Cogn Affect Behav Neurosci (2013) 13:714–724
DOI 10.3758/s13415-013-0196-0

Quantifying the reliability of image replication studies: The image intraclass correlation coefficient (I2C2)

H. Shou · A. Eloyan · S. Lee · V. Zipunnikov · A. N. Crainiceanu · M. B. Nebel · B. Caffo · M. A. Lindquist · C. M. Crainiceanu

Published online: 11 September 2013

Abstract This article proposes the image intraclass correlation (I2C2) coefficient as a global measure of reliability for imaging studies. The I2C2 generalizes the classic intraclass correlation (ICC) coefficient to the case when the data of interest are images, thereby providing a measure that is both intuitive and convenient. Drawing a connection with classical measurement error models for replication experiments, the I2C2 can be computed quickly, even in high-dimensional imaging studies. A nonparametric bootstrap procedure is introduced to quantify the variability of the I2C2 estimator. Furthermore, a Monte Carlo permutation is utilized to test reproducibility versus a zero I2C2, representing complete lack of reproducibility. Methodologies are applied to three replication studies arising from different brain imaging modalities and settings: regional analysis of volumes in normalized space imaging for characterizing brain morphology, seed-voxel brain activation maps based on resting-state functional magnetic resonance imaging (fMRI), and fractional anisotropy in an area surrounding the corpus callosum via diffusion tensor imaging. Notably, resting-state fMRI brain activation maps are found to have low reliability, ranging from .2 to .4. Software and data are available to provide easy access to the proposed methods.

H. Shou · A. Eloyan · V. Zipunnikov · B. Caffo · M. A. Lindquist · C. M. Crainiceanu (✉)
Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, 615 N Wolfe Street, Baltimore, MD 21205, USA
e-mail: ccrainic@jhsph.edu

S. Lee
Department of Psychiatry and the Department of Biostatistics, Columbia University, 722 West 168th street, New York, NY 10032, USA

A. N. Crainiceanu
Computer Science Department, United States Naval Academy, 572M Holloway Road, Stop 9F Annapolis, MD 21402-5002, USA

M. B. Nebel
Laboratory for Neurocognitive and Imaging Research, Kennedy Krieger Institute, 716 North Broadway, Baltimore, MD 21205, USA

Keywords RAVENS · DTI · fMRI · Replication studies · Intraclass correlation coefficient

Introduction

Replication is the cornerstone of science. Its absence reduces any scientific endeavor to a set of unverified beliefs. Brain imaging studies are no exception, although they have several specific characteristics that conspire to make quantification of reliability especially difficult. First, measurements are complex and idiosyncratic for each modality. Second, the definition of the actual target to be measured is often imperfect. Third, the data sets are large and not amenable to standard investigations of replication. Fourth, there is relatively little cross-pollination of research between different imaging modalities. Finally, setting up replication experiments can be difficult under many scenarios.

A variety of methods have been proposed for measuring the reliability of images, particularly in the context of functional neuroimaging studies (see Bennett & Miller, 2010, for an overview). One approach, the intraclass correlation (ICC) (Shrout & Fleiss, 1979), can be used to measure the similarity between region of interest (ROI) summaries of activation, intensity, or shape metrics in multiple subjects under two or more experimental replications. Another approach, the Dice coefficient (Rombouts, Barkhof, Hoogenraad, Sprenger, & Scheltens, 1998) measures what proportion of voxels exceed a threshold, such as one indicating activation, in both of two separate imaging sessions. A third approach, predictive modeling, measures the ability of a training data set to predict the structure of test data. One of the best established predictive modeling techniques within functional neuroimaging is the nonparametric prediction, activation, influence, and reproducibility sampling approach (NPAIRS; Strother et al., 2002), which has been used to illustrate how small changes in a functional magnetic resonance imaging (fMRI) processing pipeline can have dramatic effects on final results.

In this work, we propose a general model for brain imaging replication studies and introduce the image intraclass correlation (I2C2) as a measure of data reliability. This measure generalizes the classic (scalar) ICC to the case when the measurement target is an image. Resampling approaches are then developed to quantify I2C2 variability under the replication design and to test whether it is different from the I2C2 obtained under a random permutation of subject matching. Notably, the proposed framework is applied to three replication studies utilizing data from different brain imaging modalities. These include regional analysis of volumes in normalized space (RAVENS) imaging (a technique used to investigate localized changes in brain morphology; Davatzikos, Genc, Xu, & Resnick, 2001), seed-voxel brain connectivity maps based on resting-state fMRI (rs-fMRI), and fractional anisotropy (FA) measured using diffusion tensor imaging (DTI) in an area surrounding the corpus callosum.

The image intraclass correlation coefficient

To better understand the underlying issue, consider the most basic replication study where $J = 2$ and scalar replicate measurements are collected for each of I subjects. An example would be measuring total white matter brain volume from two imaging sessions. Yet even in such a seemingly straightforward setting, the study of and expectations for the extent of replication can vary dramatically. For example, in one study, replicate images may be collected on the same day, using the same scanner, and processed by the same technicians, while in another, replicate images may be collected weeks apart, in different laboratories, with different technicians and scanners. Using our example for context, let X_i denote the true (unknown) white matter volume and W_{ij} the white matter volume measurements from two replications. Succinctly, the observed W_{ij}'s are the measured proxies of the measurement of interest, X_i. The classical measurement error model (Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Fuller, 1987) in replication studies is

$$W_{ij} = X_i + U_{ij}, \tag{1}$$

with assumptions that the measurements, X_i, are independent across subjects and the measurement errors, U_{ij}, are independent across both subjects and replicates and are mutually independent of X_i, for $i = 1,\ldots,I$, and $j = 1,\ldots,J$. Conceptually, U_{ij} is the error that occurs during each individual measurement of the true target, X_i. The classical measurement error model further assumes that the measurement error variates U_{ij} have the same variance, . Likewise, we denote the variance of X_i by σ_X^2. This model is then equivalent to a one-way ANOVA model with random effects. Note that the observed measurements, W_{ij}, for the same subject, i, are correlated, since they share the same X_i. Specifically, the correlation is equal to

$$\mathrm{corr}(W_{i1}, W_{i2}) = \frac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} = \frac{\sigma_W^2 - \sigma_U^2}{\sigma_W^2} = 1 - \frac{\sigma_U^2}{\sigma_W^2}.$$

This is the well known ICC coefficient. Here, the "class" is the replication experiment, and the correlation is between replicated measurements for the same subject. In the measurement error literature, ICC is referred to as the reliability ratio. The ICC is a scale-free quantity between 0 and 1, where 0 corresponds to exact independence of measurements W_{i1} and W_{i2}; that is, they are unrelated, despite attempting to measure the same underlying quantity. Correspondingly, 1 indicates perfect reliability for every subject, $W_{i1} = W_{i2} = X_i$. Estimation is simple; $\sigma_W{}^2$ can be estimated as the variance of W_{ij}, and $\sigma_U{}^2$ can be estimated by the variance of $(W_{i2} - W_{i1})/2$.

Generalizations of the ICC to high-dimensional multivariate settings, such as images, are not obvious. However, a need for reliability metrics from these settings arises frequently. For example, the target of measurement might be a measure of brain morphology in a template, an rs-fMRI connectivity map, an FA map in an ROI such as the area surrounding the corpus callosum (see the Methods section), and so forth. In specific terms, let $X_i(v)$ be the (unknown) true image and $W_{ij}(v)$ be the proxy measurements of $X_i(v)$ at voxel v. The classical image measurement error can then be written as

$$W_{ij}(v) = X_i(v) + U_{ij}(v), \tag{2}$$

where all images are represented as $V \times 1$ dimensional vectors; $W_{ij} = \{W_{ij}(v): v = 1,\ldots,V\}$ are the observed proxy images; $X_i = \{X_i(v): v = 1,\ldots,V\}$ are the true images, assumed to be independent across subjects; and $U_{ij} = \{U_{ij}(v): v = 1,\ldots,V\}$ are the measurement error images, assumed to be independent across subjects and replicates and (mutually) of X_i. Here, $i = 1,\ldots,I$, and $j = 1,\ldots,J_i$. Thus, we consider a general case involving different numbers of replicates per subject, J_i of any value greater than or equal to 2.

The model further assumes that the measurement error vector, U_{ij}, has covariance K_U and X_i has covariance, K_X; that is, cov $(U_{ij},U_{ij}) = K_U$ and cov $(X_i,X_i) = K_X$. These cannot be directly estimated, since the U_{ij} and X_i are unobserved. Note that the covariance operator of the observed data K_W = cov (W_{ij}, W_{ij}), a quantity directly estimable from the data, can be written as $K_W = K_X + K_U$ via the straightforward application of the multivariate variance operator to Equation 2. Exactly, paralleling the univariate setting, K_X is interpreted as the within-subjects covariance, and K_U as the covariance of the measurement error.

On the basis of the aforementioned connection with the classical measurement error model (Equation 1), we propose the following I2C2 coefficient:

$$\rho = \frac{\text{trace}\ (K_X)}{\text{trace}\ (K_W)} = \frac{\text{trace}\ (K_W) - \text{trace}\ (K_U)}{\text{trace}\ (K_W)} = 1 - \frac{\text{trace}\ (K_U)}{\text{trace}\ (K_W)}. \tag{3}$$

One possible way of calculating I2C2 is to estimate the smoothed covariance matrices using multilevel functional principal component analysis (MFPCA; Di, Crainiceanu, Caffo, & Punjabi, 2009) or its extension to high-dimensional data (Zipunnikov et al., 2011). Alternatively, we obtain the following method of moments estimators based on formulas from Carroll et al. (2006) to reduce the computational cost,

$$\widehat{\text{trace}}(K_W) = \frac{1}{\sum_{i=1}^{I} J_i - 1} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{v=1}^{V} \left\{ W_{ij}(v) - \overline{W}_{\cdot\cdot}(v) \right\}^2,$$

and

$$\widehat{\text{trace}}(K_U) = \frac{1}{\sum_{i=1}^{I} (J_i - 1)} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{v=1}^{V} \left\{ W_{ij}(v) - \overline{W}_{i\cdot}(v) \right\}^2.$$

Here, $\overline{W}_{\cdot\cdot}(v) = \sum_{i,j,v} W_{ij}(v) / \sum_{i=1} J_i$ is the average of all images over all subjects and visits, and $\overline{W}_{i\cdot}(v) = \sum_{j=1}^{J_i} W_{ij}(v)/J_i$ is the average image for subject i over all visits j. Thus, an estimate of I2C2 can be reached by entering these estimates into Equation 3.

Calculating the I2C2 is both quick and scalable, because it does not require dealing with the $V \times V$ dimensional matrices. Indeed, the computational burden for calculating trace(K_W) and trace(K_U) is linear in V. Moreover, the formulas separate by subject, making the calculations simple and easy to implement even on very modest computational resources. Both MATLAB (MATLAB, 2010) and R (R Core Team, 2012) codes are provided for calculating I2C2 at http://www.biostat.jhsph.edu/~ccrainic/software.html. In practice, one may also be interested in the reliability of imaging in a particular ROI. The formulas for an ROI are almost identical to the ones for the whole image, except that the summation over v is done only within the ROI mask. This is especially useful when one suspects that the reliability of image measurements varies across functional or anatomical area brain regions.

To assess the variability of the I2C2 parameter, a method is proposed to calculate a confidence interval by nonparametrically bootstrapping subjects and applying the same estimation procedure for every bootstrap sample. There are multiple sources of variability for the I2C2 estimator, but the major source will be the limited number of subjects, I, and the imbalance in the number of replicates, where applicable.

Lastly, the distribution of the I2C2 under complete random sampling—that is, no reliability—is investigated. In this case, the model is $W_{ij}(v) = U_{ij}(v)$; and recall that the $U_{ij}(v)$s are independent. Draws from such a null distribution can be realized using permutation sampling. More precisely, all indexes, (i,j), are collected and relabeled as $k_{i,j}$ for $k_{i,j} = 1,\ldots,(\sum_{i=1}^{I} J_i)$. Let $\sigma(k_{i,j})$ be a random permutation obtained by sampling the k-vector without replacement. Denote the image corresponding to $\sigma(k_{i,j})$ by $\widetilde{W}_{ij}(v)$, and estimate the I2C2 coefficient for the model $\widetilde{W}_{ij}(v) = \widetilde{X}_i(v) + \widetilde{U}_{ij}(v)$. Under permutation, the (i,j) pairing does not have the same meaning as before, because the images $\widetilde{W}_{ij}(v)$ are not necessarily from the same subject. By breaking the subject associations via random permutation, a null distribution that is otherwise close to the variation in the data is obtained. Because the number of resamples must be large to minimize Monte Carlo error, for both bootstrapping and permutation testing, the speed of the proposed methods is crucial. Below, we first investigate the "reliability" of this proposed metric in the next section and then show how these quantities can be calculated and used in three different imaging applications in the Methods section.

Simulations

The I2C2 metric is developed on the basis of the assumptions that the signal and noise are independent and normally distributed across repeated measurements. Through extensive simulations, we investigate the effects of various model violations on the preformance of I2C2. In particular, we examine the performance of our algorithm when the model is correctly and incorrectly specified. When the model is misspecified, we study scenarios where (1) replication errors are non-Gaussian, (2) replication errors are correlated over repetitions, and (3) the signal is correlated with the replication errors.

Correctly specified model

Consider the data-generating mechanism $W_{ij}(v) = X_i(v) + U_{ij}(v)$, $i = 1, \cdots, I$; $j = 1, \cdots, J_i$; $v \in V$, where each subject i has J_i images repeatedly measured on a group of voxels V. Let $U_{ij}(v) = V_{ij}(v) + \varepsilon_{ij}(v)$, where $X_i(v)$ and $V_{ij}(v)$ are mutually uncorrelated with smooth covariance operators and $\varepsilon_{ij}(v)$ are the i.i.d. for each voxel, repetition, and subject. Generate $X_i(v) = \mu(v) + \sum_{k=1}^{K_1} \xi_{ik}\phi_k(v)$ and $V_{ij}(v) = \sum_{k=1}^{K_2} \zeta_{ijk}\psi_k(v)$, where $\xi_{ik} \sim N(0, \lambda_k^X)$ and $\zeta_{ijk} \sim N(0, \lambda_k^V)$. To approximate the DTI-MRI example in the Methods section, we set $\mu(v)$ to be the vector obtained by concatenating the population average of corpus callosum images. Let $V = \{v_1, v_2, \cdots, v_V\}$; then, $V = 38 \times 72 \times 11$. We set $K_1 = K_2 = 4$, $\lambda_k^X = 1400 \times 0.5^{k-1}$, and $\lambda_k^V = 840 \times 0.5^{k-1}$, $k = 1,2,3,4$. The eigenfunctions $\phi_k(v)$ and $\psi_k(v)$ are chosen to be orthonormal blocks, as in Zipunnikov et al. (2011). Data were simulated for $I = 200$ subjects, each with $J_i = 2$ replications. By definition, the theoretical I2C2 is $\sum_K^{K_1} \lambda_k^X / \left(\sum_k^{K_1} \lambda_k^X + \sum_K^{K_2} \lambda_k^V + V\sigma^2 \right)$. We show the results for the following distributions of $\varepsilon_{ij}(v)$: Gaussian, heavy-tail t, and

mixture normal with two components. For each scenario, we conduct 100 iterations.

- $\varepsilon_{ij}(v) \sim N(0, \sigma^2)$. The model is correctly specified, and results are highly reliable (see the left panel in Fig. 1). The box plots show the distribution of estimated I2C2 over 100 iterations with respect to a range of signal-to-noise ratios. The red line indicates the theoretical I2C2 values as a function of σ^2.
- $\varepsilon_{ij}(v) \sim t_3/s$, $s = 0.5 \times (1{:}20)$. Here, the t distribution generates measurement errors with a heavy tail distribution and a variance controlled by s. Results are displayed in the right panel of Fig. 1. Performance is very good, although a slight overestimation can be noted in the very low signal-to-noise setting.
- $\varepsilon_{ij}(v) \sim pN(\mu_1, {s_1}^2) + (1-p)N(\mu_2, {s_2}^2)$. This scenario corresponds to the case when measurement error has two possible sources. We simulate the case when the noise distribution is a mixture of two normal components. We consider the following three settings corresponding to three different reliability ratios: (1) $p = 0.8$, $\mu_1 = -0.2$, $\mu_2 = 0.8$, $s_1 = 0.005$, and $s_2 = 0.1$; (2) $p = 0.5$, $\mu_1 = -0.02$, $\mu_2 = 0.02$, $s_1 = 0.02$, and $s_2 = 0.1$; (3) $p = 0.3$, $\mu_1 = -1$, $\mu_2 = 0.43$, $s_1 = 0.05$, and $s_2 = 0.1$. The parameters are chosen so that the distribution of the noise has a mean of 0. The density of selected distributions and the estimated I2C2 under each setting are shown in Fig. 2, indicating excellent performance of the I2C2 estimators.

We conclude that the I2C2 is properly recovered when the model is correctly specified. This is due to the fact that we use a method of moments estimator that is insensitive to the distribution of measurement error.

Misspecified model

When the model assumptions are violated, we show that the estimated I2C2 still reflects the magnitude of reliability. Note that the theoretical I2C2 can be equivalently defined as $\text{I2C2} = \sum_{v\in V} Cov\left\{W_{ij}(v), W_{ij'}(v)\right\} / \sum_{v\in V} Var\left\{W_{ij}(v)\right\}$. xThus, I2C2 is a measure of the fraction of variability that is shared among repeated measurements, without distinguishing whether the correlation is from the signal or the noise. We consider the following scenarios where correlation among images is due not only to the signal, but also to the correlation of replication errors. This violates a basic assumption of measurement, although in the absence of gold standard measurements, it is difficult to determine whether the true errors are correlated.

- Correlated noise across replications. Consider the case where $\varepsilon_{ij}(v) \sim N(0, \sigma^2)$, and $\text{corr}\{\in_{ij}(v), \in_{ij'}(v) = \rho\}$ for every $j \neq j'$. The theoretical I2C2 is $\left(\sum_k^{k_1} \lambda_k^X + V\rho\sigma^2\right) / \left(\sum_k^{k_1} \lambda_k^X + \sum_k^{k_2} \lambda_k^V + V\sigma^2\right)$, which is larger than in the uncorrelated case. Similarly to the previous analysis, we examine the estimated I2C2 with respect to σ^2 and ρ. The mean square errors of the estimated I2C2 under a range of correlations ρ are shown in the left panel of Table 1.

The case where noise variables are not exchangeable is more difficult because defining the true I2C2 becomes tricky. For example, consider the case of AR(1) dependence: that is, $\varepsilon_{ij+1}(v) = \alpha\varepsilon_{ij}(v) + z_{ij+1}(v)$, $\varepsilon_{i1}(v) \sim N(0, \sigma^2)$, and $z_{ij}(v) \sim N(0, (1-\alpha^2)\sigma^2)$ to ensure that $\varepsilon_{ij}(v)'s$ have the same marginal distributions. A possible way to define I2C2 is to start with the pairwise correlations

$$\text{I2C2}_{jj'} = \sum_{v\in V} Cov\left\{W_{ij}(v), W_{ij'}(v)\right\} \Big/ \sum_{v\in V} Var\left\{W_{ij}(v)\right\}^{1/2} Var\left\{W_{ij'}(v)\right\}^{1/2}$$

The true I2C2 could then be defined as the average of all possible pairs $\text{I2C2} = \frac{1}{\binom{J}{2}} \sum_{j<j'} \text{I2C2}_{jj'}$. Although this is a rather contrived example, our simulations indicate good estimation of the I2C2 (results not shown).

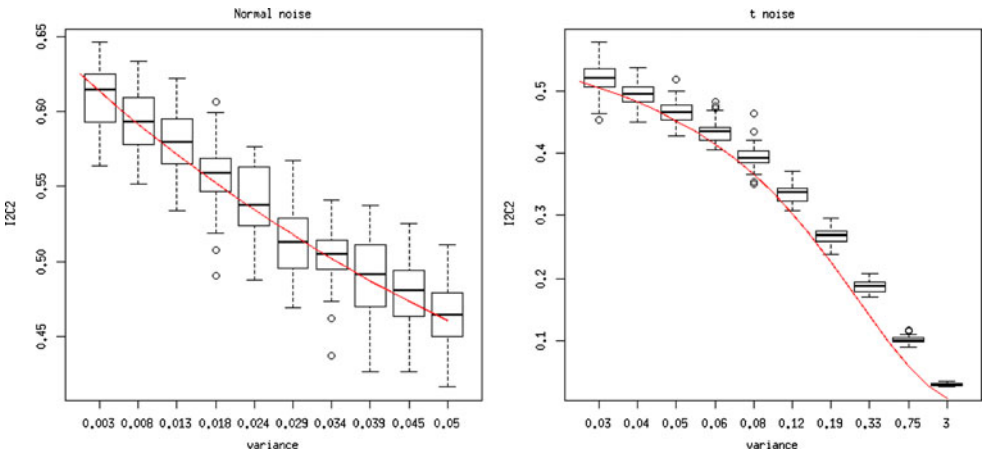

Fig. 1 Left panel: True I2C2 (red line) and estimated I2C2 (box plots over 100 simulations) for $\varepsilon_{ij}(v) \sim N(0, \sigma^2)$ and a range of σ^2. Right panel: True I2C2 (red line) and estimated I2C2 (box plots over 100 simulations) for $\varepsilon_{ij}(v) \sim t_3/s$ and a range of t distribution variances

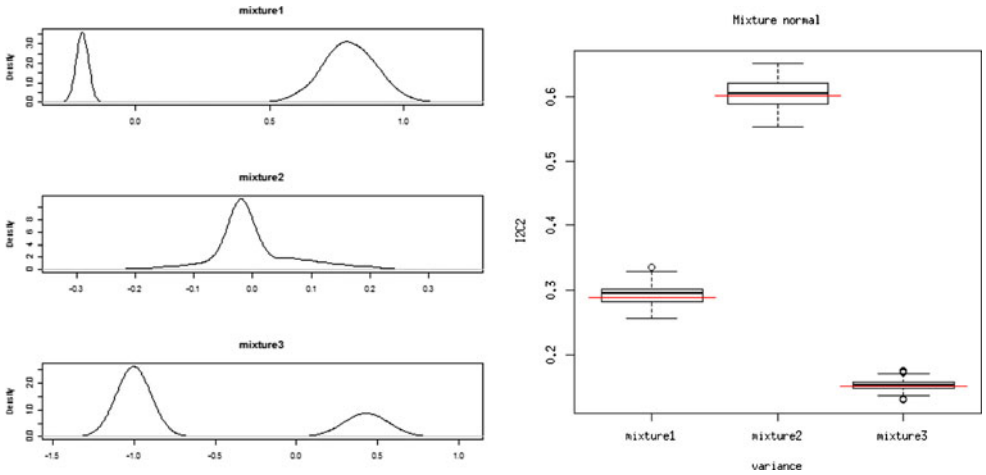

Fig. 2 Left panel: Density plots of the mixture normal distributions used for measurement noise. Right panel: True I2C2 (red lines) and estimated I2C2 (box plots) for the different mixtures of normal distributions

1. Consider the case where the true underlying image intensity is correlated with the magnitude of noise at each voxel. Consider $W_{ij}(v) = \widetilde{X}_i(v) + \widetilde{U}_{ij}(v)$, where $\widetilde{X}_i(v) = X_i(v) + z_i$ and $\widetilde{U}_{ij}(v) = V_{ij}(v) + v_{ij}$ and $X_i(v)$, $V_{ij}(v)$ are generated as in the previous sections. Correlation between signal and noise is modeled using the trivariate normal distribution $N(0,\Sigma)$ for $\{z_i, v_{i1}, v_{i2}\}$, where

$$\sum = \begin{pmatrix} \sigma_x^2 & \rho\sigma_{xu}^2 & \rho\sigma_{xu}^2 \\ \rho\sigma_{xu}^2 & \sigma_u^2 & 0 \\ \rho\sigma_{xu}^2 & 0 & \sigma_u^2 \end{pmatrix}.$$

2. We assume that $\sigma_{xu}^2 = \sigma_x^2$ and $\sigma_u^2 = 5\sigma_x^2$. In this case, the theoretical I2C2 is $\{\sum_k^{K_1}\lambda_k^X + V(1+2\rho)\sigma^2\}/\{\sum_k^{K_1}\lambda_k^X + \sum_k^{K_2}\lambda_k^V + V(6+2\rho)\sigma^2\}$. By varying the correlation ρ, we examine the estimated I2C2 in the right panel of Table 1.

In sum, simulation results demonstrate the robustness of the I2C2 estimation approach when there is a correlation among noise variables or between the signal and the noise. However, it is important to note that I2C2 is not designed to distinguish between these cases and is unbiased with respect to the true correlation; this true correlation may be different from the proportion of variability explained when model assumptions are violated. We now proceed to show how I2C2 can be calculated and used in three different imaging applications.

Methods

RAVENS acquisition

This work employs the "multimodal MRI reproducibility resource" (Landman et al., 2011), colloquially known as the Kirby21 data set, which is publicly available through the Neuroimaging Informatics Tools and Resources Clearinghouse (www.nitrc.org). The Kirby21 data set consists of test–retest structural MRI and rs-fMRI scans from 21 healthy adult volunteers with no history of neurological conditions (11 male and 10 female; 31.76±9.47 years of age) who were each scanned twice on the same day. Further details of the study can be found in Landman et al. (2011).

The structural MRI data were acquired on a 3.0T scanner (Achieva, Philips Medical Systems) using a high-resolution 3-D magnetization-prepared rapid acquisition of gradient echoes sequence with a resolution of 1.0 × 1.0 × 1.2 mm; TR,~6.7 ms; TE, 3.1 ms; TI, 842 ms; flip angle, 8°; SENSE factor, 2. All images were spatially normalized via registration of T1 maps

Table 1 Mean square errors (*MSE*s) of the estimated image intraclass correlation coefficients (I2C2s) under a range of correlations, both for correlated noise case and for correlated signal and noise

	Correlated noise				Correlated signal and noise			
ρ	0.11	0.42	0.74	0.89	0.11	0.42	0.74	0.89
True I2C2	0.41	0.54	0.67	0.74	0.27	0.33	0.37	0.40
Estimated I2C2	0.41	0.54	0.67	0.74	0.29	0.33	0.38	0.41
MSE	2.95e-4	2.08e-4	2.21e-4	1.66e-4	2.91e-3	3.35e -3	3.55e-3	2.82e-3

into the mean template generated using ANTS (Avants et al., 2011; Avants et al., 2010). Details of how the average template is generated can be found in Chen et al. (2012). All T1 images were segmented into ventricles (VNs), gray matter (GM), and white matter (WM) using Lesion-TOADS (Shiee et al., 2010). After segmentation, the final tissue maps of VNs, WM, and GM were spatially normalized using the HAMMER-SUITE (Shen & Davatzikos, 2002) to generate RAVENS images. Finally, the RAVENS maps were smoothed individually with a 4-mm FWHM Gaussian kernel using SPM8.

fMRI acquisition

The Kirby21 data set was also used to investigate the reproducibility of seed-based functional connectivity analysis as follows. In short, two 7-min resting state scans were acquired from each subject using a single-shot, partially parallel (SENSE) gradient-recalled echo planar sequence with an ascending slice order (TR/TE, 2000/30 ms; FA, 75; 3-mm axial slices with a 1-mm slice gap) and an 8-channel head coil. Subjects were instructed to relax and fixate on a cross-hair while remaining as still as possible. The two resting-state scans were separated by a short break, during which the subject exited the scanner; the T1-weighted anatomical image described in the RAVENS acquisition section was also acquired to be used as a template for spatial registration of the functional images.

Image processing was performed using SPM8 and custom MATLAB scripts. Anatomical images were registered to the first functional volume and normalized to MNI space using unified segmentation/normalization (SPM8). Functional data were adjusted for slice time acquisition, as well as subject motion, and were transformed to MNI space. Nuisance covariates from white matter and CSF were estimated using CompCor (Behzadi, Restom, Liau, & Liu, 2007) and regressed from the data along with the motion realignment estimates, their derivatives, global mean signal, and linear trends. Data were then spatially smoothed (6-mm kernel) and temporally filtered using a 0.01–0.10 band-pass filter. Data from one subject was excluded from analysis due to a misalignment of the first and second resting-state scans.

Seed voxel analysis is commonly used in fMRI studies to analyze the functional connectivity of the brain via a seed voxel from an ROI (Lindquist, 2008). Here, we investigated the reproducibility of this approach for our data set considering four different seeds, each with a 6-mm radius: the posterior cingulate cortex (labeled PCC; Fox et al., 2005), the premotor area (labeled M3) (Chouinard & Paus, 2006), and two seeds from the dorsal–ventral extremes of the motor strip, the dorsal seed representing lower limb control (labeled M1; Meier, Afalo, Kastner, & Graziano, 2008) and the ventral one corresponding to oro-motor function (labeled M5). Within each seed, fMRI time series were averaged across voxels, and a correlation map for each of the resulting four time courses was then obtained with each voxel in the brain.

DTI-MRI acquisition

The data were collected as part of an ongoing observational study being conducted at the National Institutes of Health and at Johns Hopkins University. Study subjects with multiple sclerosis (MS) were recruited from the outpatient neurology clinic and healthy volunteers from the community. Prior to MRI scanning, all subjects gave signed, informed consent, and all procedures were approved by the institutional review board. Cohort characteristics are summarized in Reich, Ozturk, Calabresi, and Mori (2010); Goldsmith, Crainiceanu, Caffo, & Reich (2011). Longitudinal analyses of the DTI-MRI substudy can be found in Greven, Crainiceanu, Caffo, and Reich (2010); Zipunnikov et al. (2012).

Scans were performed on a 3T scanner (Intera; Philips, Best, The Netherlands) over a 4.6-year period, using the body coil for transmission and either a 6-channel head coil or the eight head elements of a 16-channel neurovascular coil for reception (both coils are made by Philips). Each session included two sequential DTI scans using a conventional spin-echo sequence and a single-shot EPI readout. Whole-brain data were acquired in nominal 2.2-mm isotropic voxels with the following parameters: TE, 69ms; TR, automatically calculated (shortest); slices, 60 or 70; parallel imaging factor, 2.5; noncollinear diffusion directions, 32 (Philips overplus high scheme); high *b*-value, 700 s/mm^2; low *b*-value (*b*0), approximately 33 s/mm^2; repetitions, 2; reconstructed in-plane resolution, 0.82 × 0.82 mm. A 3-D gradient-echo magnetization-transfer sequence was also performed with segmented EPI readout (nominal acquired resolution, 1.5 × 1.5 × 2.2 mm; TE, 15ms; TR, 64 ms; parallel imaging factor, 2; EPI factor, 7; magnetization-transfer pulse, sinc-shaped, 1.5kHz off-resonance; repetitions, 3), the data from which were rigidly registered to the DTI scan before calculation of magnetization transfer ratio (MTR) maps (defined as 1 minus the voxel-wise ratio of data from this sequence to those obtained using the same sequence without the magnetization-transfer pulse). Prior to analysis, data were adjusted to account for changes in average tract-specific MRI indices that resulted from the scanner upgrades that inevitably occur over the course of a study such as this. The procedure by which this adjustment was made has been previously described (Harrison et al., 2011).

The diffusion-weighted scans were processed using CATNAP (Landman et al., 2007) to create maps of FA, mean diffusivity (MD), axial diffusivity (AD), and radial diffusivity (RD). These four quantities, together with MTR, are hereafter termed MRI indices. Whole-brain MRI indices were calculated by slice-wise averaging of all diffusion-weighted images, removal of the low-intensity voxels that are characteristic of extracerebral tissues on these images, and final removal of

voxels with MD >1.7μm^2/ms to exclude cerebrospinal fluid (Ozturk et al., 2010). The resulting brain mask was applied to all DTI maps and also to the coregistered MTR maps. The images were obtained from a natural history study where 176 MS patients were followed for up to 5.5 years, which generated a total of 446 MRI scans. The number of scans per subject varied from one to six. The scanning time is shown in Fig. 5, where time zero indicates the first scan. For illustration purposes, we focus on the measurements in a region of 30, 096 voxels that contains the corpus callosum. At each voxel, data are FA weighted by the probability of being in the corpus callosum. Images are registered using affine transformations.

Results

RAVENS replication results

RAVENS maps produce an image of the deformation of the brain necessary to fit in a given template and are proxies of brain morphology. Here, the focus is on ventricular, WM, and GM regions considered separately, segmented via Lesion-TOADS (Shiee et al., 2010). The measurement error is an uncontrollable combination of sources, including image acquisition, biological error (natural within-day brain variation), movement, magnetic field inhomogeneities, preprocessing, spatial normalization, and segmentation. Apportioning error variability is beyond the scope of this article. Instead, interest lies, first, in establishing that estimating the effect of total measurement error variability (regardless of its source) is possible and, then, in investigating its impact on image reliability.

Figure 3 displays the I2C2 estimators ($\widehat{\rho}$) as a red line with 95% equal tail probability confidence intervals obtained using the nonparametric bootstrap of subjects. The reliability in the VNs is by far the largest roughly (.9), followed by reliability in white matter (.55) and gray matter (.45). Determining the source and type of error could be done, for example, by investigating various ROIs or by inspecting the principal components of measurement error variability based on HD-MFPCA (Zipunnikov et al., 2011). The distributions of I2C2 estimators under zero reliability $\widehat{\rho}_0$ is shown in gray, with the median displayed as a black horizontal line. These results indicate strong evidence that the observed reliability values are inconsistent with zero reliability. Interestingly, the null distribution (gray histogram plot) for VNs has a long right tail, with a nontrivial probability above .3. This is somewhat unexpected and may indicate stronger between-subjects correlations of measurement error processes in the VNs. Further investigation of this postulate is left for future study.

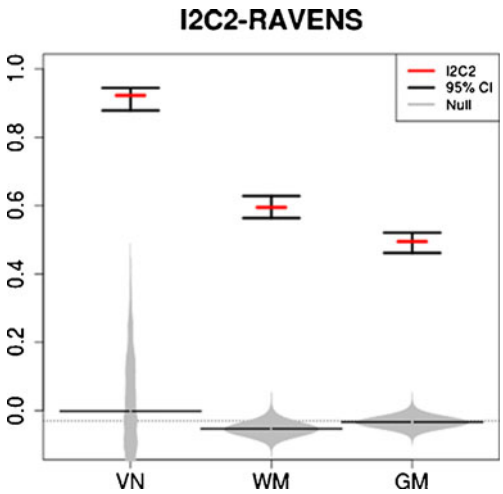

Fig. 3 Estimated I2C2 (red horizontal lines) and 95% equal tail probability confidence intervals for ventricles, white matter, and gray matter RAVENS images. Gray distributions correspond to the I2C2 estimator under the zero reliability assumption (random permutations of labels)

fMRI replication results

The I2C2 metric was used to quantify the reproducibility of the resulting connectivity map (correlation matrix) for each of the four seed regions. Results are shown in Fig. 4, using the same notation and symbols as in Fig. 3. The overall message is that the seed-voxel-based correlation maps are not reliable, with the reliability estimates varying between approximately .20 (for M1, M3, and M5) and .37 (for the PCC). These low values suggest that state-of-the-art seed-voxel-based correlation maps based on rs-fMRI data are unreliable, although the PCC seems to indicate higher (nearly double) reliability than do other regions. Thus, caution is warranted in the interpretation of these maps and in the analysis of connectivity maps obtained from thresholding unreliable fMRI resting-state correlation

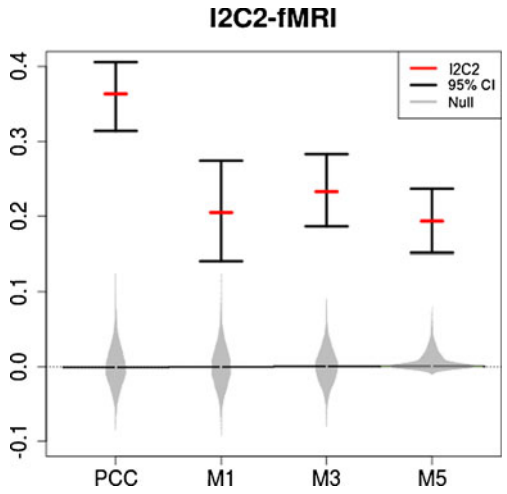

Fig. 4 Estimated image intraclass correlation coefficients (I2C2s) (red horizontal lines) and 95% equal tail probability confidence intervals for fMRI seed-voxel correlation maps for the posterior cingulate cortex (PCC), the dorsal region of the motor cortex corresponding to control of the lower limbs (M1), the premotor cortex (M3), and the ventral-most region of the motor cortex corresponding to oro-motor function (M5). Gray distributions correspond to the I2C2 estimator under the zero reliability assumption (random permutations of labels)

operators. These results are inconsistent with the large and increasing literature (Braun et al., 2012; Chen et al., 2008; Damoiseaux et al., 2006; Honey et al., 2009; Meindl et al., 2010; Schwarz & McGonigle, 2011; Shehzad et al., 2009; Wang et al., 2011; Zhang et al., 2011; Zuo, Di Martino, et al., 2010; Zuo, Kelly, et al., 2010) on rs-fMRI that reports high reliability of measurements. Much deeper investigation is needed to address these divergent findings, establish identical estimands, estimators, and evaluation procedures. Our procedure provides a clear, simple, and easy-to-use step in this direction.

DTI-MRI replication results

To highlight methods, a subset of the complete data collection (Fig. 5) consisting of subjects who have more than six visits was selected. This reduced the data set to 117 scans from 18 subjects: 14 subjects with 6 scans, 1 with 7, 2 with 8, and 1 with 10. Henceforth, the subset is viewed as the complete data set, with no further reference of the omitted subjects. We also consider four further subsets labeled as "T≤4," "T≤3," "T≤2," and "T≤1." The notation refers to the number of years since the baseline scan, as, for example, the T≤4 data set considers only images obtained within the first 4 years from the baseline scan, resulting in 110 scans from the 18 subjects (4∼5, 11∼6, 3∼8 where 4∼5 refers to 4 subjects with 5 scans). The T≤3 data set contains 88 scans broken down as 6∼4, 9∼5, 2∼6, and 1∼7. The T ≤ 2 data set contains 70 scans broken down as

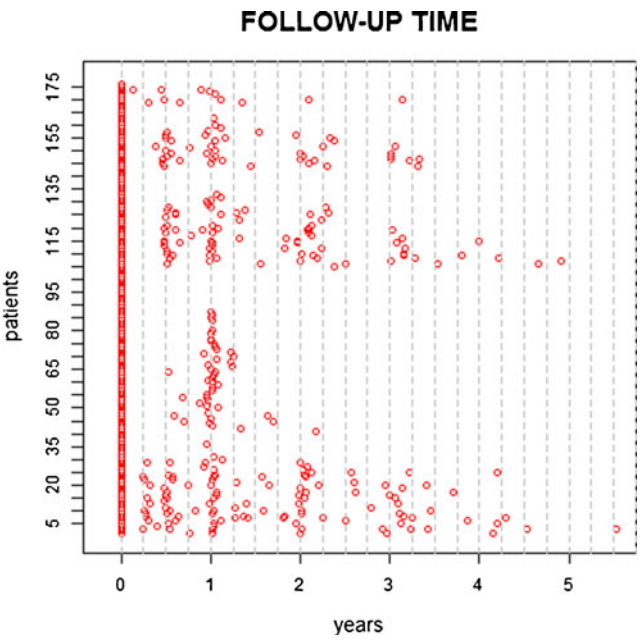

Fig. 5 Image scanning time for 176 patients. Every person has a baseline scan at time 0. The *x*-axis is time in years. The *y*-axis is patient IDs. We match visit number from different patients by rounding their scan time to quarter month, as indicated by gray dashed lines

7∼3, 7∼4, 3∼5, 1∼6, and 1∼7. Finally, the T ≤ 1 data set contains 45 scans, 1∼1, 1∼4, 8∼2, and 8∼3.

In Zipunnikov et al. (2012), the existence of a longitudinal change over time in these data was studied, with the finding that less than 1% of the variability was explained by longitudinal within-subjects changes. Thus, modeling these data as exchangeable image measurement error processes is likely a valid approximation of the underlying processes. All five data sets are unbalanced, having a different number of replicates per subject. The left panel in Fig. 6 displays the reliability estimators (red horizontal line) and the associated equal tail probability 95% confidence intervals. These results indicate that the reliability of these measurements hovers slightly below .8, which is consistent with the findings in Zipunnikov et al. (2012).

Our work investigated the reliability of the imaging studies as a function of time by selecting subjects who have at least two replications and constructing five additional replication substudies labeled "1 apart," "2 apart," "3 apart," "4 apart," and "5 apart," respectively. To be specific, each such substudy contains exactly two replicates per subject: the baseline observation and the replicate that is closest to being 1, 2, 3, 4, or 5 years apart, respectively. The number of subjects in each data set was 119, 64, 49, 31, and 18, respectively, with more subjects in data sets with shorter between-observation intervals.

The right panel in Fig. 6 displays the reliability estimators for these replication studies as a function of how many years apart images were taken. The estimated reliability of observations taken within 1 year of each other is quite high, roughly .9, which indicates that there are very few changes in the FA measurements along the corpus callosum of MS subjects within 1 year. This may be good news for individuals with MS if the lack of measured neuronal fiber integrity via FA represents actual fiber integrity. However, this finding may be disheartening to investigators searching for biomarkers of neuronal fiber degradation, if degradation is actually there. As was expected, the reliability of image replication decreases with the increased time between visits, with median reliability roughly around .8 for images collected 5 years apart. However, this decline in reliability is relatively small and likely to be indicative of small observable longitudinal changes. The variability around the estimated I2C2 also increases from the replication study "1 apart" to "5 apart," although this is most likely due to the decrease in sample size from 119 to 18 subjects with repeat samples.

Discussion

This article proposes an extension of the classical ICC coefficient to image replication studies. The resulting parameter, denoted I2C2, provides a global measurement of reliability that is intuitive and easy to calculate. Moreover, I2C2 can readily be calculated for given ROIs by simply restricting the summations in the Introduction to those voxels within the ROI

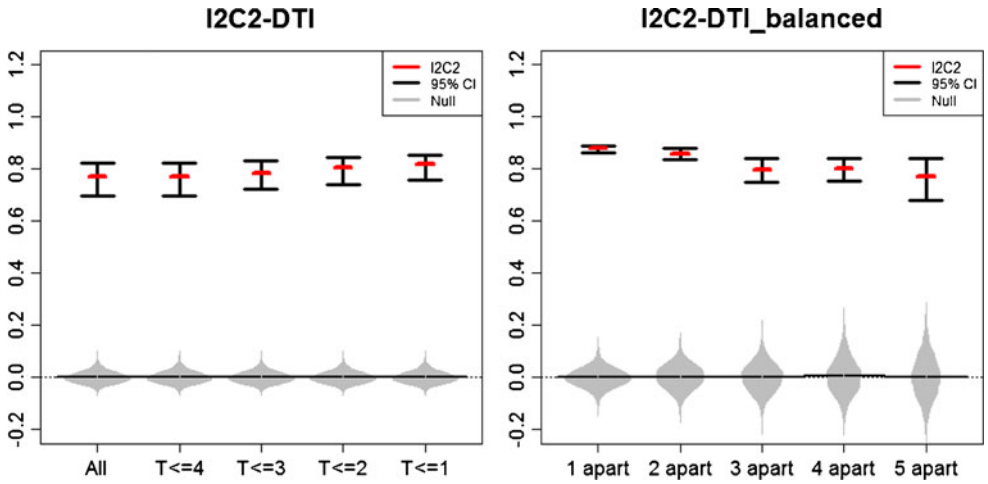

Fig. 6 Estimated image intraclass correlation coefficients (I2C2s; red horizontal lines) and 95% equal tail probability confidence intervals for fractional anisotropy in an area containing the corpus callosum. Gray distributions correspond to the I2C2 estimator under the zero reliability assumption (random permutations of labels). Left panel results are based on 18 subjects who have at least six visits ("All") and subsets of the "All" data set containing all scans within the first 4, 3, 2, and 1 year from baseline, respectively. Right panel results are based on pairs of imaging obtained, at most, 1, 2, 3, 4, and 5 years apart. The number of subjects in each data set (from left to right) was 119, 64, 49, 31, and 18, respectively

mask. In practice, one may actually report the I2C2 on a partition of the image in mutually disjoint ROIs—say, $R_1,\ldots,R_P$. Then I2C2 can be calculated for each R_p, $p = 1,\ldots,P$, and compared with the overall I2C2. Areas of unexpectedly small estimated I2C2 may further indicate the source and type of measurement error. Another practical approach would be to calculate the I2C2 hierarchically—that is, at the voxel level, then at overlapping neighborhoods of increasing size and, ultimately, at the image level. This could provide an interesting multiresolution approach to visualizing the structure of the measurement error.

An equally simple measure of reproducibility could be the average of ICC at the voxel levels. An unbiased estimator of the average ICC would then be

$$1-\frac{1}{V}\frac{\sum_i J_i-1}{\sum_i (J_i-1)}\sum_{v=1}^{V}\frac{\left\{W_{ij}(v)-\overline{W}_{i.}(v)\right\}^2}{\sum_{i=1}^{I}\sum_{j=1}^{J_1}\left\{W_{ij}(v)-W_{..}(v)\right\}^2}.$$

Irrespective of the replication estimand and estimation procedure, the subject-level bootstrap and permutation tests introduced in this article can be applied. However, there are reasonable arguments for preferring the I2C2 to the average ICC value. Indeed, the variability attributable to variation among subjects is equal to trace(K_X), whereas the variability attributable to visits is trace(K_U). Thus, I2C2 is the proportion of variability explained by subject-level variability out of the total variability of the data in the *multivariate image measurement error* model. In contrast, the average ICC is the average of the proportion of variability explained by subject-level variability out of the total variability of the data in the sequence of *univariate* (*marginal*) *measurement error* models. This distinction has practical implications. Consider, for example, the case where there are 1,000 voxels in every image. At 500 voxels, the absolute variability of the data and reliability are very low. However, at the other 500 voxels, the variability and reliability are large. In this context, the average ICC would place too much emphasis on the low-variability voxels, because it ignores the *relative variability* of the data at different voxels. A second problem occurs at locations with small visit-to-visit variability, since this variance is used in the denominator of the ICC estimator and may lead to serious computational instabilities.

While data rarely satisfy the measurement error model (Equation 2) exactly, the model is a reasonable starting point for defining the data structure under explicit assumptions. Model assumptions notwithstanding, we prefer this explicit statistical approach to an algorithmic one that obscures assumptions. Moreover, the model can easily be extended to include some obvious data-supported complications. For example, if each visit has a different mean, one can easily expand the model to include (so-called) batch or visit effects,

$$W_{ij} = B_j + X_i + U_{ij},$$

as proposed in Di et al. (2009). Here, the images B_j are visit-specific fixed effect images. Such *deterministic changes across all subjects from one visit to another* could be due to the use of different scanners, imaging parameters, scanner drift, and so forth. In quality control, agriculture, and lab sciences, such effects arise from a batch being run for measurement or assay (hence, the term "batch effect"). For subjects returning to a scanner, batches are visits. Note that the visit-specific effects can be easily estimated as $\widehat{B}_j = \sum_{i=1}^{I} W_{ij}/I$ and one can define the I2C2 for the residuals $W_{ij}-\widehat{B}_j$.

In more complex models, one may also be interested in, or worried about, the longitudinal effects of collecting the data. For example, in the DTI study, some images are taken within a few months of each other, whereas other images are collected years apart. In such situations, it is reasonable to add a term that accounts for longitudinal changes. A reasonable model for such an approach could be

$$W_{ij} = B(T_{ij}) + X_{i,0} + X_{i,0}T_{ij} + U_{ij},$$

where $B(T_{ij})$ is an effect that depends on time of the visit, T_{ij}, as in most longitudinal studies, visits are not equally spaced. In this model, $B(T_{ij}) + X_{i,0}$ is the true unobserved image at baseline ($T_{ij} = 0$), $B(T_{ij}) + X_{i,0} + X_{i,0}T_{ij}$ is the true unobserved image at time $T_{ij} > 0$, and U_{ij} is the image measurement error process. Estimation of these types of models is thoroughly discussed in Greven et al. (2010; Zipunnikov et al., 2012), but it is worth noting that reasonable assumptions about the data can easily be incorporated into statistical models.

Regardless of the model under investigation, the image error process, U_{ij}, deserves particular attention. Indeed, from all the models discussed in this article, one can estimate the covariance operator, K_U, and the first eigenvectors can be visually inspected. This provides clues into the structure of measurement error. For further reading on measurement error modeling, we recommend Carroll et al. (2006; Fuller, 1987). For the effect of image measurement error on estimating associations with outcomes, we recommend Crainiceanu, Staicu, and Di (2009), while for inference in the means of two imaging processes, we recommend Crainiceanu, Staicu, Ray, and Punjabi (2012).

Author Note This research was supported by grant R01NS060910 from the National Institute of Neurological Disorders and Stroke and by grants R01EB012547 and P41EB015909 from the National Institute of Biomedical Imaging **a**nd Bioengineering. This work represents the opinions of the researchers, and not necessarily that of the granting organizations. The authors would like to thank Dr. Daniel Reich from NIH/NINDS, Dr. Peter Calabresi, and their research teams for collecting and sharing the DTI-MRI data sets, as well as Ronald Caffo for assistance with copy editing.

References

Avants, B. B., Tustison, N. J., Song, G., Cook, P. A., Klein, A., & Gee, J. C. (2011). A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage, 54*(3), 2033–2044.

Avants, B. B., Yushkevich, P., Pluta, J., Minkoff, D., Korczykowski, M., Detre, J., & Gee, J. C. (2010). The optimal template effect in hippocampus studies of diseased populations. *NeuroImage, 49*(3), 2457–2466.

Behzadi, Y., Restom, K., Liau, J., & Liu, T. T. (2007). A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage, 37*(1), 90.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *The Year in Cognitive Neuroscience, 1191,* 133–155.

Braun, U., Plichta, M. M., Esslinger, C., Sauer, C., Haddad, L., Grimm, O., ... Meyer-Lindenberg, A. (2012). Test-retest reliability of resting-state connectivity network characteristics using fMRI and graph theoretical measures. *NeuroImage, 59,* 1404–1412.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: A modern perspective*. New York: Chapman & Hall/CRC.

Chen, M., Lee S., Carass, A., Reich, D., Pham, D., & Prince, J. (2012). High dimensional statistical deformation modeling for characterizing brain morphology in multiple sclerosis.

Chen, S., Ross, T. J., Zhan, W., Myers, C. S., Chuang, K. S., Heishman, S. J., ... Yang, Y. (2008). Group independent component analysis reveals consistent resting-state networks across multiple sessions. *Brain Research, 1239,* 141–151.

Chouinard, P. A., & Paus, T. (2006). The primary motor and premotor areas of the human cerebral cortex. *The Neuroscientist, 12*(2), 143–152.

Crainiceanu, C. M., Staicu, A. M., & Di, C. (2009). Generalized multilevel functional regression. *Journal of the American Statistical Association, 104*(488), 177–194.

Crainiceanu, C. M., Staicu, A. M., Ray, S., & Punjabi, N.M. (2012). Bootstrap-based inference on the difference in the means of two correlated functional processes. *Statistics in Medicine*, *31*(26).

Damoiseaux, J. S., Rombouts, S. A., Barkhof, F., Scheltens, P., Stam, C. J., Smith, S. M., & Beckmann, C. F. (2006). Consistent resting-state networks across healthy subjects. *Proceedings of the National Academy of Sciences of the United States of America, 103,* 13848–13853.

Davatzikos, C., Genc, A., Xu, D., & Resnick, S. M. (2001). Voxel-based morphometry using the ravens maps: Methods and validation using simulated longitudinal atrophy. *NeuroImage, 14*(6), 1361–1369.

Di, C., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Annals of Applied Statistics, 3*(1), 458–488. Online access 2008.

Fox, M. D., Snyder, A. Z., Vincent, J. L., Corbetta, M., Van Essen, D. C., & Raichle, M. E. (2005). The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proceedings of the National Academy of Sciences of the United States of America, 102*(27), 9673–9678.

Fuller, W. (1987). *Measurement error models*. New York: John Wiley & Sons.

Goldsmith, A. J., Crainiceanu, C. M., Caffo, B. S., & Reich, D. (2011). Penalized functional regression analysis of white-matter tract profiles in multiple sclerosis. *NeuroImage, 57*(2), 431–439.

Greven, S., Crainiceanu, C. M., Caffo, B. S., & Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics, 4,* 1022–1054.

Harrison, D. M., Caffo, B. S., Shiee, N., Farrell, J. A. D., Bazin, P.-L., Farrell, S. K., ... Reich, D. S. (2011). Longitudinal changes in diffusion tensor-based quantitative mri in multiple sclerosis. *Neurology*(76).

Honey, C. J., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J. P., Meuli, R., & Hagmann, P. (2009). Predicting human resting-state functional connectivity from structural connectivity. *Proceedings of the National Academy of Sciences of the United States of America, 106,* 2035–2040.

Landman, B. A., Farrell, J. A., Jones, C. K., Smith, S. A., Prince, J. L., & Mori, S. (2007). Effects of diffusion weighting schemes on the reproducibility of DTI-derived fractional anisotropy, mean diffusivity, and principal eigenvector measurements at 1.5T. *Neuroimage, 36*, 1123–1138.

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A., Farrell, J. A., ... van Zijl, P. C. (2011). Multi-parametric neuroimaging reproducibility: A 3-T resource study. *NeuroImage, 54*(4), 2854–2866.

Lindquist, M. A. (2008). The statistical analysis of fMRI data. *Statistical Science, 23*(4), 439–464.

MATLAB (2010). *version 7.10.0* (*R2010a*). The MathWorks Inc., Natick, Massachusetts.

Meier, J. D., Afalo, T. N., Kastner, S., & Graziano, M. S. A. (2008). Complex organization of human primary motor cortex: A high-resolution fmri study. *Journal of Neurophysiology, 100*(4), 1800–1812.

Meindl, T., Teipel, S., Elmouden, R., Mueller, S., Koch, W., Dietrich, O., ... Glaser, C. (2010). Test-retest reproducibility of the default-mode network in healthy individuals. *Human Brain Mapping, 31,* 237–246.

Ozturk, A., Smith, S. A., Gordon-Lipkin, E. M., Harrison, D. M., Shiee, N., Pham, D. L., ... Reich, D. S. (2010). MRI of the corpus callosum in multiple sclerosis: Association with disability. *Multiple Sclerosis*(16).

R Core Team (2012). *R*: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Reich, D. S., Ozturk, A., Calabresi, P. A., & Mori, S. (2010). Automated vs. conventional tractography in multiple sclerosis: Variability and correlation with disability. *NeuroImage, 49*(4), 3047–3056.

Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., & Scheltens, P. (1998). Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magnetic Resonance Imaging, 16,* 105–113.

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.

Shehzad, Z., Kelly, A. M., Reiss, P. T., Gee, D. G., Gotimer, K., Uddin, L. Q., ... Milham, M. P. (2009). The resting brain: Unconstrained yet reliable. *Cerebral Cortexortex, 19,* 2209–2229.

Shen, D., & Davatzikos, C. (2002). HAMMER: Hierarchical attribute matching mechanism for elastic registration. *Medical Imaging, IEEE Transactions On, 21*(11), 1421–1439.

Shiee, N., Bazin, P. L., Ozturk, A., Reich, D. S., Calabresi, P. A., & Pham, D. L. (2010). A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. *NeuroImage, 49*(2), 1524–1535.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420–428.

Strother, S. C., Anderson, J., Hansen, L. K., Kjems, U., Kustra, R., Sidtis, J., ... Rottenberg, D. (2002). The quantitative evaluation of functional neuroimaging experiments: The NPAIRS data analysis framework. *NeuroImage, 15,* 747–771.

Wang, J.-H., Milham, S., Zuo, M. P., Gohel, X.-N., & Biswal, B. B. (2011). Graph theoretical analysis of functional brain networks: Test-retest evaluation on short- and long-term resting-state functional MRI data. *PloS one, 6,* 2209–2229.

Zhang, H., Duan, L., Zhang, Y. J., Lu, C. M., Liu, H., & Zhu, C. Z. (2011). Test-retest assessment of independent component analysis-derived resting-state functional connectivity based on functional near-infrared spectroscopy. *NeuroImage, 55,* 607–615.

Zipunnikov, V., Caffo, B. S., Yousem, D. M., Davatzikos, C., Schwartz, B. S., & Crainiceanu, C. M. (2011). Multilevel functional principal component analysis for high dimensional data. *Journal of Computaional and Graphical Statistics, 20*(4), 852–873.

Zipunnikov, V., Caffo, B. S., Yousem, D. M., Davatzikos, C., Schwartz, B. S., & Crainiceanu, C.M. (2012). Longitudinal high dimensional data analysis. *Technical report*.

Zuo, X. N., Di Martino, A., Kelly, C., Shehzad, Z. E., Gee, D. G., Klein, D. F., ... Milham, M. P. (2010). The oscillating brain: Complex and reliable. *NeuroImage, 49,* 1432–1445.

Zuo, X. N., Kelly, C., Adelstein, J. S., Klein, D. F., Castellanos, F. X., & Milham, M. P. (2010). Reliable intrinsic connectivity networks: Test-retest evaluation using ICA and dual regression approach. *NeuroImage, 49,* 2163–2177.