Voxel-based mapping of lesion-behavior relationships

Daniel Y. Kimberg, PhD

Revision 1, March 2007: initial revision
Revision 2, July 2009: updated to reflect current best practices; more detail on analysis; appendix on permutation testing; new section on masking, new section on dealing with nuisance covariates

Preliminaries

Introduction

The purpose of this document is to provide a practical overview of best practices for voxel-based analysis of the relationship between brain lesion data and behavior, an approach which Bates et al. (2003) have labelled "VLSM" for voxel-based lesion-symptom mapping. Although we don't always think of the behavioral measures as symptoms *per se*, the name "VLSM" is catchy and established, so I'll use it here.

VLSM is conceptually very simple. You have behavioral data and lesion maps from a group of subjects. On a voxel-by-voxel basis, you can assess the relationship between damage in that voxel and the behavioral measure. The difficulty comes in doing the statistics and image processing well.

This document is a conceptual overview – it's intended to give you a good idea of how to go about things, or about how to understand VLSM studies that you encounter in the literature. But it will not help you with the specific details of your study, and will not (for the most part) tell you what software to use or what buttons to push. If you aren't already comfortable with using your computer to do image analysis, we hope you will still learn something from this document, but you won't necessarily be ready to get started.

VLSM contrasted with non-voxel-based lesion analysis methods

The alternatives to VLSM are mostly data reduction approaches that categorize patients according to patterns of injury at a coarser spatial scale than voxels. For example, we might divide a group of stroke patients into those with and without frontal lesions, and ask whether or not the two groups differ on some behavioral measure.

Other traditional analyses could be described as voxel-based, but are not generally described as VLSM. Overlap mapping (see Rudrauf, 2008 for a recent discussion), as carried out currently, involves calculating the number or proportion of patients with injury to each voxel. This can be done separately for groups with and without some behavioral pattern (deficit), and can be very informative. However, for historical reasons overlap mapping is not generally considered VLSM.

VLSM contrasted with fMRI

VLSM seems a lot like fMRI in that we're looking at the association between behavior and data in brain images. Here are some key differences:

	VLSM	***fMRI***
nature of image data	one image (lesion map) per subject, or perhaps a small set of clinical images	an autocorrelated time series of images from each subject
role of behavior	behavior is the dependent variable	behavior is an independent variable, manipulated by varying task demands (even though you may also collect behavioral data in fMRI studies, those data rarely if ever figure into fMRI analyses as dependent variables)
role of spatial data	the lesion map is the independent variable you expect to predict behavior; so a different model	the fMRI data are the dependent variables you expect to be predicted by task conditions; so

	VLSM	***fMRI***
	is fit to the same behavioral data for each voxel	the same model is fit to different data for each voxel

Preparing the data

In a typical VLSM analysis you will have a single behavioral score and a 3-dimensional lesion map for each subject. You may have more than one behavioral score, but we can consider each to be subject to an independent analysis, unless you want to do something more complicated, like MANOVA (not discussed here) or ANCOVA.

The tricky first step in VLSM is to get all of the lesions segmented (delineated in some way) and registered to a common reference. That reference is often but not always MNI space, as embodied in the colin27 template often used with SPM and other imaging packages (for some background on MNI space and the colin27 template, see the link to Louis Collins's web page under “References”). The advantages of having your images in this space include compatibility with the large number of studies that report coordinates in MNI space (or other closely related spaces, see Brett et al., 2002 for some helpful discussion) and compatibility with electronic atlases for structure/region labelling that are provided in MNI space (e.g., see MRIcron and AAL in the software references below). Coordinates in MNI space can be readily converted to the space of the widely used Talairach Atlas, using coordinate conversions (see Lacadie et al., 2008).

Two difficult problems lie at the heart of the segmentation/registration process. First, it's not always easy to decide where a lesion starts and ends. Second, even if you know where a patient's lesion is on that patient's MRI, it's not always easy to decide how that location maps onto your standard template of choice.

There are many other tricky issues in deciding how to segment lesions. For present purposes, I assume you are aware of these issues and have decided how to handle them. In the future, I hope to expand this section with a more detailed discussion of how to address uncertainty in lesion location and status.

Manual Registration and Segmentation in One Step

This commonly used procedure, in various forms, solves the registration and segmentation problems in one step, at the cost of expert human labor. Basically, a human rater with knowledge of neuroanatomy sits down with a template on the computer (typically in some program like MRIcro) and the patient scan(s) in some form. Then, working slice by slice, the rater uses his/her best judgment to draw the lesion onto the corresponding slice on the template. Of course, there isn't generally a clean one-to-one mapping of slices, but by rotating the template around the LR axis (i.e., changing the pitch), it's usually possible to find a reasonable match. The template volume, including the segmented lesion, can be rotated back into its original orientation later.

The major drawback of this procedure is that it's almost entirely manual – labor-intensive and subject to human error and biases. The method is time-consuming and stressful due to the uncertainty in the mapping. It's often a good idea to have at least two raters draw the same lesions, which doubles the work and usually triggers an interest in measuring inter-rater agreement. Some useful metrics of agreement are described in Fiez et al. (2000). We have also noted that spatial smoothness of lesion maps is invariably greater in the plane in which the lesions are drawn, suggesting that the process would benefit from even more labor, specifically to examine the lesions for continuity in other slice planes.

Automated Registration

In the past, registration of lesioned brains has been problematic. Automated warping procedures, which are commonly used in fMRI, might end up with lesions mapped onto ventricles (both having low signal intensity), or other anomalies. However, recent advances in warping methods

have begun to obviate these issues, at least in cases where research-quality MRI scans are available. Crinion et al. (2007) reported that SPM's unified segmentation (which incorporates the registration process) performed at a high level with lesioned brains, and importantly, did so without the need for either cost function masking (cf. Brett al al., 2001) or regularization contstraints (Tyler et al., 2005). Kim et al. (2008b) have found an approach similar to cost function masking to be helpful when using the SyN algorithm available in the ANTS software package (e.g., see Kim et al., 2008a).

The availability of high-quality registration takes a large part of the burden off the manual rater, requiring only decisions about the boundaries of the lesion in native subject space.

Automated Segmentation

Given the viability of automated registration, we lack only automated lesion segmentation for a fully automated approach to mapping lesions onto a common template. My current feeling is that there are too many idiosyncratic decisions made in this process to automate it fully, especially given that experienced raters can disagree as to what should constitute abnormal tissue. So I will defer detailed discussion of automated segmentation at this time. However, it is worth noting that automated approaches to this process have been described by Stamatakis and Tyler (2005) and by Seghier et al. (2008). Tyler et al. (2005) have also described a correlational approach that obviates the need for segmenting lesions, instead calculating the correlation between signal intensity and behavior, across subjects. While valuable, this technique generally requires scans from the same scanner for all subjects.

Analysis

The endpoint of a VLSM analysis is a brain map that includes an independently calculated statistic for each voxel, typically thresholded for some standard of statistical significance. Here I describe some of the issues involved in this process.

Masking

Researchers who use VLSM typically use a cutoff for the minimum number of lesions required for a voxel to be included in the analysis. Although the cutoff is somewhat arbitrary, it is constrained both by logic and by the requirements of the test statistics. For example, Welch's t test for unequal variances requires at least two subjects in each group, and is better behaved with more. The Brunner-Munzel test advocated by Rorden et al. (2007) is unreliable with fewer than about 10-15 subjects (the latest versions of his software use different strategies in infrequently lesioned voxels). Unfortunately, there are no absolute guidelines for how to set this cutoff. Because patient groups are often small, it's difficult to meet stringent cutoffs. For example, in a group of 24 patients the maximum number of patients with lesions of a given voxel may be 12, which means that a cutoff of 10 will remove most of the data. At the same time, being overly inclusive can hurt the statistics, particularly if you're using FDR. At present I can only offer the following very loose recommendations:

- Consider before running your analysis how small a patient group you would consider reportable if you had a traditional non-VLSM patient-control study. If you would never report a behavioral study comparing three patients to 17 control subjects (at least not as a group study), then perhaps the same reasons should apply to VLSM.
- Respect the limitations of your statistical test. If it's unreliable with fewer than n subjects, omit voxels with fewer than n subjects (or choose a different statistic).
- Eyeball your coverage with different cutoffs, so that you at least know what you're trading off when you decide which voxels to include.
- Consider a power analysis (see below) to avoid including voxels in which you would need an impossibly large effect to achieve significance.

Choosing a test statistic

The product of a VLSM analysis is a statistical map in which each voxel quantifies the difference in behavior between patients with and without a lesion in that location. Generally, each voxel will be assigned the value of some test statistic, often a Student's t statistic comparing the two groups (patients with and without lesions in that voxel). The t statistic is especially useful because it's readily available, well-understood, easily tested for significance, reasonably robust, and generally appropriate for VLSM analyses. It can be used to contrast two groups (scores for lesioned vs. intact in a given voxel) or to assess the correlation between continuous-valued lesion scores (probability of a lesion in a given voxel ranging from 0 to 1) and behavior.

Rorden et al. (2007) have characterized several other statistics that could be useful in the context of lesion analysis. The Welch's t-test, which does not assume equal variances between groups, may often be appropriate for VLSM (see Rorden et al., 2007), and is just as widely available as the regular t-test. The Brunner-Munzel test is a non-parametric replacement for the t-test. And the Liebermeister test replaces the standard chi-squared test. Presently, these tests are available for imaging data in Rorden's NPM package (distributed with MRIcron), which is in active development but already widely used.

Many others statistics could be appropriate, depending on your study. Importantly, if you use permutation testing (see below), your options are not restricted to measures that can be readily transformed to conform to a known parametric distribution.

Choosing a standard for significance

One could, by default, simply use a normal parametric test in each voxel and Bonferroni correct for multiple comparisons. However, this would be much too severe a correction for VLSM, since lesion maps have high spatial coherence (the lesion status of a voxel is well predicted by that of the neighboring voxels). I here describe three options for thresholding VLSM maps, each of which accounts for multiple comparisons in a different way.

First, although the t statistic is usually used with the t-test (i.e., its extreme-ness is tested against the parametric Student's t distribution), it doesn't have to be. We can also view it as a measure of the reliability of the relationship between lesion and behavior, and test its significance non-parametrically, with a permutation test.

Briefly, the permutation test is a non-parametric approach to significance testing that provides an intuitive solution to many of the problems associated with VLSM. The variant used with image data controls for multiple comparisons without overcorrecting for identical or highly correlated comparisons. The significance thresholds estimated by the permutation test are exact and unbiased, and the test does not depend on having a known parametric distribution. The disadvantages include a dearth of software implementations, computational cost (for typical VLSM studies it may already be negligible), and the natural reluctance of many investigators to embrace unfamiliar methods. The logic of the permutation test is described in more detail in Appendix A.

A second option, in cases where we have binomial lesion scores (each voxel is either lesioned or not in a given patient), is simpler and often nearly as effective as the permutation test: we can simply count the number of distinct patterns of lesioned patients (termed DLPs by Rorden et al., 2007). Two voxels are distinct if there is any difference between the two; specifically, if at least one patient is lesioned in one but not the other. Effectively, this approach avoids penalizing the statistical tests for regions within which the same subset of patients is lesioned. If we have a region (contiguous or not) of 1000 such voxels (i.e., 1000 voxels but only 1 DLP), we can consider that region to be a single comparison, since we have no information that would allow us to dissociate voxels within the region. We (Kimberg et al., 2007) found this approach to be nearly as good as the permutation test, and Rorden et al. (2007) found the two to be indistinguishable. Importantly, it can be used to Bonferroni correct in cases for which no permutation test is available.

A third option for thresholding is to control the false discovery rate (FDR; Benjamini and Hochberg, 1995). The false discovery rate is the expected proportion of false positives among all

reported supra-threshold voxels, and the method is easily adapted to brain imaging (Genovese et al., 2002). FDR thresholding is generally much more liberal than map-wise control of the false positive rate. If we set a threshold to control the FDR at 0.01, and 1000 voxels exceed this threshold, we expect 10 false positives. By contrast, if we control the map-wise false positive rate at 0.05, then the probability of observing any false positives at all is only 0.05. So while lower thresholds are helpful in (often under-powered) VLSM studies, it is important to remember that FDR thresholding does not offer the same assurances as Bonferroni correction or permutation testing.

If you do plan to use FDR, it is important to choose an appropriate false discovery rate. As with traditional FWER (family-wise, in this case map-wise, error rate) control, the choice is somewhat arbitrary. But unlike FWER, FDR does not have a long history during which community standards have been established. While some researchers have been using 0.05, in my view this is dangerously liberal. An FDR of 0.01 seems to provide a more reasonable balance in typical brain image data.

Nuisance covariates and other more complicated models

Even if a linear model meets your needs, you may need something a little more complex than a two-sample t-test or a regression with a single independent variable. Many such models can be solved in the most straightforward way with existing software. However, permutation tests for more complicated models are often not available, at least not for imaging data.

One of the most straightforward such needs is the inclusion of one or more nuisance covariates – variables such as age, education, motor speed, or relative humidity that might affect behavior but are not of direct interest. The easiest approach to carrying out a permutation test in this case is to regress these variables out of the data – i.e., regress your data against the nuisance covariates and then carry out your "real" model on the residuals. This general approach has been endorsed by Nichols et al. (2008) and is readily carried out in software.

In the event a permutation test is not needed, some more complex models can now be carried out within imaging software, specifically within the VoxBo and BPM packages. These tests generally yield t or F statistics that can be tested parametrically.

VLSM in fMRI software

Software for fMRI analysis usually assumes the image is your dependent measure and behavior is an independent measure. Although lesion analysis reverses this relationship logically, we can carry out the test backwards as long as there are no other covariates in the model (except an intercept term). That is, you can design the model as though the lesion score is your dependent measure and behavior is the independent measure. Doing so may be helpful if you want to use fMRI packages that have not been retrofit for lesion analysis.

Power Analysis

Power analysis is critical in lesion analysis for many reasons, including the scarcity of patients, variability in performance, and sparse spatial distribution of lesions. VLSM complicates power analysis in several ways. First, there is a correction for multiple comparisons. VLSM datasets are in general highly spatially coherent – the lesion status of a given voxel is very well predicted by the lesion status of adjacent voxels. The problem is even larger when lesions are drawn at arbitrarily high resolution, which increases the number of voxels without necessarily increasing the amount of information (i.e., the number of *independent* voxels).

The permutation test described above can be used to bootstrap a rough estimate of the number of independent voxels in a dataset. In a recent article (Kimberg et al., 2007), we carried out a permutation test using fabricated dummy behavioral data. An effective number of independent comparisons was derived by finding the number of comparisons for which Bonferroni correction would yield the same threshold as the permutation test. This at least gives us a more palatable correction for multiple comparisons. Having this correction, we can carry out power analysis on a voxel-by-voxel

basis for the statistic of interest, in the same way we would do so for a univariate test, using estimated behavioral effect size, variances, and group sizes. Of course, we could also use the number of distinct comparisons to obtain a basis for Bonferroni correction more easily.

A second complication with power analysis in VLSM is the spatial variability in power. At some level, fMRI shares this problem. However, it is likely worse in VLSM because the non-random spatial distribution of lesions (depending on the patient group) can lead to regions in which very few lesions are present – the group sizes that enter into the power analysis vary from voxel to voxel. For example, in the sample dataset described in Kimberg et al. (2007), regions of interest varied in estimated power from well over 0.8 to well below 0.4, based solely on the distribution of lesion locations in a group of 55 left hemisphere stroke patients. Low power of 0.4 may or may not be problematic, depending on how critical the under-powered regions are to the research questions of interest. It is important in VLSM to consider not just a single power measure but the expected distribution of power across the brain.

This general approach to mapping power for VLSM studies would of course need to be tailored to the demands of a given study in terms of the test procedure used and the distribution of lesions.

References

Articles

Bates, E.; Wilson, S. M.; Saygin, A. P.; Dick, F.; Sereno, M. I.; Knight, R. T. & Dronkers, N. F. (2003), 'Voxel-based lesion-symptom mapping.', *Nat Neurosci* **6**(5), 448-50.

Benjamini, Y. & Hochberg, Y. (1995), 'Controlling the false discovery rate: A practical and powerful approach to multiple testing', *Journal of the Royal Statistical Society, Series B, Methodological* **57**, 289-300.

Brett, M.; Johnsrude, I. S. & Owen, A. M. (2002), 'The problem of functional localization in the human brain.', *Nat Rev Neurosci* **3**(3), 243-9.

Brett, M.; Leff, A. P.; Rorden, C. & Ashburner, J. (2001), 'Spatial normalization of brain images with focal lesions using cost function masking.', *Neuroimage* **14**(2), 486-500.

Crinion, J.; Ashburner, J.; Leff, A.; Brett, M.; Price, C. & Friston, K. (2007), 'Spatial normalization of lesioned brains: performance evaluation and impact on fMRI analyses.', *Neuroimage* **37**(3), 866-75.

Fiez, J. A.; Damasio, H. & Grabowski, T. J. (2000), 'Lesion segmentation and manual warping to a reference brain: intra- and interobserver reliability.', *Hum Brain Mapp* **9**(4), 192-211.

Genovese, C. R.; Lazar, N. A. & Nichols, T. (2002), 'Thresholding of statistical maps in functional neuroimaging using the false discovery rate.', *Neuroimage* **15**(4), 870-8.

Kim, J.; Avants, B.; Patel, S.; Whyte, J.; Coslett, B. H.; Pluta, J.; Detre, J. A. & Gee, J. C. (2008a), 'Structural consequences of diffuse traumatic brain injury: a large deformation tensor-based morphometry study.', *Neuroimage* **39**(3), 1014-26.

Kim, J.; Avants, B.; Patel, S.; Whyte, J. (2008b), 'Spatial normalization of injured brains for neuroimaging research: An illustrative introduction of available options. Technical report available via www.ncrrn.org.

Kimberg, D. Y.; Coslett, H. B. & Schwartz, M. F. (2007), 'Power in Voxel-based lesion-symptom mapping.', *J Cogn Neurosci* **19**(7), 1067-80.

Lacadie, C. M.; Fulbright, R. K.; Rajeevan, N.; Constable, R. T. & Papademetris, X. (2008), 'More accurate Talairach coordinates for neuroimaging using non-linear registration.', *Neuroimage* **42**(2), 717-25.

Nichols, T.; Ridgway, G.; Webster, M. & Smith, S. (2008), *GLM permutation - nonparametric inference for arbitrary general linear models*, Vol. Human Brain Mapping.

Rorden, C.; Bonilha, L. & Nichols, T. E. (2007), 'Rank-order versus mean based statistics for neuroimaging.', *Neuroimage* **35**(4), 1531-7.

Rudrauf, D.; Mehta, S.; Bruss, J.; Tranel, D.; Damasio, H. & Grabowski, T. J. (2008), 'Thresholding

lesion overlap difference maps: application to category-related naming and recognition deficits.', *Neuroimage* **41**(3), 970-84.
Seghier, M. L.; Ramlackhansingh, A.; Crinion, J.; Leff, A. P. & Price, C. J. (2008), 'Lesion identification using unified segmentation-normalisation models and fuzzy clustering.', *Neuroimage* **41**(4), 1253-66.
Stamatakis, E. A. & Tyler, L. K. (2005), 'Identifying lesions on structural brain images--validation of the method and application to neuropsychological patients.', *Brain Lang* **94**(2), 167-77.
Tyler, L. K.; Marslen-Wilson, W. & Stamatakis, E. A. (2005), 'Dissociating neuro-cognitive component processes: voxel-based correlational methodology.', *Neuropsychologia* **43**(5), 771-8.

Web sites

Collins, D. L. "A short history of stereotaxic data volumes at the MNI"
http://www.bic.mni.mcgill.ca/~louis/stx_history.html

Software

The following software packages were discussed in this article:

VoxBo

www.voxbo.org
free (GPL), runs on Linux, OSX, and Windows (via cygwin)
VoxBo is a complete pacakge for fMRI analysis that is currently being extended to address lesion analysis. The front page of the web site has a link for information about lesion analysis in VoxBo.

ANTS

http://www.picsl.upenn.edu/ANTS/
free
ANTS is a flexible normalization package that includes an implementation of the SyN algorithm for symmetric diffeomorphic registration

SPM

http://www.fil.ion.ucl.ac.uk/spm/
free (GPL), requires MATLAB
SPM is the most widely used package for fMRI data analysis. Numerous add-on toolkits are available, including two listed below.

SnPM

http://www.sph.umich.edu/ni-stat/SnPM/
freely distributed, requires SPM
SnPM is an extension to SPM to provide non-parametric permutation testing.

BPM

http://www.fmri.wfubmc.edu/
free (unknown), requires SPM/MATLAB
BPM stands for biological parametric mapping. It is an SPM extension that provides for image independent variables.

VLSM

http://crl.ucsd.edu/vlsm/
free (GPL), requires MATLAB
VLSM is an independent lesion analysis package, written in MATLAB.

MRIcro

http://www.sph.sc.edu/comd/rorden/mricro.html
free, runs on Windows, Linux, and Solaris x86
Chris Rorden's MRIcro is widely used for lesion tracing and image viewing. It also has some lesion analysis tools built in.

MRIcron/NPM

http://www.sph.sc.edu/comd/rorden/mricron/
open source, runs on Linux, OSX, and Windows
Chris Rorden's MRIcron is a new program that we assume will supplant MRIcro. It's in the early stages of development, but provides exciting new anaysis functionality. The MRIcron download package includes NPM (non-parametric mapping), a program that provides an efficient implementation of permutation testing for lesion data. Rorden's web site has a lot of useful information for anyone contemplating VLSM analyses. The following page is a good place to start:
http://www.sph.sc.edu/comd/rorden/mricron/stats.html

Anatomical Automatic Labeling (AAL)

http://www.cyceron.fr/freeware/
freely available, requires SPM99 or SPM2
AAL is an atlas in MNI space.

(d)stplan

http://biostatistics.mdanderson.org/SoftwareDownload/
freely available
stplan (aka dstplan, for double-precision study planning) is a powerful command line package for power analysis

Appendix A: Permutation testing

The permutation test is a non-parametric approach to significance testing that is both intuitive and powerful. In the context of imaging, it provides a simple solution to many otherwise difficult problems, especially the multiple comparison problem. Although not a particularly new technique, the permutation test is computationally expensive, and until recently not widely available in statistical packages. For these reasons, it has only become practical for routine use in research over the past decade. The purpose of this appendix is to give a brief, non-technical introduction to the logic of the permutation test.

The key insight in understanding the logic of the permutation test is that when the null hypothesis (H_0) is true, and there is no special relationship between our dependent and independent/ grouping variable(s), we can re-order one or the other with no expected effect on the test statistic. That is, when H_0 is true, we should not expect our true test statistic to be any more extreme than those from re-ordered data. To the extent the correct arrangement produces an extreme value, we can reject H_0. In effect, permutation testing allows us to derive a null distribution that we can use in place of some parametric distribution. We can then ask if there is something special about the known correct ordering of our variables, as contrasted with all the other possible orderings.

To give a simple non-imaging example, imagine a single measure taken from twenty members of group A and twenty members of group B. We could test the group difference with a t-test, but perhaps we are unwilling to trust the assumptions dictated by the parametric test. To calculate the permutation test, we want to try all the possible ways in which those forty subjects could have fallen into two equal-sized groups. There are C(40,20) different ways of assigning the group membership labels to the 40 scores, and we can calculate a t statistic for all of them. If H_0 is true, there is nothing special about the correct labeling, and it should be no more likely than any other arrangement to produce an extreme t value – the probability of its t value being in the top 5% is 5%. To the extent the t statistic associated with the correct labelings is extreme relative to all the other possible permutations, we can describe the difference between the groups as statistically significant. Effectively, the permutations provide a distribution for our test statistic when H_0 is true, which is just what we need to determine how extreme our true observed value is.

C(40,20) is 137,846,528,820, a painfully large number of t statistics to set up and calculate, even for this small dataset. In practice, the permutation distribution for our test statistic is usually estimated by randomly sampling the space of all possible permutations. This means that the p value you get has some estimation error associated with it. The formula for the standard error on the p value is:

$$SEp = \sqrt{(p*(1-p)/N)}$$

where N is the number of permutations.

It is also important to note that although the metric of group difference used in this example is the t statistic, it is used here simply as a normalized measure of group difference. It's a t statistic, but not a t-test. This is a common procedure, although there is no reason the test statistic you choose needs to conform to a known parametric distribution. We could also have used a simple difference in means.

In a typical VLSM analysis, we can carry out a permutation test by permuting which behavioral score goes with which subject. Usually this is accomplished in software by permuting a vector that contains the behavioral scores and leaving the lesion maps intact. For each permutation of the dependent measure, we calculate a separate statistical map. From this set of 1000 or more statistical maps, we can derive our reference distribution, and from that, an appropriate statistical threshold.

The permutation test provides a natural solution for calculating a threshold that avoids the multiple comparison problem in imaging data. Consider a large brain volume and a statistic calculated for every voxel in that volume, for a large number of permutations. We can control the map-wise false positive rate by choosing a threshold that is only exceeded *anywhere* in the brain in

5% of the permutations. If we think of the permutations as replications of our experiment under null conditions (H0 is true), then it's easy to see that this provides exactly the same kind of assurance as Bonferroni correction – at our threshold, we only expect to see a false positive anywhere in the brain 5% of the time. This test is sometimes described as the “maximum test,” because it's carried out by constructing a permutation distribution from the maximum value in each permutation volume.