NeuroImage 69 (2013) 231–243

Contents lists available at SciVerse ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/ynimg

Single subject fMRI test–retest reliability metrics and confounding factors

Krzysztof J. Gorgolewski [a,b,*], Amos J. Storkey [c], Mark E. Bastin [b], Ian Whittle [d], Cyril Pernet [b]

[a] Neuroinformatics Doctoral Training Centre, University of Edinburgh, UK
[b] Brain Research Imaging Centre, a SINAPSE Collaboration centre, University of Edinburgh, UK
[c] Institute for Adaptive and Neural Computation, University of Edinburgh, UK
[d] Division of Clinical Neurosciences, University of Edinburgh, UK

ARTICLE INFO

Article history:
Accepted 30 October 2012
Available online 13 November 2012

Keywords:
fMRI
Reliability
Single subject
Test–retest
Dice
T value variance
Time-series correlation

ABSTRACT

While the fMRI test–retest reliability has been mainly investigated from the point of view of group level studies, here we present analyses and results for single-subject test–retest reliability. One important aspect of group level reliability is that not only does it depend on between-session variance (test–retest), but also on between-subject variance. This has partly led to a debate regarding which reliability metric to use and how different sources of noise contribute to between-session variance. Focusing on single subject reliability allows considering between-session only. In this study, we measured test–retest reliability in four behavioural tasks (motor mapping, covert verb generation, overt word repetition, and a landmark identification task) to ensure generalisation of the results and at three levels of data processing (time-series correlation, *t* value variance, and overlap of thresholded maps) to understand how each step influences the other and how confounding factors influence reliability at each of these steps. The contributions of confounding factors (scanner noise, subject motion, and coregistration) were investigated using multiple regression and relative importance analyses at each step. Finally, to achieve a fuller picture of what constitutes a reliable task, we introduced a bootstrap technique of within- vs. between-subject variance. Our results show that (i) scanner noise and coregistration errors have little contribution to between-session variance (ii) subject motion (especially correlated with the stimuli) can have detrimental effects on reliability (iii) different tasks lead to different reliability results. This suggests that between-session variance in fMRI is mostly caused by the variability of underlying cognitive processes and motion correlated with the stimuli rather than technical limitations of data processing.

Introduction

For the past twenty years, the tool of choice for non-invasive study of human mind/brain relationships has been functional Magnetic Resonance Imaging (fMRI). Despite the fact that it has been used in thousands of studies, many of which have been independently replicated, there is as yet no consensus on how reliable fMRI measurements are (Bennett and Miller, 2010). At the same time it is widely accepted that fMRI can provide valuable insights into the human brain even when used on the single subject level. In other words, the result of analysing fMRI time-series is not random. However, it is also accepted that there is some variability in the results that cannot be accounted for by experimental variables. Understanding this variability of fMRI is crucial to delineating limits of fMRI as a research tool.

The pursuit of scientific truth is not the only motivation behind understanding the reliability of fMRI. Shortly after its inception fMRI was adapted for clinical use. For example, presurgical mapping for tumour or epilepsy foci extraction is being performed on a regular basis in a number of medical centres (Stippich et al., 2007). Neurosurgeons appreciate the advantages of fMRI, but to be able to use this data responsibly they have to understand its limitations. It is worth noting, however, that single subject fMRI is not limited to presurgical mapping. It potentially can be used as a diagnostic tool (Raschle et al., 2012) and a way to plan and monitor rehabilitation (Dong et al., 2011). It is also being used to define individual functional regions of interest (ROIs) through functional localiser tasks (Duncan et al., 2009).

The change of focus in single subject studies is reflected in a different approach to analysing data. The Holmes–Friston (Holmes and Friston, 1998) approach discards uncertainty of the first level analysis and the within-subject variance, by using each subject's contrast maps instead of t maps. The uncertainty that influences the group level results comes from the between-subject variance. In contrast, a single subject examination relies on *t* maps, instead of beta parameter maps, and thus depends on within-subject variance. This difference between which variance is relied upon has implications for what levels and metrics of reliability are suitable for group and single subject analyses. For group studies, it is reasonable to look at the within- and between-session variance of contrast maps as well as

* Corresponding author at: Neuroinformatics and Computational Neuroscience, Doctoral Training Centre, School of Informatics, University of Edinburgh, 10 Crichton Street, Edinburgh, EH8 9AB, UK.
E-mail address: krzysztof.gorgolewski@gmail.com (K.J. Gorgolewski).

1053-8119/$ – see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.neuroimage.2012.10.085

the similarity of thresholded and unthresholded group level *t*-maps. In contrast, for single subject studies, this is the within- and between-session variance of the BOLD signal and the similarity of *t* maps that are relevant.

Volume overlap is a simple measure to quantify reliability that assesses how many of the suprathreshold voxels from many *t* maps/sessions occur in the same location. Depending on the normalisation factor there are different variants of the overlap metric; the most common are Dice (1945) and Jaccard (1901). This method has the advantage of examining the final product of the neuroimaging analysis, the t maps, and the same procedure applies to group or single subject maps. However, overlap values heavily depend on the threshold applied to the *t* maps, since the cluster overlap measures decrease with increasing threshold (Duncan et al., 2009; Fernández et al., 2003). Additionally, overlap scores are by definition dependent on the volume of activation and when used over the whole brain rather than for a specific cluster of interest, will give higher values. Worst, when different thresholds are used over a large volume different activation maps can be obtained, but similar measures of overlap can be observed. Finally, this technique is sensitive to borderline cases; two very similar t maps, one slightly above a threshold and another slightly below, would give a false impression of high variability (Smith et al., 2005). Nonetheless, thresholded maps are the typical end product of fMRI analyses and are used for ROI definitions. Furthermore, in the clinical context where single subject thresholded maps are used, their variability is a major concern.

Another popular metric to assess reliability is the Intraclass Correlation Coefficient (ICC). ICC was initially used in psychology to asses between raters variability (Shrout and Fleiss, 1979), but has been adapted to measure reliability (McGraw and Wong, 1996) by replacing judges/raters by repeated measurement sessions. One of the most commonly used ICC variants in fMRI is ICC(3,1), a two-way model (subjects vs. sessions) with no interaction and a consistency criteria; in other words allowing for a constant between-session effect such as learning. ICC(3,1) is an estimate of

$$\frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} \tag{1}$$

where σ_r^2 is between-subjects (rows) variance and σ_e^2 is the between-sessions variance (variance of the residuals after removing the subject and session effect). Since this metric combines both between-subject and between-session variance, it is suitable for providing insights into random effect group analyses. However, the same value of ICC can come from both high σ_r^2 and low σ_e^2 or low σ_r^2 and high σ_e^2, which makes the comparison between tasks harder. ICC is in fact more heavily influenced by between-subject variance than between-session variance (the variable of interest). For instance, if different tasks have the same between-session variance (σ_e^2) but different between-subjects variance (σ_r^2), ICC will be stronger for the task with the highest between-subjects variance, making its usefulness as a quality estimator for group studies debatable. From the single subject point of view, between-subject variance is irrelevant and therefore it is more informative to consider only between-session variance. Furthermore, in contrast to volume overlap, this is not the variance of contrast maps (between-subject) that must considered but the variance of t maps (contrast maps weighted by error). In the same way volume overlap is sensitive to the selected threshold, *t* value variability in ICC can be influenced by the design matrix used in GLM. This involves regressors, the hemodynamic response function (HRF) and contrasts definitions. For instance, Caceres et al. (2009) found that one can have highly correlated time-series but with a poor model fit leading to low reliability. They concluded that the wrong HRF model can lead to low reliability. However, inadequate regressors and contrast could also lead to similar results.

Apart from the issue of how to measure fMRI reliability, a further important question is what causes the lack of reliability in the first place and how this could be prevented. One of the suspected sources of variation in brain activation patterns is the possibility that different cognitive strategies and therefore different neuronal responses are produced by different subjects. These effects don't necessarily have to be task related. In a block design experiment, it would be enough that the subject consistently performs different mental tasks during the rest period to provide significantly variable results. The influence of this kind of variability is very hard to quantify because of the lack of access to the true neuronal activation patterns. It is, however, very likely that the type of task can reduce this "cognitive noise". For example, a simple finger tapping task involving primary motor cortex requires fewer possible cognitive strategies than the Iowa Gambling Task. Other possible sources of reduced reliability are easier to quantify. These include, but are not limited to, scanner noise (Bennett and Miller, 2010), subject motion (Caceres et al., 2009), and between-session coregistration errors (Fernández et al., 2003). Even though these confounds have been recognised in the literature numerous times, to our knowledge, there is no analysis on how much they contribute to reliability metrics. To date, the only study examining such effect was performed by Raemaekers et al. (2007) who showed a positive correlation between "sensitivity" (average absolute t value) and between-session volume overlap.

In the following paper, with the aim to quantify and better understand the observed fMRI reliability, we measured at the subject level and in four different behavioural tasks, the correlation between time-series, the between-session *t* value variance, and the Dice overlap coefficients between activation maps. The four tasks included motor mapping, covert verb generation, overt word repetition and landmark tasks, and were chosen because they are well established through group studies and had potential use for presurgical cortical mapping. We investigated how much the reliability measures can be explained by, the task, scanner noise, subject motion, and between-session coregistration, and how they relate to each other.

Methods

Participants and procedure

As a part of a larger study assessing suitability of different fMRI paradigms for presurgical cortical mapping in tumour resection, a group of normal healthy volunteers without contraindications to MRI scanning were recruited using flyers distributed among University of Edinburgh staff in electronic and traditional form. To match the mean age of diagnosis of the glioma patients undergoing resection surgery (Ohgaki, 2009), all volunteers were over 50 years of age. Out of 11 volunteers, data from one participant were discarded due to problems with executing the tasks. Additionally one session from the word repetition task was discarded for one of the subjects. The remaining 10 subjects included four males and six females, of which three were left-handed and seven right-handed according to their own declaration, with median age at the time of first scan of 52.5 years (min = 50, max = 58 years). The study was approved by the local Research Ethics Committee.

Tasks

All the behavioural tasks were implemented using Presentation® Software (Neuro Behavioural Systems http://www.neurobs.com/). Stimuli synchronisation and presentation were provided by NordicNeuroLab hardware (http://www.nordicneurolab.com/). During the first scanning session, each subject was trained for each task with a few trials inside the scanner. Care was taken to make sure that volunteers understood and could properly perform the tasks. For each task, the first four volumes before stimulus presentation were discarded for signal stabilisation.

Motor task

Subjects had to move a body part corresponding to a picture. The following instructions were issued: "You have to tap your index finger when you see a picture of a finger, flex your foot when you see a picture of a foot, and purse your lips when you see a picture of lips". A block design with 15 s activation periods and 15 s rest periods was employed, with four trials used for training. In every block, subjects moved the index finger of their dominant hand, or flipped their dominant foot or pouched their mouth. Movement was paced with a frequency of 0.4 Hz using visual stimuli. There were five repetitions of each activation/rest block for a total scan time of 7 min 40 s.

Covert verb generation task

Subjects were asked to think of a verb complementing a noun presented to them visually. The following instructions were used: "When a word appears it will be a noun. Think of what you can do with it and then imagine saying 'With that I can …' or 'That I can …' ". A block design with 30 s activation and 30 s rest blocks was employed, with eight trials used for training. During the activation blocks, ten nouns were presented for 1 s each followed by a fixation cross during which subject had to generate the response. The nouns were chosen at random from a set of 70 nouns (mean lexical frequency: 0.000087, min: 0.000005, max: 0.000392, std: 0.000092). Rest blocks had an analogous structure but with each word replaced by scrambled visual patterns generated by scrambling the phase of the 'picture' of each word, i.e. the control patterns were matched in the amplitude spectrum. Seven activation/rest blocks were presented for a total scan time of 7 min 12.5 s.

Overt word repetition task

Subjects had to repeat aloud words presented via headphones. The following instructions were used: "When you hear the word, repeat it immediately". A block design with 30 s activation and 30 s rest blocks was employed in conjunction with a sparse sampling data acquisition technique to present and record stimuli during the silent periods, with four trials used for training. After 2.5 s of blank screen during which the fMRI data were acquired, subjects were presented with an auditory stimulus which consisted of a pre-recorded native British English speaker reading a noun chosen at random from a set of 36 nouns (759 ms sound tracks length, mean lexical frequency: 0.000087, min: 0.000005, max: 0.000392, std: 0.000098). This was followed by a question mark prompting the subject to repeat the word. Question marks disappeared after 1741 ms and the sequence was repeated 6 times. The nouns used were randomised for every subject/session combination. A blank screen was also presented during rest periods. There were six activation/rest blocks for a total scan time of 7 min 40 s. Subject responses were recorded using an MRI compatible microphone. During the scanning session, the radiography staff listened to check if the subject was executing the task correctly.

Landmark task

Subjects performed two alternate tasks, namely tell if a horizontal line is crossed precisely in the middle (LANDMARK) and tell if a horizontal line is crossed at all (DETECTION). The following instructions were used: "Press the button with your left index finger if the line is bisected in the middle otherwise press the button with your right finger" or "Press the button with your left index finger if the line is crossed otherwise press the button with your right finger". A block design with 16.25 sec landmark/detection blocks was used, with ten trials used for training. Each task was preceded by an instruction screen which was presented for 8.25 s with a rest period of 8 s. Each block consisted of 10 lines, four correct and six incorrect. Each line was presented for 525 ms and subjects had 1100 ms to respond before the next presentation. Lines were presented in the four corners of the screen. For the landmark task and incorrect trials, the crossing line was located at three different distances from the middle, specifically 12, 40, and 62 pixels from the true middle corresponding to 0.45, 1.5, and 2.325° of visual angle. There were eight landmark/detection blocks for a total scan time of 9 min 55 s. All trials were randomised and all responses were recorded.

MRI acquisition

All scans were acquired on a GE Signa HDxt 1.5 T clinical scanner at the Brain Research Imaging Centre (http://www.bric.ed.ac.uk/), University of Edinburgh. Each volunteer was scanned twice, two (eight subjects) or three (two subjects) days apart using the same sequence. All fMRI data were acquired using a single-shot gradient-echo echo-planar imaging (EPI) sequence with the following parameters: field of view (FOV) = 256×256 mm, slice thickness 4 mm, 30 slices per volume, interleaved slices order, voxel size 4×4×4 mm, acquisition matrix 64×64, flip angle = 90°, echo time (TE) = 50 ms. The repetition time (TR) was 2.5 s for all tasks, except for word repetition where the TR was 5 s (sparse sampling 2.5 s acquisition, 2.5 s silence). In addition to the EPI data, a high-resolution 3D T1-weighted coronal scan was acquired for each session (FOV = 256×256 mm, slice thickness 1.3 mm, 156 slices, voxel size 1×1×1.3 mm and acquisition matrix 256×256).

Functional MRI pre-processing and analysis

Data was processed using SPM (http://www.fil.ion.ucl.ac.uk/spm/) and FSL (http://www.fmrib.ox.ac.uk/fsl/) within the Nipype framework Gorgolewski et al., 2011a.

Preprocessing

For every subject, the 3D T1-weighted volumes from both sessions were coregistered, resliced and averaged. A DARTEL template was created using the averaged T1-weighted volume from all subjects (Ashburner, 2007). Additionally, a brain mask was estimated from each average T1-weighted volume using BET (Smith, 2002).

As described above, the first four volumes of every EPI sequence were discarded and the remaining volumes were slice-time corrected. Finger, foot, and lips sequences of left-handed subjects (three subjects) were flipped along the Z–Y plane. For every subject, all slice time corrected volumes from all tasks and sessions were realigned and resliced to their mean volume to remove motion artefacts. The mean EPI volume was coregistered to the 3D T1-weighted between-session average volume and the resulting affine transformation was applied to headers of the realigned files. Each EPI volume was then normalised using the DARTEL template and corresponding flow field, and smoothed with 8 mm full width half maximum Gaussian kernel. Apart from the fact that smoothing improves SNR, it is necessary to maintain assumptions of the Random Field Theory which is being used for thresholding. The smoothed volumes supplemented with the previously estimated brain mask and realignment parameters were searched for artefacts using ArtifactDetection toolbox (http://www.nitrc.org/projects/artifact_detect/).

1st level analysis

Each session was analysed separately. GLM (Friston et al., 1994) was used to estimate the BOLD signal response by fitting a design matrix that consisted of an autoregressive filtering matrix (AR1), task related regressors, realignment regressors (six parameters), a high pass filter (128 Hz), and artefacts (one per artefact) regressors. Task regressors for verb generation and word repetition were simple boxcar functions convolved with a canonical HRF. For these tasks, a simple contrast including the single task regressor was used, i.e. activation vs. baseline. For the finger, foot and lips tasks, each body part was modelled with a separate boxcar regressor and three contrasts opposing each body part against the two others were obtained. The design matrix for the landmark task included five event related regressors acquired from each subject/session experiment log: landmark stimuli

with correct responses, landmark stimuli with incorrect response, detection stimuli with any response (correct or incorrect), and detection and landmark stimuli with no response. This allowed a "landmark stimuli with responses vs. only detection stimuli with responses" contrast to be estimated. Only voxels within the previously estimated brain mask were included in model fitting.

2nd level (random effect) analysis

For every subject and task, contrast volumes were averaged between the two sessions. These averages were then used in a second level group analysis following the Holmes–Friston approach (Holmes and Friston, 1998), i.e. a one sample *t* test on each contrast was run to estimate a group effect. The result of each *t* test was thresholded using the topological false discovery rate (FDR) method (Chumbley and Friston, 2009) with the cluster extent probability threshold set to 0.05 after FDR correction.

Reliability measurements and confounds

Measuring reliability

Between-session correlation on time-series. After the EPI sequences has been realigned, normalised, spatially smoothed, and detrended using second order polynomials, Pearson correlation coefficients between first and second session time-series were calculated for each voxel and then averaged. This allowed to determine the similarity of the measurements before any statistical and HRF models had been fitted. This measure, in contrast to the two described below, was calculated for each task rather than for each contrast. Because of this 39 values were entered into the analysis (10 subjects × 4 task−1; since we had to discard one run for one subject).

Between-session variance of unthresholded t maps. *t* Maps were first corrected for global effects using estimates from the adaptive cluster forming threshold method (Gorgolewski et al., 2011b, in review). The mean of the squared between-session differences was calculated

$$t_{\text{diff}} = \frac{1}{n}\sum_i (t_{i1} - t_{i2})^2 \tag{2}$$

where n is the number of voxels, t_{i1} and t_{i2} are the ith voxel t values from the first or second session respectively. This measure is equivalent to the between-session component of ICC, but adapted here for single subject analysis. For full derivation of the relation between ICC and t_{diff} see Appendix A. This as well as the following measure was calculated for every contrast resulting in 59 values entering the analysis (10 subjects × 6 contrasts−1; since we had to discard one run for one subject).

Volume overlap of threasholded t maps. Single subject *t* maps were thresholded using cluster FDR ($q = 0.05$) with an adaptive cluster forming threshold (Gorgolewski et al., 2011b, in review). This method uses a combination of Gamma–Gaussian mixture models and topological thresholding (based on RFT) and has been shown to provide results less prone to different levels of SNR and global effects, thus giving maximum overlap estimates. Using the suprathreshold maps the between-session Dice overlaps was calculated. In the case where both maps were empty (no suprathreshold voxels), a Dice overlap of zero was assumed to penalise for lack of signal. In addition, to test if the tasks were reliable, the mean Dice overlap obtained for each subject and task was compared with the between-subject Dice overlap. The between-subject Dice overlap was obtained by computing the overlap between the thresholded map of every subject in Session 1 and the thresholded maps of all the other subjects in Session 1. The procedure was repeated for Session 2 and all Dice measures were averaged for each task. This allowed the testing of whether the overlap measured within-subjects was significantly greater than the overlap measured across all subjects, given that all subjects were in standard space. A percentile bootstrap test of the Harrell–Davis (HD) median (Harrell and Davies, 1982) was used to estimate if the difference of within- and between-subject Dice overlap was statistically significant.

Results are reported for the full brain, and subsequent analyses apply to these results only. However, to also make sure results were not biased toward low values due to lack of reliability in many regions but the one targeted by the task in hand, we also report results within specific ROI. These were constructed using probability maps available in the anatomy toolbox (Eickhoff et al., 2005, 2006, 2007). For the mapping of the primary motor cortex, the whole left areas 4a and 4p were used (Geyer et al., 1996). For Broca area, Brodmann areas 44 and 45 were used (Amunts et al., 1999). For Wernicke area, area TE30 was used (Morosan et al., 2005). For the auditory cortex, we used areas TE1, 1.1 and 1.2 (Morosan et al., 2001). Finally, for the landmark task, right Inferior Parietal Cotex and Superior Parietal Lobule were used (Corbetta and Shulman, 2011). Masks were generated in the MNI space and resliced to DARTEL template dimensions.

Measuring confounding factors

For each of the above measurements, a repeated measure multiple regression approach was used. In this approach two models are fitted to the data. The 1st model included the task, scanner noise, subject motion (total displacement, stimuli/motion correlation, and interaction between task and stimuli/motion correlation), coregistration error and subjects as regressors (for the design matrix see Supplementary Fig. 1) and the 2nd model only included subjects. The R^2 of the full model is then tested by comparing the full to the reduced model, effectively testing the contribution of all regressor to the model given the presence of the repeated measure. To identify within the full model the contribution of each independent variables to the total explained variance, the relative importance bootstrap technique (Ulrike Grömping, 2006) with the Lindeman–Merenda–Gold metric (Lindeman et al., 1980) was used (performed in R using *relimpo* package). This technique estimates the relative importance by generating combinations of the given (step-wise) model and weighting contributions to the explained variance by the order of adding variables. The estimates are boot-strapped 200 times to establish confidence intervals.

Scanner noise. To estimate the noise due to scanner related fluctuations, the temporal Signal to Noise Ratio (tSNR) was measured

$$tSNR = \frac{1}{n}\sum_i \frac{\mu_1}{\sigma_i} \tag{3}$$

where n is the number of voxels, μ_i and σ_i are the mean and the standard deviation of the ith voxel across time. The average was taken across all voxels within the brain mask. Before calculating tSNR, the time-series were truncated by discarding the first four volumes, realigned to remove motion confounds and detrended using second order polynomials.

Subject motion . Two metrics were used to characterise motion: total displacement and stimulus by motion correlation. *Total displacement* (Wilke, 2012) allowed measuring in a single variable the overall motion using realignment parameters from every EPI volume. This measure has the advantage of capturing cortical voxel displacement due to both translation and rotation. Subject motion was characterised here by an average over this parameter from both sessions. Stimulus/motion correlation allowed measuring the influence of motion on regressors of interest (and thus beta values). For every design matrix (80 design matrices: 4 tasks × 10 subjects × 2 sessions), we measured the correlation between the regressors of interest and motion regressors using a multiple regression models. The dependent variable of this model were the stimuli regressors (after HRF convolution) multiplied by the contrast vector, whilst the 6 motion parameters were used as independent variables, This way, for every design matrix, we were able to calculate R^2-percentage of stimuli variance explained by motion. As for total displacement, values from the two sessions were averaged.

Coregistration error. Inaccuracies of coregistering EPI volumes between two sessions were characterised by the correlation ratio (Roche et al., 1998) between mean EPI volumes from the two sessions. This metric measures functional dependencies between voxel intensities and has been previously used as a registration cost function. The correlation ratio was calculated on brain-masked volumes.

Measuring relations between reliability metrics

To investigate the relationships between reliability metrics, robust Spearman correlations with outlier removal (Rousselet and Pernet, 2012; Wilcox, 2005) were computed between each pair of measurements before and after fitting the multiple regression models accounting for confounds.

For each subject, the HD estimate of the median of t_{diff} and each time-series was also computed for three different ROIs: the area activated in both sessions (overlap), the area activated in one (either the first or second) of the sessions, and the area not activated in any of the sessions. Correlations were then computed to test whether the voxelwise reliability measures (t_{diff} and time-series correlations) were significantly different between these regions.

Results

Random effect results

Regions activated by each task followed previously reported patterns of activation. For the motor tasks, strong activations of the left precentral gyrus were observed respecting the known motor homunculus: (1) foot contrast revealed activations near the top end of the contralateral precentral gyrus extending to left Supplementary Motor Area (SMA) and also showing activation in ipsilateral crebellum and ipsilateral precentral sulcus; (2) finger contrast produced activation in the middle/lateral contralateral precentral gyrus and ipsilateral cerebellum; (3) lips contrast produced bilateral activation in the inferior part of the precentral gyrus, but also the cerebellum. Activations were also observed in the visual cortex over inferior occipital/fusiform gyri in response to the stimulus presentation. For the verb generation task, activations were observed in left Broca's area (BA 44 and 45), left temporal gyrus, left inferior parietal lobule, SMA and left thalamus. For the word repetition task, activations were observed over the superior temporal gyrus, mostly in the left and right primary auditory cortex (areas TE 1.1, TE 1.2 and TE 3; Morosan et al., 2001) and left Wernicke's area (Caspers et al., 2006). Additional activations were found in the SMA, Brodmann area (BA) 6, the postcentral gyrus (BA 3b) and the cerebellum. Finally, for the landmark task, activations were observed mainly in the right superior and inferior parietal lobule, left fusiform gyrus, left cerebellum, left postcentral gyrus (BA 2), right inferior temporal gyrus, precentral gyrus (BA 6—bilaterally), SMA (bilaterally), right inferior frontal gyrus (BA 44 and 45), and left calcarine gyrus (BA 17). The group level activations therefore confirmed that the stimuli used in all of the tasks were correct.

Reliability

Low mean correlation values were observed on voxel time-series across the four tasks (range 0.07 to 0.17). Time-series correlations were not homogenous through the whole brain and higher values were observed within ROI (range 0.12 to 0.23) compared to the whole brain (Table 1). This indicates that for 'activated' regions, time-series were more similar than for not activated regions.

The opposite pattern of results was observed with T_{diff} (the between session variance of T values). We observed lower T_{diff} values for the whole brain (range 1.36 to 6.1) than within ROI, (range 1 to 8.4), but high T_{diff} values indicate lower reliability. However, as we show later, there was no clear relation between absolute *t* values ('activated' area) and T_{diff}.

Dice overlap values show relatively high reliability (average over all tasks for the whole brain 0.41) especially compared to the low correlations observed on time-series. However, as for time-series, higher Dice coefficients were observed within ROI (range 0 to 0.93) than the whole brain (range 0 to 0.76). Here one has to be aware of a bias related to the previously mentioned thresholding issue. Indeed, restricting the suprathreshold voxels just to a smaller ROI is necessarily biased towards higher Dice values.

Using the Dice metric, we also tested for the full brain if a given task was significantly reliable by comparing within- to between-subject Dice overlap coefficients. Analyses revealed that the motor task had a higher reliability within- than between-subjects. Similar results were obtained for the verb and word generation tasks (see Table 1). In contrast, despite group level analyses showing right intra-parietal activations, no consistent activation was observed, with the within-subject (0.17) not being significantly different than the between-subject (0.11; bootstrap difference [−0.01 0.19] p = 0.74). Thus, despite showing similar time-series correlations as the other tasks and even a lower t_{diff}, the landmark task did not perform well in practice. The technique we are proposing here (comparing within- and between-subjects Dice overlap) is analogous to ICC for thresholded maps. Often reliability studies state a given amount of reliability but it is not known if it is 'good' or not (Bennett and Miller, 2010). Our results show that the answer depends on the task at hand as the same amount of reliability could be good enough in one case (because it is higher than measuring the reliability between subjects) but not the other. For an example map showing the reliability measures of one subject see Fig. 2.

Table 1
Reliability measurements obtained across the full brain and within ROI. Significant differences in within and between subjects Dice overlaps are marked in bold.

		Mean time-series correlations	Mean t_{diff}	Mean within-subject Dice	Mean between-subject Dice
Finger	Full brain	0.101 ± 0.040	1.98 ± 0.505	**0.574 ± 0.189**	**0.352 ± 0.115**
Foot			2.10 ± 0.524	**0.517 ± 0.125**	**0.321 ± 0.101**
Lips			2.44 ± 1.15	**0.454 ± 0.161**	**0.286 ± 0.145**
(All)			(2.17 ± 0.80)	**(0.515 ± 0.168)**	**(0.319 ± 0.124)**
	Motor cortex	0.230 ± 0.067	2.03 ± 0.78	**0.751 ± 0.146**	**0.566 ± 0.148**
			2.50 ± 0.53	**0.724 ± 0.098**	**0.526 ± 0.160**
			1.85 ± 0.70	**0.837 ± 0.086**	**0.629 ± 0.157**
			(2.13 ± 0.73)	**(0.771 ± 0.123)**	**(0.574 ± 0.161)**
Verb generation	Full brain	0.070 ± 0.084	3.58 ± 1.15	**0.502 ± 0.216**	**0.250 ± 0.129**
	BA 44/45	0.120 ± 0.086	4.39 ± 2.28	**0.595 ± 0.238**	**0.346 ± 0.214**
Word repetition	Full brain	0.090 ± 0.031	2.83 ± 0.68	**0.452 ± 0.097**	**0.218 ± 0.128**
	Auditory cortices	0.255 ± 0.066	3.42 ± 1.24	**0.537 ± 0.199**	**0.303 ± 0.210**
Landmark	Full brain	0.135 ± 0.054	1.69 ± 0.28	0.173 ± 0.177	0.115 ± 0.114
	Right IPL	0.173 ± 0.063	1.97 ± 0.37	0.150 ± 0.209	0.203 ± 0.182

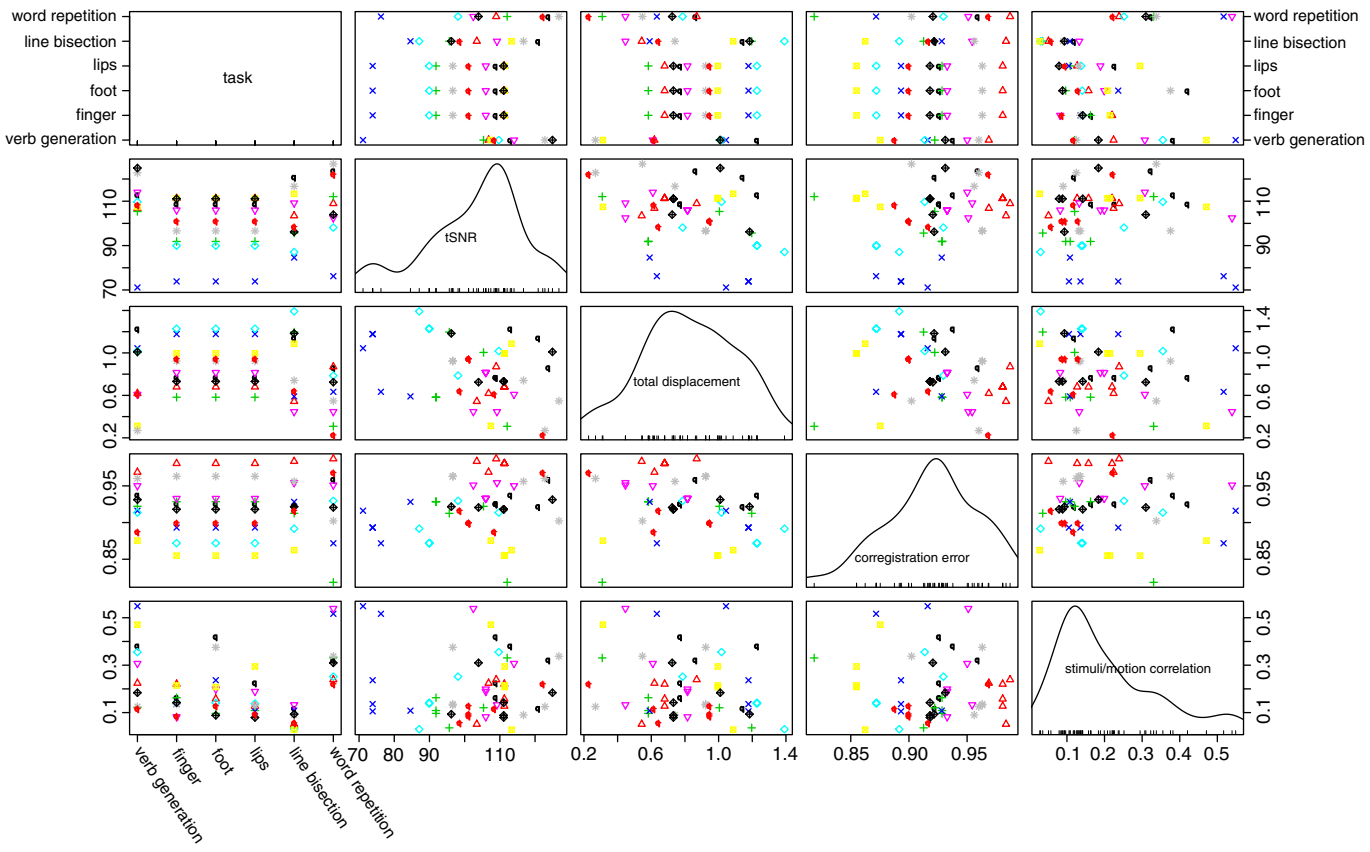

Fig. 1. Distributions of modelled explanatory factors to reliability. Combinations of symbols and colours of points represent different subjects.

Contribution of scanner noise, subject motion and coregistration errors to between-session variance

Data modelling

Since multiple regression is based on correlations, it requires a non-zero variance of the explaining factors. As shown by the correlogram between all of the confounding factors (Fig. 1), the explaining factors have a reasonable spread of values (for example total displacement ranges from 0.2 to 1.4 mm). We also looked at the contributions of the number of artefacted volumes found by the ArtDetect algorithm used in preprocessing. These volumes are selected based on the signal intensity and motion signals and added as a confounding regressor (one per artefact) to the single subject design matrix. On average there were 1.75 artefacts in motor tasks, 0.27 in word repetition, 1 in verb generation, and 2.95 in line bisection. Despite the fact that the tasks differed significantly in terms of those numbers ($F(5,53) = 4.121$ $p = 0.003$) adding them to the multiple regression model used to analyse reliability did not yield significant improvements in the model fit (similar adjusted R^2). Similarly, the model used here was the most parsimonious among a set a models where motion regressors were modelled either as a single parameters, split per task or both (see Supplementary table 2).

Model results

Fitting task, scanner noise, subject motion and coregistration error to the time-series correlation values led to a not statistically significant R^2 of 48% ($F(29,10) = 1.3939$, $p = 0.255$,[1] adjusted $R^2 < 0\%$) with a large contribution of the task (17.53%) and subject motion 20%). When tested on t_{diff} (the between-session differences of t values—a component of the ICC measure), the model yielded a higher R^2 of 76% ($F(35,14) = 4.7986$, $p = 8.106e-05$, adjusted $R^2 = 60\%$), with again a large contribution of the task (40.32%) but also of motion (24%) and scanner noise (11.02%). Finally, when fitted to the Dice values, the model produced an R^2 value of 75% ($F(35,14) = 6.3365$, $p = 4.597e-06$, adjusted $R^2 = 59\%$) with again a major contribution of the task (42.68%) and motion (23%).

Overall task-induced variations are a major single contributor to reliability (18%, 31%, and 43% respectively). This could be explained by the high variability of the landmark task compared to others. If we sum up contributions from all motion related regressors (total displacement, stimuli/motion correlation, and interaction between task and stimuli/motion correlation) it also explains a large portion of the variance (20%, 24%, and 23% respectively). Interestingly, this was not the actual amount of motion that mattered the most (i.e. total displacement) but the correlation between the stimulus presentation (paradigms) and motion. Scanner noise and coregistration confounds add little to the equation, accounting only for 6%, 2% and 6% respectively (see Table 2 and Supplementary Fig. 11).

No matter how we measured reliability, out of the most commonly reported in previous reliability studies confounds (scanner noise, subject motion, and coregistration error) only subject motion has a high contribution. To further verify these findings, we reran the reliability analysis on data acquired using the same pipeline but without motion correction (no realignment with runs, no motion parameter regressors and artefact detection in the design matrix). Turning off those corrections decreased the Dice overlap by 20% ($t(58) = 3.0795$, $p = 0.003166$), increased t_{diff} by 28% ($t(58) = -4.4787$, $p = 3.578e-05$) but did not influence time-series correlation significantly (8% decrease; $t(38) = 1.6644$, $p = 0.1043$). It is worth noticing that for Dice and t_{diff}, turning off motion lead to changes equivalent to the amount of variance that can be explained by motion regressors, that is motion lead to a decreases in T value reliability and thus a decrease in map overlap (for percentages of variances of each factor on un-realigned data see Supplementary table 2).

[1] Reported F and p values correspond to the repeated measures test: full vs. subject model comparison.

Relationships between reliability metrics

No significant correlations were observed between time-series correlations, *t* value variance and Dice coefficients. Weak negative correlations were observed between time-series correlations and t_{diff}, however these weak effects disappeared once confounds were accounted for (see Supplementary Fig. 11). At the same time regressing out the confounds strengthens the relation between t_{diff} and Dice making it statistically significant ($\rho = 3.11$ vs. $\rho_{0.05} = 2.43$). The direction of the relation ($r = -0.44$) makes conceptual sense (smaller differences in t values lead to higher overlaps). To investigate further a possible (non-monotonic) relationship between these variables, all voxels from each task/contrast were pooled together to create a series of scatter plots between *t* values and time-series correlations and t_{diff}.

Most voxels with high *t* values show increased time-series correlation for all tasks. The same is true for negative *t* values, which indicates that even though the negative *t* values are not usually of interest, they are stable between-sessions even on the time-series level (Fig. 3a). It is also worth noting that there were many voxels with high correlation but low *t* value. These indicate a reliable signal not captured by the design matrix. When restricting the analyses to overlapping vs. non-overlapping activated areas, the highest time-series correlation values were observed in the overlapping (those that by definition will have high *t* values) rather than non-overlapping areas (Fig. 4), confirming that high *t* values relates to reliable voxels (time-series).

No relationship was observed between mean *t* values and the variance (t_{diff}). The highest t_{diff} values (poorest reliability) were observed for *t* values close to zero, but one has to bear in mind that those values were also the most common. There were, however, differences in the observed patters between tasks. The distribution of t_{diff} across mean *t* values was almost uniform for verb generation. This was in contrast to the lips task for which the highest t_{diff} values were observed almost exclusively for the voxels with *t* values close to zero. The landmark task, on the other hand, showed a smaller spread in both *t* values and their between-session variance. When restricting the analyses to overlapping vs. non-overlapping activated areas, we noticed that mean t_{diff} in the overlapping area was no different than in the parts of the brain that were not active in either of the two session, but there was a significant increase of t_{diff} for non overlapping active areas (Fig. 4). This relation can even be observed on the individual subjects maps (see Fig. 2 and Supplementary Figs. 2–10). In other words, t_{diff} is bigger in regions that were active in one of the sessions, but not in both of them. These are usually the borders of suprathreshold clusters.

Discussion

Studies involving fMRI are complex and easily influenced by many factors. This is not only because the subject in question, the human brain, has intricate and not fully understood hemodynamics. The data acquisition and processing is a multilevel complicated process (Savoy, 2005). In this study, we investigated how different factors can contribute to between-session variance. We found that about 30–40% of the observed single subject reliability (unthresholded or thresholded T-maps) can be explained by the task used and that among confounding factors, motion is the main problem accounting for about 20% of the variance.

Choosing the right metric

One important aspect of this study is the application of different methods of measuring reliability. Specifically, we assessed three different ways of measuring reliability, from the correlation of time-series, to *t* values and thresholded *t* maps. In addition, compared to many previous studies (e.g. Caceres et al., 2009; Raemaekers et al., 2007), we have not restricted our measurements to a predefined ROI or split analyses between different ROIs. This decision was motivated by the fact that reliability and activations are not strictly related (see e.g. Caceres et al., 2009) and in some cases like Dice, ROI analysis introduces a selection bias. It is therefore misleading to assess a task only by the reliability within a predefined ROI.

Our decision to use t_{diff} as the measure of between session variability of unthresholded maps was mostly driven by ability to relate it to Dice overlap measure. First, we decided to use t-values instead of beta values because t-values are influenced by residual noise and thus reflect better acquisition (scanner) related variance. Second, t_{diff} captures the variance of t-values that translate directly into the extents of suprathreshold regions. Finally, t_{diff} can be related to ICC. As mentioned in the introduction, this choice is of course only relevant if one is looking at single subject reliability, and $beta_{diff}$ could be more appropriate for group reliability.

Despite the fact that Dice overlap was previously being criticised as a reliability measure (Smith et al., 2005), we have still included it in our analysis. Thresholding as any form of dimensionality reduction can introduce biases and we agree that calculating overlaps of thresholded maps is a rough estimate of reliability. However, let us not forget that the thresholded maps are what the end result of an fMRI analysis is. Papers describing group studies are presenting and making claims about thresholded maps. The same applies to the single subject domain. Neurosurgeons plan and execute procedures based on thresholded maps. Functional localisers produce ROIs which are nothing less than thresholded statistical maps. We acknowledge problems with analysing thresholded maps (that is why we have included two other reliability metrics) and at the same time we try to minimise their influence. Importantly, we have used the same method as Smith et al., 2005 for correcting for global effect (a t value distribution shift derived from a Gamma-Gaussian model).

Indeed, global effects in context of single subject test–retest reliability have also been a topic of a recent work by Raemaekers et al., 2012. In their approach they fitted a line to session 1 vs. session 2 scatter plots. This allowed them to estimate between session variance as the variance orthogonal to this line. This is a variant of global effect correction used in our work. Their approach allowed the amount of shift applied to the t values to be in linear relation to them. In other words, in our model this line can be shifted from the centre of the data cloud, but keeps the 45 degrees angle. However, the approach we used (Smith et al., 2005) is more flexible as it allows applying the correction to one session without knowing anything about the other (i.e. the model is fitted using single session distribution, not the joint scatter plot). Additionally, when applied to the full brain, a linear fit to the joint distribution of values from two sessions would be driven by values close to zero and thus not capturing the shape of the tails which are the activated voxels (see Fernández et al., 2003).

Finally, we found that good time-series reliability is a necessary but not sufficient condition for good *t* map reliability. For example, one could observe a good correlation between time-series of two sessions, but a large difference in *t* values, a case that may correspond to a poor model fit (i.e. some regions may activate similarly in both sessions, in relation to the task, but not with the stimulus or block onsets described in the design matrix – such regions can be captured by e.g. ICA analyses). More intriguingly, we have observed a similar effect in the relationship between *t* values and thresholded maps. Small between-session differences in *t* values are necessary for a good suprathreshold overlap, but not sufficient, because a high threshold can lead to low Dice overlap. For instance, the task that performed the worst (landmark) in terms of Dice, was the one showing the lowest t_{diff} values. This brings us to a paramount question, namely what makes a good task/analysis? The task should be reliable, but this is not the full answer, because it can be reliable in not measuring any meaningful activation. In other words we don't only want low between-session *t* value variance, but also high *t* values consistently across sessions. Dice overlap captures this property due to thresholding, since only high *t* values that survive thresholding contribute to the overlap. We showed here that using Dice, one can

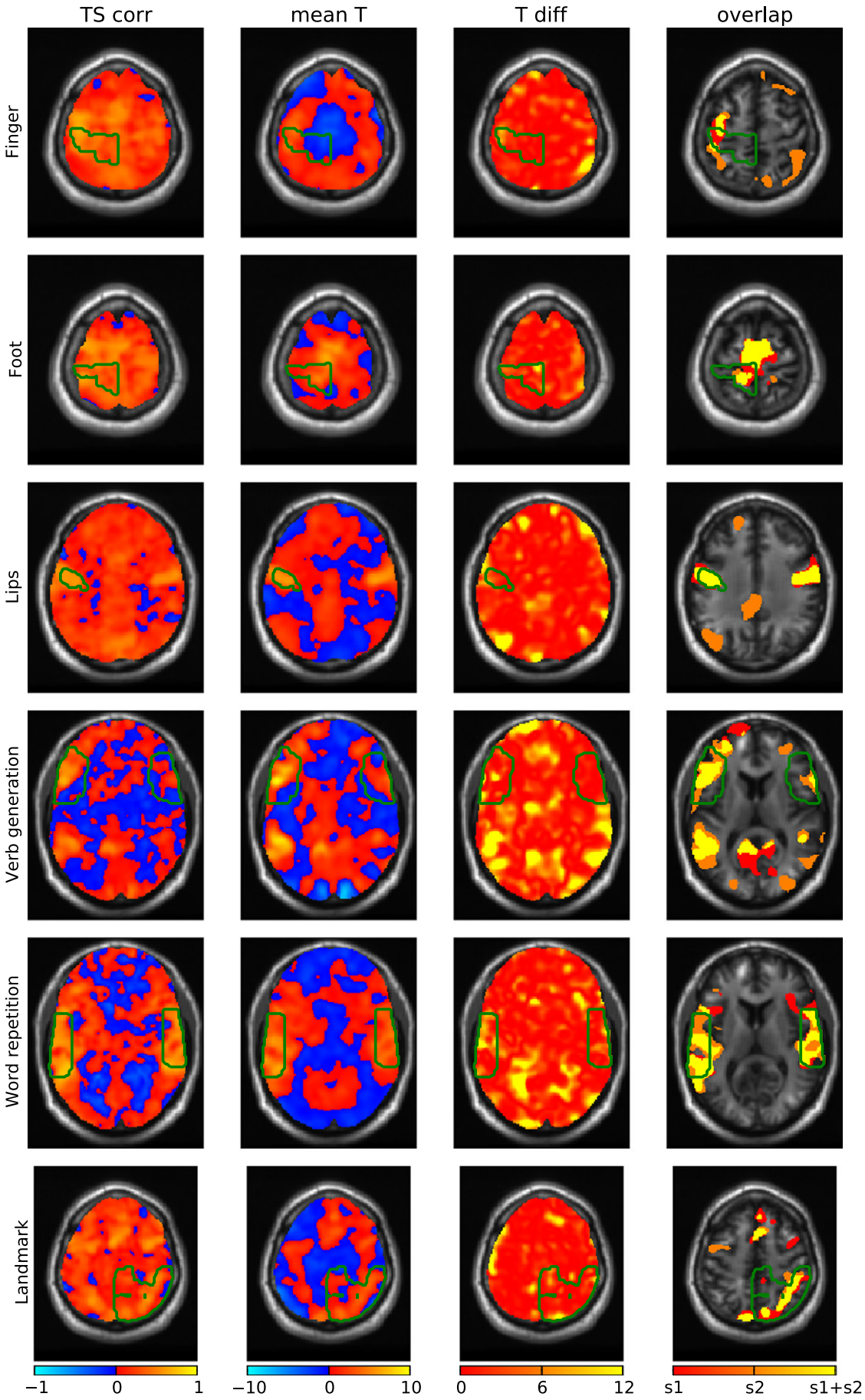
TS corr
mean T
T diff
overlap
Finger
Foot
Lips
Verb generation
Word repetition
Landmark
−1
0
1
−10
0
10
0
6
12
s1
s2
s1+s2

Table 2
Relative contribution in percentage (with 95% confidence intervals) of task, scanner noise, subject motion and coregistration error to time-series correlation, between session variance and Dice overlap.

	Time-series correlation	Between-session variance	Dice
Task	17.54%	31%	42.68%
	[7.31 49.19]	[20.4 48.25]	[22.39 63.76]
Scanner noise	1.57%	11.84%	4.48%
	[0.57 22.66]	[1.67 25.95]	[0.5 16.23]
Subject motion (total displacement)	7.18%	0.48%	4.95%
	[0.63 23.44]	[0.03 4.46]	[1.14 15.56]
(stimuli/motion correlation)	4.5%	17.96%	3.84%
	[1.14 26.35]	[4.35 40.08]	[1.23 14.21]
(task*stimuli/motion correlation)	8.4%	5.35%	14.52%
	[1.84 36.4]	[1.60 17.17]	[4.2% 25.81%]
Coregistration error	1.28%	0.66%	0.96%
	[0.18 10.95]	[0.16 3.42]	[0.24 6.67]

compare within- vs. between-subject overlap, and a reliable task can be defined as having a significantly higher degree of overlap within- than between-subjects.

Explanatory factors

The type of task was the main explanatory factor on our reliability metrics, which can explain the large variance observed across different studies (Bennett and Miller, 2010). Here, one can argue that the large effect observed depends essentially on the landmark detection task which failed to produce any suprathreshold clusters more often than the other tasks. This indeed can explain the effect over Dice overlap measurements, but not on t_{diff}. The observed between-session t value differences were actually lower for the landmark task than for the other tasks. It is therefore a case where one can observe differences between small t values not yielding any statistically significant activation. As already mentioned in the introduction, this result also highlights the need to differentiate reliability of the BOLD signal (single subjects) from reliability of contrast maps (group studies) since a small BOLD signal but with a low between-subject variance gives significant group results.

The fact that the type of task can have such a big influence on reliability should perhaps not be surprising. First of all, the tasks in our study were not only different in terms of the behavioural paradigms (or in other words what the subject was meant to do during the scan), but also in terms of acquisition parameters. Word repetition used sparse sampling which in theory should improve SNR (Hall et al., 1999), although at the cost of the number of volumes acquired. Scanning time and therefore the number of volumes acquired ranged from seven up to almost ten minutes. All the tasks were executed in blocks, but the landmark task used event related regressors to restrict the response to correct answers only. All these factors can influence reliability on a purely data acquisition level. Further studies with systematic variation of these parameters, for example sparse/non-sparse, block/event related and number of volumes acquired, would be necessary to establish their exact contribution to reliability.

Apart from the data acquisition aspect of different tasks there is one more important reason explaining the observed influence of the task type on reliability. Different tasks involve different neuronal populations and can incorporate different cognitive strategies. For a given task the same observed behavioural response, such as generating a verb, can be achieved by different neuronal subsystems, hence eliciting different BOLD reaction. We hypothesise that this "cognitive freedom" is different for different tasks. For example, a simple finger tapping task is most likely to be executed in a similar fashion each time. In contrast, a more sophisticated task involving language generation or spatial attention could involve different neuronal subsystems each time. This might be part of the explanation why in our study the landmark task did not perform well in terms of single subject reliability.

Scanner noise, and coregistration errors have previously been suggested to contribute to reliability (Bennett and Miller, 2010; Caceres et al., 2009; Fernández et al., 2003). Even though we have found such relationships, their magnitude was surprisingly small. Both of the confounding factors we investigated have been accounted and corrected for in the data processing pipeline. Scanner noise, for example, can be influenced by signal dropouts due to failing coils. Smoothing can mitigate this to a certain extent by improving tSNR (see Supplementary Fig. 12), although this is achieved by the loss of spatial accuracy. Volumes with sudden signal dropout are also either removed or accounted for in the design matrix during the artefact detection step. As for the coregistration step, it is perhaps not surprising that modern algorithms managed to realign brain volumes of the same person scanned using the same sequence on the same scanner. Our results therefore suggest that thanks to advances in data processing methods, issues such as scanner noise and coregistration errors are not the most important contributing factors to between-session variance. This is, however, true only within normal working conditions. For example a serious scanner malfunction would inevitably result in poor reliability.

Subject motion on the other hand had non-negligible influence on reliability. It was the largest confounding factor (the 2nd largest explanatory variable) and for time-series correlation, even explained more than the task. Comparison of realigned vs. non realigned data confirmed those results by showing equivalent changes in t_{diff} and Dice. Only correlations on time-series were not significantly affected by turning off motion correction (−8%) despite a large portion of the variance explained by motion regressors on realigned data. In the present context this is difficult to explain. One possibility is that using Pearson correlation is not efficient enough to fully capture changes in reliability given the various limitations related to data range restriction, curvature, or heteroscedasticity (Wilcox, 2005). Additionally after correcting for repeated measures, the fitted model did not explain time-series correlations with statistical significance. Because of this time-series correlations results presented here should be treated with lower confidence. Time-series signal is richer than t_{diff} or Dice and therefore failure to explain the variance of correlation through a handful of regressors is not surprising.

Importantly for planning fMRI experiment, we have found that motion correlated with the stimuli explains the lack of reliability much better than absolute motion. On time-series correlation (i.e. before model fitting) both total displacement and motion correlated

Fig. 2. Brain statistical maps from a representative subject (subject 2). There is a spatial correspondence between time-series correlation (TS corr) and mean t maps, but there is no correspondence between T_{diff} and mean t. Additionally most heat points of the T_{diff} maps overlap with non-overlapping active areas (orange and red colours in the overlap column). The anatomical ROIs are marked in green. For remaining subjects see Supplementary Material Figs. 1–9. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

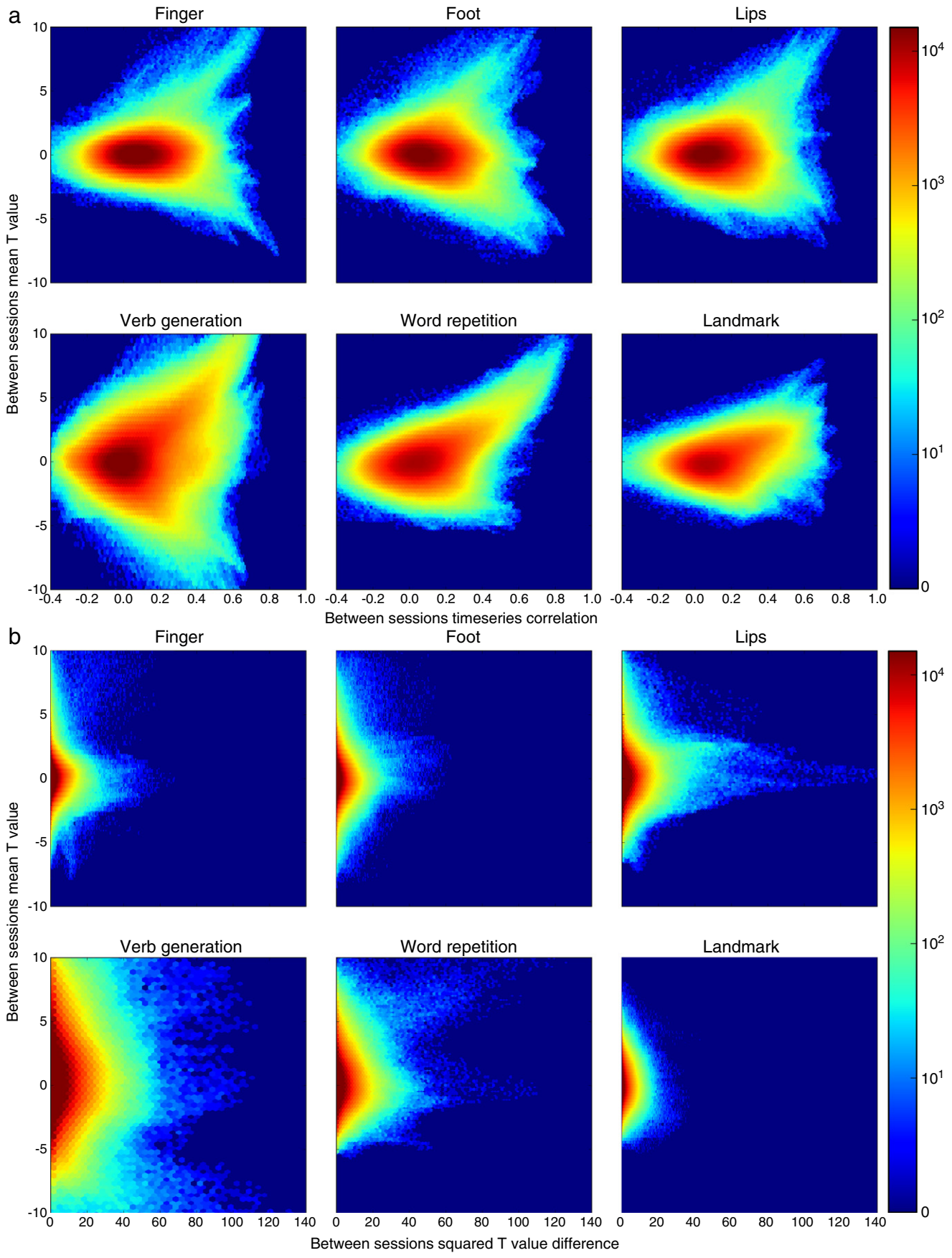

Fig. 3. Joint distributions of mean t values and time-series correlations (a) and mean t values and t_{diff} (b). Voxels from all subjects were pooled together.

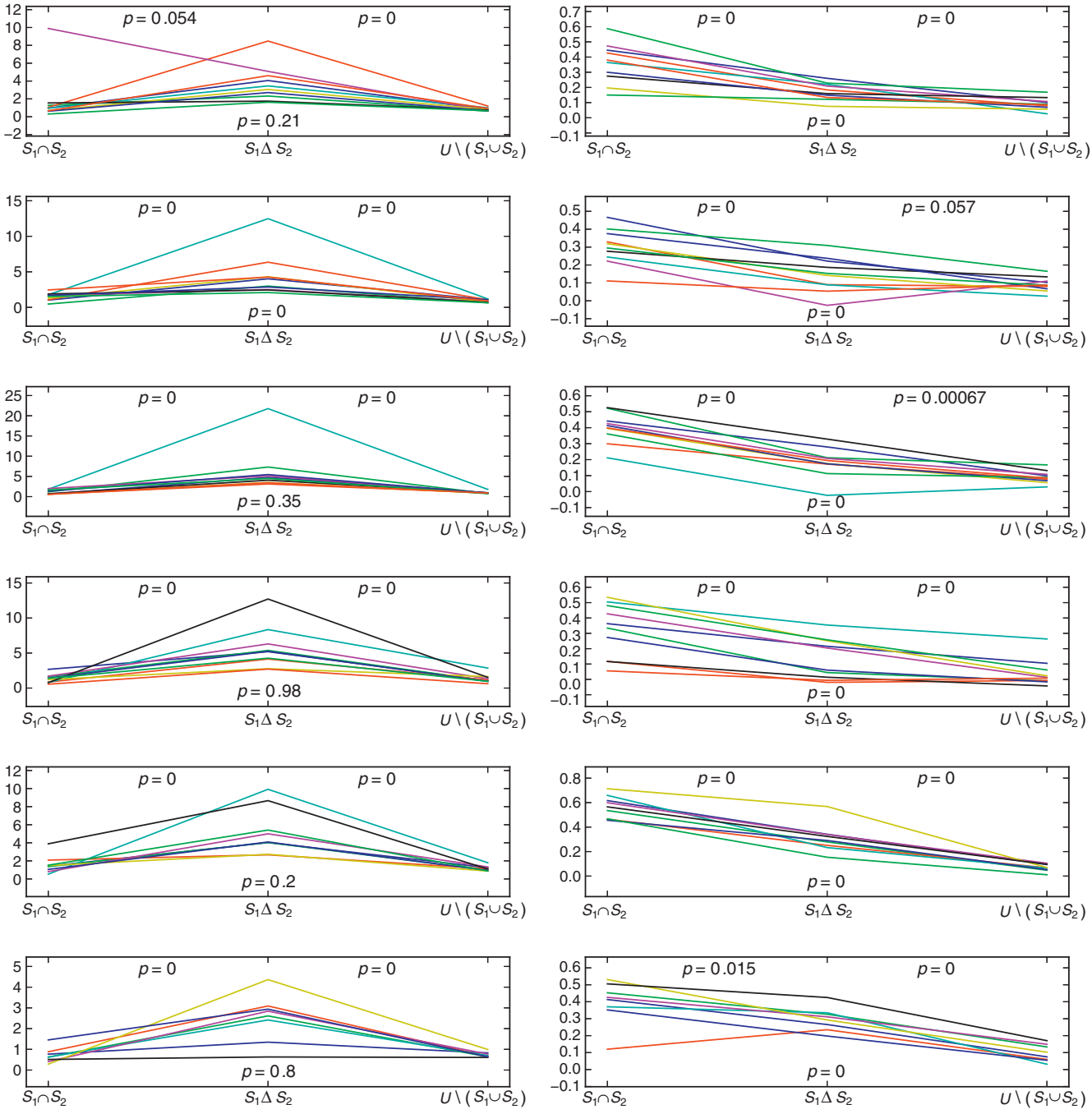

Fig. 4. ROI analysis of voxel-wise reliability metrics. HD median of t_{diff} and time-series correlation for three different ROIs: area activated in both sessions ($S_1 \cap S_2$), area activated in either the first or second session ($S_1 \Delta S_2$), area not activated in any session ($U/(S_1 \cup S_2)$). P values were estimated using pair-wise (within-subject) one sample bootstrap test. Each colour represents a different subject. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

with the stimuli mattered whilst for t_{diff} and Dice, only motion correlated with the stimuli mattered. This can be explained by the fact that t-values depends strongly on the signal correlated with the task (beta value) while the whole time-series correlation is also affected by the overall motion. This finding has implication towards behavioural task design and poses a question of theoretical upper limit on single subject reliability of motion related tasks. It is however important to also acknowledge the limitations of our modelling approach. We did not control explicitly the levels of confounding factors. In this study we have relation between reliability and measured (but not induced experimentally) confounds. What we are reporting is how much those factors explain the variability in reliability measures. This approach has some obvious limitations—for example if all of the subjects were expressing substantial motion but of an identical level there would be no variance within the confounding factor and it would yield no explanatory value. However, as we shown in Fig. 1 we have a reasonable spread of combinations of values of confounds across subjects and tasks, which allows concluding reasonably on the contribution of each factor.

In conclusion, we have shown that task and motion are the major contributor to single subject reliability whilst scanner noise, coregistration errors have little influence. Additionally we have found that the relationship between time-series correlation, t values difference and Dice overlap is not simply linear. We are also recommend using a within- vs. between-subject Dice overlap difference as a way for evaluating single subject fMRI paradigms.

Acknowledgements

We would like to thank the reviewers for insightful comments. Cyril Pernet is partly funded by SINPASE. This study was funded by the Edinburgh Experimental Cancer Medicine Centre. We would like to thank Prof. Joanna Wardlaw for securing the funding which made this study possible.

Appendix A. Relation between t_{diff} and ICC

The t_{diff} metric which we have used to measure the variance of unthresholded T-maps relates to the between-subjects variance of the ICC metric. The mean t_{diff} across subjects is inversely proportional to ICC, assuming constant between-subjects variance across sessions. The derivations for this relation depend on the assumptions made while calculating ICC (for k = 2 sessions case):

1. ICC(1) assumes no session (learning) effects and is defined for one voxel as following

$$ICC(1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_w^2}$$

where σ_r^2 is between subjects variance and σ_w^2 is defined in the following manner

$$\sigma_w^2 = \frac{\sum_{j=1}^{n}\left(t_{1j} - t_{2j}\right)^2}{n} = \frac{\sum_{j=1}^{n} t_{\mathrm{diff}(j)}}{n}$$

where n is the number of subjects, t_{1j} and t_{2j} are t values for subject j for first and second sessions respectively, $t_{diff(j)}$ is t_{diff} for subject j as defined in the paper. Therefore:

$$ICC(1) \propto \frac{1}{\sigma_w^2} U\sigma_w^2 \propto t_{\mathrm{diff}} \rightarrow ICC(1) \propto \frac{1}{t_{\mathrm{diff}}}$$

2. ICC(3,1) assumes session effects (learning) and is defined as

$$ICC(3,1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}$$

where σ_e^2 is defined in the following manner

$$\sigma_e^2 = \frac{\sum_{j=1}^{n}\left(\left(t_{1j} - \overline{t_1}\right) - \left(t_{2j} - \overline{t_2}\right)\right)^2}{n-1}$$

Where $\overline{t_1}$ and $\overline{t_2}$ are across subjects mean t values for first and second sessions respectively. Since:

$$\begin{aligned}\left(\left(t_{1j} - \overline{t_1}\right) - \left(t_{2j} - \overline{t_2}\right)\right)^2 &= \left(\left(t_{1j} - t_{2j}\right) - \left(\overline{t_1} - \overline{t_2}\right)\right)^2 \\ &= \left(t_{1j} - t_{2j}\right)^2 - 2\left(t_{1j} - t_{2j}\right)\left(\overline{t_1} - \overline{t_2}\right) + \left(\overline{t_1} - \overline{t_2}\right)^2 \\ &= \left(t_{1j} - t_{2j}\right)^2 + \left(\overline{t_1} - \overline{t_2}\right)\left(\overline{t_1} - \overline{t_2} - 2\left(t_{1j} - t_{2j}\right)\right)\end{aligned}$$

Therefore:

$$\begin{aligned}\sigma_e^2 &= \frac{\sum_{j=1}^{n}\left[\left(t_{1j} - t_{2j}\right)^2 + \left(\overline{t_1} - \overline{t_2}\right)\left(\overline{t_1} - \overline{t_2} - 2\left(t_{1j} - t_{2j}\right)\right)\right]}{n-1} \\ &= \frac{\sum_{j=1}^{n}\left[t_{\mathrm{diff}(j)} + \left(\overline{t_1} - \overline{t_2}\right)\left(\overline{t_1} - \overline{t_2} - 2\left(t_{1j} - t_{2j}\right)\right)\right]}{n-1}\end{aligned}$$

The relation between σ_e^2 and $t_{diff(j)}$ depends on the contributions of the sessions effects (mainly through the $(\overline{t_1} - \overline{t_2})$ part of the equation). However for small or no session effects the relation still holds:

$$ICC(3,1) \propto \frac{1}{\sigma_w^2} U\sigma_w^2 \propto t_{\mathrm{diff}} \rightarrow ICC(1) \propto \frac{1}{t_{\mathrm{diff}}}$$

Appendix B. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.neuroimage.2012.10.085.

References

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H.B., Zilles, K., 1999. Broca's region revisited: cytoarchitecture and intersubject variability. J. Comp. Neurol. 412, 319–341.
Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38, 95–113.
Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. 1191, 133–155.
Caceres, A., Hall, D.L., Zelaya, F.O., Williams, S.C.R., Mehta, M.A., 2009. Measuring fMRI reliability with the intra-class correlation coefficient. NeuroImage 45, 758–768.
Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., Zilles, K., 2006. The human inferior parietal cortex: cytoarchitectonic parcellation and interindividual variability. NeuroImage 33, 430–448.
Chumbley, J.R., Friston, K.J., 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. NeuroImage 44, 62–70.
Corbetta, M., Shulman, G.L., 2011. Spatial neglect and attention networks. Annu. Rev. Neurosci. 34, 569–599.
Dice, L., 1945. Measures of the amount of ecologic association between species. Ecology 26, 297–302.
Dong, Y., Dobkin, B.H., Cen, S.Y., Wu, A.D., Winstein, C.J., 2011. Motor cortex activation during treatment may predict therapeutic gains in paretic hand function after stroke, pp. 1552–1555 (October).
Duncan, K.J., Pattamadilok, C., Knierim, I., Devlin, J.T., 2009. Consistency and variability in functional localisers. NeuroImage 46, 1018–1026.
Eickhoff, S.B., Stephan, K.E., Mohlberg, H., Grefkes, C., Fink, G.R., Amunts, K., Zilles, K., 2005. A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. NeuroImage 25, 1325–1335.
Eickhoff, S.B., Heim, S., Zilles, K., Amunts, K., 2006. Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. NeuroImage 32, 570–582.
Eickhoff, S.B., Paus, T., Caspers, S., Grosbras, M.-H., Evans, A.C., Zilles, K., Amunts, K., 2007. Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. NeuroImage 36, 511–521.
Fernández, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., Klaver, P., Ruhlmann, J., Reul, J., Elger, C.E., 2003. Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. Neurology 60, 969–975.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.-P., Frith, C.D., Frackowiak, R.S.J., 1994. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.
Geyer, S., Ledberg, A., Schleicher, A., Kinomura, S., Schormann, T., Bürgel, U., Klingberg, T., Larsson, J., Zilles, K., Roland, P.E., 1996. Two different areas within the primary motor cortex of man. Nature 382, 805–807.
Gorgolewski, K., Burns, C.D., Madison, C., Clark, D., Halchenko, Y.O., Waskom, M.L., Ghosh, S.S., 2011a. Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in Python. Front. Neuroinformatics 5, 13.
Gorgolewski, K., Storkey, A., Bastin, M., Pernet, C., 2011b. Using a combination of a mixture model and topological FDR in the context of presurgical planning. 17th Annual Meeting of the Organization for Human Brain Mapping.
Gorgolewski, K.J., Storkey, A.J., Bastin, M.E., Pernet, C.R., 2012. Adaptive thresholding for reliable topological inference in single subject fMRI analysis. Front. Hum. Neurosci. 6, 245. http://dx.doi.org/10.3389/fnhum.2012.00245.
Hall, D.A., Haggard, M.P., Akeroyd, M.A., Palmer, A.R., Summerfield, A.Q., Elliott, M.R., Gurney, E.M., Bowtell, R.W., 1999. "Sparse" temporal sampling in auditory fMRI. Hum. Brain Mapp. 7, 213–223.
Harrell, F., Davies, C., 1982. A new distribution-free quantile estimator. Biometrika 69, 635–640.
Holmes, A., Friston, K.J., 1998. Generalisability, random effects & population inference. NeuroImage S754.
Jaccard, P., 1901. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. Bull. Soc. Vaudoise Sci. Nat. 37, 547–579.
Lindeman, R.H., Merenda, P.F., Gold, R.Z., 1980. Introduction to bivariate and multivariate analysis. Scott, Foresman, and Company, Glenview, IL.
McGraw, K.O., Wong, S.P., 1996. Forming inferences about some intraclass correlation coefficients. Psychol. Methods 1, 30–46.
Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., Zilles, K., 2001. Human primary auditory cortex: cytoarchitectonic subdivisions and mapping into a spatial reference system. NeuroImage 13, 684–701.

Morosan, P., Schleicher, A., Amunts, K., Zilles, K., 2005. Multimodal architectonic mapping of human superior temporal gyrus. Anat. Embryol. 210, 401–406.
Ohgaki, H., 2009. Epidemiology of brain tumors. Methods Mol. Biol. 472, 323–342.
Raemaekers, M., Vink, M., Zandbelt, B., van Wezel, R.J.A., Kahn, R.S., Ramsey, N.F., 2007. Test–retest reliability of fMRI activation during prosaccades and antisaccades. NeuroImage 36, 532–542.
Raemaekers, M., du Plessis, S., Ramsey, N., Wuesten, J., Vink, M., 2012. Test retest variability underlying fMRI measurements. NeuroImage 60, 717–727.
Raschle, N.M., Zuk, J., Gaab, N., 2012. Functional characteristics of developmental dyslexia in left-hemispheric posterior brain regions predate reading onset. Proc. Natl. Acad. Sci. U. S. A. 109, 2156–2161.
Roche, A., Malandain, G., Pennec, X., 1998. The correlation ratio as a new similarity measure for multimodal image registration. Med. Image Comput. 1496.
Rousselet, G.A., Pernet, C.R., 2012. Improving standards in brain-behavior correlation analyses. Front. Hum. Neurosci. 6.
Savoy, R.L., 2005. Experimental design in brain activation MRI: cautionary tales. Brain Res. Bull. 67, 361–367.
Shrout, P.E., Fleiss, J.L., 1979. Intraclass correlations: uses in assessing rater reliability. Psychol. Bull. 86, 420–428.
Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17, 143–155.
Smith, S.M., Beckmann, C.F., Ramnani, N., Woolrich, M.W., Bannister, P.R., Jenkinson, M., Matthews, P.M., McGonigle, D.J., 2005. Variability in fMRI: a re-examination of inter-session differences. Hum. Brain Mapp. 24, 248–257.
Stippich, C., Blatow, M., Krakow, K., 2007. Presurgical functional MRI in patients with brain tumours. In: Stippich, C. (Ed.), Clinical Functional MRI. Springer, Berlin, pp. 87–134.
Ulrike Grömping, 2006. Relative importance for linear regression in R: the package relaimpo. J. Stat. Softw. 17.
Wilcox, R., 2005. Introduction to robust estimation and hypothesis testing, 2nd ed. Academic Press, Burlington, MA, USA.
Wilke, M., 2012. An alternative approach towards assessing and accounting for individual motion in fMRI timeseries. NeuroImage 59, 2062–2072.