Journal of Serendipitous and Unexpected Results

# Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction

Craig M. Bennett [1*], Abigail A. Baird [2], Michael B. Miller [1] and George L. Wolford [3]

[1]Department of Psychology, University of California at Santa Barbara, Santa Barbara, CA 93106
[2]Department of Psychology, Blodgett Hall, Vassar College, Poughkeepsie, NY 12604
[3]Department of Psychological and Brain Sciences, Moore Hall, Dartmouth College, Hanover, NH 03755

With the extreme dimensionality of functional neuroimaging data comes extreme risk for false positives. Across the $130,000$ voxels in a typical fMRI volume the probability of at least one false positive is almost certain. Proper correction for multiple comparisons should be completed during the analysis of these datasets, but is often ignored by investigators. To highlight the danger of this practice we completed an fMRI scanning session with a post-mortem Atlantic Salmon as the subject. The salmon was shown the same social perspective-taking task that was later administered to a group of human subjects. Statistics that were uncorrected for multiple comparisons showed active voxel clusters in the salmon's brain cavity and spinal column. Statistics controlling for the family-wise error rate (FWER) and false discovery rate (FDR) both indicated that no active voxels were present, even at relaxed statistical thresholds. We argue that relying on standard statistical thresholds ($p < 0.001$) and low minimum cluster sizes ($k > 8$) is an ineffective control for multiple comparisons. We further argue that the vast majority of fMRI studies should be utilizing proper multiple comparisons correction as standard practice when thresholding their data.

**Keywords:** fMRI statistics FDR FWER

**Contact:** bennett@psych.ucsb.edu

## 1 INTRODUCTION

Fifty years ago few researchers ever thought of doing thousands of statistical tests on the same contrast. Completing the required calculations by hand would have been impractical and computers were not powerful enough to store and operate on that quantity of data. The situation is quite different today, as the capacity for data acquisition and analysis has evolved considerably. A prime example of this evolution is the ability to record in vivo images of brain anatomy and function. With $130,000$ voxels in a single functional neuroimaging volume it is now common practice to do tens of thousands of tests per contrast over multiple contrasts. While this extreme dimensionality offers dramatic new opportunities in terms of analysis it also comes with dramatic new opportunities for false positives in the results. As a result the nagging issue of multiple comparisons has been thrust to the forefront of discussion in a diverse array of scientific fields, including cognitive neuroscience. More and more researchers have realized that correcting for chance discoveries is a necessary part of imaging analysis. This is a positive trend, but it overlooks the fact that a sizable percentage of results still utilize uncorrected statistics. An unknown quantity of these results may be false positives.

There are well-established techniques that can and should be used for the correction of multiple comparisons in fMRI. When they are applied these methods hold the probability of a false positive to a specified, predetermined rate. Two widely utilized approaches are to place limits on the family-wise error rate (FWER) and the false discovery rate (FDR). The family-wise error rate represents the probability of observing one or more false positives after carrying out multiple significance tests. Using a familywise error rate of $\mathrm{FWER} = 0.05$ would mean that there is a $5\%$ chance of one or more false positives across the entire set of hypothesis tests. The Bonferroni correction is probably the most widely known FWER control and is the correction method that most investigators are familiar with. In functional imaging the control of FWER is most often done through the use of Gaussian Random Field Theory or permutation methods. There are excellent articles by Brett *et al*. (2004) and Nichols and Hayasaka (2003)

*to whom correspondence should be addressed

that provide greater detail on the control of familywise errors in the analysis of fMRI data.

Controlling the FWER does the best job of limiting false positives but also comes at the greatest cost of statistical power. A second approach to multiple comparisons correction is to place limits on the false discovery rate. Using a false discovery rate of $\mathrm{FDR} = 0.05$ would mean that at most $5\%$ of the detected results are expected to be false positives. See Benjamini and Hochberg (1995), Benjamini and Yekutieli (2001), and Genovese *et al*. (2002) for a more in-depth discussion of false discovery rate in fMRI. FDR is a less conservative approach relative to FWER methods, but it may represent a more ideal balance between statistical power and multiple comparisons control.

Sadly, while methods for multiple comparisons correction are included in every major neuroimaging software package these techniques are not always invoked in the analysis of functional imaging data. For the year 2008 only $74\%$ of articles in the journal NeuroImage reported results from a general linear model analysis of fMRI data that utilized multiple comparisons correction ($193/260$ studies). Other journals we examined were Cerebral Cortex ($67.5\%$, $54/80$ studies), Social Cognitive and Affective Neuroscience ($60\%, 15/25$ studies), Human Brain Mapping ($75.4\%$, $43/57$ studies), and the Journal of Cognitive Neuroscience ($61.8\%$, $42/68$ studies). A list of these studies is available in the online supplemental materials. The issue is not limited to published articles, as proper multiple comparisons correction is somewhat rare during neuroimaging conference presentations. During one poster session at a recent neuroscience conference only $21\%$ of the researchers used multiple comparisons correction in their research ($9/42$). A further, more insidious problem is that some researchers would apply correction to some contrasts but not to others depending on the results of each comparison.

Many researchers who report uncorrected statistics tend to rely on increased significance thresholds ($0.001 < p < 0.005$) and minimum cluster sizes ($6 < k < 20$ voxels) to restrict the rate of false positives. While this does increase the effective significance threshold, it is an inadequate approach to address the multiple comparisons problem. These same threshold values are used in contrasts testing across $15,000$ voxels and $45,000$ voxels. The same cutoff value simply cannot be accurate in all cases. Simulation data has shown that a significance threshold of $p < 0.005$ combined with a $10$ voxel minimum cluster size is likely to yield significant voxel clusters almost $100\%$ of the time in data comprised of random noise (Vul *et al*., 2009). It remains the case that high significance thresholds with predefined minimum cluster sizes are an unknown, soft control to the multiple comparisons problem.

For some situations a cutoff value of $p < 0.001$ might be too conservative while in other cases it will be too liberal. Still, in every case it is an unprincipled approach. The reader can't possibly know what percentage of the reported results might be false positives, seriously impairing the interpretability of the findings. To illustrate the magnitude of the problem we carried out a real experiment that demonstrates the danger of not correcting for chance properly.

## 2 METHODS

One mature Atlantic Salmon (Salmo salar) participated in the fMRI study. The salmon measured approximately $18$ inches long, weighed $3.8$ lbs, and was not alive at the time of scanning. It is not known if the salmon was male or female, but given the post-mortem state of the subject this was not thought to be a critical variable.

Image acquisition was completed on a $1.5$-tesla GE Signa MR scanner (General Electric Medical Systems, Milwaukee, WI). A quadrature birdcage head coil was used for RF transmission and reception. Foam padding was placed within the head coil as a method of limiting salmon movement during the scan, but proved to be largely unnecessary as subject motion was exceptionally low. Scanning parameters for the $T_2*$ echo-planar imaging (EPI) sequence were: $25$ slices ($4$mm thick, $1$mm gap), $\mathrm{TR} = 2500$ms, $\mathrm{TE} = 30$ms, $\mathrm{flipangle} = 90^\circ$, and $256$x$256$ field of view. Only a subset of slices were necessary to ensure whole-brain coverage in the salmon. Dummy shots were used during the first $10$ seconds of scanning to ensure magnetization equilibrium. Stimuli were projected onto a ground glass screen located at the head of the magnet bore by an LCD projector. A mirror directly above the head coil allowed the salmon to observe experiment stimuli.

The task administered to the salmon involved completing an open-ended mentalizing task. The salmon was shown a series of photographs depicting human individuals in social situations with a specified emotional valence, either socially inclusive or socially exclusive. The salmon was asked to determine which emotion the individual in the photo must have been experiencing. The photo stimuli were presented in a block design, with each block consisting of four photos presented

individually for $2.5$ seconds each ($10$ seconds total) followed by $12$ seconds of rest. A total of $12$ blocks of photo presentation were completed with $48$ photos presented during the run. Photos were presented with the experiment-scripting program Psyscope (Cohen *et al*., 1993) and advanced by a TTL voltage trigger from the scanner. Total scan time for the task was $5.8$ minutes, with $140$ acquired image volumes.

Image processing was completed using the program SPM2 (Wellcome Department of Imaging Neuroscience, London, UK) in the MATLAB 6.5.1 environment (Mathworks Inc, Natick, MA). Preprocessing steps for the functional imaging data included a 6-parameter rigid-body affine realignment of the functional timeseries, coregistration of the functional data to a T1-weighted anatomical image, and $8$ mm full-width at half-maximum (FWHM) Gaussian smoothing. Spatial normalization was not completed as there is currently no standardized MRI atlas space for the Atlantic Salmon.

Voxelwise statistics on the salmon data were calculated through an ordinary least-squares estimation of the general linear model. Predictors of the hemodynamic response were modeled by a boxcar convolved with a canonical hemodynamic response function. A temporal high pass filter with a cutoff period of $128$ seconds was included to account for low frequency drift in the functional imaging data. No autocorrelation correction was applied.

# 3 RESULTS

A $t$-contrast was used to test for regions with significant BOLD signal change during the presentation of photos as compared to rest. The parameters for this comparison were $t(131) > 3.15, p(\text{uncorrected}) < 0.001$, 3 voxel extent threshold. The relatively low extent threshold value was chosen due to the small size of the salmon's brain relative to voxel size. Several active voxels were observed in a cluster located within the salmon's brain cavity (see Fig. 1). The size of this cluster was $81$ mm$^3$ with a cluster-level significance of $p = 0.001$. Another, smaller region was observed in the dorsal spinal column. Due to the coarse resolution of the echo-planar image acquisition and the relatively small size of the salmon brain further discrimination between regions could not be completed.

Identical $t$-contrasts were also completed that controlled for multiple comparisons. The first additional contrast controlled for the proportion of false discoveries in the results. This method, titled the Benjamimi-Hochberg correction but commonly referred to as FDR, allows an investigator to set the expected proportion of false discoveries in the results to a desired value (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001). For the purposes of this contrast the proportion of false discoveries was set at $\text{FDR} = 0.05$. A second additional contrast controlled for the familywise error rate in the results. The selected method controls the FWER through the use of Gaussian Random Field Theory (Friston *et al*., 1996; Worsley *et al*., 1996, 2004). Using this strategy the spatial smoothness of the results is estimated and the probability of a false positive in a random field of similar Gaussians is calculated. For the purposes of this contrast the probability of a familywise error was set at $\text{FWER} = 0.05$. Both of the additional contrasts controlling for multiple comparisons indicated that no significant voxels were present in the dataset. This was true even at the relaxed thresholds of $\text{FDR} = 0.25$ and $\text{FWER} = 0.25$.

# 4 DISCUSSION

Either we have stumbled onto a rather amazing discovery in terms of post-mortem ichthyological cognition, or there is something a bit off with regard to our uncorrected statistical approach. Could we conclude from this data that the salmon is engaging in the perspective-taking task? Certainly not. By controlling for the cognitive ability of the subject we have thoroughly eliminated that possibility. What we can conclude is that random noise in the EPI timeseries may yield spurious results if multiple testing is not controlled for. In a functional image volume of $130,000$ voxels the probability of a false discovery is almost certain. Even in the restricted set of $60,000$ voxels that represent the human brain false positives will continue to be present. This issue has faced the neuroimaging field for some time, but the implementation of statistical correction remains optional when publishing results of neuroimaging analyses.

What, then, is the best solution to the multiple comparisons problem in functional imaging? The Bonferroni correction (Bonferroni, 1936) is perhaps the most well-known formula for the control of false positives. The Bonferroni correction is quite flexible as it does not require the data to be independent for it to be effective. However, there is some consensus that Bonferroni may be too conservative for most fMRI data

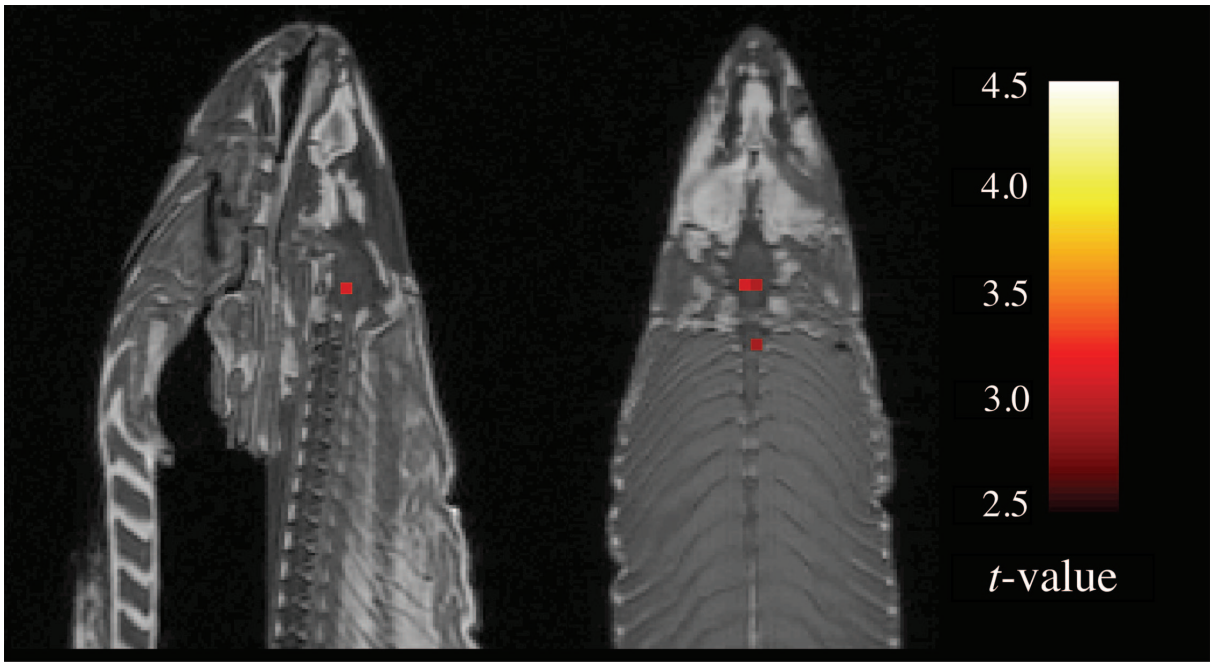


**Fig. 1.** Sagittal and axial images of significant brain voxels in the $\mathrm{task} > \mathrm{rest}$ contrast. The parameters for this comparison were $t(131) > 3.15$, $p(\mathrm{uncorrected}) < 0.001$, 3 voxel extent threshold. Two clusters were observed in the salmon central nervous system. One cluster was observed in the medial brain cavity and another was observed in the upper spinal column.

sets (Logan *et al*., 2008). This is because the value of one voxel is not an independent estimate of local signal. Instead, it is highly correlated with the values of surrounding voxels due to the intrinsic spatial correlation of the BOLD signal and to Gaussian smoothing applied during preprocessing. This causes the corrected Bonferroni threshold to be unnecessarily high, leading to Type II error and the elimination of valid results. More adaptive methods are necessary to avoid the rejection of true signal while controlling for false positives.

The other methods mentioned earlier use aspects of the data itself to determine the optimal corrected statistical threshold. For functional imaging there are strategies such as Benjamini and Hochberg's FDR, resampling FWER, and Gaussian Random Field FWER estimation that have proven to be effective options. All of them provide multiple comparisons correction with increased statistical power relative to Bonferroni. One or more of these methods are available in all major fMRI analysis packages, including SPM, AFNI, FSL, FMRISTAT, and BrainVoyager. The only decision an investigator has to make is what kind of balance to strike between the detection of legitimate results and presence of false positives. In the future other methods such as topological FDR (Chumbley and Friston, 2009) have the potential to further improve false positive control while minimizing the impact on statistical power.

It is important to note that correction for multiple comparisons does not address other important statistical issues in fMRI. Specifically, a distinction should be drawn between the multiple comparisons problem and the 'non-independence error' highlighted by Vul *et al*. (2009) and Kriegeskorte *et al*. (2009). The non-independence error refers to the inflation of cluster-wise statistical estimates when the constituent voxels were selected using the same statistical measure. For example, the correlation value of a voxel cluster will be inflated if the voxels were originally selected based on the criteria that they have a high correlation. Voxels with beneficial noise that increases their correlation value will be selected during the first stage, inflating the apparent cluster-wise correlation during the second stage. This stands in contrast to the multiple comparisons problem, which is related to the prevalence of false positives present across the set of selected voxels at the first stage. Other statistical issues, such as temporal autocorrelation and low frequency drift, are also separate statistical problems that are best addressed with their own set of corrections (Nandy and Cordes, 2007). It is also important to recognize that there are

some situations where a lower statistical threshold can still be used. For example, in a split-half analysis a researcher may use a more liberal threshold to define a region-of-interest (ROI) for later testing in a separate, independent set of data. These are special cases though, and as currently conducted the vast majority of neuroimaging studies require some form of multiple comparisons correction.

The multiple testing problem is not unique to neuroimaging. Instead, it is an issue that most scientific fields face as data analysis is completed. Anytime that multiple tests are completed without proper correction it has the potential to impact the conclusions drawn from the results. See Austin *et al*. (2006) for an example from clinical epidemiology of how multiple testing can lead to spurious associations between astrological sign and health outcome. This commentary is not intended as a specific accusation against functional imaging, but rather an argument in favor of continued evolution in the standards of fMRI analysis. There have been several in-depth articles regarding the multiple testing problem in neuroimaging, but a sizable fraction of published research still report results using uncorrected statistics. The control of false positives is not a matter of difficulty, as all major analysis packages for fMRI include some form of multiple comparisons correction. Rather it seems to be the case that investigators do not want to jeopardize their results through a reduction in statistical power. While we must guard against the elimination of legitimate results through Type II error, the alternative of continuing forward with uncorrected statistics cannot be an option.

# REFERENCES


Austin, P., Juurlink, M., Juurlink, D., and Hux, J. (2006). Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, **59**, 964–969.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, **57**, 289–300.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependence. *Ann. Statist.*, **29**, 1165–1188.

Bonferroni, C. (1936). Teoria statistica delle classi e calcolo delle probabilit. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, **8**, 3–62.

Brett, M., Penny, W., and Kiebel, S. (2004). *An introduction to Random Field Theory*. Academic Press.

Chumbley, J. and Friston, K. (2009). False discovery rate revisited: FDR and topological inference using gaussian random fields. *Neuroimage*, **44**(1), 62–70.

Cohen, J., MacWhinney, M., Flatt, M., and Provost, J. (1993). PsyScope: A new graphic interactive environment for designing psychology experiments. *Behavioral Research Methods, Instruments Computers*, **25**(2), 257–271.

Friston, K., Holmes, A., Poline, J., Price, C., and Frith, C. (1996). Detecting activations in PET and fMRI: levels of inference and power. *Neuroimage*, **4**(3 Pt 1), 223–235.

Genovese, C., Lazar, N., and Nichols, T. (2002). Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage*, **15**(4), 870–878.

Kriegeskorte, N., Simmons, W., Bellgowan, P., and Baker, C. (2009). Circular analysis in systems neuroscience: the dangers of double dipping. *Nat Neurosci*, **12**(5), 535–540.

Logan, B., Geliazkova, M., and Rowe, D. (2008). An evaluation of spatial thresholding techniques in fMRI analysis. *Hum Brain Mapp*, **29**(12), 1379–1389.

Nandy, R. and Cordes, D. (2007). A semi-parametric approach to estimate the family-wise error rate in fMRI using resting-state data. *NeuroImage*, **34**(4), 1562–1576.

Nichols, T. and Hayasaka, S. (2003). Controlling the familywise error rate in functional neuroimaging: a comparative review. *Stat Methods Med Res*, **12**(5), 419–446.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, **4**(3), 274–290.

Worsley, K., Marrett, S., Neelin, P., Vandal, A., Friston, K., and Evans, A. (1996). A unified statistical approach for determining significant signals in images of cerebral activation. *Hum Brain Mapp*, **4**, 58–73.

Worsley, K., Taylor, J., and Tomaiuolo, F. (2004). Unified univariate and multivariate random field theory. *NeuroImage*, **23**, 189–195.