NeuroImage 114 (2015) 287–293

Contents lists available at ScienceDirect

NeuroImage

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

Improving the reliability of single-subject fMRI by weighting intra-run variability

F. de Bertoldi [a], L. Finos [b], M. Maieron [c], L. Weis [a], M. Campanella [a], T. Ius [d], L. Fadiga [a,e,*]

[a] Department of Robotics, Brain and Cognitive Sciences, Istituto Italiano di Tecnologia, Genoa, Italy
[b] Department of Statistical Sciences, University of Padova, Italy
[c] Fisica Medica, Azienda Ospedaliero Universitaria Santa Maria della Misericordia, Udine, Italy
[d] Department of Neurosurgery, Azienda Ospedaliero Universitaria Santa Maria della Misericordia, Udine, Italy
[e] Section of Human Physiology, University of Ferrara, Italy

ARTICLE INFO

Article history:
Received 7 November 2014
Accepted 27 March 2015
Available online 8 April 2015

Keywords:
Reliability
Single subject
Clinical applications
Intra-run variability
fMRI analysis

ABSTRACT

At present, functional magnetic resonance imaging (fMRI) is one of the most useful methods of studying cognitive processes in the human brain in vivo, both for basic science and clinical goals. Although neuroscience studies often rely on group analysis, clinical applications must investigate single subjects (patients) only. Particularly for the latter, issues regarding the reliability of fMRI readings remain to be resolved. To determine the ability of intra-run variability (*IRV*) weighting to consistently detect active voxels, we first acquired fMRI data from a sample of healthy subjects, each of whom performed 4 runs (4 blocks each) of self-paced finger-tapping. Each subject's data was analyzed using single-run general linear model (GLM), and each block was then analyzed separately to calculate the *IRV* weighting. Results show that integrating *IRV* information into standard single-subject GLM activation maps significantly improved the reliability ($p = 0.007$) of the single-subject fMRI data. This suggests that taking *IRV* into account can help identify the most constant and relevant neuronal activity at the single-subject level.

Introduction

Over the last 20 years, functional magnetic resonance imaging (fMRI) has become a standard tool for mapping human brain function. It is also widely used for surgical planning (De Benedictis et al., 2010; Skrap et al., 2012) and clinical diagnostics (Matthews et al., 2006). Many experimental investigations have attempted to quantify the reliability of activation maps based on fMRI data (Rombouts et al., 1997; McGonigle et al., 2000; Loubinoux et al., 2001; Marshall et al., 2004; Yoo et al., 2005; Aron et al., 2006; Vul et al., 2009), but this still remains a largely problematic issue (Bennett and Miller, 2010). Problems with the reliability of fMRI data are derived from the nature of the fMRI signal. For example, the blood-oxygenation-level dependent (BOLD) fMRI technique uses the hemodynamic changes that accompany neuronal activity to infer brain (electrical) activity. However, observed local vascular changes are very small, and the recorded signal exhibits a very low signal to noise ratio. Therefore, BOLD is susceptible to several imaging artifacts. To quantify neuronal activity in fMRI data accurately, all known factors that could affect activation patterns in the fMRI study must be taken into account. Such factors can include scanner noise, in addition to physiological variations, such as breathing and heart rate (Hu et al., 1995), and patient motion (Lund et al., 2005).

Statistics can play a critical role in identifying and interpreting the activation patterns generated in fMRI studies (Lindquist, 2008). An easy way of enhancing the statistical power of fMRI results is to increase either the number of subjects or the number of trials. However, this is not an option in single-subject studies in the clinical environment, and increasing the duration or number of runs is often not feasible because it would inflict additional discomfort to the patient. Thus, clinical findings are generally based on a single fMRI run, and it is important to be able to localize reliable brain activity in this run. Unfortunately, very little research has focused on single-subject fMRI (Fadiga, 2007) mainly because of the lack of appropriate analytical methods to achieve an adequate reliability at the single-subject level and, secondly because single subject studies require a strong neuroanatomical background. An important limitation of model-based analytical methods is their dependence on a number of assumptions (Lu et al., 2003; Zang et al., 2004). In the general linear model (GLM) with block design, a (possibly generalized) linear model is fitted to model the BOLD activity along all blocks using one or more predictors that describe the design of the experiment (e.g., a zero/one vector for a task/resting design). This statistical approach assumes that the neural activity associated with a task does not change over time. However, the measure of signal stability over time (i.e., over blocks) is an interesting information which, in our

* Corresponding author at: Department of Robotics, Brain and Cognitive Sciences, Italian Institute of Technology, Via Morego 30, 16163 Genoa, Italy.
E-mail address: luciano.fadiga@iit.it (L. Fadiga).

http://dx.doi.org/10.1016/j.neuroimage.2015.03.076

view, has not been exploited enough. For example, it is possible that, in the same voxel, the stability of the BOLD signal across blocks of the same run, relate with high reliability across different runs/session. In this paper, we examined the possibility to make use of this information in order to select those voxels that are more consistently activated across the blocks forming a run to assess the reliability of fMRI data in single-subject tests. The idea was to fit a model separately for each block to define intra-run variability (*IRV*) as a function of the variance of the estimated coefficients of all blocks (i.e., increased heterogeneity of the estimated coefficients among the blocks implies larger *IRV*). A previous study used the inter-trial consistency of neuronal activity between different trials to estimate the functional relevance of different areas in a complex cognitive task (Windischberger et al., 2002). Likewise, we set out to determine whether *IRV* could be similarly effective in identifying more reliable voxels. To this end, we evaluated the impact of weighting standard GLM analysis data for *IRV* (i.e., the *p*-value from the GLM is divided by weight, which is a function of the *IRV*) on the reliability of the fMRI results. We ran a simulation and came to some conclusions based on the features of the method. More remarkably, we collected fMRI data from 7 healthy subjects and adopted a specific experimental protocol to deal with potential clinical constraints. The 7 subjects underwent 2 fMRI sessions, each consisting of 2 short runs (4 blocks each) of self-paced finger-tapping to provide data with a high level of variability and low statistical power. The data from each subject were first processed using a single-run GLM analysis, and the reliability index across the 4 runs was calculated. We then calculated the respective *IRV* values for each voxel of the resulting *t*-value maps and quantified the reliability of the proposed *IRV*-weighted maps in detecting activation at the single-subject level, as compared to standard GLM activation maps. This allowed us to assess the impact of integrating *IRV* into fMRI data analysis.

Methods

Subjects and fMRI protocol

Seven healthy subjects (4 males, 3 females, aged 19–24), all previously fMRI naive, were scanned while performing 4 identical runs with a block design (15/15 s of task/rest conditions, with 45 volumes recorded; 4 blocks for each run). The subjects performed a self-paced finger-tapping task with their right hand (all the subjects were right-handed). To minimize subject error, all underwent specific training before the start of the experiment. Four runs were acquired in 2 sessions of 2 runs each, with a 15 min gap between one session and the next. All subjects gave their informed consent and the experimental protocol was approved by the Udine Hospital Ethics Committee.

MRI acquisition

The MR images were acquired using a Siemens 1.5 T MRI whole-body scanner (Siemens Avanto, Erlangen, Germany), a 12-channel matrix head coil, and a custom-built head restrainer to minimize head movements. Both structural and functional images were recorded during the MRI sessions. High-resolution T1-weighted structural images were acquired using the following parameters: number of volumes = 1, repetition time (*TR*) = 2300 ms, time of echo (*TE*) = 2.86 ms, flip angle = 20, orientation = sagittal, number of slices = 160, volume thickness = 1 mm, voxel size = 0.5 × 0.5 × 1 mm, and field of view = 448 × 512. T2-weighted functional scans were performed using the following parameters: number of volumes = 45, *TR* = 3000 ms, *TE* = 60 ms, flip angle = 90, orientation = transversal, number of slices = 30, volume thickness = 1 mm, voxel size = 3.4375 × 3.4375 × 5 mm, field of view = 64 × 64, and acquisition order = interleaved. The subjects wore special MR-compatible glasses while receiving instructions.

Data preprocessing and analysis

The fMRI data analysis was carried out using the software packages MATLAB 7.6 (MathWorks Inc. Natick, Ma, USA) and SPM8 (http://www.fil.ion.ucl.ac.uk/spm/).

Preprocessing

In every subject, for each run, all volumes were realigned with respect to the first scan of the same run and co-registered with each participant's anatomical data. Functional time series were then smoothed with a Gaussian kernel width at a half-maximum of 10 mm, and high-pass temporal filtering (filter width, 45 s) was applied.

Single-run t-value maps

A 1st-level GLM analysis was performed separately for each run. The reference function applied was created by convolving a box-car with a canonical hemodynamic response function of the same on–off period as the stimulus (Friston et al., 1995, 2006a,b). Two *t* maps of each run was obtained for each subject after thresholding all single-run *t* maps at $p < 0.05$ and $p < 0.01$.

Single-block t-value maps

We also conducted a separate GLM analysis of each block in every run to define the estimated value of the model separately for every block. Sixteen single-block *t* maps (4 in each run) were obtained for each subject.

IRV index

The method presented here is for a simple task/rest design. The prototypical model for the task/rest design can be described as follows:

$$Y = a + bX + \varepsilon, \tag{1}$$

where *X* is a 0/1 indicator function — i.e. 0 at rest and 1 under task condition — possibly convolved with a canonical hemodynamic response function, *a* is the average effect at rest scan, *b* is the extra activation induced by the stimulus, and ε is the error term (i.e., the homoscedastic white noise with null mean and variance σ). After fitting the model to the recorded data, 2 quantities were computed: the deviance explained by the model (*modelDev*) and the residual deviance (*residDev*). The *F* statistic, which is typically used to infer the effect of a stimulus, was derived from the ratio between these 2 quantities:

$$F = modelDev/residDev * (N-2) \tag{2}$$

where *N* is the number of observations. The *t* statistic was also based on these quantities, seeing as:

$$t = \text{sign}(b) * \text{sqrt}(F). \tag{3}$$

The present study assumes that the effect (i.e. the true coefficient of the model) of the task may change over time (i.e., between the 4 blocks). Thus, we defined a slightly more complex model:

$$Y = a_1Z_1 + a_2Z_2 + a_3Z_3 + a_4Z_4 + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + \varepsilon, \tag{4}$$

where Z_i has a value of 1 in trial *i* and 0 otherwise, and X_i has a value of 1 under the stimulus in trial *i* and 0 otherwise. Likewise, the a_i coefficients represent the effects at rest, and b_i denotes the effect of the stimulus in trial *i*. The white noise ε is assumed homoscedastic, but it can be trivially extended to noise which has constant variance only within the block (i.e., heteroscedastic errors). We defined *modelDevB* as the explained deviance of model (4) minus *modelDev*, the explained deviance of model (1). It can be interpreted as the gain in the explained deviance when moving from model (1) to model (4) (i.e., moving from a common stimulus effect among the trials to a different effect in each trial).

If the effect of the stimulus in each block was constant, then *modelDevB* would be (about) equal to zero.

Normalizing this quantity in a zero-one interval, we obtained an initial raw indicator of the *IRV* as follows:

$$IRV = modelDevB/(modelDevB + residDevB) = modelDevB/residDev \quad (5)$$

with $residDevB = residDev - modelDevB$ being the residual deviance of the model (4). The *IRV* index resembled the definition of the R^2 index of regression analysis. Indeed, *IRV* index can be interpreted as the proportion of residual deviance due to the difference in the estimated effects between blocks. Low *IRV* values (i.e., close to 0) indicate that the coefficients vary little between the blocks, whereas *IRV* values close to one suggest that there is large variation between the blocks.

The *IRV* index has been introduced under the assumption of independent observations.However, this is not the typical assumption of the fMRI signal that is characterized by temporal autocorrelation. Within the GLM framework, there are many approaches that deal with this dependence among error terms, such as temporal filtering, correlation removal or modeling of the correlation structure (an excellent review is given in Woolrich et al., 2001). Despite these methods provide only asymptotically valid inferential results (Worsley and Friston, 1995), they have been shown to be generally valid (Purdon and Weisskoff, 1998; Ashburner et al., 2003) and constitute the most widely used approach to fMRI data analysis. The extension of the *IRV* index to the case of auto-correlated errors is straightforward. Since we set the *IRV* index within the general framework of the GLM — which deals with correlated errors — the index can be applied without further modification.

IRV maps

For balanced designs, the *IRV* can be computed by fitting a single-block analysis for each block. The numerator in the *IRV* formula (5) is equal to the variance of the estimated parameters (*SPM con_000*. image*) for each block, multiplied by the number of blocks minus one (3 in the current). The denominator is the residual sum of the mean squared error of a single-run analysis (*ReMS. Image*), multiplied by its degrees of freedom.

Correlation between the IRV and the overlap score

The measure of reliability of a given voxel was defined as the "overlap score," which denoted the proportion of runs in which the significance of a given voxel was $p < 0.05$ (the index assumes values 0, 0.25, 0.5, 0.75, or 1 when there are four valid runs and 0, 0.33, .66, or 1 when there are only three). We then explored the relationship between *IRV* and the reliability of the overlap score (i.e., inter-runs). For each map of the seven subjects, we computed the correlation between the overlap score and the *IRV*. To confirm the statistical significance of our findings, the seven correlations are used to test the hypotheses of null correlation by means of a one-sample *t*-test with two-tailed alternatives.

The IRV-weighting method

For each *t*-map, we calculated a correspondent *IRV*-weighted map in order to assess the potential improvement of the standard GLM analysis after the introduction of the *IRV* parameter.

IRV-weighted t maps and p maps

The values of the computed *IRV* maps were standardized using the overall mean for each subject:

$$w = (1-IRV)/\,\mathrm{mean}(1-IRV) \quad (6)$$

where *mean*(1 − *IRV*) means the average *1 − IRV* among all *m* voxels. Subsequently, the *t* maps were converted to *p* maps (i.e., maps of *p* values), and the value of each voxel was divided by the correspondent *w* value. In this way, the voxels with stable signal among blocks — i.e., small *IRV*, hence high *w* — are favored, while voxels signal changing over blocks — i.e., high *IRV* and small *w* — are penalized. The resulting *IRV*-weighted *p* map was then thresholded with the same threshold used in the other steps of the analysis ($p < 0.05$ or $p < 0.01$). Note that it could be convenient to transform back the weighted *p* maps to weighted *t* maps.

Theorem 1, which formalizes the procedure, is stated below and an informal discussion of the consequences of this theorem is also provided. The theorem's proof is given in the Appendix A.

Theorem 1. Let be p_i, $i = 1,\ldots,m$ the *p*-values derived from the test statistics T_i (2) or T_i (3), that is, testing the null hypothesis H_i of coefficient $b = 0$ in model (1) for the voxel *i*.

i) Define w_i for each voxel as in Eq. (6) and $p_i^w = p_i / w_i$. The average Type I error among all *m* tests is bounded by α for each α in (0,1], that is:

$$\sum_{i=1,\ldots,m} P(p_i^w \le \alpha | H_i)/m \le \alpha.$$

ii) More, generally, for any choice of w_i subject to
   a) the weights w_i are a function of the observed data only through IRV_i and
   b) $\sum_{i\,=\,1,\ldots,m} w_i = m$ (i.e., their sum equals *m*).

The same property holds.

Theorem 1 states that the Type I error (i.e., 1 − Specificity) is controlled at level α "on average" among all voxels. This mean that, considering all voxels together, the proportion of false positives is the same as in the standard analysis. We can make additional comments about any single voxel. The condition $p_i^w = p_i / w_i \le \alpha$ used to select the active voxels can be restated as $p_i \le \alpha w_i$. This means that weighting the *p*-values is equivalent to thresholding the (unweighted) *p*-values at different levels, i.e. αw_i, which depends on the IRV_i of the single voxel.

The proposed approach is a special case of a more general one. From thesis ii), it is made clear that two conditions are sufficient to make any weighted method valid. Therefore, the one proposed here is just one among infinitely many other possible choices. The optimal definition of weights depends on many parameters and is not an aim of this work. Finally, it can be noted that the method shares many aspects with weighted multiplicity control methods (Benjamini and Hochberg, 1997), in particular from the data-driven weighted methods (Westfall et al., 2004). If desired, the control of the Familywise Error Rate (FWER) can be easily reached by changing condition ii.b) $\sum_{i\,=\,1,\ldots,m} w_i = 1$ (instead of *m*), which is equivalent to thresholding the actual p_i^w at level α/m instead of α (the proof is trivial from Theorem 1). Furthermore, using the same definition of weights we can control of the False Discovery Rate following the guidelines proposed by Genovese et al. (2006). In this case, however the main goal is the reliability and not the control of the multiple Type I error.

Simulation study

We explored the behavior of the proposed method through a simulation study. The aim of this simulation is to highlight the main features of the method and it is certainly not intended to be an exhaustive exploration of all possible scenarios. We simulate the data from a linear model under the following setting: we consider 4 blocks, each with 4 scans (2 rests, 2 tasks). Errors are standard normal. The number of voxels (i.e. tested hypotheses) is 1000. Ten percent of the voxels are active (i.e. false null hypothesis), hence with non-null effect (i.e. the coefficient of the linear model). In this setting, we assume that the "reliable" voxel will have stable signal (i.e., a constant coefficient)

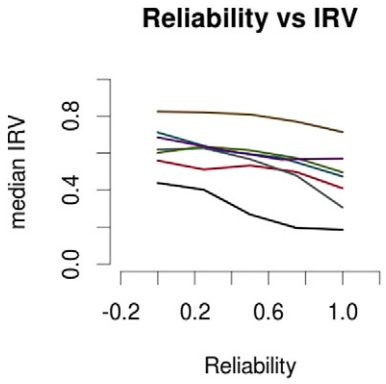

Fig. 1. Median *IRV* (ordinate) as a function of the reliability (abscissa) for the 7 subjects. The reliability was determined by the fraction of the runs in which a given voxel reached a significance level of $p < 0.05$ (as described in the section on Improving the reliability by weighting the intra-run variability on real data).

among the blocks, while the "unreliable" voxels will have block-specific coefficients, which are drawn from a normal with mean 2 and standard deviation SigmaB. We explored the following scenarios: SigmaB = 1, 2, and 3; when SigmaB is larger the effects among the blocks are more variable; therefore, "reliable" and "unreliable" voxels share the same effect in mean, but the "unreliable" voxels have non-constant coefficients among the blocks with an intra-run variability driven by SigmaB. We now identify the active voxels using a *p*-map with threshold alpha = 0.05. The proportion of truly active voxels identified by the method is computed for "reliable" and "unreliable" voxels. The same proportions are computed for the *IRV*-weighted *p*-map (alpha = 0.05). We replicate the data generation and the estimate of the proportion 1000 times (i.e., 1000 Monte Carlo iterations) and compute the average scores.

Improving the reliability by weighting the intra-run variability on real data

Reliability index

From the single-run *t* maps of each subject, the index of reliability (*Irel*) between all possible pairs of runs (e.g., run 1 vs. run 2, run 1 vs.

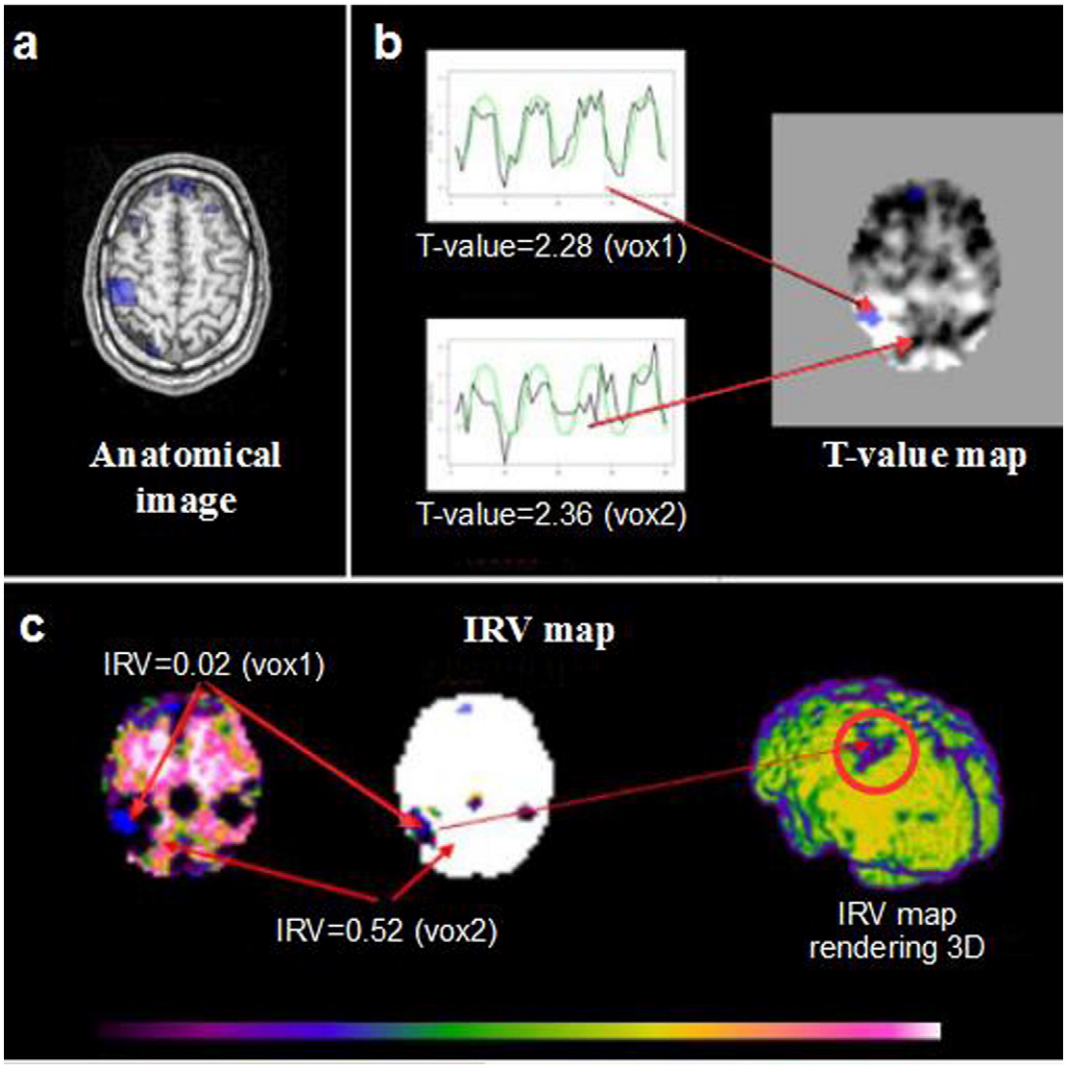

Fig. 2. Schematic depiction of the relationship between the *IRV* index and the overlap area. (a) Axial slice of a single subject's anatomical image. The blue area represents the overlap areas (i.e., the most reliable areas). (b) The single-run *t* map of the same subject (on the right). The time course of 2 exemplificative voxels is shown on the left side. Despite the comparable *t* value, the 2 time courses are very different: voxel 1, which falls inside the most reliable area, shows a more regular time profile compared to voxel 2, which falls outside. (c) *IRV* maps, unthresholded on the left ($0 < IRV < 1$) and thresholded ($0 < IRV < 0.1$) in the center. The *IRV* value enabled us to identify the most reliable voxel between the 2 selected voxels. On the right, the left hemisphere is visualized by a 3D rendering of the thresholded map. On the surface, the primary motor cortex (red circle) is well-delineated from the surrounding area of higher variability.

run 3, etc.) is computed using the overlap method (Rombouts et al., 1997). More specifically, for each pair of runs, we considered the size of the activated area in the first (*Va*) and second (*Vb*) runs at the 2 chosen thresholds ($p < 0.01$ and $p < 0.05$). Next, as a measure of reproducibility, we computed the size of the areas activated in both runs (V overlap). Finally, the reliability of each pair of runs is defined as $2 \times V\ overlap\ /\ (Va + Vb)$. To obtain a single value of *Irel* for each subject, we averaged the various indices calculated for the different pairs of runs. The *Irel* ranged between zero and one.

Measure of gain in reliability

At both the thresholds adopted ($p < 0.01$ and $p < 0.05$), we calculated the *Irel* index for the standard *t* maps (*Irel std*) and the *IRV*-weighted *t* maps (*Irel wgt*). Subsequently, for each subject, we quantified the gain in reliability reached with the proposed *IRV*-weighting method (*gain Irel*), as follows:

$$gain\ Irel = (Irel\ wgt / Irel\ std) - 1. \quad (7)$$

Finally, to assess the statistical significance of the results obtained, we performed a two-tailed *t*-test using the 2 scores of each single subject as paired samples.

Results

Correlation between the IRV and the overlap score

The median *IRV* as a function of the voxel overlap score across runs in the 7 subjects is shown in Fig. 1. The 7 correlations were tested with a *t*-test as described in the section on Correlation between the *IRV* and the overlap score. The resulting $p < 0.001$ significance level obtained from the correlation strongly supports the hypothesis of a negative correlation between these two variables. Therefore, the stable voxels tend to have lower *IRV* values. In sense, this justifies the simplification we have made in the simulation section (Simulation study) to identify the reliable voxels, in which one has a stable signal (i.e., low *IRV*).

Fig. 2 provides a schematic depiction of the relationship between the *IRV* index and the overlap area. Fig. 2b displays the time courses of 2 voxels, one falling inside and one falling outside the most reliable area, here called the "overlap area" (i.e., the overlap area of a voxel with $p < 0.05$ in all runs). As can be seen in the figure, the voxels have very similar *t* values in the first single-run activation. However, the signal of the more reliable voxel is visibly more constant than that of the other voxel. Furthermore, the first voxel has a lower *IRV* value (Fig. 2c), with the estimated effect in every block being similar (i.e., small variance among the estimated coefficients). This finding supports the idea that *IRV* maps can help to identify more reliable voxels (Figs. 2b and c). The *IRV* maps revealed patterns of low values of variability between blocks in specific regions of the brain area under study, suggesting that their activation was more consistent across different runs (see the example in Fig. 3).

A comparison of the IRV-weighted t maps and the t maps

Simulation study results

Several noteworthy findings can be observed from the results presented in Fig. 4. The sensitivity of the reliable voxels does not change when the intra-run variability SigmaB increases. There is no practical difference between standard and *IRV*-weighted method. The sensitivity of unreliable voxels, on the contrary, is heavily affected by SigmaB. Bigger is the variability between blocks, bigger is the loss of power of the test (i.e. the sensitivity). The reduction in sensitivity is markedly stronger for *IRV*-weighted method. This is the result that we expected since our aim was not to detect effects that are not reliable. The control at level α of the mean proportion of the false null *IRV*-weighted method has been proved in Theorem 1 and confirmed by the simulation

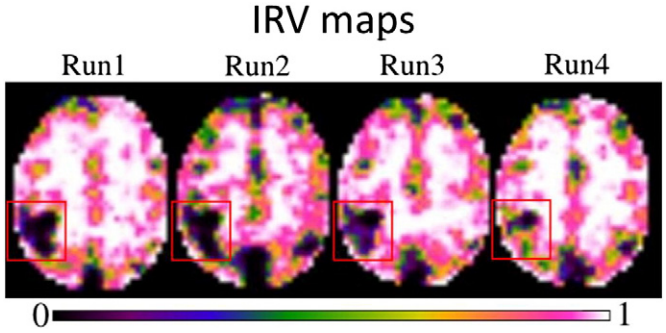

Fig. 3. *IRV* maps of the 4 runs in one subject. The maps reveal a spatial pattern of low *IRV* values, indicating reliability across the 4 runs consistent with the left primary motor area of the hand (red square).

(i.e., bounded by 0.95 both for the standard method and the *IRV*-weighted method for every value of SigmaB). We also run other simulations varying the setting such as the number of blocks and the number of scan per block. Despite the magnitude of the effect of *IRV*-weighted method may change from setting to setting, the overall picture remains very similar to Fig. 4 (data not shown).

Real data results

Comparison of the 2 sets of results showed a gain in reliability in all subjects with the proposed *IRV*-weighting method at both the thresholds we adopted ($p < 0.05$ and $p < 0.01$), as shown in Tables 1 and 2. The estimated gain in reliability was greater at $p < 0.01$ (mean gain +15.6%), with respect to $p < 0.05$ (mean gain +12.9%), but both these figures were statistically significant ($p = 0.009$ and $p = 0.032$, respectively). Fig. 5 shows an example of strong consistency between the area of overlap and the single-run *IRV*-weighted map, highlighting the increased precision achieved by our approach with respect to an equivalent standard single-run *t* map.

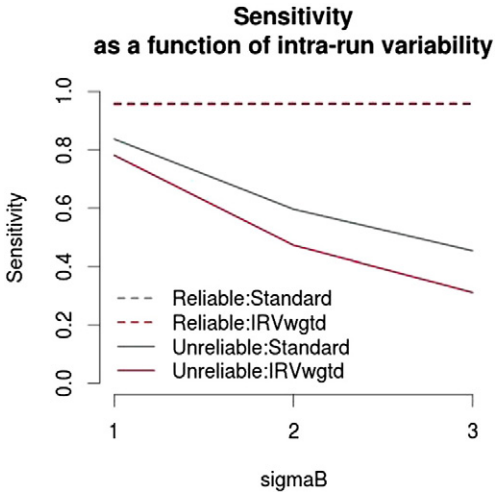

Fig. 4. Here we report the results of the simulation study. The sensitivity (ordinate) is plotted as a function of the intra-run variability SigmaB (abscissa). The sensitivity is the probability of detecting an active voxel. Therefore, it is strictly connected to the reliability (defined in this work as the probability of detecting an active voxel in two subsequent runs) that can approximately be computed as the square of the sensitivity. A remarkable result is the reduction in sensitivity of the unreliable voxels markedly stronger by using *IRV*-weighted method.

Tables 1, 2
Indices of reliability of the standard single-run *t* maps (*Irel std*) and the single-run *IRV*-weighted *t* maps (*Irel wgt*). Analysis of all subjects at $p < 0.05$ (left) and $p < 0.01$ (right). The last column reports the estimated gain in reliability for each subject, and the last row presents the mean gain in *Irel* (Gain *Irel*).

$p < 0.05$				$p < 0.01$			
Subject	*Irel std*	*Irel wgt*	% Change	Subject	*Irel std*	*Irel wgt*	Gain *Irel*
S1	0.16	0.19	+18.1%	S1	0.089	0.110	+23.4%
S2	0.18	0.19	+4.5%	S2	0.125	0.130	+3.6%
S3	0.29	0.32	+8.4%	S3	0.171	0.206	+20.1%
S4	0.46	0.46	+0.4%	S4	0.325	0.335	+2.9%
S5	0.27	0.33	+22.1%	S5	0.150	0.193	+28.8%
S6	0.59	0.60	+2.6%	S6	0.472	0.504	+6.7%
S7	0.09	0.12	+34.1%	S7	0.021	0.025	+23.7%
Mean			+12.9%	Mean			+15.6%

Discussion and conclusion

In this paper, we explored the hypothesis that the most reliably activated voxels should have the lowest *IRV* of the BOLD signal (i.e., high stability between blocks). First, we established a statistical relationship between the *IRV* index and the reliability between runs. Furthermore, we developed a method that integrates this information and produced statistical maps that are more reliable than standard *t* maps. Both these findings support our initial hypothesis. The efficacy of the method depends on the fact that the most reliable areas maintain stable activity over time when subjected to the same stimulation for a short period.

Previous fMRI studies demonstrated the same phenomena for a BOLD signal, showing a high degree of reproducibility of the hemodynamic response in the same subject, cortical area, and day (Aguirre et al., 1998). However, some authors previously questioned whether nervous activity invariance occurs over time (Waites et al., 2007). During experiments, various factors, such as cognitive state, task strategy and, in the clinical context, the compliance and the functional level of the subject have been reported to induce variation in task performance, thus affecting the stability of the BOLD signal over time (Waites et al., 2007; Baldo et al., 2001).

To test the robustness of the proposed *IRV* method in various experimental conditions, including clinical settings, we adopted a specific protocol with short runs of self-paced finger tapping. In this task, many sources of intra-individual variability, such as a change in force/velocity or errors in task execution, can affect the signal. Despite these potential sources of variability, our results show that the core of the activation, reliable between runs, is characterized by higher stability of the signal within a single run. The findings, highlight the weakness of the standard GLM approach in fMRI data analysis. The latter only generates an average result by considering all the blocks together. Thus, when the signal changes over the blocks (i.e., over time), the hypotheses assumed for the statistical models are incorrect. Our approach tends to penalize the unstable voxels rather than affect the stables ones. Fig. 5 shows an example of an unreliable voxel (i.e., not active in the overlap map) that was obtained with a single-run standard analysis appearing inactive in the *IRV*-weighted map. The simulation study confirms all these results. The power (i.e., sensibility) of the tests on unreliable voxels decreases with increasing *IRV*. Our results point also to the possibility of using *IRV* maps to localize brain functions by identifying core areas of activation associated with a given function. Previous works have discussed the possibility of using the functional response to a specific task to define distinct brain regions (Friston et al., 2006a,b; Saxe et al., 2006; Duncan et al., 2009). The preliminary observations of the present study suggest that the spatial patterns, defined by a decrease in *IRV* values, correspond to areas in the functional motor network, and that these patterns are well delineated from the surrounding area of higher variability (i.e., with higher *IRV*). As shown in Fig. 2, not all of the regions with *IRV* decrement correspond to the most reliable areas (blue areas). In our view, the *IRV* index tends to highlight all the "real" stable activations, not necessarily the reliable activations between the runs; this is the case for example of "real" activations with high signal stability degree but run-related. Moreover, previous studies reported that the baseline signal intensity can vary across areas and subjects due to vascular compliance or vascular density, thus affecting the validity of the fMRI analysis (Harrison et al., 2002). On the contrary, the *IRV* should be independent from the signal intensity level and it should be effective even for voxels where the signal amplitude is low, but constant. Such aspects suggest that the independent use of the *IRV* map could represent a more sensitive and specific approach to mapping brain functions than the standard map.

In summary, in this paper, we showed that *IRV* can be used as an additional parameter for detecting the activation of the most relevant areas associated with a given function. We demonstrated that integration of *IRV* maps into standard GLM analysis produces more reliable results than using GLM analysis alone, thus supporting our initial hypothesis. These findings indicate that *IRV* may aid characterization of "core" brain activations in single-subject fMRI, even in a single run. It is essential however, to further validate our method with larger cohorts of subjects performing other tasks, perhaps related to higher cognitive functions, such as language and sensory perception, that are characterized by larger inter- and intra-individual variability. In this study, we presented the method for a simple task/rest design. However, future studies that apply the same method to more complex designs (e.g. event-related) should be conducted. However, *IRV* map could

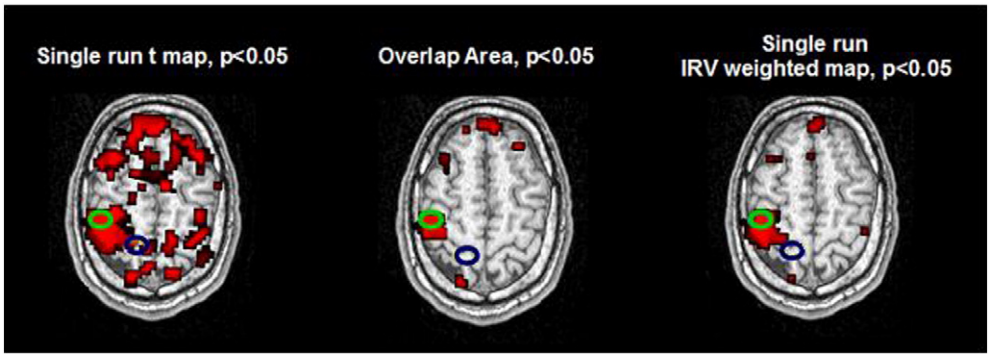

Fig. 5. Qualitative evaluation of functional activation maps. Axial anatomical image of the subject in Fig. 2. Three different maps are shown: a single-run *t* map ($p < 0.05$) on the right, the corresponding single-run *IRV*-weighted *t* map ($p < 0.05$) on the left, and the area activated in all single-run activations ($p < 0.05$), here called the "Overlap Area", in the center. The same 2 voxels shown in Fig. 2 are circled in all 3 images: voxel 1 (green circle) and voxel 2 (blue circle). The similarity between the overlap area and the single-run *IRV*-weighted map with respect to the standard single-run *t* map is evident. Moreover, the least reliable voxel (voxel 2), despite its slightly higher *t* value with respect to voxel 1, is not active in either the overlap area or the *IRV*-weighted map.

potentially serve as a stand-alone method to identify core areas of specific brain functions. Future studies could include an assessment of the consistency of the various spatial patterns detected by both *IRV* and *t* maps in different runs and in different tasks. Another important area that could be explored in future works is the functional relevance of the stability of signals over time. The inability to distinguish between essential and nonessential areas is one of the main limitations of fMRI. Including transcranial magnetic stimulation or direct cortical stimulation in patients undergoing neurosurgical operations could provide insightful information to better define both the functional relevance and the sensitivity/specificity of *IRV* spatial patterns.

Acknowledgments

This work has been supported by the BMI Project of the IIT RBCS Department and by the EU (Robotcub FP6-004370, Poeticon FP7-215843, Poeticon++ FP7-288382), E-R Region-University Area1a and Italian Ministry of University (PRIN 2010MEFNF7_003) grants to LF. We deeply thank Dr. Miran Skrap for his support and suggestions.

Appendix A

Proof

Let us recall that $modelDev_i$, $modelDevB$, and $residDevB_i$ are independent random variables by the Cochran's theorem. $residDev_i$ is a sufficient complete statistic for σ_i (i.e., the variance of ε_i in Eq. (1)). By definition, IRV_i is ancillary statistic to σ_i (which essentially cancels out in the ratio in the same way as any *F* statistic does). Therefore, IRV_i is independent of the test $residDev_i$ through Basu's theorem. Since test statistic T_i, as defined in Eq. (2) or (3), is a function of $modelDev_i$ and $residDev_i$, while IRV_i is a function of $modelDevB_i$ and $residDevB_i$, T_i and IRV_i are independent random variables. This holds for all *p*-values calculated by transformation of the null CDF of test statistics T_i. The same independence holds between T_i and w_i, since T_i is independent of all other w_j with $j \neq i$ (see also the proof for Theorem 3 in Farcomeni and Finos, 2013). Therefore, the condition p_i w $= p_{ii} / w_i \leq \alpha$ can be restated as $p_i \leq \alpha w_i$ and, thus, $P(p_i \leq \alpha w_i | H_i) \leq \alpha w_i$. Now $\Sigma_{i=1,\ldots,m} P(p_i \leq \alpha\, w_i | H_i) \leq \alpha \Sigma_{i=1,\ldots,m}$ $I(H_i) * w_i \leq \alpha m$($I(H_i) = 1$ or 0, depending on whether or not H_i is a true null hypothesis, and result i) holds true. The proof of thesis ii) is a trivial generalization of the proof of thesis i).

References

Aguirre, G., Zarahn, E., D'Esposito, M., 1998. The variability of human, BOLD hemodynamic responses. NeuroImage 8 (4), 360–369.
Aron, A.R., Gluck, M.A., Poldrack, R.A., 2006. Long-term test–retest reliability of functional MRI in a classification learning task. NeuroImage 29 (3), 1000–1006 (Feb 1).
Ashburner, J., Friston, K., Penny, W., 2003. Human Brain Function. In: Frackowiak, Richard S.J., Friston, Karl J., Frith, Christopher D., Dolan, Raymond J., Price, Cathy J., Zeki, Semir, Ashburner, John T., Penny, William D. (Eds.), 2nd edition. ISBN: 978-0-12-264841-0.
Baldo, J.V., Shimamura, A.P., Delis, D.C., Kramer, J., Kaplan, E., 2001. Verbal and design fluency in patients with frontal lobe lesions. J. Int. Neuropsychol. Soc. 7 (5), 586–596 (Jul).
Benjamini, Y., Hochberg, Y., 1997. Multiple hypotheses testing with weights. Scand. J. Stat. 24 (3), 407–418 (Sep.).
Bennett, C.M., Miller, M.B., 2010. How reliable are the results from functional magnetic resonance imaging? Ann. N. Y. Acad. Sci. Mar. 1191, 133–155.
De Benedictis, A., Moritz-Gasser, S., Duffau, H., 2010. Awake mapping optimizes the extent of resection for low-grade gliomas in eloquent areas. Neurosurgery 66 (6), 1074–1084 (Jun, discussion 1084).
Duncan, K.J., Pattamadilok, C., Knierim, I., Devlin, J.T., 2009. Consistency and variability in functional localisers. NeuroImage 46 (4), 1018–1026 (Jul 15, Epub 2009 Mar 14).
Fadiga, L., 2007. Functional magnetic resonance imaging: measuring versus estimating. NeuroImage 37 (4), 1042–1044 (Oct 1).
Farcomeni, A., Finos, L., 2013. FDR control with pseudo-gatekeeping based on a possibly data driven order of the hypotheses. Biometrics 69 (3), 606–613.
Friston, K.J., Holmes, A.P., Worsley, K.J., Poline, J.B., Frackowiak, R.S.J., 1995. Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. 2, 189–210.
Friston, K.J., Ashburner, J.T., Kiebel, S., Nichols, T.E., Penny, W.D., 2006a. Statistical Parametric Mapping: The Analysis of Functional Brain Images. Elsevier, London.
Friston, K.J., Rotshtein, P., Geng, J.J., Sterzer, P., Henson, R.N., 2006b. A critique of functional localisers. NeuroImage 30 (4), 1077–1087 (May 1).
Genovese, C.R., Roeder, K., Wasserman, L., 2006. False discovery control with *p*-value weighting. Biometrika 93 (3), 509–524. http://dx.doi.org/10.1093/biomet/93.3.509.
Harrison, R.V., Harel, N., Panesar, J., Mount, R.J., 2002. Blood capillary distribution correlates with hemodynamic-based functional imaging in cerebral cortex. Cereb. Cortex 12 (3), 225–233 (Mar).
Hu, X., Le, T.h, Parrish, T., Erhard, P., 1995. Retrospective estimation and correction of physiological fluctuation in functional MRI. Magn. Reson. Med. 34, 201–212.
Lindquist, M., 2008. The statistical analysis of fMRI data. Stat. Sci. 23, 439–464.
Loubinoux, I., Carel, C., Alary, F., Boulanouar, K., Viallard, G., Manelfe, C., Rascol, O., Celsis, P., Chollet, F., 2001. Within-session and between-session reproducibility of cerebral sensorimotor activation: a test–retest effect evidenced with functional magnetic resonance imaging. J. Cereb. Blood Flow Metab. 21 (5), 592–607.
Lu, Y.L., Jiang, T.Z., Zang, Y.F., 2003. Region growing method for the analysis of fMRI data. NeuroImage 20, 455–465.
Lund, T.E., Norgaard, M.D., Rostrup, E., Rowe, J.B., Paulson, O.B., 2005. Motion or activity: their role in intra- and inter-subject variation in fMRI. NeuroImage 26, 960–964.
Marshall, I., Simonotto, E., Deary, I.J., Maclullich, A., Ebmeier, K.P., Rose, E.J., Wardlaw, J.M., Goddard, N., Chappell, F.M., 2004. Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. Radiology 233 (3), 868–877 (Dec).
Matthews, P.M., Honey, G.D., Bullmore, E.T., 2006. Applications of fMRI in translational medicine and clinical practice. Nat. Rev. Neurosci. 7 (9), 732–744 (Sep).
McGonigle, D.J., Howseman, A.M., Athwal, B.S., Friston, K.J., Frackowiak, R.S., Holmes, A.P., 2000. Variability in fMRI: an examination of intersession differences. NeuroImage 11, 708–734.
Purdon, P.L., Weisskoff, R.M., 1998. Effect of temporal autocorrelation due to physiological noise and stimulus paradigm on voxel-level false-positive rates in fMRI. Hum. Brain Mapp. 6 (4), 239–249.
Rombouts, S.A., Barkhof, F., Hoogenraad, F.G., Sprenger, M., Valk, J., Scheltens, P., 1997. Test–retest analysis with functional MR of the activated area in the human visual cortex. AJNR Am. J. Neuroradiol. 18, 1317–1322.
Saxe, R., Brett, M., Kanwisher, N., 2006. Divide and conquer: a defense of functional localizers. NeuroImage 30 (4), 1088–1096 (May 1, discussion 1097-9).
Skrap, M., Mondani, M., Tomasino, B., Weis, L., Budai, R., Pauletto, G., Eleopra, R., Fadiga, L., Ius, T., 2012. Surgery of insular nonenhancing gliomas: volumetric analysis of tumoral resection, clinical outcome, and survival in a consecutive series of 66 cases. Neurosurgery 70 (5), 1081–1093 (May, discussion 1093-4).
Vul, E., Harris, C., Winkielman, P., Pashler, H., 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. Perspect. Psychol. Sci. 4, 274–290.
Waites, A.B., Mannfolk, P., Shaw, M.E., Olsrud, J., Jackson, G.D., 2007. Flexible statistical modelling detects clinical functional magnetic resonance imaging activation in partially compliant subjects. Magn. Reson. Imaging 25 (2), 188–196 (Feb, Epub 2006 Nov 28).
Westfall, P.H., Kropf, S., Finos, L., 2004. Weighted FWE-controlling methods in high-dimensional situations. Recent in Multiple Comparison Procedures. Institute of Mathematical Statistics Lecture Notes Monograph Series vol. 47, pp. 143–154.
Windischberger, C., Lamm, C., Bauer, H., Moser, E., 2002. Consistency of inter-trial activation using single: assessment of regional differences. Brain Res. Cogn. Brain Res. 13 (1), 129–138 (Feb).
Woolrich, M.W., Ripley, B.D., Brady, M., Smith, S.M., 2001. Temporal autocorrelation in univariate linear modeling of FMRI data. NeuroImage 14 (6), 1370–1386 (Dec).
Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—again. NeuroImage 2 (3), 173–181 (Sep).
Yoo, S.S., Wei, X., Dickey, C.C., Guttmann, C.R., Panych, L.P., 2005. Long-term reproducibility analysis of fMRI using hand motor task. Int. J. Neurosci. 115 (1), 55–77 (Jan).
Zang, Y., Tianzi, J., Lu, Y., He, Y.A., Tiana, L., 2004. Regional homogeneity approach to fMRI data analysis. NeuroImage 22, 394–400.