7-2007

Power in Voxel-based Lesion–Symptom Mapping

Daniel Y. Kimberg
University of Pennsylvania, kimberg@mail.med.upenn.edu

H. Branch Coslett
University of Pennsylvania, hbc@mail.med.upenn.edu

Myrna F. Schwartz
Moss Rehabilitation Research Institute

Suggested Citation:
Kimberg, D.Y., Coslett, H.B. and Schwartz, M.F. (2007). Power in Voxel-based Lesion–Symptom Mapping. *Journal of Cognitive Neuroscience*. Vol. 19(7). pp. 1067-1080.

Power in Voxel-based Lesion–Symptom Mapping

Abstract
Lesion analysis in brain-injured populations complements what can be learned from functional neuroimaging. Voxelbased approaches to mapping lesion–behavior correlations in brain-injured populations are increasingly popular, and have the potential to leverage image analysis methods drawn from functional magnetic resonance imaging. However, power is a major concern for these studies, and is likely to vary regionally due to the distribution of lesion locations. Here, we outline general considerations for voxel-based methods, characterize the use of a nonparametric permutation test adapted from functional neuroimaging, and present methods for regional power analysis in lesion studies.

Disciplines
Medicine and Health Sciences

Comments
Suggested Citation:
Kimberg, D.Y., Coslett, H.B. and Schwartz, M.F. (2007). Power in Voxel-based Lesion–Symptom Mapping. *Journal of Cognitive Neuroscience.* Vol. 19(7). pp. 1067-1080.

Power in Voxel-based Lesion–Symptom Mapping

Daniel Y. Kimberg[1], H. Branch Coslett[1], and Myrna F. Schwartz[2]

Abstract

■ Lesion analysis in brain-injured populations complements what can be learned from functional neuroimaging. Voxel-based approaches to mapping lesion–behavior correlations in brain-injured populations are increasingly popular, and have the potential to leverage image analysis methods drawn from functional magnetic resonance imaging. However, power is a major concern for these studies, and is likely to vary regionally due to the distribution of lesion locations. Here, we outline general considerations for voxel-based methods, characterize the use of a nonparametric permutation test adapted from functional neuroimaging, and present methods for regional power analysis in lesion studies. ■

INTRODUCTION

Functional brain imaging and the study of brain-injured patients are complementary methods for investigating brain–behavior relationships using brain images (Chatterjee, 2005; Rorden & Karnath, 2004; Shallice, 2003). Despite the many substantive differences between them, the two approaches are in a strong position to share methods and tools for data analysis due to their shared concern with characterizing the regional distribution of cognitive function in the brain.

Voxel-based lesion-symptom mapping (VLSM; Bates et al., 2003) has increasingly been advocated as an approach to measuring the role of regional injury in patterns of behavior. Briefly, the method involves mapping the relationship between brain injury and behavioral performance on a voxel-by-voxel basis.[1] That is, the statistical relationship between damage and behavior (across patients) is calculated separately for each voxel. This may be contrasted with less fine-grained techniques that involve forms of data reduction such as grouping patients according to the involvement of some region or stratifying on behavioral scores. These approaches will continue to be appropriate (and in some cases preferable) to the extent that data reduction abstracts over incidental features of the data while capturing meaningful regularities. However, where in the past data reduction has been necessary due to the burden of computing large numbers of statistical tests, modern computing combined with VLSM makes a voxel-based approach practical, and, in some cases, preferable.

VLSM represents an especially important advance in methods for patient-based cognitive neuroscience research, given the recently dominant role of functional neuroimaging. VLSM puts lesion analysis in a position to take advantage of many of the tools originally developed for functional neuroimaging, most of which are voxel-based. This potentially helps researchers who study patients make optimal use of their data either independently or in connection with functional magnetic resonance imaging (fMRI) studies. As well, VLSM makes patient-based research both more attractive and more accessible to imaging researchers already comfortable with image analysis methods.

One of the drawbacks of VLSM when compared to traditional lesion analysis methods is power. Power is often problematic when planning patient studies in general, due to the heterogeneity of patient groups, the high variability in patient performance, and difficulties in recruiting. VLSM adds to this list a potentially severe correction for multiple comparisons. Given the expense and difficulty in carrying out patient-based research, it is important to assess power quantitatively as early as possible in planning a study.

In this article, we articulate some basic considerations for processing and analysis in VLSM, with a focus on statistical power. In particular, we consider the power for two kinds of tests: univariate tests of the relationship between lesion status and some behavioral measure carried out separately for each voxel; and tests meant to discriminate between voxels or locations. We describe the former as useful in making *spatially localized* inferences, in the sense that inferences drawn from massively univariate tests provide reliable information about specific brain regions without providing useful information about differences between regions. We describe tests designed specifically to differentiate between regions as *spatially discriminating*. For example, in a population with damage to the motor cortex, we might

[1]University of Pennsylvania, [2]Moss Rehabilitation Research Institute, Philadelphia, PA

test the association between left motor cortex damage and right-hand performance, supporting a spatially localized inference that concerns the left motor cortex in isolation. Or we might test whether the association between damage and right-hand performance is greater for the left motor cortex than for the right motor cortex, supporting a spatially discriminating inference that compares the roles of the left and right motor regions.

We proceed by first reviewing the similarities and differences between lesion and functional imaging methods, and methods for registration and segmentation of lesion maps. We then describe procedures for correction for multiple comparisons, including a resampling approach that improves dramatically on Bonferroni correction. We further characterize power for statistical tests on the basis of lesion data, including power to make both spatially localized and spatially discriminating inferences. We conclude with a brief discussion of inferential problems difficult to address using strictly voxel-based methods.

LESION ANALYSIS COMPARED TO fMRI

Functional imaging and lesion[2] studies are both imperfect methods for discovering the functional organization of the brain. Each presents difficulties in localization stemming from the highly interactive nature of neural systems (Farah, 1994), and both present inferential weaknesses that can make interpreting results difficult. There are also cases for which one method or the other presents special difficulties. For example, imaging studies of memory function in the hippocampus are complicated by both susceptibility-related sensitivity loss and difficulty in constructing an appropriate cognitive subtraction (Stark & Squire, 2001; Binder et al., 1999). Patient studies are complicated by the nonrandom distribution and heterogeneous nature of brain injuries. Although there are often points of convergence, studies using these two methods do not always implicate exactly the same regions.

Although neither method is perfect, the fact that they suffer from *different* inferential limitations underscores the importance of convergent evidence for arguments concerning the functional architecture of the brain. Despite this, imaging studies have had a disproportionate impact on cognitive neuroscience since the advent of widely available BOLD (blood oxygen level dependent) fMRI in the mid to late 1990s (Fellows et al., 2005). The reasons for this state of affairs are complex (Chatterjee, 2005), but likely include the intrinsic appeal of the technology associated with fMRI and the ease with which fMRI data may be collected once the technology is available.

As noted by Chatterjee (2005), we can ask questions about the functional architecture of the brain in two ways: for a given region, we may ask what it does, or for a given function, we may ask where it is located. How we go about answering these questions depends on a causal structure that is strikingly different between fMRI and lesion studies. In typical lesion studies, brain function is varied (nonexperimentally) and behavioral performance is measured. In fMRI studies, task performance is manipulated experimentally and magnetic resonance imaging (MRI) signal is measured (as an index of neural activity).

fMRI data provide high-quality evidence that observed differences in signal are due to the experimental manipulation, but little evidence concerning the role of the observed region in cognitive processes. Although fMRI activation is often presumed to be best explained by a causal role of the observed region in some cognitive process, the data generally admit a variety of other explanations, often including the possibility that the activity is epiphenomenal to the process of interest. Lesion data, by contrast, provide high-quality evidence that observed differences in behavior are due to differences in brain injury. When groups are differentiated solely by lesion location and otherwise randomly sampled, this can support spatially discriminating inferences. These inferences must often be qualified—even if damage to a certain region is associated with difficulty in a certain task, we cannot always infer that the region is functionally involved in the normal performance of the task. For example, the region may be adjacent to (and therefore, naturally confounded with) functionally unrelated areas, or may cause remote impairment through diaschisis. We review some additional inferential weaknesses of lesion studies in the Discussion.

The difference in the direction of causality between lesion data and fMRI results in differently structured statistical models. In the SPM (see Appendix) approach to fMRI analysis, the signal value in each voxel is considered the dependent variable for a separate test, and behavior is usually considered the independent variable. In VLSM analyses, each voxel's lesion status is considered to be an independent variable for a separate test, and behavioral scores are the dependent outcome measure. Thus, fMRI involves a dependent measure that varies from voxel to voxel, and a single model that is fitted separately to each voxel's data. A typical lesion analysis involves an independent variable—lesion status—that varies from voxel to voxel, and therefore, a separate model for each voxel, but the same dependent (behavioral) measure. When there is only a single independent variable, we can reverse the relationship (modeling damage as a linear function of some behavioral index) without disturbing the logic of our statistical tests. That is, whether we model lesion score as a function of behavior or vice-versa, we will still get the same t scores and significance levels. However, when additional variables are used to explain behavioral scores (e.g., nuisance covariates in addition to voxel lesion status), the model needs to be structured canonically, with

behavior as the dependent measure and other variables including lesion status as independent measures. This is because we typically hypothesize an effect of the covariate on behavior, but not on the location of the lesion.

Having a model that varies from voxel to voxel has practical consequences for the software used to carry out VLSM-style analyses. fMRI analysis packages do not typically support using image data as an independent variable, and may be nontrivial to retrofit due to the way intermediate products are stored (because they may only be reused when the model remains the same from voxel to voxel, as in fMRI). At the same time, lesion analysis involves image data with no temporal structure (i.e., autocorrelation), which avoids a major complicating factor in methods for image data analysis.

Finally, we note that lesion studies are complicated by marked heterogeneity in both the imaging and behavioral performance. Most functional imaging studies of healthy subjects combine data from subjects scanned on the same scanner, using the same scanning protocol. Even meta-analyses and multisite studies generally have sizable groups from each site/study. Without minimizing the other sources of intersubject variability in fMRI, it is safe to say that this kind of uniformity is rarely achieved in studies of patients with focal lesions, which often rely on clinical imaging. Variation in image quality may be due to the nature of the imaging protocol (including a vast array of imaging parameters), the quality of the imaging hardware, differences in head movement, and when the images were taken relative to the injury or injuries. Studies often must include patients who cannot undergo MRI, but who may have lower quality computed tomography (CT) images available. In principle, variation in the quality of images complicates the process of analyzing combined data, although in practice these differences are often ignored. Similarly, VLSM studies often have discretely identifiable subpopulations with potentially different variances in behavioral measures. This may make it inappropriate to use ordinary least squares solutions to linear models.

REGISTRATION AND SEGMENTATION

VLSM-type statistical analyses presuppose methods for reliable registration of structural maps of patients' brains, as well as methods for reliable segmentation of lesions. There are several viable options for both these procedures, varying in the type of data required, their performance under different conditions, and in their labor-intensiveness. A detailed evaluation of viable approaches to these two processes is beyond the scope of this article, but we here provide a brief review of prominent approaches.

The purpose of intersubject registration to a common template (''spatial normalization'') is to align the brains such that voxels at the same coordinate in two subjects' brain images have signal intensities drawn from structurally corresponding regions (which we typically hypothesize will also correspond functionally). The accuracy of this process can vary from subject to subject, or even from region to region. A number of automated and semi-automated algorithms are available for MRI data (e.g., Avants, Schoenemann, & Gee, 2005; Shen & Davatzikos, 2002; Ashburner & Friston, 1999; Woods, Grafton, Watson, Sicotte, & Mazziotta, 1998; see also Brett, Johnsrude, & Owen, 2002 for discussion of this issue) and are routinely used in group analysis for both structural and fMRI.

The procedures used to accomplish this normalization can be problematic even with structurally intact brains from homogeneous populations of healthy subjects, all acquired on the same scanner. The impact of small defects in registration is reduced somewhat by the inherent spatial smoothness of the data of interest (BOLD fMRI data or lesion maps), and typically by explicit smoothing of the data, which trades spatial resolution for robustness to misalignment. But with the diverse populations typical of lesion studies, and the presence of (often large) structural defects, fully automated normalization to a common template can easily fail. Modes of failure are varied, but a typical undesirable result might map a large lesion onto a ventricle or a small lesion onto a wide sulcus, where in neither case is the lesion properly mapped onto locations associated with the damaged underlying tissue.

We can identify four prominent approaches to avoiding these problems. First is the cost function masking approach described by Brett, Leff, Rorden, and Ashburner (2001). Within a class of normalization algorithms, finding the best transformation involves searching a space of transformations for the one that minimizes some cost function, essentially a measure of mismatch between the lesioned brain and the template. Cost function masking involves identifying a region to be excluded from the cost function calculation, in this case, the lesion itself, which must be identified manually. By excluding the lesion from the cost function calculation, the algorithm will search only for the best fit for the remaining tissue.

A second approach adopted by Tyler, Marslen-Wilson, and Stamatakis (2005) is to use the standard SPM normalization algorithm with a high regularization constraint to penalize ''unlikely'' deformations. This essentially constrains the transformations considered in searching for the best fit. Using this approach within SPM99 (www.fil.ion.ucl.ac.uk/spm/), they reported satisfactory results in 16 of 19 patients with focal lesions.

Landmark-based registration is still a viable approach for bringing brain images into rough alignment (cf. Talairach & Tournoux, 1988). In this approach, manually chosen landmarks are identified on both volumes, which constrains the transformation at these points. The mapping for intermediate points is determined by assuming a smooth, continuous function between the landmarks.

A fourth approach is manual transformation. Skilled raters here transcribe the lesion onto a standard template (typically first rotated to at least coarsely match the pitch of the subject's images). Although various parts of the process may still be automated, the rater is responsible for mapping the contour of the lesion onto homologous structures on the template image. This approach combines registration and segmentation into one step.

Several approaches have been used for lesion segmentation. As noted above, manual segmentation by skilled raters, which has been used in lesion studies since long before computers became commonplace, continues to be a viable option. Without a true gold standard method for assessing damage status on a voxel-by-voxel basis, manual segmentation draws on the knowledge of a skilled rater that may be as yet poorly captured by automated algorithms. Interrater reliability between skilled segmenters has been shown to be high (Fiez, Damasio, & Grabowski, 2000), and we have collected pilot data demonstrating a similar degree of reliability between a skilled rater and specially trained research assistants.

Tyler et al. (2005) have used a markedly different approach, looking for relationships between the intensity values from structural MRI images and behavioral measures. This approach avoids the need to delineate lesions at all, but requires images that have common signal qualities (i.e., were collected on the same scanner) in order to ensure that damage-related signal differences are large compared to other sources of interscan signal variability. Because the brain images themselves are coregistered, the coregistration must be automated or semiautomated. A chief advantage of this approach is that it will be sensitive to patterns of structural damage that might be difficult to detect by eye, or that might not have sharply delineated boundaries. At the same time, it may be more vulnerable to artifactual findings due to misregistration.

Stamatakis and Tyler (2005) have described a similar approach just for segmentation, which involves first registering the lesioned brain to a template (using SPM normalization with the regularization constraint), and then using SPM to compare the signal intensity to a reference group of healthy subjects. Again, this approach depends on reliable registration of the lesioned brains and on having a set of reference images with the same signal qualities.

Finally, we note that voxel-based morphometry (VBM; Ashburner & Friston, 2000) has been widely used in assessing subtle structural changes—differences in the balance of gray and white matter density within a voxel—associated with neurodegenerative and psychiatric conditions. VBM is not designed to identify discrete focal brain lesions, but conceivably could be useful in identifying these less subtle structural differences. Stamatakis and Tyler (2005), in describing their GLM-based method, note that gray–white segmentation (upon which VBM depends) can fail easily in the presence of large structural lesions. They do, however, note that VBM may be useful in the case of small lesions or atrophy, where gray–white segmentation is more likely to succeed. Mehta, Grabowski, Trivedi, and Damasio (2003), in explicitly evaluating the utility of VBM in segmenting lesions, found that VBM was markedly inferior to manual segmentation. Rorden and Karnath (2004) have also noted the shortcomings of VBM in the analysis of focal lesions.

It is worth noting that for both registration and segmentation, some of the methods are conceivably useful in combining data from widely varied sources (e.g., MRI-based and CT-based lesion maps), whereas others may require more closely matched images from all subjects. Even among the automated registration methods, some but not others are appropriate for intermodality registration (e.g., registering a subject's CT image to a study's MRI template might work well with mutual information-based registration, but not with an approach based on matching raw signal intensity). Notwithstanding concerns about systematic differences between CT and MRI, the flexibility to combine data from different modalities is generally important for lesion studies, which typically include patients who cannot undergo MRI.

CORRECTION FOR MULTIPLE COMPARISONS

Without correcting for multiple comparisons, the probability of making a Type I error goes up as we carry out more statistical tests. This can be a significant concern in voxel-based studies, which typically involve tens of thousands of voxels. The prevailing standard (across many scientific disciplines) is to control ''family-wise error rate'' (FWER), or the probability of making one or more Type I errors among the entire set of tests. Achieving this control normally entails accepting an increased risk of Type II error.

Bonferroni correction, the most common procedure used for FWER control, is overly conservative when the comparisons are not independent. That is, it accepts higher Type II error than is necessary to guarantee at least the desired FWER. VLSM is liable to be a particularly bad case for Bonferroni correction because of the inherent spatial coherence of lesion maps. At typical resolutions, lesions tend to be formed of contiguous voxels, and the lesion status of a voxel is well predicted by that of its neighboring voxels. This lack of spatial independence may be especially exaggerated to the extent the resolution of the lesion maps exceeds the resolution at which meaningful decisions about lesion tracing are made. Delineating three-dimensional lesions at 1 mm instead of 2 mm resolution incurs a correction for eight times the number of tests, regardless whether this additional resolution is used meaningfully.

In fMRI, alternatives to Bonferroni correction include Gaussian Random Field Theory (RFT; Worsley, Taylor,

Tomaiuolo, and Lerch, 2004), False Discovery Rate (FDR) control (Benjamini & Hochberg, 1995), and permutation testing (Nichols & Holmes, 2002). RFT takes advantage of the approximately Gaussian spatial structure of the signal to borrow an elegant solution from topology (Worsley, 1996). Although widely applicable, RFT solutions are most appropriate with high degrees of freedom, high spatial smoothness, and ideally require Gaussian smoothness over a large, regularly shaped region. These are overly restrictive conditions for typical lesion studies. With low *df* and/or spatial smoothness, the RFT threshold will be even more restrictive than Bonferroni correction, and therefore, not useful. With non-Gaussian spatial structure, the RFT solution will be invalid and will not offer appropriate control of the false-positive rate.

FDR has been recently advocated as an alternative standard for scientific reportability, and has been specifically proposed for lesion analysis (Rorden & Karnath, 2004). As its name implies, FDR provides a method for controlling the expected proportion of false positives among suprathreshold voxels (Genovese, Lazar, & Nichols, 2002). It provides improved power by relaxing control of false positives—even a typical FDR criterion of 0.01 (1 false discovery expected per 100 positive results) typically provides a more inclusive threshold than FWER of 0.05 (5% chance of any false positives at all). FDR may turn out to be valuable in lesion studies, and is already available in MRIcron, VLSM, and VoxBo (the package used for the present work), as well as SPM. However, because it does not control FWER (as do Bonferroni correction, the RFT-based method, and the permutation test described below), it should be compared to these other methods with caution.

Permutation testing is a nonparametric resampling approach to significance testing that provides elegant solutions for numerous vexing statistical problems (Good, 2004). Briefly, permutation testing is a procedure whereby a test statistic may be compared to a null distribution derived from the dataset of interest rather than from some parametric distribution. The permutation null distribution is typically derived by permuting how the dependent and independent variables are paired. When the null hypothesis is true, there is nothing special about the correct pairings, and the ''correct'' pairings should be no more likely to generate an extreme test statistic than any other. If the correct ordering does generate an extreme value relative to the other permutations, we may reject the null hypothesis. Although the details of a specific permutation test may vary, the key point is that the method allows us to derive a reference null distribution for just about any statistic of interest, not just those statistics that can be coerced into conformity with some parametric distribution.

Permutation tests rely on few assumptions, and asymptotically approach exact significance levels. The two most commonly cited disadvantages are the computational cost and the lack of widespread availability. As faster computers become available, computational cost becomes more and more irrelevant. And although permutation tests are still not widely available in general statistical packages, they are supported in several packages for fMRI analysis (see Appendix), and should become increasingly prevalent.

It is important to note that although many researchers are rightly uncomfortable with radical new statistical procedures, the permutation test is neither radical nor particularly new. In making this point, Nichols and Holmes write, ''Had R. A. Fisher and his peers had access to similar [computing] resources, it is possible that large areas of parametric statistics would have gone undeveloped.'' Similarly, Kempthorne (1955, cited by Good, 2004) wrote, ''Tests of significance in the randomized experiment have frequently been presented by way of normal law theory, whereas their validity stems from randomization theory.'' We can adopt the perspective that, for many purposes, parametric statistics are a compromise that we have been forced to live with solely due to the cost of computing. That cost has been dropping steadily for the past 50 years, and is no longer a meaningful impediment for most purposes.

In the present context, permutation testing provides an elegant solution to the multiple comparison problem that, in effect, corrects exactly for the number of independent comparisons in a volume, without making assumptions about the spatial structure of the data. This approach, first described for fMRI by Holmes, Blair, Watson, and Ford (1996), has been widely leveraged precisely because of the overly conservative nature of RFT-based and Bonferroni correction. In brief, if we generate 1000 VLSM maps from permutations of our data, we can examine the distribution of the *maximum* statistic across the brain volume. If our statistic is the *t* statistic,[3] then the 95% percentile maximum *t* statistic is the value of *t* that is exceeded somewhere in the brain in only 5% of the permutations. This is precisely the assurance we would expect from Bonferroni correction—when H_0 is true, we expect to observe so much as a single false positive across the brain only 5% of the time.

Permutation testing using the maximum statistic should have some particular advantages for VLSM. First, it corrects for multiple comparisons in a way that is sensitive to the independence of the observations. Adding a completely redundant voxel will never affect the maximum statistic in the volume. Similarly, to the extent that additional voxels are highly correlated with existing voxels, they will be unlikely to affect the maximum statistic. Resampling the lesion maps to extremely high resolution, which increases the number of comparisons but not the number of *independent* comparisons, would incur no additional penalty.

Second, permutation testing does not depend on any specific spatial structure of the lesion maps. Although visual inspection of the pattern of correlations among

voxels suggests a spatial structure that is smooth and may perhaps be at least coarsely Gaussian, the pattern is likely influenced by vascular patterns, and may be influenced in some studies by inclusion/exclusion criteria (e.g., in the case that behavioral exclusion criteria exclude patients with specific patterns of damage). This makes lesion data a problematic case for methods that assume a specific spatial structure a priori.

The use of the maximum statistic with the permutation test does not solve a separate problem that is apparent with fMRI data, and that may be of even greater concern in lesion data: nonuniform power across the volume. We may be able to improve the overall sensitivity of the permutation test using either step-down or step-up procedures (Holmes et al., 1996) to avoid having small effects in one region masked by large effects elsewhere.

We evaluated the permutation test alongside four other methods for thresholding VLSM maps: counting distinct patient–lesion patterns, Bonferroni correction, RFT, and FDR. A ''patient–lesion pattern'' as used here simply means the list of which patients are lesioned versus intact in a given voxel. If 10 voxels are lesioned in exactly the same subset of patients, they contribute only a single distinct pattern. This measure provides a straightforward improvement over Bonferroni correction, as it takes into account observations that are not only highly correlated, but in fact, completely collinear. Counting distinct patient–lesion patterns does not count the number of independent observations in an information theoretic sense, as it ignores the likelihood of correlated (but not identical) observations, and may therefore still be more conservative than the permutation test. It is especially unhelpful when a continuous measures of lesion status is used, as neighboring voxels will have highly correlated but nonidentical lesion patterns, and the correction will be the same as Bonferroni correction. For these reasons, permutation testing with the maximum statistic (which provides strong control of the false positive rate, see Holmes et al., 1996) should still be preferable in general. However, with 0/1 lesion scores, counting only unique patient–lesion patterns removes a major source of nonindependent comparisons, making a Bonferroni correction less dramatically overconservative. When applicable, it provides an easily obtained improvement over strict Bonferroni correction in reducing unnecessary correction without the computational burden of permutation testing.

We evaluated these methods for three datasets: two small datasets of typical size for patient studies ($n = 12$ and 13) and a larger dataset ($n = 55$) lacking a meaningful behavioral measure. Patients in all three datasets had damage restricted to the left hemisphere. Lesions were segmented and coregistered using a manual procedure with MRIcro (Rorden & Brett, 2000). A T1-weighted MNI template image was first rotated (pitch only) into correspondence with the patients' scans as well as possible. An experienced researcher outlined the lesions on the rotated template, resulting in a map in which each voxel was labeled either 0 (intact) or 1 (lesioned). Finally, the lesion maps were rotated back into a canonical orientation, using nearest-neighbor interpolation to restrict the map values to 0 and 1. For most of the subjects, lesions were drawn on a $2 \times 2 \times 2$ mm template. For some that were originally drawn at higher resolution (Dataset 1), we first resampled the lesions to $2 \times 2 \times 2$ mm just for the purposes of this test.

All of the reported Bonferroni-corrected and RFT thresholds for the three datasets correspond to an alpha criterion of .05. Note that the RFT thresholds were carried out without regard for the likely violation of assumptions. In particular, we used the method of Kiebel, Poline, Friston, Holmes, and Worsley (1999) to estimate smoothness for the region enclosing all the lesioned voxels, even though the smoothness is not necessarily Gaussian. The FDR thresholds reported in Table 1 were calculated with $q = 0.01$ (the expected false discovery rate) and $c(V) = 1$ (see Genovese et al., 2002).

Dataset 1: 12 Patients with Left Hemisphere Lesions

Schnur, Lee, Coslett, Schwartz, and Thompson-Schill (2005) segmented the lesions of 12 aphasic patients

Table 1. All t Thresholds Are Selected to Meet a Criterion of $p \leq .05$

		Bonferroni Correction		*Distinct Patient–Lesion Patterns*		*Permutation Test*	*RFT*	*FDR*
	Patients	*Voxels*	*t Threshold*	*Patterns*	*t Threshold*	*t Threshold*	*t Threshold*	*t Threshold for q = 0.01*
Dataset 1	12	62,990	10.0	680	5.92	5.4	11.07	3.14
Dataset 2	13	51,410	8.53	295	4.94	3.17	8.95	3.49
Dataset 3	55	96,210	5.51	17,258	5.03	4.35	5.56	3.25

For the permutation test, the standard error on the p value is .0069.

(from Schnur, Schwartz, Brecher, & Hodgson, 2006) willing to undergo MRI or CT. In aggregate, lesions from the 12 subjects covered 62,990 left hemisphere voxels. Bonferroni correction based on this count would yield a significance threshold of $t > 10.0$. However, only 680 patient–lesion patterns were observed, which with Bonferroni correction yields a more reasonable threshold of $t > 5.92$. Our permutation test, using the maximum statistic to control for multiple comparisons, yielded a threshold of $t > 5.4$, which is slightly more inclusive than counting voxel patterns.

Dataset 2: 13 Patients with Left Hemisphere Lesions

We performed the same comparisons in a separate group of 13 patients with left hemisphere lesions, in connection with a study examining action representations. The lesions of patients in this study included 51,410 voxels. The Bonferroni-corrected threshold would be $t > 8.53$. The number of patient–lesion patterns was 295, which with Bonferroni correction yields a threshold of $t > 4.94$. By contrast, the permutation test yields a threshold of $t > 3.17$.

Dataset 3: 55 Patients with Left Hemisphere Lesions

Finally, we evaluated the permutation test using a larger group of 55 patients (the 12 patients from Dataset 1 and an additional 43 patients) for which no experimental data of interest were available. This is the typical situation for study planning. In these subjects, a total of 96,210 voxels were implicated, in which were represented 17,258 patient–lesion patterns. Bonferroni correction on the total number of voxels would give a threshold of $t > 5.51$. Note that despite the much larger number of voxels than in the previous dataset, the much larger *df* gives us a more lenient threshold. Using the distinct voxel count reduces the threshold to $t > 5.03$.

We created normally distributed random experimental data to evaluate the permutation test, which yielded a threshold of $t > 4.35$. As an approximation of the degree of spatial dependence in the data, we note that this threshold would correspond to roughly 1607 comparisons under Bonferroni correction for an alpha of .05 (we use this estimate below as an approximation of the number of independent comparisons in the dataset).

Table 1 summarizes the results from the three datasets, including the RFT and FDR thresholds. Clearly, the degree of spatial coherence of lesion maps makes Bonferroni correction on the basis of the number of voxels grossly overconservative, no less so for lesion analysis than for fMRI. Counting distinct voxel patterns is a dramatic improvement in cases with discrete-valued damage measures. The RFT-derived thresholds fail to improve on Bonferroni correction, although they would likely do so if the maps had been spatially smoothed. Permutation testing provides a more general solution, and should provide an additional advantage over counting in cases where lesion patterns are less randomly distributed, or where continuous valued lesion scores are used. Finally, FDR control provides a more liberal test in two of the three datasets, and should be generally more powerful when strict control of FWER is not needed.

Note that because we do not sample the permutations exhaustively, the permutation p value is subject to sampling error. The standard error on the p value of .05, due to sampling error, is .0069. The risk of being overly liberal can be reduced by running more permutations than the 1000 used here, and/or by using a lower alpha criterion. For small datasets, it may sometimes be possible to get an exact p value by running the permutations exhaustively.

POWER

Power, in the hypothesis testing framework, is the probability of rejecting a false null hypothesis (i.e., the probability of obtaining a significant result for a true effect). Underpowered studies (by common consensus, acceptable power is 0.8 or greater) are a poor investment of resources—in a sense, the researcher is hoping to get lucky. Adequately powered studies not only stand a reasonable chance of rejecting the null hypothesis but also lend more weight to null findings.

Power is of special importance in the context of patient studies because it suffers in studies with small groups and noisy measurements. Patients are a scarce and closely protected resource, and the rate of recruitment is typically a major concern in planning patient-based projects. Patient performance may be highly variable, both within- and between-subjects. Lesion size and location, as well as baseline demographics, may be highly variable as well. Problems with excessive variability are sometimes exacerbated by overly inclusive recruiting to develop adequate sized groups. In the context of VLSM analyses that incur severe corrections for multiple comparisons, understanding power is especially important.

At the same time, VLSM studies are in a better position for power analysis than fMRI because the measures being compared often have real-world meaning (unlike BOLD signal values). For example, we may decide that adequate power in a voxel means detecting lesion-associated differences in digit span of 0.25 digits or more, where it would be difficult to make a comparable decision regarding BOLD signal units. Null results are easier to interpret because we can place confidence intervals around effect sizes measured in meaningful units, instead of simply reporting the association as nonsignificant.

We here consider power for two types of tests: univariate tests of the relationship between lesion status

and some behavioral measure carried out separately for each voxel; and tests meant to discriminate between voxels or locations. As noted earlier, we can refer to these as supporting ''spatially localized'' versus ''spatially discriminating'' inferences. Although we consider the two separately below, it is helpful to bear in mind that they are closely interconnected. Power for localized inference is a prerequisite to power for spatially discriminating—the less we can learn about a given voxel, the less we can learn about how it differs from other voxels. Although many studies are reported in such a way to support localized but not discriminating inferences, understanding cortical organization ultimately requires both.

Spatially Localized Inference: Power to Detect Lesion–Behavior Correlations in the Entire Brain

The power to detect a true relationship between damage in some region and some behavioral pattern of interest varies as a function of the proportion of patients with damage in that region. When no patients are lesioned in a given voxel, we can learn nothing about the correlation between damage there and some behavior of interest (although we may learn that a lesion there is not necessary to produce a deficit of interest). In principle, the same is true when all patients in a group are lesioned in a given voxel, although we may sometimes make implicit comparisons to a control group of healthy patients (no lesions anywhere) or some other reference population.

For a fixed-size patient group, without knowledge of the behavioral scores, power is maximized when the variance in the lesion score is maximized. For 0/1 lesion scores, this occurs when a voxel is lesioned in half the population. This ideal proportion may be rarely met in typical lesion studies. In our sample of 55 patients with left hemisphere lesions, the maximum number of patients lesioned in any voxel was 25 (45% of the group), with a roughly normal distribution centered around 10–11 (18–20% of the group).

It is worth noting that nonuniform sensitivity is a problem for fMRI as well, albeit one often overlooked. Trivially, static susceptibility artifacts make specific regions especially difficult to image with BOLD contrast. Typical models of hemodynamic responses to neural activity may provide better fits in some regions than others. Local neural architecture and proximity to major blood vessels, combined with spatial smoothing, may also contribute.

With lesion studies, however, we can carry out a first approximation analysis of the regional distribution of power. To do so, we need the following: an estimate of the behavioral effect size and its variance; estimates of the number of lesioned versus intact patients in each voxel; and an estimate of the alpha criterion appropriately corrected for multiple comparisons. In the present example, all but the last item can be estimated from previously collected data. Behavioral effect sizes and variability are drawn from Schwartz et al. (2006), who reported quantitative estimates of lexical–semantic word production individually for a group of patients roughly comparable to the 55 described earlier. The balance of lesioned versus intact subjects in each voxel is estimated from the scans for these 55 patients.

To correct our alpha criterion appropriately, we need an estimate of the number of independent comparisons. To do this, we carried out a permutation test using the 55 lesion volumes and a fabricated behavioral score of normally distributed random values. From the distribution of the maximum statistic, we took the 95% percentile maximum t score (4.35, $df = 53$) and identified the number of independent comparisons for which that t score would yield a Bonferroni-corrected p value of .05. This was used as a coarse estimate of the number of independent comparisons. In our sample dataset, this produced an estimate of 1607.

Using these data, as well as the DSTPLAN software for power analysis (see Appendix), we can map power to detect the effect size derived from Schnur et al. (2006) across the brain. Only the lesioned/intact balance varies on a voxel-by-voxel basis—given that no more than 25 patients were lesioned in any given voxels, we can identify 25 strata of power. We constructed color power maps, as in Figure 1 below.

The maps show that in this dataset, there is adequate power to detect effects in the behavioral measure across most of the left peri- and extra-Sylvian cortex, except for the inferior temporal gyrus (ITG), in which too few of the patients had lesions. This distribution of power is clearly important to understand both in planning the study and in interpreting its results. If a region of low power (such as the ITG in this example) is important to the study, then remedial measures might include designing a more sensitive behavioral measure, recruiting a larger or different patient sample, or restricting our hypotheses more severely (to reduce the correction for multiple comparisons), for instance, by using region-of-interest (ROI) analyses. Power analysis for ROI analyses would generally be more straightforward, considering the small number of regions involved and the reduced impact of interregion correlation. If the study is carried out without changes and yields null effects in the ITG, the power analysis helps us conclude that the study contributes very little to our knowledge of the effects of ITG damage on behavior, due to lack of power in that region.

Assigning each voxel an expected power value is just one type of power calculation. We might also decide to fix power at 0.8 and map the size of the detectable effect. Or we might fix power, effect size, and the proportion of patients in the two groups, and map the required total group size as a function of the proportion of patients lesioned in each voxel. The DSTPLAN package used for the power calculations here is flexible both in

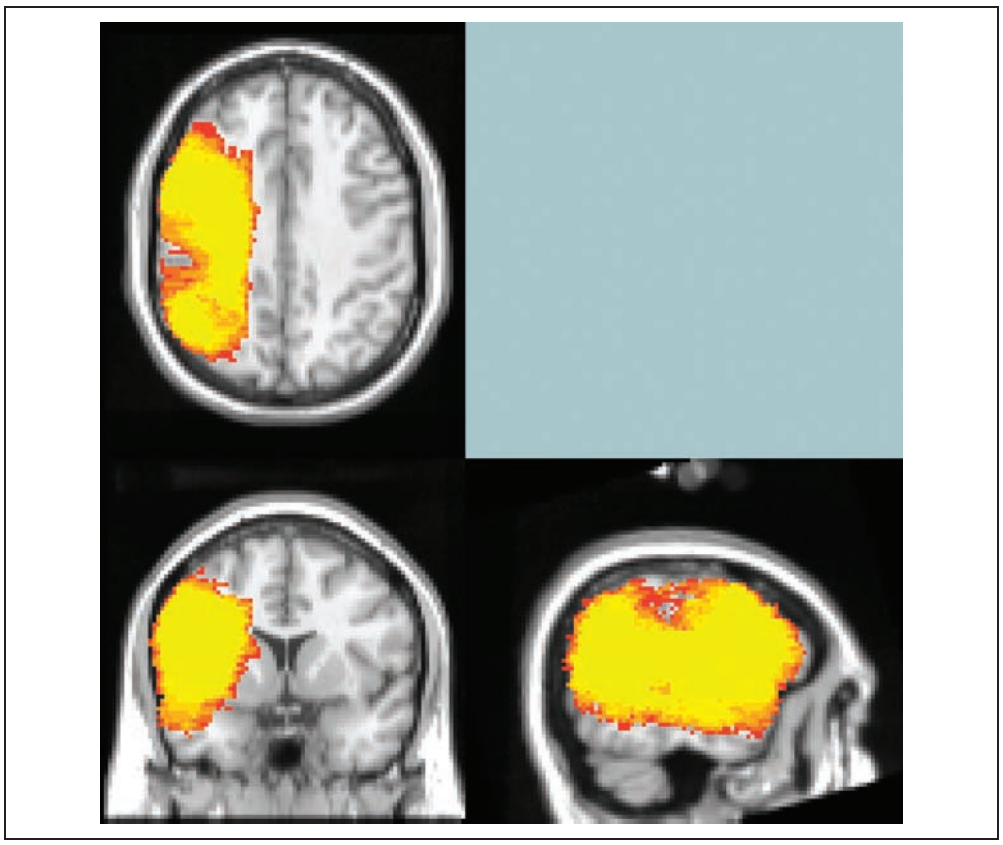

Figure 1. Power mapped on selected slices in three orientations. All plotted voxels have power >0.4. Voxels in red are barely above 0.4, whereas voxels in yellow are 0.8 or above.

providing for a variety of statistical tests (not just two-sample *t* tests), and in allowing the user to choose which quantities are fixed, solving for a single unknown.

Spatially Discriminating Inference: Power for Interregion Comparisons

It will be difficult to resolve functional differences on a millimeter scale with much power if patients' lesions are typically centimeters across (barring unreasonably large patient groups). Even in the presence of good regional power, the power to detect differences between two regions may be low when the two regions are positively correlated. If two regions or voxels are always lesioned together, then it would be impossible to learn from VLSM that only one is causally related to some deficit of interest. This kind of spatial coherence in lesion data arises both from the spatial extent of lesions (effectively defining the resolution of lesion data) and from inter-region correlations due to the physical processes that produce the damage, especially including vascular organization in stroke.

It may be helpful to think of this problem in a multiple linear regression framework, in which we posit that the behavioral measure may be explained by some linear combination of lesion scores from two regions. When the lesion scores are independent, we get independent measures of the contribution of each, and we can specifically contrast the two to see which has a greater impact on behavior (given meaningfully scaled covariates). When the lesion scores are positively correlated, differences between the two parameter estimates will reflect the differences between the two, although the greater the correlation, the less sensitive the design will be to genuine differences between the regions. At the extreme, when the two covariates are perfectly correlated, the model will be unsolvable, which sensibly reflects the fact that we can learn nothing independently about the two voxels.

When the lesion scores in two voxels are negatively correlated, the voxels may provide completely or partially redundant information about differences between the regions. At the extreme, perfect inverse correlations (patients are lesioned in one region or the other but never both) can arise trivially when patients are recruited with damage to one of two structures but not to both. They can also arise systematically when large lesions are either intentionally excluded or are unusual in the patient group. For example, in a group of patients selected for having single focal lesions of limited size, frontal and occipital lesions will be negatively correlated, and lesion–behavior correlations in the two regions will be non-independent. This will be common in stroke studies that are not restricted to a single vascular territory.

With negatively correlated lesions, we do not have independent measures of the effect of damage in one region and the effect of damage in the other region, and in principle, can say little about differences between the two. That is, if damage is positively correlated with our

behavioral score in Region A, we cannot, in principle, differentiate between damage in Region A causing an increase in scores and damage in Region B causing a decrease in scores. In practice, when scores are clearly valenced and/or we can identify ''normal'' scores, we can interpret these differences more sensibly. In such cases, we implicitly decorrelate the two by reference to an imagined reference population.

For the remainder of this section, we consider the common situation of mostly positive correlations. In this case, even if we have good power within individual voxels, the power to detect differences between voxels or regions depends not only on our ability to estimate some effect efficiently in both regions but also on the ability to do so independently. Knowing the expected relationship between damage in different regions is therefore a prerequisite to carrying out a study to compare those regions. This kind of power analysis can be carried out using either the patient group of interest or a comparable group. In some cases, it may be possible to recruit patients selectively to maximize power.

A complete map of interregion correlations would be difficult to represent graphically in limited space (it would require one full brain map for each voxel's correlations). However, we can identify a ''seed'' voxel in an ROI, and plot the correlation with all other voxels in the brain. This kind of map provides a sense of what kinds of differences might be detected in the dataset. The precise correlation between two voxels can, of course, be queried from this map.

The maps below present correlation maps for two seed voxels selected from the group of 55 patients described earlier. These maps are preliminary to considering the power of a specific test, which must consider the variability of the behavioral score as well.

The apparent smoothness of the lesion maps suggests that this type of analysis is not necessarily vulnerable to a poorly selected seed voxel, although, clearly, care must be taken to ensure that this is the case in a given dataset (Figure 2).

These interregion correlations prove useful in planning studies of lexical access. Although we may have theoretical reasons to be interested in the differences between the anterior insula and the posterior inferior frontal gyrus (pIFG), lesions in typical voxels in these regions are highly correlated ($r = .7$). By contrast, damage in the pIFG is almost perfectly uncorrelated with damage in the pITG ($r = .03$).

These seed voxel maps provide one view of the degree of spatial coherence in the lesion maps. Another way to consider spatial resolution is in terms of the spatial smoothness of the data, which reflects how well a voxel's score is predicted by those of its neighbors. We used the method described by Kiebel et al. (1999) to estimate the Gaussian smoothness of this set of lesion maps. Although the lesion data are certainly not Gaussian in general, and the smoothness is not uniform, this may be a useful informal measure for estimating the degree of spatial dependence in lesion data. In the current dataset, we estimated a smoothness of 8.84 mm, 7.69 mm, and 10.59 mm full width at half maximum for the three datasets. These estimates are averaged over the three dimensions. In our test datasets, despite isotropic ($2 \times 2 \times 2$) voxels, smoothness was least in the z dimension, presumably because the drawings were done one z slice at a time.

DISCUSSION

Lesion analysis is important to the enterprise of brain mapping, and is well positioned to take advantage of (and build on) methods developed for fMRI. These voxel-based methods have the potential to support both spatially localized inferences (concerning a specific voxel or region) and spatially discriminating inferences (concerning differences between voxels or regions). These two modes are often confused in functional imaging,

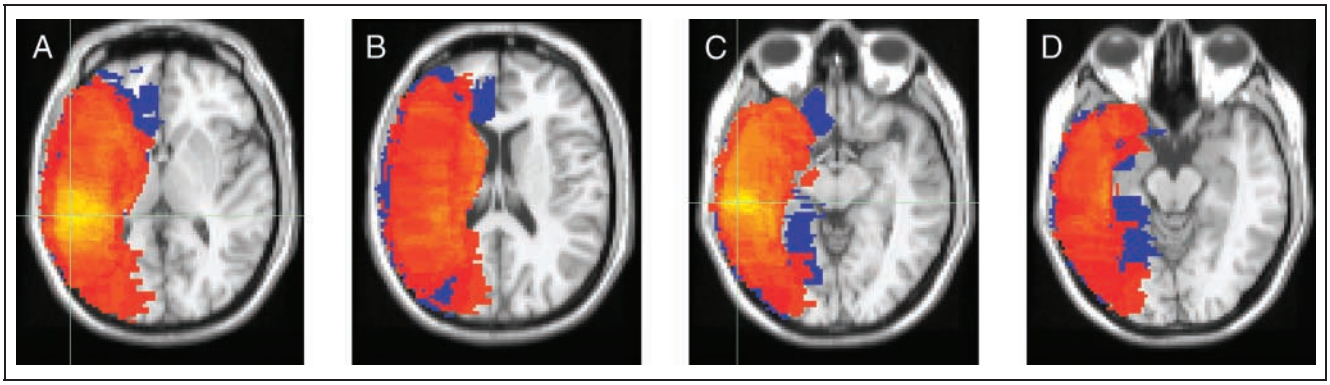

Figure 2. Correlations with seed voxels in BA 21 and BA 22. All voxels with lesion data are depicted. (A) Area 21, including the seed voxel at the crosshairs. (B) Correlations with the area 21 seed voxel in a slice 14 mm superior. (C) Area 22, same slice as the seed voxel. (D) Correlations with the area 22 seed voxel in a slice 9 mm inferior. All red/orange/yellow voxels have correlations >0, with yellow closer to 1 and red closer to 0. Blue voxels have small negative correlations with the seed voxel.

where it is tempting to assume that significant activity in one region (a spatially localized inference) and not another implies some difference between the two (a spatially discriminating inference). This kind of misinterpretation is highly problematic in fMRI, where it is typical to conduct thousands of spatially localized tests and no spatially discriminating ones. The problem is liable to be more apparent in VLSM, where the differences in power between regions are even more pronounced. Even for studies reporting only localized inferences, careful reporting requires some characterization of regional power. This may be somewhat easier to do, as a first pass, in lesion studies than in fMRI, because the dependent measures are often behavioral scores that have meaningful magnitudes.

In this article, we have described an approach for calculating regional power to detect localized effects of brain lesions, as well as a preliminary step toward mapping power to make spatially discriminating inferences. Although different studies require different statistical tests, and therefore, different power calculations, we hope this provides a general framework for characterizing regional differences in power.

We have also described the application of permutation testing to VLSM to maximize power while maintaining control of the FWER. Again, depending on the structure of the study, the approach to permutation testing will vary. We have covered a simple and common case here, but of course, each study requires careful consideration of the proper test.

Decisions about the statistical model can also have a dramatic effect on regional power. For example, it may often be important to covary for lesion size in order to consider separately the specific effects of damage to a particular voxel and the more general effects of the amount of damage (see Rorden & Karnath, 2004). This potential confound is especially important given the likelihood that, in typical studies, patients with larger lesions are more likely to have damage in a given voxel. Covarying for lesion size makes it possible to look for effects of damage in a particular voxel beyond what can be attributed to overall lesion size. However, doing so can dramatically reduce power in voxels in which damage is highly correlated with lesion size. Removing the effects of lesion size effectively reduces the variance in the voxel's lesion status that could be useful in predicting behavioral outcome.

Similarly, it may sometimes be useful in the context of specific hypotheses to examine the predictive value of a given voxel or region above and beyond that of another voxel or region, in which case the latter could be included as a covariate in the model, an approach described by Bates et al. (2003). Or we might model the interaction between regions in predicting behavior. These models can only be tested effectively when the patient–lesion patterns in the two voxels are reasonably orthogonal. Models that involve nuisance covariates can complicate the permutation test described earlier. A viable approach in this case, covered by Good (2004), is to regress out the effects of these covariates, and then to perform the permutation test on the residuals.

Quality of Lesion Data

We have largely sidestepped the issue of varying quality in lesion maps until now, although it is an important issue in VLSM. Variation in quality can occur readily when mixing modalities, typically CT and MRI, when mixing imaging protocols within a modality, and as a function of the type of injury. Mixing modalities is never ideal but is often necessary due to the scarcity of patients. Transcribing lesions by hand often forces the rater into a 0/1 decision on each voxel, effectively throwing away information about the reliability of the evidence in each voxel. But a more sensitive approach would take into account the weight of evidence for damage in a particular location, mapping real values onto each voxel. We can easily imagine a continuous scale for confidence that a voxel is damaged, ranging from 0 (no lesion) to 0.5 (no evidence either way) to 1 (certain lesion). Although it would be impractical to have raters produce these confidence ratings for each voxel, we could conceivably combine the 0/1 maps with information about the imaging modality and the registration and segmentation techniques to produce continuous valued maps, in standard space, of the probability that each voxel is lesioned. For example, a voxel marked intact, surrounded by other intact voxels, in a region that is well-imaged by the method used, and that is subject to minimal registration error, could be assigned a lesion score of close to 0, reflecting high confidence in its intactness. A voxel marked intact that is close to a lesion boundary and in a region that is highly subject to registration error might be given a score closer to 0.5. This approach would be potentially more sensitive than using discrete lesion scores, although it would require more information about each scan, including its relative sensitivity to the types of damage of interest in each region, as well as the magnitude of expected registration error (as a function of registration method and region), which may also vary spatially.

In the discussion of power earlier, we focused on the number of patients lesioned versus intact in a given voxel as a critical variable in assessing power on a voxel-by-voxel basis. The balance is really a surrogate for variance in lesion score, which is maximized for discrete lesion scores when exactly half the patients are lesioned in a given voxel. A slightly different approach to power would be required for continuous-valued lesion scores. When the image quality is systematically poorer in some brain regions than others, variance in lesion status will be correspondingly lower in those regions, resulting in lower power. When imaging modalities are mixed (e.g., in the case of MRI and CT), however, power calculations that presume a normally distributed lesion score may be inappropriate.

Finally, it is worth noting that the use of a continuous lesion score means there are no discrete groups, and a simple two-sample *t* test cannot be performed. Regression models provide a more general solution that is widely available in imaging packages and equally amenable to power analysis.

Inferential Problems Not Addressed by VLSM

Rorden and Karnath (2004) review a number of general limitations of lesion analysis that are not specific to VLSM, including the assumption of modularity, variability in functional organization, the nonrandom spatial distribution of lesions, and the potential disconnect between structural and functional intactness. Voxel-based methods do not offer an immediate solution to any of these problems, although it is possible that methods drawing on voxel-based representations of brain lesions will be more amenable to testing and validating solutions.

In this section, we discuss some additional inferential problems that are as yet poorly addressed by VLSM (analogous to similar problems with the SPM approach to fMRI), some of which are better addressed by traditional methods for lesion analysis.

The massively univariate SPM approach advocated here can be a weak tool when variability in location among lesions responsible for a particular deficit is large compared to the resolution of the lesion data. This variability can be due to interindividual variation in structure, idiosyncratic functional organization, and registration error. For example, consider two patient groups, with and without some behavioral deficit of interest, all with relatively small lesions. If the lesions of the impaired patients are tightly clustered in a small region, but not generally overlapping, whereas the lesions of the intact patients are widely distributed elsewhere in the brain, we would likely be justified in drawing an inference about that region. However, VLSM would be a poor choice for discovering this regularity because it considers each voxel independently. A potential solution is to use spatial smoothing to reduce the impact of anatomic variability, although smoothing can wash out highly localized effects.

Although voxel-based methods are vulnerable to registration errors, VLSM works in part because the spatial extent of lesions tends to overwhelm anatomical variability. However, it is more likely to fail when variability in functional organization reduces the likelihood that patients with the same functional deficit have nearby lesions, even within the same structure. Multivariate techniques may be more appropriate in this case. However, region-based analyses, such as stratifying patients on the basis of whether the lesion impinges some structure, may capture regularities that would not be readily apparent in single voxel analyses. This is a form of data reduction that takes into account information about structural boundaries that is not well captured by blind averaging strategies such as Gaussian smoothing.

Note that although variability in functional organization may be a problem for VLSM as currently implemented, multivariate techniques that have already been applied to fMRI data may overcome some of these limitations. Such techniques include machine learning techniques such as Support Vector Machines (Mourao-Miranda, Bokde, Born, Hampel, & Stetter, 2005) and MVPA (Norman, Polyn, Detre, & Haxby, 2006), Canonical Correlation Analysis (Nandy & Cordes, 2003), and Partial Least Squares (McIntosh & Lobaugh, 2004). These techniques have the potential to discover regularities that would be more difficult to find with univariate methods, including discovery of voxels that predict behavior better collectively than individually.

At the same time, in a highly interactive brain, the boundaries of a lesion likely understate the extent of the abnormally functioning cortex in brain injury. Regions that appear structurally intact may be effectively deafferented, or disconnected from the input needed in a given context. This adds noise to the analysis in that it creates variance in the location and extent of lesions that may affect a given function.

Lastly, we note that VLSM does not guarantee that the association between damage and behavior will be meaningful or easy to interpret. In addition to problems in formulating meaningful cognitive subtractions, and in disentangling the roles of interacting brain regions, studies comparing different tasks in brain-injured populations are vulnerable to artifactual differential deficits (Strauss, 2001; Chapman & Chapman, 1978)—whereas it may be easy to demonstrate that patients with a lesion in a particular location are more likely to be impaired at a task than patients without such a lesion, additional work is required to establish that the task is differentially impaired and that the impairment is related to a specific cognitive process.

Reporting Power for VLSM Studies

As with fMRI, it is easy to overinterpret null results in VLSM studies—if there is a suprathreshold blob in Region A and not in Region B, it is tempting to assume that there is no effect in Region B and that there is a difference between the two regions. Within the hypothesis testing framework, power calculations are critical to meaningful discussion of null findings. It would be even better, in many cases, to develop a graphical representation of the confidence intervals surrounding the effects at voxels or ROIs, to carry out lesion analyses in an estimation rather than a hypothesis testing mode.

Because the distribution of power across the volume is of particular interest in VLSM (and not just the power to detect a relationship anywhere in the brain), reporting power can be a graphical challenge. The color map presented earlier divides the brain into three strata of interest (below 0.4, between 0.4 and 0.8, and above 0.8), and represents our first attempt to present power for lesion data in an immediately useful way for study planning.

Conclusion

We have tried to provide some critically needed groundwork for power analysis and improved statistical methods in VLSM. Although more sophisticated approaches, tailored to the needs of a given study, are certainly possible, we believe that addressing power—and thereby, the limits of discoverability via the lesion method—is a fundamental step upon which improved methods for voxel-based lesion–behavior analysis can be built. We further argue that permutation testing, already widely used in fMRI, can be critically valuable for VLSM analyses.

APPENDIX: SOFTWARE

Most of the software for the analyses described in this article is available as part of the VoxBo package (www.voxbo.org) VoxBo is a self-contained package for image analysis with a focus on fMRI, but which is actively being extended to include lesion analyses. VoxBo is released in both source code and binary form under the GNU General Public License, and runs under Linux, OSX, and Windows (via Cygwin).

The DSTPLAN software used for power (sample size) calculations is available for download (as ''STPLAN'') with accompanying article, from:

http://biostatistics.mdanderson.org/SoftwareDownload/

The VLSM toolkit (version 1.6, as of this writing) is available from:

http://crl.ucsd.edu/vlsm/

VLSM is implemented in MATLAB and includes functionality that depends on the statistics and image processing toolboxes. The package includes templates for display of lesion maps. VLSM is released under the GNU General Public License, and may be freely redistributed or modified under the terms of that license.

MRIcro is widely used for visualization and lesion tracing (Rorden & Brett, 2000), and was used to trace lesions for the three sample datasets described in this article. MRIcro is made freely available in binary form for Windows, Linux, and Solaris. MRIcro is in the process of being supplanted by MRIcron, and both are available via:

www.mricro.com/

SPM is the most widely used package for fMRI analysis, and is the foundation for a large number of toolboxes and extensions. It is available from:

www.fil.ion.ucl.ac.uk/spm/

Permutation testing for image data is available in VoxBo, as well as in SPM (via the SnPM toolbox, see Nichols & Holmes, 2002) and FSL (Smith et al., 2004). FDR thresholding is also available in each of these packages.

Acknowledgments

This work was supported by the Human Brain Project via R01MH073529 and R01DA014418, by the Neuro-cognitive Rehabilitation Research Network (ncrrn.org) via R24HD050836, P30NS045839, and R01DC000191. We also gratefully acknowledge the contributions of Tatiana Schnur, Esther Lee, Laura Barde, Laurel Buxbaum, and Katie Kyle in lesion delineation and analysis, and Cris Hamilton and Olu Faseyitan, as well as two anonymous reviewers for constructive comments on earlier drafts.

Reprint requests should be sent to Daniel Y. Kimberg, Department of Neurology, 3W Gates, Hospital of the University of Pennsylvania, 3400 Spruce Street, Philadelphia, PA 19104, or via e-mail: kimberg@mail.med.upenn.edu; Web: www.voxbo.org, http://cfn.upenn.edu.

Notes

1. Although VLSM is also a software package (http://crl.ucsd.edu/vlsm/), we use the term more generically here to refer to methods for mapping the relationship between behavior and injury that depend on voxel-level representations. This may be a more inclusive definition than originally intended. VLSM is essentially a form of Statistical Parametric Mapping (SPM), another technique that shares its name with the software package that embodies it.
2. Note that throughout this article we use the term ''lesion'' generically to refer to structural characteristics that might be related meaningfully to behavior. The analysis approaches described here would, in many cases, be appropriate or even better suited for nonfocal brain injuries, or perhaps for understanding the correlates of normal anatomic variation.
3. In practice, the t statistic is often used in permutation tests, but its significance is tested nonparametrically, not against the t distribution.

REFERENCES

Ashburner, J., & Friston, K. J. (1999). Nonlinear spatial normalization using basis functions. *Human Brain Mapping, 7,* 254–266.

Ashburner, J., & Friston, K. J. (2000). Voxel-based morphometry—The methods. *Neuroimage, 11,* 805–821.

Avants, B. B., Schoenemann, P. T., & Gee, J. C. (2005). Lagrangian frame diffeomorphic image registration: Morphometric comparison of human and chimpanzee cortex. *Medical Image Analysis, 10,* 397–412.

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., et al. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience, 6,* 448–450.

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.

Binder, J. R., Frost, J. A., Hammeke, T. A., Bellgowan, P. S., Rao, S. M., & Cox, R. W. (1999). Conceptual processing during the conscious resting state. A functional MRI study. *Journal of Cognitive Neuroscience, 11,* 80–95.

Brett, M., Johnsrude, I. S., & Owen, A. M. (2002). The problem of functional localization in the human brain. *Nature Reviews Neuroscience, 3,* 243–249.

Brett, M., Leff, A. P., Rorden, C., & Ashburner, J. (2001). Spatial normalization of brain images with focal lesions using cost function masking. *Neuroimage, 14,* 486–500.
Chapman, L. J., & Chapman, J. P. (1978). The measurement of differential deficit. *Journal of Psychiatric Research, 14,* 303–311.
Chatterjee, A. (2005). A madness to the methods in cognitive neuroscience? *Journal of Cognitive Neuroscience, 17,* 847–849.
Farah, M. J. (1994). Neuropsychological inference with an interactive brain: A critique of the locality assumption. *Behavioral and Brain Sciences, 17,* 43–104.
Fellows, L. K., Heberlein, A. S., Morales, D. A., Shivde, G., Waller, S., & Wu, D. H. (2005). Method matters: An empirical study of impact in cognitive neuroscience. *Journal of Cognitive Neuroscience, 17,* 850–858.
Fiez, J. A., Damasio, H., & Grabowski, T. J. (2000). Lesion segmentation and manual warping to a reference brain: Intra- and interobserver reliability. *Human Brain Mapping, 9,* 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,* 870–878.
Good, P. (2004). *Permutation, parametric, and bootstrap tests of hypotheses* (3rd ed.). New York: Springer.
Holmes, A. P., Blair, R. C., Watson, J. D., & Ford, I. (1996). Nonparametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism, 16,* 7–22.
Kiebel, S. J., Poline, J. B., Friston, K. J., Holmes, A. P., & Worsley, K. J. (1999). Robust smoothness estimation in statistical parametric maps using standardized residuals from the general linear model. *Neuroimage, 10,* 756–766.
McIntosh, A. R., & Lobaugh, N. J. (2004). Partial least squares analysis of neuroimaging data: Applications and advances. *Neuroimage, 23(Suppl 1),* S250–S263.
Mehta, S., Grabowski, T. J., Trivedi, Y., & Damasio, H. (2003). Evaluation of voxel-based morphometry for focal lesion detection in individuals. *Neuroimage, 20,* 1438–1454.
Mourao-Miranda, J., Bokde, A. L., Born, C., Hampel, H., & Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *Neuroimage, 28,* 980–995.
Nandy, R. R., & Cordes, D. (2003). Novel nonparametric approach to canonical correlation analysis with applications to low CNR functional MRI data. *Magnetic Resonance in Medicine, 50,* 354–365.
Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping, 15,* 1–25.
Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences, 10,* 424–430.
Rorden, C., & Brett, M. (2000). Stereotaxic display of brain lesions. *Behavioural Neurology, 12,* 191–200.
Rorden, C., & Karnath, H. O. (2004). Using human brain lesions to infer function: A relic from a past era in the fMRI age? *Nature Reviews Neuroscience, 5,* 813–819.
Schnur, T. T., Lee, E., Coslett, H. B., Schwartz, M. F., & Thompson-Schill, S. L. (2005). When lexical selection gets tough, the LIFG gets going: A lesion analysis study of interference during word production. *Brain and Language, 95,* 12–13.
Schnur, T. T., Schwartz, M. F., Brecher, A., & Hodgson, C. (2006). Semantic interference during blocked-cyclic naming: Evidence from aphasia. *Journal of Memory and Language, 54,* 199–227.
Shallice, T. (2003). Functional imaging and neuropsychology findings: How can they be linked? *Neuroimage, 20(Suppl. 1),* S146–S154.
Shen, D., & Davatzikos, C. (2002). HAMMER: Hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging, 21,* 1421–1439.
Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., et al. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *Neuroimage, 23(Suppl. 1),* S208–S219.
Stamatakis, E. A., & Tyler, L. K. (2005). Identifying lesions on structural brain images—Validation of the method and application to neuropsychological patients. *Brain and Language, 94,* 167–177.
Stark, C. E., & Squire, L. R. (2001). When zero is not zero: The problem of ambiguous baseline conditions in fMRI. *Proceedings of the National Academy of Sciences, U.S.A., 98,* 12760–12766.
Strauss, M. E. (2001). Demonstrating specific cognitive deficits: A psychometric perspective. *Journal of Abnormal Psychology, 110,* 6–14.
Talairach, J., & Tournoux, P. (1988). *Co-planar stereotactic atlas of the human brain*. Stuttgart, Germany: Thieme.
Tyler, L. K., Marslen-Wilson, W., & Stamatakis, E. A. (2005). Dissociating neuro-cognitive component processes: Voxel-based correlational methodology. *Neuropsychologia, 43,* 771–778.
Woods, R. P., Grafton, S. T., Watson, J. D., Sicotte, N. L., & Mazziotta, J. C. (1998). Automated image registration: II. Intersubject validation of linear and nonlinear models. *Journal of Computer Assisted Tomography, 22,* 153–165.
Worsley, K. J. (1996). The geometry of random images. *Chance, 9,* 27–39.
Worsley, K. J., Taylor, J. E., Tomaiuolo, F., & Lerch, J. (2004). Unified univariate and multivariate random field theory. *Neuroimage, 23(Suppl. 1),* S189–S195.