

Multiple testing corrections, nonparametric methods, and Random Field Theory

Thomas Nichols¹

1. Warwick Manufacturing Group & Department of Statistics,
University of Warwick, Coventry, CV4 7AL, U.K.

March 21, 2012

Abstract

I provide a selective review of the literature on the multiple testing problem in fMRI. By drawing connections with the older modalities, PET in particular, and how software implementations have tracked (or lagged) behind theoretical developments, my narrative aims to give the methodological researcher a historical perspective on this important aspect of fMRI data analysis.

Keywords: .

Running title: Multiple testing in fMRI

1 Introduction

In the whimsically titled letter “Holmes & Watson reply to Sherlock” (Holmes *et al.*, 1998) my colleagues and I made a serious critique of Halber *et al.* (1997), a paper evaluating thresholding methods for PET activation data. The paper directly compared a nonparametric permutation method (named “Sherlock”), which provided familywise-error corrected inferences, to uncorrected $P < 0.05$ inference, finding that the latter method was to be preferred for its power. In a response to our letter, the paper’s authors’ defended the uncorrected approach as the (then) default setting in the SPM¹ software and claimed that it had been used in “approximately 1,200 publications”.

Over a decade later, and one “Voodoo correlations” (Vul *et al.*, 2009) imbroglio and post-mortem ichthyological fMRI study (Bennett *et al.*, 2011) later, it seems everyone agrees that (a) correcting inferences for the search over the brain is essential and (b) such corrections are not consistently utilized in fMRI. Hopefully some historical perspective can strengthen the discipline’s resolve to uphold good statistical practice.

What follows is a highly selective review of the literature on the multiple testing problem in fMRI and its antecedents (PET & M/EEG). I have tried to capture the major landmark publications, and while this selection is inevitably quirky and personal, I hope it will provide a useful perspective in this important aspect of fMRI data analysis. See Holmes (1994, Chapter 3) and Petersson *et al.* (1999) for more careful and detailed reviews of early work in this area.

2 The Problem

Whether studying brain structure or brain function, using MRI, PET or M/EEG modalities, the end result of an experiment is typically a set of statistic values (e.g. T or F values) that comprise an image. This “image” may be a 2D surface, a 3D volume, or even a 4D movie of statistics over time. Call $T = \{T_i\}$ the statistic image, with T_i the value at voxel i . Before even mentioning “multiple testing” we must define the objects under inference. There are in fact a variety of ways of summarizing a statistic image, including voxel-wise, cluster-wise, peak-wise

¹<http://www.fil.ion.ucl.ac.uk/spm>

and others.

2.1 Assessing statistic images: Voxels, clusters & peaks

Voxel-wise inference uses a threshold u and classifies voxel i as “active” if $T_i \geq u$; inference is made on each voxel individually. Cluster-wise inference uses a cluster-forming threshold arc to define blobs, i.e. contiguous suprathreshold regions. If S is the size of a cluster, cluster inference consists of retaining all clusters with $S \geq k$ for some cluster size threshold k . For voxel-wise inference, when $T_i \geq u$ we can make a statement about the signal at voxel i . For cluster-wise inference, when $S \geq k$ we’re making a statement about random set, the collection of voxels in the cluster. With replication of our experiment, voxel i still means the same thing, but a cluster will comprise different voxels if it exists at all.

So what exactly is the interpretation of a significant cluster? I usually answer “that one or more voxels within that cluster have evidence against the null” (Poldrack *et al.*, 2011); that is, the test can localize the effect to somewhere within the cluster². This lack of precise spatial specificity is a shortcoming, but voxel-wise inference has its critics too. Friston & colleagues have argued against voxel-wise inference (Chumbley & Friston, 2009; Chumbley *et al.*, 2010), saying that for smoothed data a voxel is ill-defined, and only topological quantities are interpretable, like peaks (local maxima) or clusters. As peaks, like clusters, are randomly located, and as voxels have reasonable finite support in practice, I counter that voxel-wise inference remains a useful approach.

Peak-wise inference is based only on local maximum above a given screening threshold u_p . Peak-wise is not the same as voxel-wise inference³ and the inference will depend on the chosen u_p threshold. Finally, set-wise methods, based on just the count of clusters, and other omnibus measures can be defined, but they do not have any localizing power.

²This said, I know of no formal proof that cluster inference has such strong control of Familywise error. I will make ample use of footnotes to make such pedantic comments.

³Jumping ahead, FWE-corrected peak-wise P-values equal FWE-corrected voxel-wise P-values at the peaks. This is because FWE is determined by the distribution of the maximal statistic, and the maximum voxel-wise is the maximum peak-wise.

2.2 Statistics, P's & corrected P's

Once a method for assessing the statistic image is chosen, a test statistic needs to be defined. For voxel-wise or peak-wise inference, the statistic value is obvious (just the value of T at the voxel, or the peak), and for cluster-wise inference this is naturally the number of voxels in the clusters (though there are other ways; see below). Based on a test statistic an *uncorrected* P-value P can be defined. For example, suppose we are performing voxel-wise inference; for a randomly (or *a priori*) selected voxel i , the P-value P_i is the chance of observing a test statistic T_i as or more extreme, assuming that the null hypothesis is true. For voxel-wise statistic T_i this is a trivial computation, even possible with a table in the back of a textbook. For a peak value or cluster size, however, no standard results are available. Before reviewing the tools imagers used to find uncorrected P-values, let me first introduce an even greater challenge, the multiple testing problem.

Whether voxel-wise or cluster-wise, there is a huge multiplicity. Searching over 100,000 voxels in the brain we expect to find 5,000 P_i 's smaller than 0.05. Likewise, searching over 100 clusters will on average produce 5 uncorrected cluster P-values less than 0.05. To account for the multiplicity, we have to define a measure of error when searching the brain. The standard measure is the Familywise Error Rate (FWE), the chance of one or more false positives (Nichols & Hayasaka, 2003). FWE is the quantity controlled by the well-known Bonferroni procedure, and while it is a sensible measure of false positives, many find it lacks power⁴.

The False Discovery Rate (FDR) is a more lenient measure of false positive risk, defined as the expected proportion of false positives among detections (Benjamini & Hochberg, 1995)⁵. My colleagues and I (Genovese *et al.*, 2002) introduced FDR to functional neuroimaging, and I see its wide embrace as a sign of how hungry users were for calibrated multiple testing procedures that are more powerful than FWE.

Another less-used alternative to FWE is the expected number of false positives (Bullmore *et al.*, 1996). This measure is used in the CamBA software⁶ to control the expected number of

⁴People often say "FWE is conservative", but that's like saying a meter is too short. FWE is just a measure of false positive risk, a stringent one.

⁵The work was circulating in statistics circles well before 1995; see Benjamini (2010) for some history.

⁶<http://www-bmu.psychiatry.cam.ac.uk/software/>

false positive clusters at just below 1.0.

For either FWE or FDR, you can define *corrected* P-values for a particular T_i (or peak value or cluster size): The smallest FWE (or FDR) α level that will *just* reject the null hypothesis for T_i .

And what about poor old uncorrected $P < 0.001$, with perhaps some cluster threshold like $S > 10$ voxels? In principal, the false positive risk of any fixed heuristic could be validated with a sufficient amount of real null data, and then the heuristic could safely applied to data *with the very same characteristics*. But if any aspect of the data changes, say different voxels sizes, different smoothing kernels, number of slices or orientation (which will affect to analysis mask size), then the false positive risk will vary in some undetermined way. Hence, best practice (Poldrack *et al.*, 2008) and the need for reproducible science dictates multiple-testing corrected inferences that have the same interpretation for all data.

2.3 A preview of solutions

The crux of methodological research in neuroimaging inference has been how to find corrected P-values, or, equivalently, thresholds on test statistics that control a specified error rate. Before a historical tour of this research, it's helpful to lay out the three broad types of approaches that have been used. The best known (if least understood) approach is Random Field Theory (RFT). In rough terms, RFT uses the smoothness of the image noise to predict the behavior of extreme values of voxel-, peak- and cluster-wise statistics. The underlying theory is elegant and has connections to topology but requires that, in addition to the usual Gaussian assumption, the image data behave like a continuous random process. The other frequently used approach is Monte Carlo (MC). By estimating basic features of the data under the null hypothesis, like image smoothness, MC repeatedly simulates null replicates of the data. The observed test statistics (peak, cluster, whatever) can then be compared to the simulated null distributions, creating P-values. Just like RFT, Gaussianity has to be assumed and the smoothness has to be estimated, but MC doesn't depend on the accuracy of RFT approximations. Finally, there is the permutation test. Using the data itself, empirical null distributions are created by permuting (or otherwise altering) the data under the null hypothesis. This approach has the weakest

assumptions and is growing in use, but has limitations, in particular in dealing with time series autocorrelation and general experimental designs.

3 A Tour of Solutions

3.1 Early Days

Many “fMRI inference methods” are generic procedure developed first for PET. Hence we start with seminal work by Fox & Mintun (1989), who showed that non-quantitative H_2^{15}O PET⁷ could be used to map brain function. As part of that paper they proposed “Change Distribution Analysis” to determine if there were any effects in the image. They used the distribution of all local extrema, that is, the value of local maxima for $T_i > 0$ and all local minima for $T_i < 0$ (no screening threshold u_p). Defining global skew and kurtosis statistics on the distribution of peak values, and using conventional standard errors⁸ they produced an omnibus test for activation in the brain.

3.2 Random Field Theory

Change Distribution Analysis lacked any localization power, and of course there was a need for methods that would assign significance locally, to each voxel, while still controlling FWE. Friston *et al.* (1991) solved this problem using general theory of Gaussian processes, working in 2D and assuming equal smoothness in X & Y directions. Shortly afterwards, Worsley *et al.* (1992) produced a more general 3D solution that would define a class of methods: Random Field Theory. By drawing a connection between the voxel-wise FWE and the expected Euler characteristic, Worsley created inferences that accounted for both the volume and smoothness of the search region. He created the notion of a Resolution Element, or RESEL, a virtual voxel with dimensions equal to $\text{FWHM}_x \times \text{FWHM}_y \times \text{FWHM}_z$ ⁹.

⁷Quantitative PET required blood-draws and difficult-to-fit compartmental models

⁸Peak statistics are reasonably assumed independent

⁹Contrary to intuition and some publications that I shall refrain from citing, a RFT voxel-wise P-value cannot be seen as a Bonferroni correction based on the number of RESELS. See Eqns. 30 & 31 of Nichols & Hayasaka

In the PET data Worsley & colleagues were using, there seemed to be no evidence for spatially varying variance. Hence the initial 1992 work assumed the variance estimate could be pooled over the entire brain, producing a Z statistic image. Others groups found PET data to have spatially varying standard-deviation, and in particular Friston *et al.* (1991) used a voxel-wise variance estimate, the resulting T image were Gaussianized to apply the Z results. Worsley & coauthors soon generalized his results to account for voxel-wise variance estimation, for T , and F images (Worsley *et al.*, 1993; Worsley, 1994), though these results didn't make their way into Friston's SPM until "SPM99" and FSL¹⁰ still uses the Gaussianization. See Table 1 for a tale how Worsley & Friston came to be collaborators after this potentially fractious beginning.

3.3 Monte Carlo

Voxel-wise thresholding couldn't detect low-intensity, spatially extended effects. In lieu of theoretical results, a Monte Carlo simulation approach was proposed first for PET (Poline & Mazoyer, 1993; Roland *et al.*, 1993) and then for fMRI (Forman *et al.*, 1995). Using an estimate of the smoothness of the data, simulated statistic images under the null hypothesis generate an empirical estimate of the maximum cluster size, from which cluster size statistics can be converted to FWE-corrected P-values. This approach is still used today in the AFNI¹¹ software's `alphasim`¹². In their first joint work, Worsley & Friston (& colleagues) used Random Field Theory to produce closed-form FWE P-values for cluster size statistics (Friston *et al.*, 1993).

An entire separate review paper is needed to track all the RFT work produced, but a few highlights include: A unified result for Z , T , χ^2 and F images (Worsley *et al.*, 1996); a solution for the conservativeness found at low smoothness (Worsley & Taylor, 2005); and a unified multivariate result from which all other results are special cases (Worsley *et al.*, 2004). These

(2003).

¹⁰<http://www.fmrib.ox.ac.uk/fsl>

¹¹<http://afni.nimh.nih.gov>

¹²Care must be taken that smoothness parameter is set from the estimated noise smoothness, using say `3dFWHMx`, and not just taken to be equal to the smoothing kernel applied to the data

“Keith and I first met in 1990 at a workshop at Harvard Medical School. I was chaperoned by Richard Frackowiak and Keith by Alan Evans. Alan had famously recruited Keith after finding him collecting maple leafs on the campus of McGill University—in the fond hope of finding something interesting to study! Keith had seen the potential of random field theory and had been sent a final draft of my 1991 paper (Friston *et al.*, 1991). I remember him being very excited by the prospect of applying random field theory to neuroimaging data. He was also bemused and intrigued by the convergence of the general theory of stochastic processes and level sets (my 1991 paper) and random field theory proper (his paper, Worsley *et al.*, 1992). I also recall him being exercised by a mild discrepancy between the two formulations; the discrepancy boiled down to a square root two factor that he could not resolve, and remains unresolved two decades years later.

We became firm friends over the ensuing months, or more exactly ‘pen-pals’. Getting emails from Keith was a bit like playing Russian roulette. Most of the time they were insightful, reassuring and helpful but—occasionally—he would start with ‘I think there’s a small problem...’. What he meant was that there was a substantial conceptual or technical problem that would take at least six months hard work to resolve.

The substantial exchanges between us often weren’t reflected in publications or the rhetoric we each developed, perhaps to underscore the distinct contributions of our respective groups. It is worth remembering that we were separated not just by the Atlantic but by some aesthetic and pragmatic differences. For example, we always assumed that error variance was regionally specific, but Keith never liked this, because is destroyed some of the simple beauty of implementing the theory. One the other hand, Keith loved the most advanced graphics software that he could find, whereas we stuck religiously to Matlab despite its very limited graphics support (at that time).

Years later, the intellectual collaboration rested on shared students and fellows, like Jean-Baptiste Poline and Stefan Kiebel. Much of that work is embodied in SPM and has remained the mainstay of topological inference using random field theory to date.”

Table 1: KJF on KJW. Keith Worsley and Karl Friston authored foundational papers in the 1990’s on inference for neuroimaging. We lost Keith suddenly in 2009, so I asked Karl to comment on how it was that a psychiatrist and a statistician came to be friends and collaborators.

methods and more are implemented in surfstat¹³, a program Worsley was actively developing until his death in 2009.

Whether Monte Carlo or RFT, the estimation of smoothness is crucial. Poline *et al.* (1995) found that if smoothness was estimated from but a single image (as was done in SPM95), RFT P-values should have confidence intervals of about $\pm 40\%$! This uncertainty affects Monte Carlo P-values to the same or greater degree. Standard practice now is to estimate smoothness from standardized residuals images (Kiebel *et al.*, 1999), but there remains two different approaches. Forman *et al.* (1995) estimated the smoothness based on a discretized Gaussian kernel, where as the estimator of Kiebel *et al.* (1999) is based on partial differences approximating a continuous random field's derivatives. While the later makes no assumption about the shape of the autocorrelation function—except the existence of 2 derivatives at the origin—it has greater bias at low smoothness.¹⁴

Cluster-wise inference captures the spatial nature of the signals, and suffers from less multiplicity than voxel-wise inference. However it is not always more sensitive, and Friston *et al.* (1996) showed that the power of cluster inference depends on the spatial scale of the signal relative to the noise smoothness: Focal, intense signals will be better detected by voxel-wise inference. Thus there is a natural temptation to compute both cluster-wise and voxel-wise results and take the better of the two. This of course forms a new multiple testing problem, which will yield more false positives¹⁵. To address this, Poline *et al.* (1997) proposed a RFT-based joint cluster size, cluster peak-height test.

3.4 Permutation.

Inspired by Blair & Karniski's EEG permutation work (1994), Holmes *et al.* (1996) proposed a permutation test for PET that controlled FWE with few assumptions. Based on that work,

¹³<http://galton.uchicago.edu/faculty/InMemoriam/worsley/research/surfstat>

¹⁴SPM and AFNI use the Kiebel approach, though SPM only uses up to 64 images by default; FSL uses a version of the Forman approach on the standardized residuals (Flitney & Jenkinson, 2000; Jenkinson, 2000).

¹⁵The SPM software encourages profligate exploration of results, showing all possible types of inferences, while the FSL software only provides users one of voxel-wise or cluster-wise inferences

Holmes and I created the SnPM¹⁶ software, which we thought would quickly become irrelevant as fMRI came to dominate neuroimaging. The problem was that fMRI times series' autocorrelation violates a basic assumption needed by permutation, exchangeability. Others had tackled this problem, by decorrelating the fMRI data using the fit of a parametric autocorrelation model (Bullmore *et al.*, 1996; Locascio *et al.*, 1997), however we found this mix of parametric and nonparametric modeling unsatisfactory¹⁷. However fMRI analysis quickly came to focus on group analysis using a summary statistic approach (Holmes & Friston, 1999; Mumford & Nichols, 2009), meaning our PET 1-scan-per-subject permutation methods remained relevant.

Despite some compelling results on the tremendous power gains of voxel-wise FWE permutation inference over RFT (Nichols & Hayasaka, 2003; Nichols & Holmes, 2001), SnPM did not become an integral SPM tool (in part because it remained difficult to use). In FSL, however, the randomise software¹⁸ has become a central tool for all voxel-based anatomical analyses. Aside from overcoming any RFT conservativeness, permutation inference works in nonstandard settings like Track-Based Spatial Statistics (Smith *et al.*, 2006) where tracks are highly irregular and vary from 1-D to 2-D. Permutation also allows consideration of new test statistics, where no parametric result is available. Examples include: The smoothed variance T-test (Holmes *et al.*, 1996), cluster-mass (Bullmore *et al.*, 1999), different peak-cluster combining tests (Hayasaka & Nichols, 2004), and a completely new cluster-inspired method, Threshold-Free Cluster Enhancement (Smith & Nichols, 2009). Permutation even feeds-back into RFT research: We developed a RFT cluster-mass test (Zhang *et al.*, 2009) only after extensive experience with permutation showed that it dominated alternate peak-cluster combining methods (Hayasaka & Nichols, 2004).

¹⁶<http://go.warwick.ac.uk/tenichols/software/snpm>

¹⁷More flexible wavelet decorrelation can whiten better (Bullmore *et al.*, 2001), but can have problems with simple block designs (Friman & Westin, 2005). Also note that a randomized experimental design justifies a randomization test with any data (Raz *et al.*, 2003), though this has limited application.

¹⁸Initially an exercise for Tim Behrens to teach Steve Smith C++; I gave instructions from the sidelines.

4 Looking Ahead

Looking ahead, there is renewed enthusiasm for resampling-based test as GPU's make order-of magnitude speed-ups (Eklund *et al.*, 2011), and in particular which make local multivariate methods attractive (Eklund *et al.*, 2011; Nandy & Cordes, 2007). Predictive analyses and “brain reading” distill inference to a single accuracy number (Haynes & Rees, 2006) seem to be a step away from “brain mapping”. But in practice investigators wish to determine which brain regions are responsible for the predictive power, and thus we return to a spatial mapping exercise (Kriegeskorte *et al.*, 2006). And perhaps the most promising direction, is the application of explicit spatial models to brain image data, for both original fMRI data (Keller *et al.*, 2008; Xu *et al.*, 2009; Weeda *et al.*, 2009; Thirion *et al.*, 2010; Kim *et al.*, 2010; Gershman *et al.*, 2011) and meta-analysis data (Neumann *et al.*, 2008; Kang *et al.*, 2011). These methods can provide spatial confidence intervals on effects of interest and more flexible and interpretable model fits.

Finally, I apologize to the authors of scores of papers on fMRI inference that I have not cited. Sometime in the next 20 years I hope I can make a more comprehensive review.

References

- Benjamini, Y., & Hochberg, Yosef. 1995. Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B, Methodological*, **57**(1), 289–300.
- Benjamini, Yoav. 2010. Discovering the false discovery rate. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **21**(4), 3771–416.
- Bennett, Craig M, Baird, Abigail A, Miller, Michael B., & Wolford, George L. 2011. Neural Correlates of Interspecies Perspective Taking in the Post-Mortem Atlantic Salmon: An Argument For Proper Multiple Comparisons Correction. *Journal of Serendipitous and Unexpected Results*, **1**(1), 1–5.

- Blair, R C, & Karniski, W. 1994. *Functional Neuroimaging: Technical Foundations*. San Diego: Academic Press. Pages 19–28.
- Bullmore, Edward T., Brammer, Michael J, Williams, Steven C R, Rabe-Hesketh, Sophia, Janot, Nicolas, David, Anthony S, Mellers, John, Howard, Robert, & Sham, Pak. 1996. Statistical Methods of Estimation and Inference for Functional MR Image Analysis. *Magnetic Resonance in Medicine*, **35**, 261–277.
- Bullmore, Edward T., Suckling, John, Overmeyer, S, Rabe-Hesketh, Sophia, Taylor, E, & Brammer, Michael J. 1999. Global, voxel, and cluster tests, by theory and permutation, for a difference between two groups of structural MR images of the brain. *IEEE Transactions on Medical Imaging*, **18**, 32–42.
- Bullmore, Edward T., Long, C, Suckling, John, Fadili, J., Calvert, G., Zelaya, F., Carpenter, T.A., & Brammer, Michael J. 2001. Colored noise and computational inference in neurophysiological (fMRI) time series analysis: resampling methods in time and wavelet domains. *Human Brain Mapping*, **12**(2), 61–78.
- Chumbley, J.R., & Friston, Karl J. 2009. False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage*, **44**(1), 62–70.
- Chumbley, J.R., Worsley, Keith J, Flandin, Guillaume, & Friston, Karl J. 2010. Topological FDR for neuroimaging. *NeuroImage*, **49**(4), 3057–64.
- Eklund, Anders, Andersson, Mats, & Knutsson, Hans. 2011. Fast random permutation tests enable objective evaluation of methods for single-subject FMRI analysis. *International journal of biomedical imaging*, **2011**(Jan.), 627–47.
- Flitney, David E., & Jenkinson, Mark. 2000. *Cluster Analysis Revisited*.
- Forman, S D, Cohen, J D, & Fitzgerald, M. 1995. Improved assessment of significant activation in functional magnetic resonance imaging (fMRI): use of a cluster-size threshold. *Magnetic Resonance in Medicine*, **33**, 636–647.

- Fox, Peter T, & Mintun, MA A. 1989. Noninvasive functional brain mapping by change-distribution analysis of averaged PET images of H215O tissue activity. *Journal of nuclear medicine : official publication, Society of Nuclear Medicine*, **30**(2), 141–9.
- Friman, Ola, & Westin, Carl-Fredrik. 2005. Resampling fMRI time series. *NeuroImage*, **25**, 859–867.
- Friston, Karl J., Frith, C D, Liddle, P F, & Frackowiak, R. S. J. 1991. Comparing functional (PET) images: the assessment of significant change. *Journal of cerebral blood flow and metabolism*, **11**(4), 690.
- Friston, Karl J., Worsley, Keith J, Frackowiak, R. S. J., Mazziotta, John C, & Evans, Alan C. 1993. Assessing the significance of focal activations using their spatial extent. *Human Brain Mapping*, **1**(3), 210–220.
- Friston, Karl J., Holmes, Andrew P, Poline, Jean-Baptiste, Price, Cathy J, & Frith, C D. 1996. Detecting activations in PET and fMRI: levels of inference and power. *NeuroImage*, **4**(3), 223–235.
- Genovese, Christopher R, Lazar, Nicole A, & Nichols, Thomas E. 2002. Thresholding of Statistical Maps in Functional Neuroimaging Using the False Discovery Rate. *NeuroImage*, **15**(4), 870–878.
- Gershman, Samuel J, Blei, David M., Pereira, Francisco, & Norman, Kenneth a. 2011. A topographic latent source model for fMRI data. *NeuroImage*, **57**(1), 89–100.
- Halber, Marco, Herholz, Karl, Wienhard, Klaus, Pawlik, Gunter, & Heiss, Wolf-Dieter. 1997. Performance of a Randomization Test for Single-Subject 15-O-Water PET Activation Studies. *Journal of Cerebral Blood Flow and Metabolism*, **17**, 1033–1039.
- Hayasaka, Satoru, & Nichols, Thomas E. 2004. Combining voxel intensity and cluster extent with permutation test framework. *Neuroimage*, **23**(1), 54–63.
- Haynes, John-Dylan, & Rees, Geraint. 2006. Decoding mental states from brain activity in humans. *Nature reviews. Neuroscience*, **7**(7), 523–34.

- Holmes, Andrew P. 1994. *Statistical Issues in Functional Brain Mapping*. Ph.D. thesis, University of Glasgow.
- Holmes, Andrew P, & Friston, Karl J. 1999. Generalisability, Random Effects & Population Inference. *NeuroImage*, **7**(4 (2/3)), S754.
- Holmes, Andrew P, Blair, RC, Watson, G, & Ford, I. 1996. Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow & Metabolism*, **16**(1), 7–22.
- Holmes, Andrew P, Watson, J D G, & Nichols, Thomas E. 1998. Holmes and Watson on 'Sherlock'. *Journal of Cerebral Blood Flow & Metabolism*, **18**(6), 697–698.
- Jenkinson, Mark. 2000. *Estimation of Smoothness from the Residual Field*.
- Kang, Jian, Johnson, Timothy D, Nichols, Thomas E., & Wager, Tor D. 2011. Meta Analysis of Functional Neuroimaging Data via Bayesian Spatial Point Processes. *Journal of the American Statistical Association*, **106**(493), 124–134.
- Keller, Merlin, Roche, Alexis, Tucholka, Alan, & Thirion, Bertrand. 2008. Dealing with spatial normalization errors in fMRI group inference using hierarchical modeling. *Statistica Sinica*, **18**, 1357–1374.
- Kiebel, S, Poline, Jean-Baptiste, Friston, Karl J., Holmes, Andrew P, & Worsley, Keith J. 1999. Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *NeuroImage*, **10**, 756–766.
- Kim, Seyoung, Smyth, Padhraic, & Stern, Hal. 2010. A Bayesian mixture approach to modeling spatial activation patterns in multisite fMRI data. *IEEE transactions on medical imaging*, **29**(6), 1260–74.
- Kriegeskorte, Nikolaus, Goebel, Rainer, & Bandettini, Peter. 2006. Information-based functional brain mapping. *Proceedings of the National Academy of Sciences*, **103**(10), 3863–8.

- Locascio, J J, Jennings, P J, Moore, C I, & Corkin, S. 1997. Time series analysis in the time domain and resampling methods for studies of functional magnetic resonance brain imaging. *Human Brain Mapping*, **5**(3), 168–193.
- Mumford, Jeanette A, & Nichols, Thomas E. 2009. Simple group fMRI modeling and inference. *NeuroImage*, **47**(4), 1469–75.
- Nandy, Rajesh, & Cordes, Dietmar. 2007. A semi-parametric approach to estimate the family-wise error rate in fMRI using resting-state data. *NeuroImage*, **34**(4), 1562–76.
- Neumann, Jane, von Cramon, D Yves, & Lohmann, Gabriele. 2008. Model-based clustering of meta-analytic functional imaging data. *Human brain mapping*, **29**(2), 177–92.
- Nichols, Thomas E., & Hayasaka, Satoru. 2003. Controlling the familywise error rate in functional neuroimaging: a comparative review. *Statistical Methods in Medical Research*, **12**(5), 419–446.
- Nichols, Thomas E., & Holmes, Andrew P. 2001. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, **15**(1), 1–25.
- Petersson, Karl Magnus, Nichols, Thomas E., Poline, Jean-Baptiste, & Holmes, Andrew P. 1999. Statistical limitations in functional neuroimaging II. Signal detection and statistical inference. *Philosophical Transactions of the Royal Society B: Biological Sciences*, **354**(354), 1261–1281.
- Poldrack, Russell A., Fletcher, Paul C, Henson, Richard N A, Worsley, Keith J, Brett, Matthew, & Nichols, Thomas E. 2008. Guidelines for reporting an fMRI study. *NeuroImage*, **40**, 409 – 414.
- Poldrack, Russell A., Mumford, Jeanette A., & Nichols, Thomas E. 2011. *Handbook of fMRI Data Analysis*. Cambridge University Press.
- Poline, Jean-Baptiste, & Mazoyer, B M. 1993. Analysis of individual positron emission tomography activation maps by detection of high signal-to-noise-ratio pixel clusters. *Journal of Cerebral Blood Flow & Metabolism*, **13**(3), 425–437.

- Poline, Jean-Baptiste, Worsley, Keith J, Holmes, Andrew P, Frackowiak, R. S. J., & Friston, Karl J. 1995. Estimating smoothness in statistical parametric maps: variability of p values. *Journal of Computer Assisted Tomography*, **19**(5), 788–796.
- Poline, Jean-Baptiste, Worsley, Keith J, Evans, Alan C, & Friston, Karl J. 1997. Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, **5**(2), 83–96.
- Raz, Jonathan, Zheng, Hui, Ombao, Hernando, & Turetsky, Bruce. 2003. Statistical tests for fMRI based on experimental randomization. *NeuroImage*, **19**, 226–232.
- Roland, P. E., Levin, B., Kawashima, R., & Kerman, S. 1993. Three-dimensional analysis of clustered voxels in 15O-butanol brain activation images. *Human Brain Mapping*, **1**(1), 3–19.
- Smith, Stephen M., & Nichols, Thomas E. 2009. Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, **44**(1), 83–98.
- Smith, Stephen M., Jenkinson, Mark, Johansen-Berg, H., Rueckert, Daniel, Nichols, Thomas E., Mackay, Clare E, Watkins, K.E., Ciccarelli, Olga, Cader, M.Z., Matthews, Paul M., & Behrens, Timothy E J. 2006. Tract-based spatial statistics: voxelwise analysis of multi-subject diffusion data. *Neuroimage*, **31**(4), 1487–1505.
- Thirion, Bertrand, Varoquaux, Gaël, & Poline, Jean-Baptiste. 2010. Accurate definition of brain regions position through the functional landmark approach. *Medical image computing and computer-assisted intervention (MICCAI)*, **13**(Pt 2), 241–8.
- Vul, Edward, Harris, C, Winkielman, P, & Pashler, H. 2009. Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science*, **4**, 274–290.
- Weeda, Wouter D, Waldorp, Lourens J, Christoffels, Ingrid, & Huizenga, Hilde M. 2009. Ac-

- tivated region fitting: a robust high-power method for fMRI analysis using parameterized regions of activation. *Human brain mapping*, **30**(8), 2595–605.
- Worsley, Keith J. 1994. Local Maxima and the Expected Euler Characteristic of Excursion Sets of χ^2 , F and t Fields. *Advances in Applied Probability*, **26**(1), 13–42.
- Worsley, Keith J, & Taylor, Jonathan E. 2005. An Improved Theoretical p Value for SPMs Based on Discrete Local Maxima. *NeuroImage*, **28**, 1056–1062.
- Worsley, Keith J, Evans, Alan C, Marrett, S, & Neelin, P. 1992. A three-dimensional statistical analysis for CBF activation studies in human brain. *Journal of Cerebral Blood Flow and Metabolism*, **12**(6), 900–900.
- Worsley, Keith J, Evans, AC, Marrett, Sean, & Neelin, Peter. 1993. Detecting and estimating the regions of activation in CBF activation studies in human brain. *Pages 535–547 of: Uemura, K. (ed), Quantification of brain function. Tracer Kinetics and brain PET*, vol. 2.
- Worsley, Keith J, Marrett, S, Neelin, P, Vandal, A C, Friston, Karl J., & Evans, Alan C. 1996. A unified statistical approach for determining significant signals in images of cerebral activation. *Human Brain Mapping*, **4**, 58–73.
- Worsley, Keith J, Taylor, Jonathan E, Tomaiuolo, Francesco, & Lerch, Jason P. 2004. Unified univariate and multivariate random field theory. *NeuroImage*, **23 Suppl 1**, S189–95.
- Xu, Lei, Johnson, Timothy D, Nichols, Thomas E., & Nee, Derek E. 2009. Modeling Inter-Subject Variability in fMRI Activation Location: A Bayesian Hierarchical Spatial Model. *Biometrics*, **65**(4), 1041–1051.
- Zhang, Hui, Nichols, Thomas E., & Johnson, Timothy D. 2009. Cluster mass inference via random field theory. *Neuroimage*, **44**(1), 51–61.