Sensitivity of Random Effects Analyses to Group Size and Individual Outliers: A Jackknife Study

P. C. Venkat[1], T. Johnstone[2], A. L. Alexander[2,3], and T. R. Oakes[2]

[1]Electrical and Computer Engineering, University of Wisconsin-Madison, Madison, WI, United States, [2]Waisman Center for Brain Imaging and Behavior, Madison, WI, United States, [3]Medical Physics, University of Wisconsin-Madison, Madison, WI, United States

Background

It is well known that functional MRI data from individual subjects suffer from low contrast-to-noise ratio, and reliable detection of activations entails multi-subject analysis. Furthermore, a random effects analysis must be carried out if the inferences drawn on the subject group are to be generalized onto the whole population. The validity of such a generalization hinges on the assumption that the random effects model is robust with respect to group size and to outlier subjects, provided the subjects were drawn randomly from the population at large. In the current study, we investigate the validity of these assumptions by trying to answer the following questions: 1) How does the number of subjects in a study impact the inferences drawn from a multi-subject random effects analysis? 2) How sensitive are these inferences with respect to outlier subjects? 3) How does the size of the group affect this sensitivity? Three different measures were considered, viz. the strengths of activations, the sizes of clusters of activation and lastly the overall pattern of activations.

Methods

Functional MRI data for this study were drawn from 44 subjects ranging in age from 18-50 with number and sex balanced within each decade. Functional tasks included: 1) a rapid event related task, the Go/Nogo task [Garavan et al., 1999; Liddle et al., 2001] (consisting of random ITI between 1.5 and 3.5 sec., 120 Go trials, 30 Nogo trials) and 2) a multi-condition block design task, the N-back task [Casey et al., 1998; Cohen et al., 1997; Smith et al., 1996] (consisting of 0-, 1- and 2-back trials, 51 s per block with 10 s of rest between each block and three blocks of each condition). Voxel-wise data analysis was performed using the FMRISTAT package [Worsley 2002] in MATLAB. For the Go/Nogo data, a GLM-based deconvolution was performed to allow for flexibility in terms of the actual time course of response both within and across subjects. The time courses for 'Go' and 'Nogo' conditions were modeled independently over an 18s duration using sinusoidal basis functions modulated by an exponential damping function. Additional regressors were used to account for motion related spikes as well as baseline drift. Analysis of the N-back data used a GLM based on an ideal hemodynamic response function with no motion covariates. A group of 10 subjects was initially chosen, and a group level random effects analysis was performed to obtain a voxelwise T-statistic. Multi-subject jacknife analysis was then performed by sequentially dropping each subject. This process was repeated on groups of 15, 20,25,30,35,40 and 44 subjects where each group was formed by adding subjects to the previous one. T-statistic values at each voxel were converted to Z-scores to allow for comparison across different groups. Generous clusters of activation were chosen from the 44-subject T-statistic maps with a p<0.05 threshold. Cluster specific measures for all groups were obtained by considering the voxels that lie within these 'global' or gold standard clusters.

Results and Conclusions

Mean Z-Score Effects: The significance levels of activated clusters were quantified by the mean Z-score of all voxels within the corresponding cluster obtained from the 44-subject group T-map. Fig. 1 shows the impact of group size and individual subjects on mean Z-scores in the largest and most robust cluster for the Go-Nogo contrast, situated in motor cortex. It may be seen that the mean Z-score in the cluster increases monotonically with group size, going from 1.7 for the 10-subject group, up to 2.5 for the 44-subject group. The corresponding values for the largest N-back cluster were seen to be 1.3 and 2.2. A similar pattern was observed for smaller clusters in both experiments. What is striking is that the mean Z-score is in fact highly sensitive to individual subjects, though the impact is reduced in larger groups. Dropping one subject caused the mean Z-score to fall by up to 30% in a 10-subject group and up to 15% in a 44-subject group. The impact was amplified in smaller clusters where a 30% drop was observed even in the largest groups in both experiments. It was seen that a single subject in the group consistently caused the largest drops in mean Z-score when excluded from the analysis and could be termed as an outlier. However, who the outlier was could vary from cluster to cluster. *Cluster Size Effects:* Cluster-size was estimated by counting the number of voxels lying within the global cluster, that individually survived the p<0.05 threshold. Fig. 2 shows the results for the cluster-size measure on the motor cortex cluster for the Go-Nogo contrast. It may be seen that cluster size behaves similarly as mean Z-score. The size of the cluster went from 1800 voxels in a 10-subject group to nearly 3600 voxels in a 44-subject group. The impact of individual subjects was striking. Dropping one of the subjects caused a 70% drop in cluster size for the 10-subject group. Even the 44-subject group saw a 20% drop in cluster-size when the same subject was dropped. In the smallest clusters, dropping a single subject caused almost an order of magnitude drop in cluster-size in both experiments across all group sizes, leading to the conclusion that these clusters were false positives, driven entirely by one subject. *Spatial Correlation of Activation Patterns:* Patterns of activations were quantified by measuring their correlation coefficient with the Z-statistic map obtained for the 44-subject group. Since underlying anatomical variations could affect this correlation, we only considered voxels with P>0.6 in the MNI-supplied gray matter probability map. Fig. 3 shows the impact of group size and individual subjects on the masked correlation coefficients. While the general pattern of activations diverges from the gold standard for small group sizes, the change is not as drastic as in the previous measures. In fact, the Z-map from the 10-subject group had a correlation coefficient of over 0.7 with the gold standard, indicating that the overall pattern of activation is robust with respect to individual subjects. Even in the smallest group, dropping one subject caused a change of at most 5% in the correlation coefficient. The robustness of the correlation coefficient is largely due to the absence of an exclusionary threshold. Similar behavior was observed in the N-back case as well. *Summary:* The results demonstrate that threshold-based inferences from random effects analyses can be highly susceptible to the impact of individual subjects. This is especially true when the group size is small, or when the cluster in consideration is small. It is imperative in such cases that studies exclude outlier subjects prior to performing group level analyses. The jackknife analysis method is sensitive to the effect level generated by each subject and can be used to identify outlier samples that either significantly increase or decrease the level of statistical significance. The results demonstrate that this method is a powerful approach for assessing the statistical robustness of fMRI cluster measurements.

References

1.FMRISTAT: http://www.math.mcgill.ca/keith/fmristat 2.Garavan, et al. Proc Natl Acad Sci USA 1999; 96:8301-8306
3. Liddle, et al. Hum Brain Map 2001; 12:100-109 4. AFNI: http://afni.nimh.nih.gov/afni

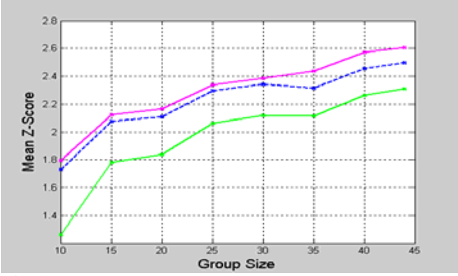

Fig.1. *Mean Z-scores in motor cortex cluster vs. Group Size for Go/Nogo*

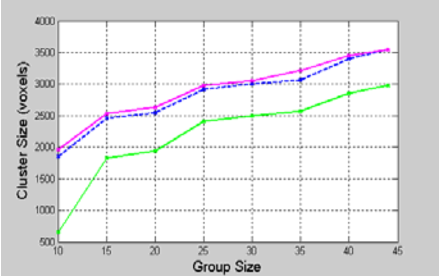

Fig.2. *Proxy cluster-size measures of motor cortex cluster vs. Group Size for Go/Nogo task.*

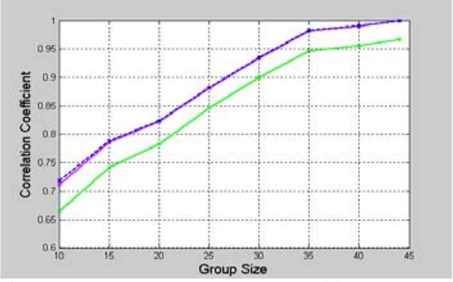

Fig.3. *Correlation coefficient (with 44-subject Z-map) vs. Group Size for Go/Nogo task*

For all figures, the value with all subjects is shown by dashed blue line. The solid magenta and green curves correspond to the maximum and minimum values, respectively, from the jackknife analysis.