1998–2005. Available from: www.ncbi.nlm.nih.gov/books/NBK11930/ Last accessed December 15, 2014.
Whitlow CT, Casanova R, Maldjian JA. 2010. Effect of resting-state functional MR imaging duration on stability of graph theory metrics of brain network connectivity. Radiology 259: 516–524.
Yang GJ, et al. 2014. Altered global brain signal in schizophrenia. Proc Natl Acad Sci U S A 111:7438–7443.
Zuo X-N, Xing X-X. 2014. Test-retest reliabilities of resting-state FMRI measurements in human brain functional connectomics: a systems neuroscience perspective. Neurosci Biobehav Rev 45:100–118.

Address correspondence to:
Robert A. Dineen
Sir Peter Mansfield Imaging Centre
University of Nottingham
Room W/B 1441, Queen's Medical Centre
Derby Road
Nottingham NG7 2UH
Nottinghamshire
United Kingdom

E-mail: rob.dineen@nottingham.ac.uk