Chapter number

Reliability Maps in Event Related Functional MRI Experiments

Aleksandr A. Simak[1,2], Michelle Liou*[1], Alexander Yu. Zhigalov[1],
Jiun-Wei Liou[2], Phillip E. Cheng[1]
[1]Institute of Statistical Science, Academia Sinica
[2]Department of Computer Science and Information Engineering,
National Taiwan University
Taiwan, R.O.C.

1. Introduction

In functional magnetic resonance imaging (fMRI) studies, the blood oxygen level-dependent (BOLD) signal change, in contrast to noise, is typically small (< 5%; e.g., Chen & Small, 2007). Although the quality of acquired image data may be improved by pre-processing images with low- or high-pass filters, classification of voxels into the active/inactive status could vary from one study to the next even when the same experimental paradigm is implemented (Maitra, 2009). Reliability assessment would contribute significantly to the knowledge on noise structures in image data, as a function of stimulus sequences, ethnic groups, imaging techniques and scanner differences (Biswal et al., 1996; Genovese et al., 1997; Maitra et al., 2002).

In the literature, there have been two main approaches to quantifying reliability of activation. The first involves the analysis of fMRI data acquired in a group of subjects (or more than one group) performing the same task in different days under multiple experimental conditions. The noise structure can be assessed by the intra-class correlation (ICC) analysis (Brennan, 2001; McGraw & Wong, 1996), which provides individual sources of noise associated with experiment-specific conditions (Aron et al., 2006; Fernandez et al., 2003; Franco et al., 2009; Friedman et al., 2008; Manoach et al., 2001; Miezin et al., 2000; Raemaekers et al., 2007; Specht et al., 2003; Zuo et al., 2010). The second approach considers the same group of subjects in multiple experimental replications, and evaluates test-retest reliability by modeling the number of times out of all replications, that a voxel is consistently classified as active given a decision threshold, as a mixture of binomial random variables (Genovese et al., 1997; Noll et al., 1997). This statistical approach has been extended by incorporating more accurate mixtures distributions and optimization procedure for estimating test-retest reliability (Gullapalli et al., 2005; Maitra et al., 2002).

Other than studying noise structures, reliability analysis would also provide information on invariant brain activity during the experimental session as a useful addition to the conventional measurement of response amplitudes. In a study using the forward-backward viewing movies paradigm, for instance, Hasson et al. (2010) have shown that brain responses in the visual cortex are highly reliable between subjects for both forward and backward presentations; responses in other cortical regions such as the precuneus, lateral sulcus, temporal-parietal junction tend to be less reliable in the backward presentation. However, disrupting the viewing order has no effect on response amplitudes in major cortical regions; markedly though, reliability magnitude varies in these regions. This type of studies has introduced dissociation between persistency and amplitude in brain activity.

The event-related paradigm was originally proposed for detecting transient BOLD responses to brief stimuli or tasks, but its potential use is not limited to short-term stimulation (Josephs et al., 1997). In statistical analysis of event-related fMRI data, a few design contrasts must be specified to estimate stimulus and task effects (Friston et al., 2002; Strother et al., 2004; Worsley et al., 2002). Conventionally, the ICC or test-retest reliability indices have been computed across experimental replicates using the t- or F-values, which are standardized parameter estimates in a linear model with BOLD responses as the dependent variable and a few design contrasts as regressors. The design contrast constitutes a hypothesis on temporal behavior in the brain following the stimulus or task onset.

In this chapter, we outline a reliability analysis procedure applicable directly to BOLD responses (i.e., image intensity) without a prior specification of design contrasts. In a sense, the procedure assesses the persistency in BOLD responses during the experimental session. Nonpersistency implies that a brain region is either heavily contaminated by noise or possibly contains a transient response, the onset of which is not reproducible between replicates. In applications, the procedure would suggest a collection of stable brain responses and their spatial distributions that may or may not be easily modeled or detected by using a weighted linear sum of a few basis functions (Lindquist et al., 2009). For instance, it might not be immediately clear how to specify design contrasts for a relatively longer duration of stimuli (> 40 sec per event) or for analysis of spontaneous brain activity under the eyes-closed and –open states. Stable BOLD responses can be further classified into distinct types featuring the time to response peak, amplitude, duration and sign (increased or decreased responses).

In the method section, we will elaborate the step-by-step procedure for assessing reliability of BOLD responses, testing reliability indices for statistical significance, and constructing reliability maps. For illustration, the method will be applied to an empirical dataset collected in a change detection task using the event-related paradigm (Huettel et al., 2001). Empirical results will show that the criterion of persistency is more sensitive to activity in the grey matter in contrast to that in the white matter. Finally, we will discuss the neurophysiological basis and clinical usage of reliability maps.

2. Hemodynamic Response Functions

In statistical analysis of fMRI data, it is important to model the BOLD responses as a function of the external stimulus (Buxton et al., 1997; Friston et al., 2000; Obata et al., 2004). By convention, a canonical hemodynamic response function (HRF) can be convolved with the external stimulus function to estimate the responses. The HRF can be formulated using one or two gamma functions to model a slight intensity dip after the response has fallen back to zero (Friston et al., 1998; Lange & Zeger, 1997). The estimated response at a particular scan is then subsampled from the response function specified at the scan acquisition time (Bandettini et al., 1993; Worsley et al., 2002). Canonical HRF assumes an instantaneous short stimulus with a few parameters determined by empirically observing activity in the primary visual cortex (Boynton et al., 1996; Glover, 1999); the function has been well fitted to experimental data in many fMRI studies involving healthy subjects, and is suitable for testing hypotheses on the strength and location of brain activation. However, the function may be ineffectual with experiments involving younger children or clinical patients.

There is an increasing amount of literature showing significant variability in HRFs between brain regions and subjects (Aguirre et al., 1998; Handwerker et al., 2004). The HRF variability also appears between experimental sessions recorded in different days on the same subject (Neumann et al., 2003). If variability in HRFs is known to be quite large, an empirically derived HRF can be used instead of the theoretical one in the generalized linear model (GLM) analysis of fMRI data (Handwerker et al., 2004). However, the variability problem cannot be precisely resolved by inserting an empirical HRF into the GLM because the HRF onset time and latency also varies seriously between brain regions especially in event-related experiments with long-term stimuli. In addition to microvasculature disturbances to variability in HRFs, the temporal behavior of BOLD signal has been found stable in repeated trials recorded in a single session (Aguirre et al., 1998; Miezin et al., 2000; de Zwart et al., 2005). Conditional on a fixed brain region, the HRF shape and amplitude can be highly reproducible within a subject.

Numerous studies in recent years have reported the relative efficiency of different HRF models, including finite impulse response models using basis functions and extension of the canonical HRF to more complicated situations with possible temporal and dispersion derivatives (Lindquist et al., 2009; Stephan et al., 2007). On the other hand, localization of brain activity can be done by data driven methods such as independent component analysis (ICA) or group ICA methods (Gu & Pagnoni, 2008; Varoquaux et al., 2010), which can extract reproducible components between experimental trials or between subjects. Most data driven methods assume non-Gaussian distributions for unknown temporal behaviors, which would make thresholding more difficult in constructing activation maps based on voxel-wise component scores. As was mentioned, BOLD responses can be highly reproducible in a fixed brain region within each subject. In this chapter we introduce a simple procedure for research into stable temporal behaviors in the brain, which can be further classified into different response patterns for selecting regions of interest (ROIs) or for GLM analysis.

3. Measures of Reliability

Reliability analysis requires assessment data to be structured in similar events or replicates. Event-related fMRI experiments are normally conducted over a period of time which is split into smaller segments or experimental runs to allow subjects to rest. Different runs can be considered as experimental replicates implemented under the same condition for evaluation of between-run reliability. Although the notation used in this section has been designed for between-run reliability analysis, a generalization of the method to other types of situations can be easily made by analogy (e.g., between trials or between subjects). The ICC index is a prominent statistic for measuring reliability of image data between runs. Here we specify the assumption used for computing the index, and its potential competitors. Let S denote the variance–covariance matrix of image data between M runs in a single voxel. The ICC index can be expressed as

$$\text{ICC} = \frac{\text{M}}{\text{M}-1}\left[1 - tr(\mathbf{S})(\underline{1}'\mathbf{S}\underline{1})^{-1}\right], \tag{1}$$

where $\underline{1}$ is the summing vector of order M, and $tr(\boldsymbol{S})$ is the trace of $\boldsymbol{S}$. The index has additionally assumed that the assessment data between replicates can be expressed as an additive equation except for random measurement errors.

The index is particularly sensitive to the variance within each run; if variances of image intensity vary from one run to the next, the size of ICC becomes smaller. However, the index is unaffected by adding a constant to image intensity within each run. As an alternative to ICC, the agreement index is sensitive to all aspects of between-run variation including the mean image intensity. An interested reader may refer to McGraw & Wong (1996) for a detailed comparison between the ICC and agreement index. In general, the two indices give comparable results when the number of scan volumes increases in each run. In some experimental paradigms, the scale of image intensity is allowed to vary between runs, and the additive assumption could be too restrictive for general applications. For example, tasks with high and low working memory loads are implemented in different runs. There are also reliability indices robust to scale changes; that is, image data in one run can be expressed as a linear combination of those in another run.

In our empirical studies, reliability indices with lesser restrictive assumptions always yield greater index values, but the ordering of voxels according to index values remains unchanged especially with fMRI data. Interested readers may refer to Liou (1989) for a review on robust reliability indices. An alternative approach is to specify a common factor model underlying image data and to compute the reliability index based on factor loadings (McDonald, 1999). If the common factor model is misspecified, the reliability estimate using factor loadings could be seriously biased (Yang & Green, 2010). An empirical comparison between the ICC and factor analysis models for estimating reliability can be found in Luke (2005).

We now assume that the maximum autocorrelation coefficient in a fMRI time series decreases toward zero as the time-lag between the correlated observations increases toward infinity. It follows from a standard result for a weakly dependent sequence of random variables (Peligrad, 1996, Theorem 2.1) that the asymptotic distribution of elements in $\boldsymbol{S}$ can be assumed to be multivariate Gaussian as the number of scan volumes $n \to \infty$,

$$\sqrt{n}(\text{vech}\boldsymbol{S} - \text{vech}\boldsymbol{\Phi}) \to \text{N}(0, 2\text{H}_\text{M}(\boldsymbol{\Phi}\otimes\boldsymbol{\Phi})\text{H}_\text{M}'),$$

where $\otimes$ is the Kronecker product and $\boldsymbol{\Phi}$ is the population counterpart of S; vech$\boldsymbol{S}$ denotes the vector of those non-duplicated elements in S, and the operational matrix $\mathrm{H_M}$ satisfies the identity vech$\boldsymbol{S}$ =$\mathrm{H_M}$vec$S$ (Henderson & Searle, 1979). Because ICC is a differentiable function of a multivariate Gaussian vector, the asymptotic variance of ICC can be derived as follows

$$\mathrm{Var(ICC)} \approx 2\mathrm{n}^{-1}\mathrm{d}'\mathrm{H_M}(\boldsymbol{\Phi}\otimes\boldsymbol{\Phi})\mathrm{H}'_\mathrm{M}\mathrm{d}, \tag{2}$$

Where d′ is the derivative of ICC with respect to vech′S as follows:

$$\mathrm{d}' = \frac{\mathrm{M}}{(\mathrm{M}-1)\left(\underline{1}'\boldsymbol{\Phi}\underline{1}\right)^2}\left[-\left(\underline{1}'\boldsymbol{\Phi}\underline{1}\right)\mathrm{vec}'I_\mathrm{M} + \, tr(\boldsymbol{\Phi})\left(\underline{1}'\otimes\underline{1}'\right)\right]\mathrm{G_M}$$

and vec$\boldsymbol{S}$ = $\mathrm{G_M}$vech$\boldsymbol{S}$. With a moderate size of n (> 100), it is reasonable to assume that

$$\mathrm{Z = (ICC - \mu_{ICC})/Std(ICC)} \tag{3}$$

is distributed as a standard Gaussian distribution with mean 0 and variance 1, where μ_{ICC} is the mean value in the population and Std(ICC) denotes the square root of Var(ICC) in (2). In applications, one may hypothesize that $\mu_{\mathrm{ICC}} = 0$ and test an observed ICC for statistical significance against this hypothesis.

4. Multiple testing

The ICC value is computed by using temporal information within each individual voxel, and the resulting Z value in (3) can be tested against a nominal Type-I error rate α. In this chapter, we assume that only positive ICC values are acceptable. With α = 0.05 for a one-tailed test, for example, voxels with $Z \geq 1.64$ can be selected for constructing reliability maps. As the number of voxels to be evaluated increases, the likelihood of having at least one Type-I error out of all tests also increases in the experiment. By the Bonferroni inequality, $p_B \leq \alpha/\mathrm{V}$, where V is the total number of voxels with positive Z values, the familywise error (FWE) rate can be controlled at α by selecting error rate at p_B for each test. Because of the spatial dependence among image voxels, the Bonferroni procedure tends to be too conservative in general (Nichols & Hayasaka, 2003).

Alternatively, the multistep test uses a sequence of ordered p-values which are compared against different thresholds. The false discovery rate (FDR) is a step-up test widely applied in neuroimaging studies (Chumbley et al., 2010; Genovese et al., 2002; Langers et al., 2007). The FDR procedure considers $p_{\mathrm{FDR}} \leq (i/\mathrm{V})[\alpha/\mathrm{C(V)}]$ as the critical value for the i-th test in the ordered sequence to control the false positive rate at α, where $\mathrm{C(V)}$ is a predetermined constant. The choice of the constant depends on the joint distribution of p-values in the sequence. For instance, $\mathrm{C(V)} = 1,$ in case of the Gaussian noise with nonnegative correlation across voxels, and $\mathrm{C(V)} = \sum_{i=1}^{V} i^{-1}$, in case of no assumption on dependence (Genovese et al., 2002). The FDR procedure is easily implemented even for large data sets, and more powerful than the Bonferroni procedure (Langers et al., 2007; McNamee & Lazar, 2004; Nichols & Hayasaka, 2003).

The random field theory (RFT) methods account for spatial dependence in the data, as captured by the maximum of a random field (Nichols & Hayasaka, 2003; Worsley, 1996). It has been shown that the probability of observing a cluster of voxels exceeding a threshold in a smooth Gaussian RF can be approximated by the expected Euler characteristic (EC). The EC counts the number of clusters above a sufficiently high threshold in a smoothed Gaussian RF. Methods based on the RFT comprise a flexible framework for neuroimaging inference, but RFT relies on the assumptions of stationarity and smoothness (Nichols et al., 2003). Without spatial smoothing on ICC values, the RFT methods yield similar results to the Bonferroni procedure in our applications. In the empirical example, we will only present results based on the FDR procedure with C(V) = 1. An interested reader may refer to Nichols and Hayasaka (2003) for a detailed comparison between different approaches to controlling the FWE.

5. Reliability Maps

In fMRI experiments, it is reasonable to assume that environmental, physiological, and psychological factors fluctuate randomly throughout the experimental period on a moment-by-moment basis, and these random effects occur equally likely in all runs. Imaging techniques, such as pulse sequences, imaging parameters and scanner performance, also affect the quality of observations. Those artifacts, however, may be systematic and non-random as long as the same scanners, sequences, and parameters are implemented in the experiments. In addition to regular prewhitening procedures (i.e., slice timing, motion correction, and adjustment for autocorrelation), the fMRI time series in reliability analysis must be corrected for major trend effects in order to account for magnetic field drifts.

There are qualitative descriptions of reliability indices, for instance, poor (<0.00), slight (0.00-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80), and almost perfect (0.81-1.00) (Landis & Koch, 1977). The size of ICC also depends on the number of runs and number of scan volumes. However, the standardized ICC in (3) can take into account the sample size effects. In order to construct the reliability maps, the Z value is computed by substituting sample estimate $\boldsymbol{S}$ for population $\boldsymbol{\Phi}$ in (3) for each individual voxel using the preprocessed fMRI time series. Voxels with Z values significantly greater than zero can be selected to construct the reliability maps (e.g., $Z \geq 1.64$ with Type I error controlled at $\alpha = 0.05$). In order to control the FWE, the Bonforroni correction, random field theory, and FDR control methods can be applied (Hochberg & Tamhane, 1987; Worsley et al., 1996). In the empirical example, we only present results based on the FDR method which yields most reasonable findings as compared with the other two methods.

The reliability analysis can be applied to each individual subject as well as to a group of subjects. After normalization of each subject's fMRI scans to a standard brain atlas (e.g., the MNI brain), the group standardized ICC for K subjects can be computed as follows:

$$Z_{\mathrm{G}} = \left(\sum_j^K \mathrm{ICC}_j\right) \Big/ \sqrt{\sum_j^K \mathrm{Var}(\,\mathrm{ICC}_j)}\;, \tag{4}$$

where ICCj denotes the ICC index corresponding to the j-th subject. The Z_G values can be tested for significance against a standard Gaussian distribution with FDR control of FWE in the normalized space.

6. Empirical Example

We illustrate the use of the reliability analysis procedure in an example using the long-term stimulus in the experiment (42 sec per event). The dataset was collected in an event-related fMRI experiment involving 10 subjects for investigating brain functions in a change-detection task (Huettel et al., 2001; fMRIDC Accession No: 2-2001-111T9). There were 10 stimulus trials in each run, and each subject completed 10-12 runs in his/her experimental session. In each trial, a pair of images was presented with difference in either the presence/absence of a single object or color of the object. The subjects made the behavioral response by pressing a button when they felt that there was something changing on the trial. Each trial began with 2 sec fixation cross at the center of the screen as a warning signal, followed by the first 30 sec of the trial, during which two images were presented for 300 ms, separated by a 100-ms mask. The mask was removed during the last 10 sec of the trial, and the stimuli alternated every 400ms. During the experiment, the subjects were instructed to keep their eyes on the display at all times.

Figure 1 lists the frequency distribution of categorized trials in four response-time intervals, namely, 0-10, 10-20, 20-30 and 30-40 sec according to the time point at which a subject made the change identification responses. All distributions in the figure show two frequency peaks in the first and last bins. On the average, the between-subject variation is small by inspecting the behavioral data. Figure 2 plots the frequency distribution of voxel-wise Z values in (3) for each of the 10 subjects. Theoretically, the ICC values lie within the range of $(-\infty, 1]$, and the plots suggest that all subjects have similar ICC distributions except for Subjects 7 and 8, whose ICC values are mainly distributed in the negative direction (smaller proportions of positive ICC values). The two subjects have average shorter reaction times as compared with other subjects. During the experimental sessions, instantaneous BOLD responses could occur in the two subjects, which may or may not be related to the task. The image data of these two subjects were later eliminated from the group reliability analysis. In applications, visual inspection on the ICC distributions would suggest removing irregular subjects from the group analysis.

Reliability analysis was applied to each individual subject's data without normalization of image to the MNI template. Figure 3 shows the reliability maps for four subjects whose behavioral data and ICC values have comparable distributions in Figures 1 and 2, respectively. The maps for each subject were constructed by computing Z values in (3) for every voxel, and then testing these values for significance against a standard Gaussian distribution using the FDR control of FWE at $\alpha = 0.05$. The colored overlays in Figure 3 are those voxels exceeding the FDR threshold, and were shown by using each subject's own anatomy in the background. The time series data in the colored regions were further clustered into different patterns using the k-mean method (MacQueen, 1967). Figure 3 also plots the response patterns that are reproducible between the four subjects derived from the k-mean method.

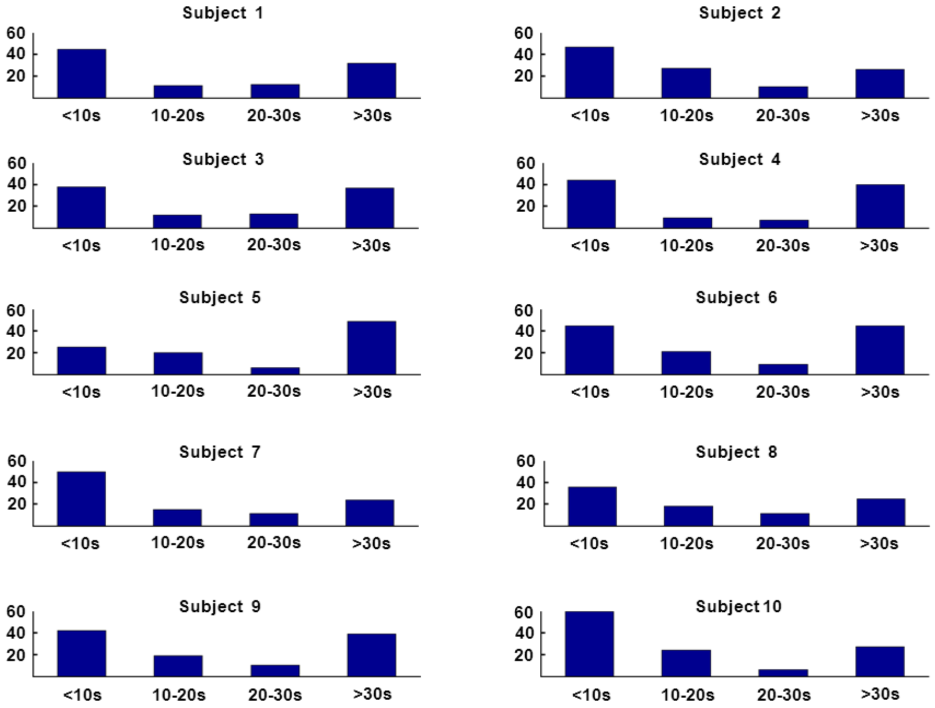

Figure 1: The frequency distribution of categorized trials in four response-time intervals, namely 0-10, 10-20, 20-30 and 30-40 sec, according to the time point at which a subject made the change identification responses.

The response functions in Figure 3 suggest that there are at least two types of increased BOLD responses (blue and yellow) time locked to the stimulus onset, and one type of decreased responses (green) also time locked to the stimulus onset. The onset time of one increased response (red) is earlier than the stimulus onset. The time to the response peak varies between the three increased BOLD responses and between the four subjects. The time to the dip in the decreased response also varies between the four subjects. The response function in yellow in the figure shows a minor peak in the last 10 sec of the trial during which the mask was removed between images in the change detection task. In applications, these response functions can be smoothed and inserted into the design matrix of GLM in SPM or FSL for advanced statistical analysis, such as comparing stimulus or task effects in different groups.

In order to illustrate group reliability maps, functional and anatomical images of eight subjects (1, 2, 3, 4, 5, 6, 9 and 10) with more reliable data were normalized to the MNI template. The ICC values corresponding to the same voxel in the normalized brain were computed for each of the eight subjects. The average ICCs across the eight subjects and Z_G values were computed for all voxels in the normalized brain. Figure 4 shows the group reliability maps in different brain regions. Brain regions with higher Z_G values imply that there were stable temporal behaviors during the experimental session in these regions. The reliability maps might suggest potential ROIs for probing high-level brain functions.

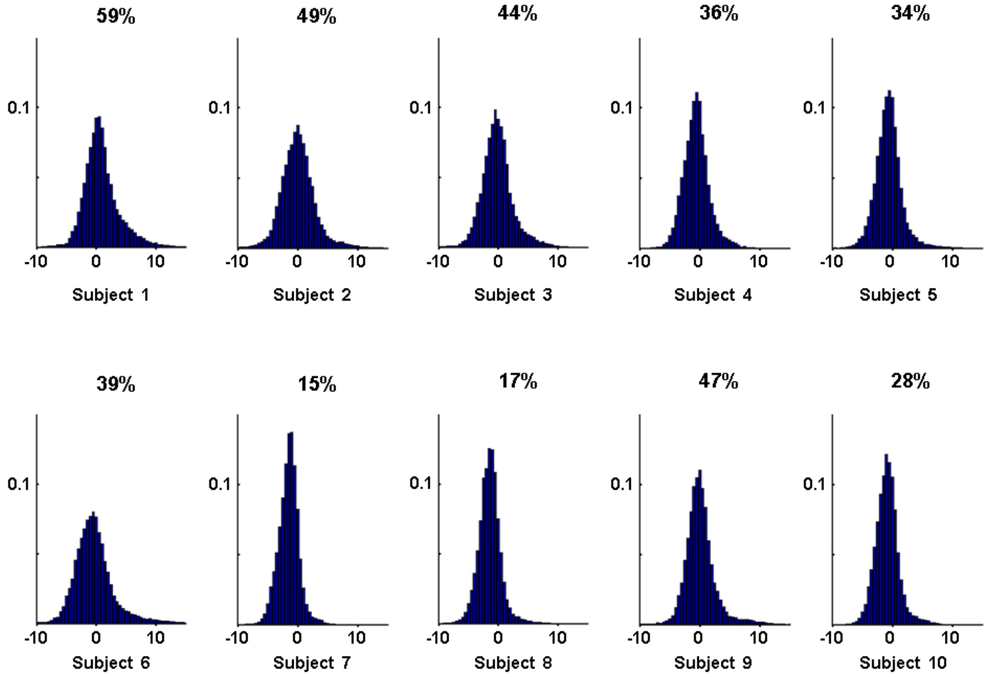

Figure 2: The frequency distributions of standardized ICC values for each of the 10 subjects. The percentage above each individual's histogram shows the proportion of voxels with positive ICC values.

The average time series across the eight subjects in different voxels with significant Z_G values were clustered using the k-mean method. Figure 6 shows the BOLD responses based on the average time series. According to the figure, there are at least two types of decreased responses (green and brown) in the group reliability maps. Figure 7 shows the brain regions corresponding to different BOLD response patterns in Figure 6. The color overlays in Figure 5 and 7 have the same spatial locations with the normalized anatomy in the background. It is interesting to note that the onset time of responses in the parahippocampal gyrus, posterior cingulate and precuneus (the red plot) is slightly earlier than the stimulus onset. This suggest that a mechanism could be carried over from the previous trial to a new trial. The parahippocampal gyrus participates in novelty perception, and the posterior cingulate and precuneus, in attentional shift. The early-onset response could be induced by a preparatory mechanism in anticipation of an expected task (Sirotin & Das, 2009).

Responses in the parahippocampal gyrus, middle frontal gyrus, precuneus and superior/inferior parietal lobule (the yellow plot) show a second peak in the last 10 sec of each trial during which the mask was removed between two images. The second peak could be induced by the task change (with/without the mask). The early decreased responses (in green) in the insula, supramarginal, angular gyrus and precuneus could be coupled with the early increased responses (in red) carried over from the previous trial. The late decreased responses (in brown) in the cingulate gyrus, superior temporal gyrus and medial frontal gyrus could be coupled with increased responses (blue and yellow) for performing the change detection task.

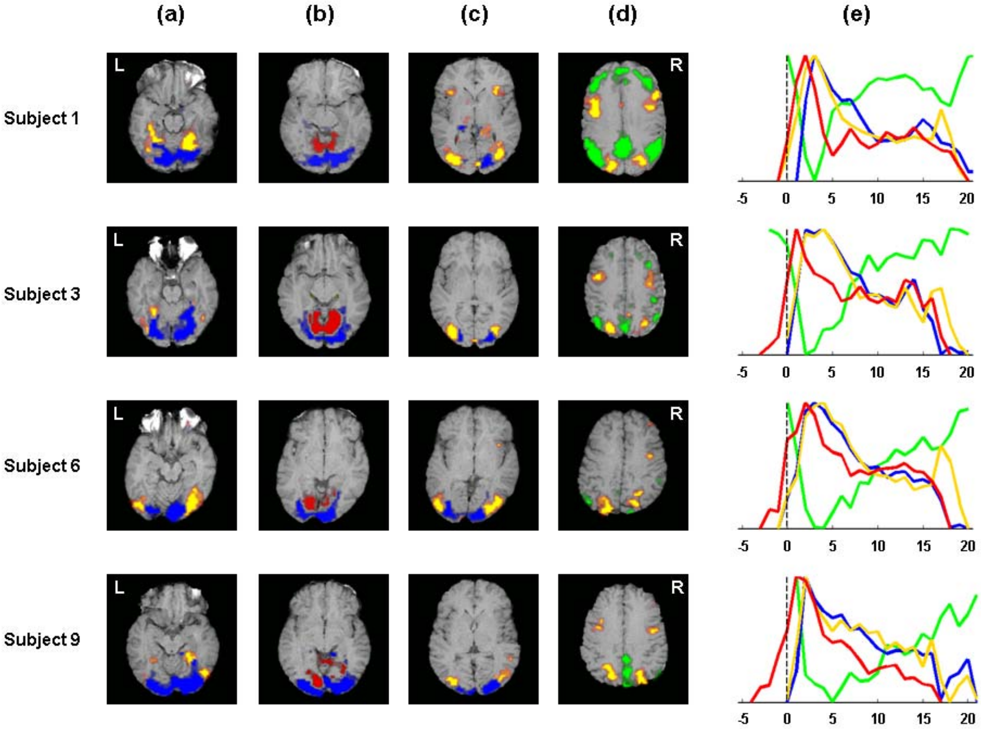

Figure 4: The reliability maps of four subjects participating in the change detection task. The colored voxels have ICC values that are significantly greater than zero by a standard Gaussian test with the FDR control of FWE at α = 0.05. The selected slices are all in the axial sections; the slices in columns (a), (b), (c) and (d) locate approximately at z = -9, -4, +5 and +33, respectively. The BOLD response plots in column (e) are those stable response patterns associated with each subject from the k-mean method. The spatial distribution of each response pattern in the brain is highlighted using the same color. The reliability maps are reported by showing in a subject's own anatomy in the background with colored overlays indicating those voxels with Z values exceeding the FDR thresholds.

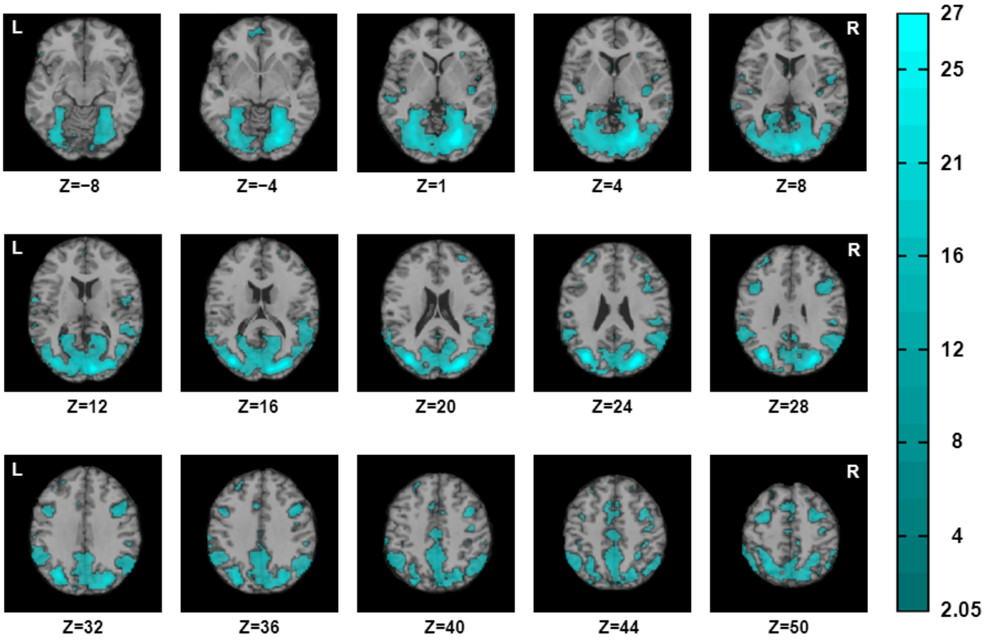

Figure 5: The group reliability maps for eight subjects participating in the change detection task. The colored voxels have Z_G values significantly greater than zero by a standard Gaussian test with the FDR control of FWE at α = 0.05. Coordinates are in the normalized space of the Talairach and Tournoux 1988 brain atlas. The intensity of the color indicates the size of Z_G values.

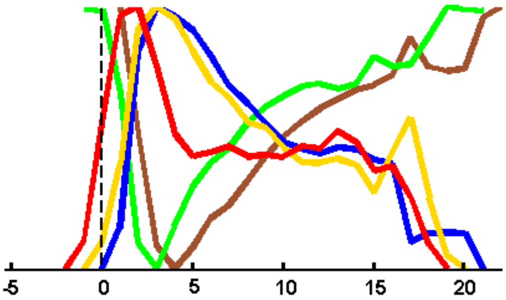

Figure 6: Stable response patterns from the k-mean method; the patterns are found by averaging time series across 8 subjects in each voxel in the normalized space.

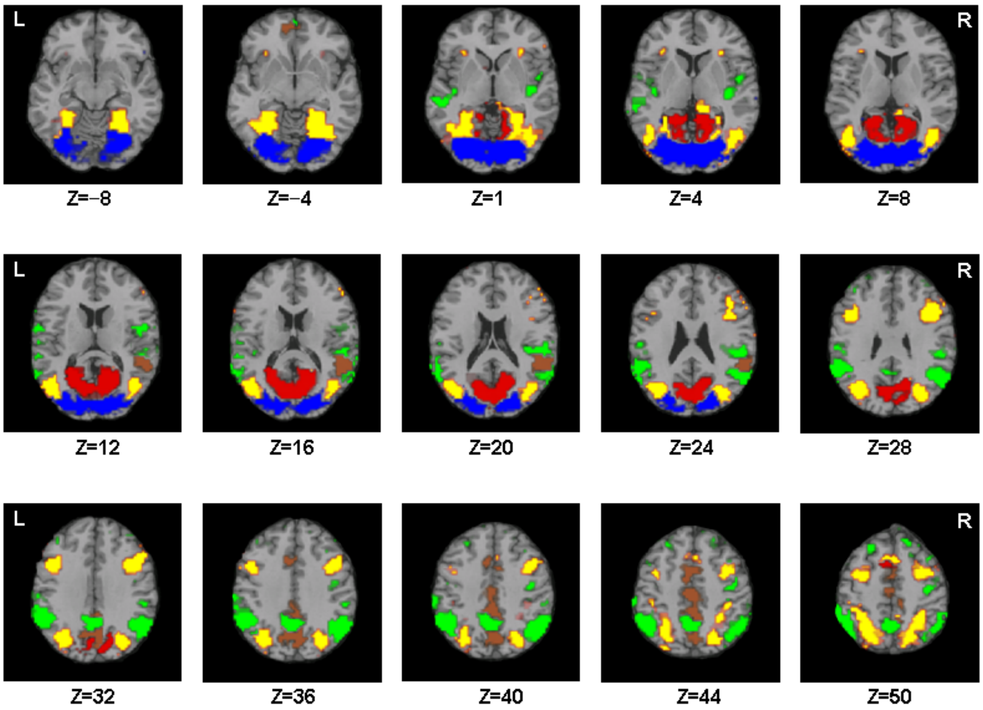

Figure 7: The group reliability maps for eight subjects using the same colors corresponding to the response plots in Figure 6. The increased responses in blue are mainly distributed in the fusiform gyrus, lingual gyrus, middle occipital gyrus and cuneus; those in yellow are distributed in the parahippocampal gyrus, middle frontal gyrus, precuneus, and superior/inferior parietal lobule; those in red are distributed in the parahippocampal gyrus, posterior cingulate and precuneus. The decreased responses in green are distributed in the insula, precuneus, and inferior parietal lobule; those in brown are distributed in the cingulate gyrus, superior temporal gyrus and medial frontal gyrus.

7. Discussion

Comparing with other existing methods, reliability maps are simple in construction, and offer rich information on temporal behaviors of different brain regions. The method is applicable to each individual subject as well as to a group of subjects, and has potential use for investigating the HRF variability between subjects and between brain regions. Without reliability assessment, some functional mechanisms could be overlooked in the analysis such as novelty perception in the last few seconds of each trial in the change detection task. One restriction of reliability assessment is that the fMRI time series must be partitioned into replicates (e.g., experimental runs). This might not pose difficulty in applications. For instance, we applied the same procedure to the resting state fMRI with alternate eyes-closed and –open periods (3 min per state with two replications). It was found that the thalamus showed most reliable results, but BOLD responses in this region were quite different from those in other regions. Regions located in the default mode network (e.g., precuneus) also had moderate sizes of reliability. By the conventional approaches to analyzing resting state data (e.g., independent component analysis, and the seed correlation approach), one would mainly find those regions located in the default mode network. We conclude that persistency in BOLD responses as measured by the ICC index is an important indicator of temporal behaviors in fMRI experiments.

The asymptotic theory supporting reliability maps need not require a large number of scan volumes or any stationarity assumption. In our experience, the procedure works well in datasets involving two runs with 160 scan volumes each. In clinical studies involving patients, the number of experimental replicates might be small (a few runs), and the reliability maps would support information on functional distortions in task execution, especially for those decreased responses in Figure 6. Although the k-mean methods can be applied directly to image data without using reliability maps, those clusters of larger size may give ambiguous response patterns for each subject. As was mentioned, nonpersistency might suggest random noise or transient BOLD responses. In event-related experiments with long-term stimuli, there might be several kinds of transient responses occurring in each trial with different onset times. For example, novelty in perceived images may depend on the order of stimulus presentation. The proposed reliability analysis procedure can only assess those temporal behaviors that regularly occur across trials and runs, which could be the major limitation of the procedure.

8. Conclusion

There have been an increasing number of event-related fMRI studies in cognitive, psychological, and medical research. The procedure for constructing reliability maps is proposed mainly for experiments using event-related designs involving a relatively longer stimulus trial. Its potential use is not limited to event-related designs, however. The method had successfully identified reliable regions in a block-design experiment with two runs involving alternate blocks of high- and low-load working memory tasks. Reliability maps suggest stable temporal behaviors in the experimental session that vary between brain regions among individual subjects. Reliability maps would assist researchers to select ROIs for further analysis, or to insert the obtained response functions into GLM for testing stimulus and task effects. In clinical studies with a small number of replicates, reliability maps can assist for detecting functional disorders in the brain for each individual patient. We conclude that the proposed procedure would support a stream of research probing more complicated BOLD responses in fMRI studies such as the early-onset mechanism in the change detection task.

9. Acknowledgment

The authors are indebted to the fMRIDC at the University of California, Santa Barbara for supporting the data sets analyzed in this study. This research was supported by grants NSC98-2410-H-001-012-MY3 and NSC099-2811-H-001-004 from the National Science Council, Taiwan.

10. Reference

Aguirre, G. K.; Zarahn, E. & D'Esposito, M. (1998). The variability of human BOLD hemodynamic responses. *NeuroImage*, Vol.8, No.4, (November 1998), pp. 360-369, ISSN 1053-8119

Aron, A. R.; Gluck, M. A. & Poldrack, R. A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *NeuroImage*, Vol.29, No.3, (February 2006), pp. 1000-1006, ISSN 1053-8119

Bandettini, P. A.; Jesmanowicz A.; Wong E. C. & Hyde J. S. (1993). Processing strategies for time-course data sets in functional MRI of the human brain. *Magnetic Resonance in Medicine*, Vol.30, No.2, (August 1993), pp. 161 – 173, ISSN 1522-2594

Biswal, B.; DeYoe, E. & Hyde, J. (1996). Reduction of physiological fluctuations in fMRI using digital filters. *Magnetic Resonance in Medicine*, Vol.35, No.1, (January 1996), pp. 107-113, ISSN 1522-2594

Boynton, G. M.; Engel, S. A.; Glover, G. H. & Heeger, D. J. (1996). Linear systems analysis of fMRI in human V1. *Journal of Neuroscience*, Vol.16, No.13, (July 1996), pp. 4207-4221, ISSN 0270-6474

Brennan, R. L. (2001). Generalizability theory. New York: *Springer Verlag*

Buxton, R. B. & Frank, L. R. (1997). A model for the coupling between cerebral blood flow and oxygen metabolism during neural stimulation. *Journal of Cerebral Blood Flow Metabolism*, Vol.17, pp. 64–72, ISSN 0271-678X

Chen, E. E. & Small, S. L. (2007). Test–retest reliability in fMRI of language: group and task effects. *Brain and Language*, Vol.102, No.2, (August 2007), 176–185, ISSN 0093-934X

Chumbley, J.; Worsley, K. J.; Flandin, G. & Friston, K. J. (2010). Topological FDR for neuroimaging. *NeuroImage*, Vol.49, No.4, (February 2010), pp. 3057-3064, ISSN 1053-8119

de Zwart, J. A.; Silva, A. C.; van Gelderen, P.; Kellman, P.; Fukunaga, M.; Chu, R.; Koretsky, A. P.; Frank, J. A. & Duyn, J. H. (2005). Temporal dynamics of the BOLD fMRI impulse response. *NeuroImage*, Vol.24, No.3, (February 2005), pp. 667–677, ISSN 1053-8119

Fernandez, 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*, Vol.60, No.6, (March 2003), pp. 969– 975, ISSN 0028-3878

Franco, A. R.; Pritchard, A.; Calhoun, V. D. & Mayer, A. R. (2009). Interrater and intermethod reliability of default mode network selection. *Human Brain Mapping*, Vol.30, No.7, (July 2009), pp. 2293–2303, ISSN 1097-0193

Friedman, L.; Stern, H.; Brown, G. G.; Mathalon, D. H.; Turner, J.; Glover, G. H.; Gollub, R. L.; Lauriello, J.; Lim, K. O.; Cannon, T.; Greve, D. N.; Bockholt, H. J.; Belger, A.; Mueller, B.; Doty, M. J.; He, J.; Wells, W.; Smyth, P.; Pieper, S.; Kim, S.; Kubicki, M.; Vangel, M. & Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Human Brain Mapping*, Vol.29, No.8, (August 2008), pp. 958-972, ISSN 1097-0193

Friston, K. J.; Josephs, O.; Rees, G. & Turner, R. (1998). Nonlinear event-related responses in fMRI. *Magnetic Resonance in Medicine*, Vol.39, No.1, (January 1998), pp. 41-52, ISSN 1522-2594

Friston, K. J.; Mechelli, A.; Turner, R. & Price, C. J. (2000). Nonlinear responses in fMRI: The Balloon model, Volterra kernels and other hemodynamics. *NeuroImage*, Vol.12, No.4, (October 2000), pp. 466-477, ISSN 1053-8119

Friston, K. J.; Penny, W.; Phillips, C.; Kiebel, S.; Hinton, G. & Ashburner, J. (2002). Classical and Bayesian inference in neuroimaging: Theory. *NeuroImage*, Vol.16, No.2, (June 2002), pp. 465-483, ISSN 1053-8119

Genovese, C. R.; Lazar, N. A. & Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, Vol.15, No.4, (April 2002), pp. 870-878, ISSN 1053-8119

Genovese, C. R.; Noll, D. C. & Eddy, W. F. (1997). Estimating test-retest reliability in functional MR imaging I: Statistical methodology. *Magnetic Resonance in Medicine*, Vol.38, No.3, pp. 497-507, ISSN 1053-8119

Glover, G. H. (1999). Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, Vol.9, No.4, (April 1999), pp. 416–29, ISSN 1053-8119

Gu, Y. & Pagnoni, G. (2008). A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage*, Vol.42, No.3, (September 2008), pp. 1078-1093, ISSN 1053-8119

Gullapalli, R. P.; Maitra, R.; Roys, S.; Smith, G.; Alon, G., & Greenspan, J. (2005). Reliability Estimation of grouped functional imaging data using penalized maximum likelihood. *Magnetic Resonance in Medicine*, Vol.53, No.5, (May 2005), pp. 1126–1134, ISSN 1053-8119

Handwerker, D.; Ollinger, J. & D'Esposito M. (2004). Variation of BOLD hemodynamic responses across subjects and brain regions and their effects on statistical analyses. *NeuroImage*, Vol.21, No.4, (April 2004), pp. 1639–1651, ISSN 1053-8119

Hasson, U.; Malach, R. & Heeger, D. J. (2010). Reliability of cortical activity during natural stimulation. *Trends in Cognitive Sciences,* Vol.14, No.1, (December 2009), pp. 40-48, ISSN 1364-6613

Henderson, H. V. & Searle, S. R. (1979). Vec and Vech operators for matrices, with some uses in Jacobians and multivariate statistics. *The Canadian Journal of Statistics*, Vol.7, No.1, pp. 65-81, ISSN 0319-5724

Hochberg, Y. & Tamhane, A. C. (1987). Multiple Comparison Procedures. New York: *John Wiley,* ISBN 978-0471470151

Huettel, S. A., Guzeldere, G., & McCarthy, G. (2001). Dissociating neural mechanisms of visual attention in change detection using functional MRI. *Journal of Cognitive Neuroscience*, Vol.13, No.7, (October 2001), pp. 1006-1018, ISSN 0898-929X

Josephs, O.; Turner, R. & Friston, K. J. (1997). Event-Related fMRI, *Human Brain Mapping*, Vol.5, No.4, pp. 243-248, ISSN 1097-0193

Landis, J. R. & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, Vol.33, No.1, (March 1977), pp. 159-174, ISSN 0006-341X

Lange, N. & Zeger, S. L. (1997). Non-linear fourier time series analysis for human brain mapping by functional magnetic resonance imaging (with discussion). *Applied Statistics*, Vol.46, No.1, pp. 1-29, ISSN 1467-9876

Langers, D. R.; Jansen, J. F. & Backes, W. H. (2007). Enhanced signal detection in neuroimaging by means of regional control of the global false discovery rate. *NeuroImage*, Vol.38, No.1, (October 2007), pp. 43-56, ISSN 1053-8119

Lindquist, M. A.; Meng Loh J.; Atlas, L. Y. & Wager, T. D. (2009). Modeling the hemodynamic response function in fMRI: efficiency, bias and mis-modeling. *NeuroImage*, Vol.45, No.1, Suppl.1, (March 2009), pp. S187-198, ISSN 1053-8119

Liou M. (1989). A note on reliability estimation for a test with components of unknown functional length. *Psychometrika*, Vol.54, No.1, (March 1989), pp. 153-163, ISSN 0033-3123

Lucke, J. F. (2005). The α and the ω of congeneric test theory: An extension of reliability and internal consistency to heterogeneous tests. *Applied Psychological Measurement*, Vol.29, No.1, (January 2005), pp. 65-81, ISSN 1552-3888

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol.1, pp. 281-297, Berkeley, Calif.: University of California Press, 1967, ISSN 0097-0433

Maitra, R. (2009). Assessing certainty of activation or inactivation in test-retest fMRI studies. *NeuroImage*, Vol.47, No.1, (August 2009), pp.88-97, ISSN1053-8119

Maitra, R.; Roys, S. R. & Gullapalli, R. P. (2002). Test–retest reliability estimation of functional MRI data. *Magnetic Resonance in Medicine*, Vol.48, No.1, (July 2002), pp. 62–70, ISSN 1053-8119

Manoach, D. S.; Halpern, E. F.; Kramer, T. S.; Chang, Y.; Goff, D. C. & Rauch, S. L. (2001). Test–retest reliability of a functional MRI working memory paradigm in normal and schizophrenic subjects. *American Journal of Psychiatry*, Vol.158, No.6, (June 2001), pp. 955–958, ISSN 1535-7228

McDonald, R. P. (1999). Test theory: A unified approach. Mahwah, NJ: Lawrence Erlbaum, ISBN 0805830758

McGraw, K. & Wong, S. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, Vol.1, No.1, (March 1996), pp. 30–46, ISSN 1082-989X

McNamee, R. L. & Lazar, N. A. (2004). Assessing the sensitivity of fMRI group maps. *NeuroImage*, Vol.22, No.2, (June 2004), pp. 920-931, ISSN 1053-8119

Miezin, F. M.; Maccotta, L.; Ollinger, J. M.; Petersen, S. E. & Buckner, R. L. (2000). Characterizing the hemodynamic response: Effects of presentation rate, sampling procedure, and the possibility of ordering brain activity based on relative timing. *NeuroImage*, Vol.11, No.6, (June 2000), pp. 735-759, ISSN 1053-8119

Neumann, J.; Lohmann, G.; Zysset, S. & von Cramon, D. Y. (2003). Within-subject variability of BOLD response dynamics. *NeuroImage*, Vol.19, No.3, (July 2003), pp. 784-796, ISSN 1053-8119

Nichols, T. & Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, Vol.12, No.5, (October 2003), pp. 419-446, ISSN 0962-2802

Noll, D. C.; Genovese, C. R.; Nystrom, L. E.; Vazquez, A. L.; Forman, S. D.; Eddy, W. F. & Cohen, J. D. (1997). Estimating test–retest reliability in functional MR imaging: II. Application to motor and cognitive activation studies. *Magnetic Resonance in Medicine*, Vol.38, No.3, (September 1997), pp. 508–517, ISSN 1053-8119

Obata, T.; Liu, T. T.; Miller, K. L.; Luh, W. M.; Wong, E. C.; Frank, L. R. & Buxton, R. B. (2004). Discrepancies between BOLD and flow dynamics in primary and supplementary motor areas: Application of the Balloon model to the interpretation of

BOLD transients. *NeuroImage*, Vol.21, No.1, (January 2004), pp. 144–153, ISSN 1053-8119

Peligrad, M. (1996). On the asymptotic normality of sequences of weak dependent random variables. *Journal of Theoretical Probability*, Vol.9, No.3, (July 1996), 703-715, ISSN 1572-9230

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*, Vol.36, No.3, (July 2007), pp. 532–542, ISSN 1053-8119

Sirotin, Y. B.; Das, A. (2009). Anticipatory haemodynamic signals in sensory cortex not predicted by local neuronal activity. Nature, 457, (January 2009), pp. 475-479, ISSN 0028-0836

Specht, K.; Willmes, K.; Shah, N. J. & Jancke, L. (2003). Assessment of Reliability in functional imaging studies. *Journal of Magnetic Resonance Imaging*, Vol.17, No.4, (April 2003), pp. 463–471, ISSN 1522-2586

Stephan, K. E.; Weiskopf, N.; Drysdale, P. M.; Robinson, P. A. & Friston, K. J. (2007). Comparing hemodynamic models with DCM. *NeuroImage*, Vol.38, No.3, (November 2007), pp. 387–401, ISSN 1053-8119

Strother, S.; LaConte, S.; Hansen, L. K.; Anderson, J.; Zhang, J.; Pulapura, S. & Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrices: I. A preliminary group analysis. *NeuroImage*, Vol.23, Suppl.1, (2004), pp. S196-S207, ISSN 1053-8119

Varoquaux, G.; Sadaghiani, S.; Pinel, P.; Kleinschmidt, A.; Poline, J. B. & Thirion, B. (2010). A group model for stable multi-subject ICA on fMRI datasets. *NeuroImage*, Vol.51, No.1, (May 2010), pp. 288-299, ISSN 1053-8119

Worsley, K. J. (1996). The geometry of random images, *Chance*, Vol.9, No.1, pp. 27-40, ISSN 0933-2480

Worsley, K. J.; Liao, C. H.; Aston, J.; Petre, V.; Duncan, G. H.; Morales, F. & Evans, A. C. (2002). A general statistical analysis for fMRI data. *NeuroImage*, Vol.15, No.1, (January 2002), pp. 1-15, ISSN 1053-8119

Worsley, K. J.; Marrett, S.; Neelin, P.; Vandal, A. C.; Friston, K. J. & Evans, A. C. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, Vol.4, No.1, 58–73, ISSN 1097-0193

Yang, Y. & Green, S. B. (2010). A note on structural equation modeling estimates of reliability. *Structural Equation Modeling*, Vol.17, No.1, (January 2010), pp. 66-81, ISSN 1070-5511

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*, Vol.49, No.3, (February 2010), pp. 2163-2177, ISSN 1053-8119