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 $\sigma_r^2$ is between-subjects (rows) variance and $\sigma_e^2$ 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 $\sigma_r^2$ and low $\sigma_e^2$ or low $\sigma_r^2$ and high $\sigma_e^2$, 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 ($\sigma_e^2$) but different between-subjects variance ($\sigma_r^2$), 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.