Improving Lesion–Symptom Mapping

Chris Rorden[1], Hans-Otto Karnath[2], and Leonardo Bonilha[1]

Abstract

■ Measures of brain activation (e.g., changes in scalp electrical potentials) have become the most popular method for inferring brain function. However, examining brain disruption (e.g., examining behavior after brain injury) can complement activation studies. Activation techniques identify regions involved with a task, whereas disruption techniques are able to discover which regions are crucial for a task. Voxel-based lesion mapping can be used to determine relationships between behavioral measures and the location of brain injury, revealing the function of brain regions. Lesion mapping can also correlate the effectiveness of neurosurgery with the location of brain resection, identifying optimal surgical targets. Traditionally, voxel-based lesion mapping has employed the chi-square test when the clinical measure is binomial and the Student's *t* test when measures are continuous. Here we suggest that the Liebermeister approach for binomial data is more sensitive than the chi-square test. We also suggest that a test described by Brunner and Munzel is more appropriate than the *t* test for nonbinomial data because clinical and neuropsychological data often violate the assumptions of the *t* test. We test our hypotheses comparing statistical tests using both simulated data and data obtained from a sample of stroke patients with disturbed spatial perception. We also developed software to implement these tests (MRIcron), made freely available to the scientific community. ■

INTRODUCTION

Much of our classical understanding of brain function comes from observing the consequences of focal disruption of the brain. For example, if a neurological patient has difficulty speaking, we can hypothesize that the damaged regions are required for speech. Recently, there has been a rapid development in techniques that measure brain activation (e.g., scalp electrical recordings of event-related potentials [ERPs] or functional magnetic resonance imaging [fMRI] that can detect task-related changes in blood-oxygenation levels).

Activation techniques offer many advantages to classic studies of brain injury, for example, they can allow us to observe the timing of neural activation so one can infer the sequence of neural processes. Nevertheless, lesion mapping continues to offer a unique and powerful tool for understanding of brain function, complementing recent methods that measure brain activation in healthy adults (Fellows et al., 2005; Rorden & Karnath, 2004). Specifically, lesion mapping can identify which regions are critical for a task, whereas regions observed using activation techniques may not actually be essential for that task. For example, whereas lesion studies demonstrate that profound chronic language deficits are rarely seen following right-hemisphere injury, activation studies tend to detect changes in both hemispheres, potentially reflecting the strong connection between these homologous brain areas. Furthermore, examining the consequences of brain injury is vital for treatment of neurological disorders.

The recent popularity of brain activation techniques has been driven by the development of statistical tools that help identify which brain regions show task-related responses. In contrast, lesion mapping has historically been conducted without the use of objective statistics. For example, many studies have simply reported which region of the brain is most commonly damaged in individuals who exhibit a specific neurological symptom. There are two primary problems with this approach. First, it is not an objective test: Any random collection of patient lesions will necessarily show some region of maximal overlap, regardless of whether this location has any influence on symptoms or behavior. Second, brain injury is not random: Because of vasculature and architecture, certain regions of the brain are particularly vulnerable to injury. Therefore, specific brain regions are commonly injured regardless of their influence on symptoms; for example, the centrosylvian region is frequently damaged following middle cerebral artery stroke (Caviness et al., 2002). In order to properly infer a direct relationship between the location of a brain lesion and a given clinical manifestation, scientific investigations of the consequences of brain injury need to include variability in the dependent measure, for example, a comparison between patients with and without major symptoms as well as objective statistics. Here we describe traditional statistical approaches and suggest that different techniques can improve the power of lesion analysis.

[1]University of South Carolina, [2]University of Tuebingen, Germany

Statistical lesion analyses are either conducted by examining damage to predefined anatomical regions of interest or on a voxelwise basis (with an independent test computed for every 3-dimensional pixel of the brain image). Region of interest studies can offer better statistical power because fewer tests are computed simultaneously and therefore there is less correction for multiple comparisons. On the other hand, voxelwise analysis can offer much better spatial resolution and is typically better suited when there is not a strong a priori prediction regarding the spatial location of the critical lesion. The improvements on lesion analysis suggested in this article focus on voxelwise statistical analyses, although the methods are applicable to both.

The basic principle underlying lesion analyses is the question of whether the spatial location of the brain lesion can predict the individual's behavioral symptoms. The test used in lesion analysis is driven by the nature of the symptoms. Binomial tests are performed when the symptoms are binary (e.g., the presence or absence of a disorder), whereas the *t* test is traditionally used if the severity of symptoms lies on a continuum (with patients exhibiting a graded range of performance) and all patients included should be investigated at the same time point after stroke onset (otherwise differences in behavior reflect different stages of recovery rather than differences due to different locations of brain lesions).

The Yates-corrected chi-square test is a popular measure for examining binomial data, for example, presence or absence of a deficit. For example, the free MRIcro software (Rorden & Brett, 2000) includes a Yates-corrected chi-square test (which closely approximates the Fisher exact test). The Fisher exact test is often thought to offer a precise probability of an observed outcome, and therefore offer optimal power (Siegel & Castellan, 1988). However, the Fisher exact test is only accurate in situations when fixed marginals are expected.[1]In contrast, lesion mapping is a situation in which unconstrained marginals are expected.[2]A number of alternatives to the Fisher exact test have been suggested (Seneta & Phipps, 2001; Lancaster, 1961; Tocher, 1950; Liebermeister, 1877). The quasi-exact tests of Liebermeister and Lancaster tests are appealing, as they always provide the same *p* value in response to a set of observations. In practice, these tests closely approximate each other (as later demonstrated in Figure 1), so we focus on the Liebermeister test (which is in general more liberal and powerful than the Lancaster test). Computationally, these tests are very similar to the Fisher exact test, although offering better sensitivity (Seneta & Phipps, 2001). The additional statistical power of the Liebermeister test comes at the cost of a higher false alarm rate, but the crucial point is that the Leibermeister measure is giving a more accurate probability of an event occurring by chance (i.e., the Fisher test is always conservative, whereas the Liebermeister test reports near-nominal probabilities).

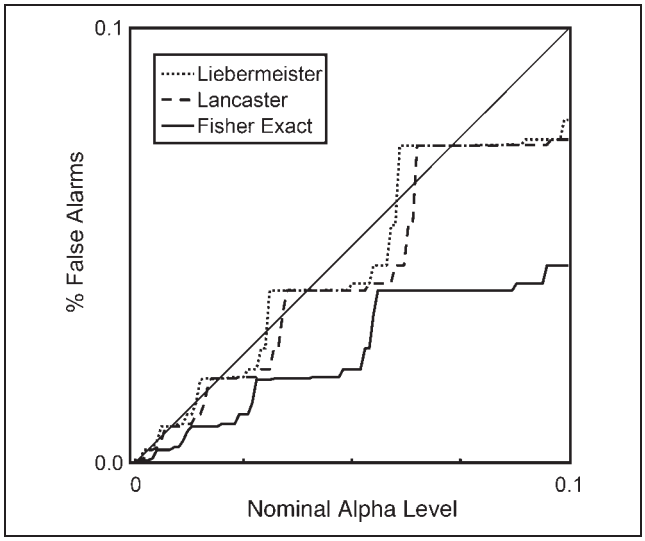

Figure 1. False alarm rate of the Liebermeister, Fisher–Lancaster Mid-P, and Fisher exact tests. The vertical axis shows the observed rate of false alarms, and the horizontal axis shows the statistical threshold. Optimally, a test should generate false alarms at the same rate as the alpha level, as shown by the diagonal line. Note that the Fisher exact test is always conservative; for example, at a 5% alpha level the Fisher exact test only reports false alarms in 2.15% of the simulations. The Liebermeister and Mid-P tests provide substantially more power. Note that the discrete nature of binomial data creates jagged *p* values, in particular when the sample size is small.

The *t* test and analysis of variance (ANOVA) are popular tests for analyzing lesion data when the symptoms can be represented as continuous data (Bates et al., 2003; Frank, Damasio, & Grabowski, 1997). This test assumes that the data are normally distributed, that the two groups have similar variance, and that the data represent an interval measurement. Although these measures are very robust (when the assumptions are violated, they tend to lose sensitivity rather than generate false alarms), there are three important limitations to their use with voxelwise lesion analysis. First, due to the multiple comparisons problem, it is difficult to test the underlying assumptions of these tests across the thousands of comparisons conducted during voxelwise analysis. Second, these tests directly measure differences in the mean, and will not offer optimal sensitivity when the mean is not a good measure of central tendency (e.g., when outliers are present or when the data are skewed). For example, on neuropsychological tests, patients without a deficit may show negatively skewed performance (a ceiling effect), whereas patients with a deficit may exhibit a positively skewed distribution (a floor effect). In situations where the data are skewed, the median is a more accurate measure of central tendency than the mean. A third criticism of these tests is that they explicitly assume that the dependent measure represents an interval measurement, and therefore they are not applicable to situations where performance is

measured using an ordinal scale. Many measures of neuropsychological deficits are ordinal, for example, the NIH Stroke Scale (NIHSS), which ranges from a normal healthy examination (score = 0) to severe stroke (score = 42). Although a patient with a lower score is performing better than a patient with a higher score, the magnitude of difference between a score of 8 and 12 is not necessarily identical to the difference between a score of 12 and 16. As noted by Heinemann et al. (1997), the ordinal nature of the NIHSS precludes the uses of parametric statistics.

Historical alternatives to the *t* test have included the Wilcoxon-Mann-Whitney test (Mann & Whitney, 1947; Wilcoxon, 1945) and the robust rank order test (Fligner & Policello, 1981). These tests do not assume that the data are normally distributed or based on an interval measurement. However, these tests are not entirely nonparametric. Specifically, the Wilcoxon-Mann-Whitney test assumes that the two groups of data have a similar shape and similar range. On the other hand, the robust rank order test assumes that the data are drawn from symmetrical distributions. Unfortunately, both of these tests tend to generate false alarms when their assumptions are violated (Reiczigel, Zakariás, & Rózsa, 2005; Feltovich, 2003; Zumbo & Coulombe, 1997; Pratt, 1964). Such situations are not uncommon when measuring lesion symptoms, for example, if one group is showing a floor or ceiling effect.

Brunner and Munzel (2000) have recently described a rank order test that is essentially assumption free. This test should provide better power than the *t* test when the assumptions of the *t* test have been violated. It is also applicable in situations where there are many ties or in situations where the data are based on an ordinal rather than interval scale. This test is relatively rapid to compute, and generates a statistic that is approximately normal for situations with at least 10 observations in each group. For smaller groups, one can either compute all possible more extreme permutations (to derive a precise *p* value) or use a permutation test to approximate the precise *p* value (Neubert & Brunner, 2007). Rorden, Bonilha, and Nichols (2007) have recently suggested that this test is suitable for voxel-based morphometry, albeit their implementation does not implement the permutation test for small groups. This small group correction is vital for lesion analysis, as the size of each group varies with lesion density (e.g., any voxel where only a few people have a lesion or almost all people have a lesion will require a small group permutation test).

Our aim in the present study was to develop and validate software employing the Liebermeister and Brunner–Munzel (BM) tests for lesion analysis. We wanted to validate these tests using both simulated and real data to evaluate their performance. Simulated data allow us to identify the potential benefit of a test and assess the false alarm rate. Real-world data help establish whether the benefit is meaningful.

A major issue for voxelwise analysis (regardless of whether one is looking at activation maps or lesion data) is the multiple comparisons problem. Specifically, voxelwise analyses computes thousands of statistical tests (one for each voxel in the brain). Computing thousands of tests dramatically increases the chance of making false alarms. For example, if we compute 1000 *t* tests on random data, we would expect about 50 comparisons to exceed a 5% statistical threshold. Historically, Bonferroni correction was a popular method for controlling the familywise error. This correction is both simple and intuitive: If we conduct 1000 tests and want to keep the overall chance of an error at 5%, we would only identify effects that had only a 1/20,000 probability of being due to chance (1000/0.05). Although this method offers very strong protection against false alarms, it tends to be exceptionally conservative, in particular in situations such as lesion mapping where the number of truly independent tests is much smaller than the number of tests computed. Therefore, we evaluated these tests using two proven methods for thresholding data in order to control for this familywise error effect. Therefore, we compared permutation thresholding as well as false discovery rate (FDR) thresholding, which were both implemented in our software. Permutation thresholds are based on randomly relabeling and resampling the data, computing the maximum observed statistic within the entire 3-D volume for each permutation. Therefore, with a 5% permutation threshold, the chance of a false alarm anywhere in the data set would be 1 in 20. Permutation thresholding is particularly suited to lesion analysis: lesions are formed from large contiguous regions, where each voxel is not truly independent. Therefore, permutation thresholding typically offers much better power than Bonferroni correction while offering strong control against familywise error. This technique has been previously described for voxelwise analysis in general (Holmes, Blair, Watson, & Ford, 1996) as well as voxelwise lesion analysis in particular (Kimberg, Coslett, & Schwartz, 2007; Frank et al., 1997). In contrast, FDR thresholding controls the ratio of false alarms to hits, rather than familywise error. In situations where there is little signal, this technique converges toward Bonferroni correction. However, FDR is much more sensitive in situations where a signal is present in a substantial portion of the data. FDR has also been previously employed both in voxelwise analysis in general (Genovese, Lazar, & Nichols, 2002) as well as lesion analysis (Rorden & Karnath, 2004; Bates et al., 2003).

METHODS

Binomial Data

We used both simulated and real lesion data to contrast the Fisher exact test to the Liebermeister test. The simulated data allowed us to examine the false alarm rate of

each test and to determine whether each test tends to provide liberal, nominal or conservative probabilities. The simulation compared two groups of 20 observations each. Every observation had an equal probability of being set to a one or a zero. We repeated this simulation 10,000 times, successively measuring the reported p value. Note that the two groups have equal probabilities, allowing us to test the null distribution. A test provides optimal power when the frequency of false alarms matches the nominal p value, as illustrated in Figure 1 by a diagonal line.

In order to examine the Liebermeister test with real lesion data, we conducted an analysis on a subset of 63 patients reported in Karnath, Fruhmann Berger, Küker, and Rorden (2004). Specifically, this subset was selected to have a very similar time between brain injury and behavioral testing. Each individual was asked to complete the letter cancellation task (Weintraub & Mesulam, 1985), where the participant is required to mark each occurrence of a target (letter ''A'') in a cluttered display of other letters (with a total of 60 targets). The performance of the patients is shown in Figure 2. In order to examine these data with a binomial test, patients who missed more than four items on this task were considered to have a deficit. A total of 1000 permutations were generated to compute the 5% permutation threshold.

Continuous Data

In order to examine performance on continuous measures, we compared the BM test to the conventional pooled-variance t test. We conducted a simulation study to examine the power of the two tests when faced with real-world neuropsychological data. Data were generated for two equal-size groups, with data drawn from the letter cancellation data (shown in Figure 2) with replacement. This distribution was sampled in a biased manner, creating an effect size between the two groups. For each observation, a linear random number R was selected, and then biased using Schlick's (1994) formula $R/[(1/b - 2)(1 - R) + 1]$. For one group, the value b was set to 0.65 and for the other group the value was 0.35. This bias meant that samples from one group *tended* to have higher scores than the other group, but individual samples from either group *could* occur throughout the entire range of observed data. The groups were then compared using the BM test (with 2000 permutations generated to establish accurate Z scores despite the small sample size) and the t test. This simulation was then repeated 5000 times. An advantage to analyzing simulated data is that we can also compute the rate of false alarms. Therefore, an identical simulation was run with the only modification being that both groups were drawn from the distribution shown in Figure 2 in an unbiased manner.

We also compared these tests using real lesion data. Specifically, we used the lesion maps from the 63 patients described for the binomial data, except using the number of cancellations (see Figure 2) as our dependent variable. Statistical maps were generated for both the t test and the BM test (with 10,000 permutations of this test generated for each possible group size in order to provide accurate, small group p values). Five percent permutation thresholds were generated based on 1000 iterations for both the t test and BM test.

Our lesion data set was based on a total of 63 patients; however, it should be noted that most lesion studies are based on smaller sample sizes. In addition, it is possible that with this large sample size, statistical power has essentially saturated. Therefore, we conducted a Monte Carlo simulation to compare the performance of the t test versus the BM test when assessing the data from 30 patients. The simulation was repeated 20 times, with each simulation drawing 30 patients randomly from the population of 63 patients used in the previous analysis. The data from each simulation were thresholded with a 1% FDR, with identical data assessed by both the t test and BM test.

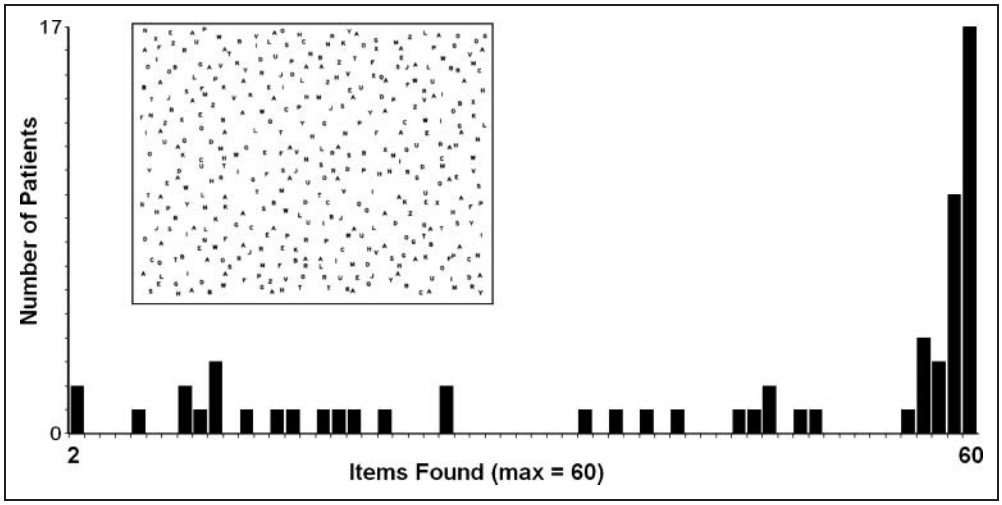

Figure 2. Performance of 63 right-hemisphere patients on the letter cancellation task (Weintraub & Mesulam, 1985). The task is displayed as an inset—the patients are asked to find the target letters ''A'' in a cluttered display of other letters. Note that this task does not result in a normal distribution, with many individuals finding all or nearly all of the items.

RESULTS

Binomial Data

We conducted a simulation of random data to contrast the false alarm rate of the Fisher exact test to the Liebermeister test. The results are shown in Figure 1. Note that the Fisher exact test (dotted line) is always conservative. For example, the 5% criterion is only achieved in 2.15% of our trials. In contrast, the quasi-exact Liebermeister test provided near-nominal false alarm rates.

The results for the binomial analysis of lesion data are shown in Figure 3A. Note that the Liebermeister test always returns a higher significance level than the Fisher exact test (as well as the Yate's-corrected chi-square test). Therefore, with a threshold of 1% FDR the Liebermeister test is more sensitive at detecting differences ($Z = 2.845$ for Fisher exact test, $Z = 2.701$ for Liebermeister measure). However, these tests offer identical power using a permutation threshold (as the Fisher's conservative nature is seen both in the observed as well as in the randomly generated permutations). Therefore, the 5% familywise permutation threshold was $Z = 2.584$ for the Fisher test, and 2.859 for the Liebermeister test.

Overall, the Fisher test is consistently conservative, whereas the Liebermeister test tends to perform near the nominal false alarm rate. This effect is clearly seen when FDR thresholding is conducted. However, these tests offer identical performance when permutation thresholding is applied (as the Fisher test is consistently conservative for all samples).

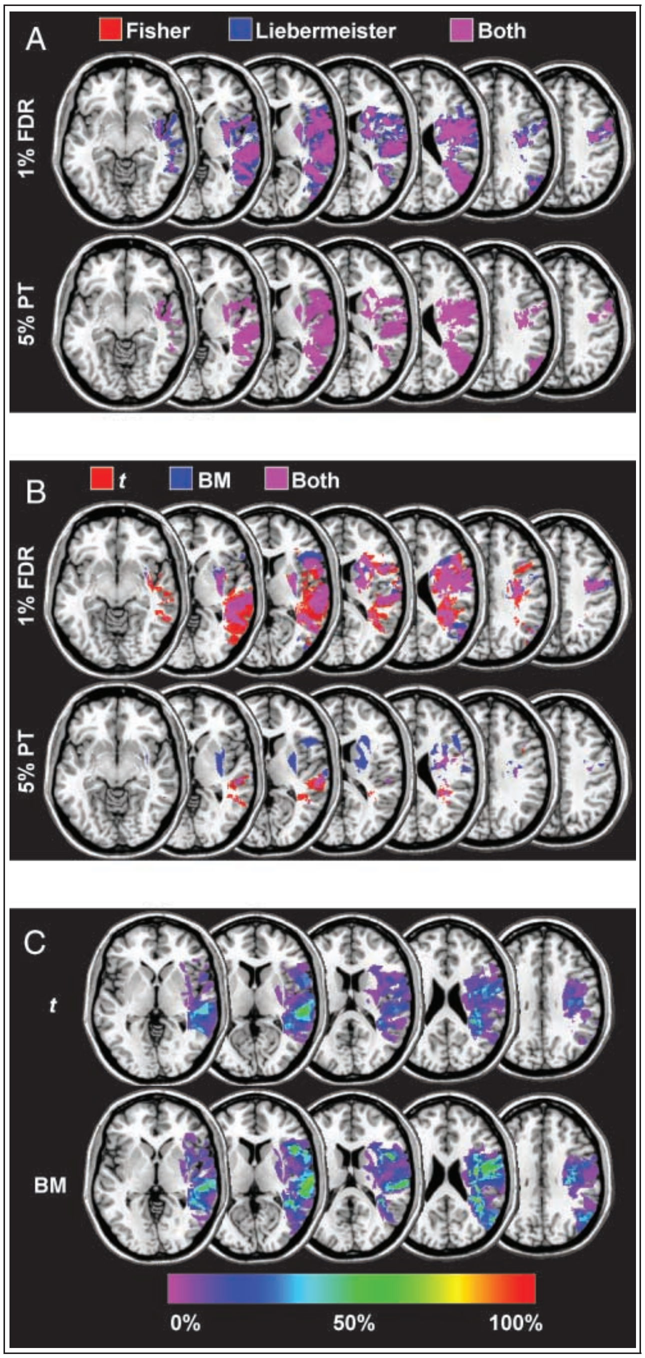

Figure 3. Comparison of tests for voxelwise lesion analysis. (A) Binomial data: comparison of the Fisher exact test and the Liebermeister measure for detecting differences on the letter cancellation task. Patients who missed at least five left-sided targets were considered to have a deficit at this task (n = 28), and individuals who missed fewer were considered to be unimpaired (n = 35). Regions in red were detected only by the Fisher exact test, regions in blue were detected only by the Leibermeister test, and regions in violet were detected by both tests. The top row shows data that survived a 1% false discovery rate (1% FDR) threshold, and the lower row shows data that survived the more stringent 5% permutation threshold (5% PT). (B) Continuous data: comparison of the t test (t) and the BM test, comparing lesion location and performance on the cancellation task. Regions in red were detected only by the t test, regions in blue were detected only by the BM test, and regions in violet were detected by both tests. The top row shows data that survived a 1% FDR, and the lower row shows data that survived the more stringent 5% PT. (C) Monte Carlo simulations of continuous data: comparing the t test (t) and BM test. Twenty simulations were run, with 30 individuals selected from the population of 63 patients. Each simulation was thresholded at a 1% FDR. The color scheme indicates each voxel's frequency of exceeding the statistical threshold. For example, voxels shown in green exceeded the statistical threshold in approximated 60% of the simulations. Note that the BM test typically detected more voxels than the t test.

Continuous Data

We used simulated skewed data to compare the power of the t test to the BM test. The results of this analysis are displayed in Figure 4. Note that for this data set the BM test proved slightly more powerful than the t test at both the $p < .05$ and $p < .01$ thresholds.

Next we estimated false alarm rates by sampling the distribution shown in Figure 2 in an unbiased manner. This analysis suggested that both the t test and the BM test offered similar protection against false alarms (with the t test generating false alarm rates of 0.0458–0.0584 and 0.0098–0.0118 for the 5% and 1% alpha thresholds,

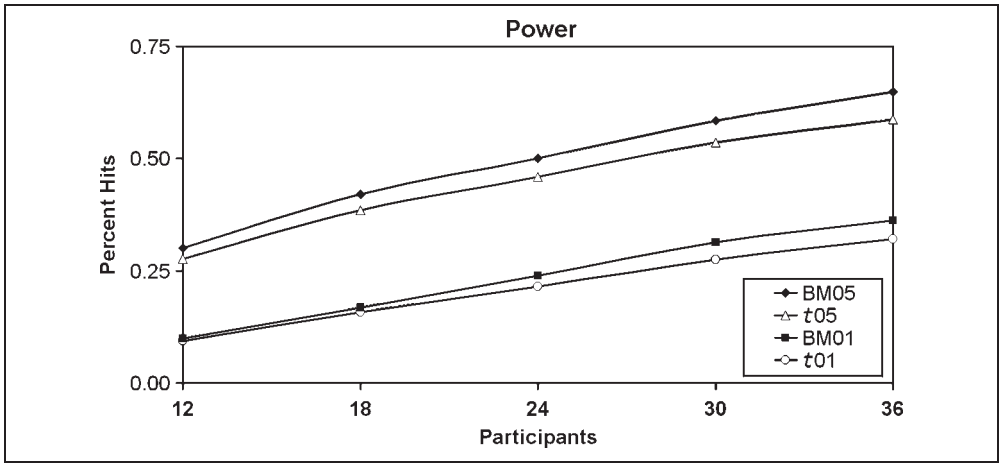

Figure 4. Results from a simulation comparing the sensitivity t test (t) and the BM test. Data for this simulation were drawn from the distribution shown in Figure 3. The vertical axis shows the tests' sensitivity, and the horizontal axis shows the influence of increasing sample size. Individual lines plot the performance of each test at both the $p < .05$ (BM05 and $t05$) and $p < .01$ thresholds (BM01 and $t01$).

whereas the same data analyzed with the permuted BM test generated false alarm rates ranging from 0.0448–0.0556 and 0.0082–0.0116).

The results for the continuous-measure lesion analysis for the 63 stroke patients are depicted in Figure 3B. The two tests identified a similar proportion of the voxels, although the precise pattern differed between the two tests, reflecting that these tests are sensitive to different aspects of the data.

Finally, we report the results of the Monte Carlo simulation. This test estimates our probability of detecting differences if we had only tested 30 patients instead of 63 individuals. By iteratively testing a random sample of 30 of the 63 patients, we can gauge the relative power of each test for this population. The t test detected between 0 and 7623 voxels per simulation, with a mean of 1718 and a median of 696 voxels detected. The BM test detected between 52 and 7497 voxels, with a mean of 2846 and a median of 2803 voxels detected. A repeated measures t test revealed that the BM test detected significantly more voxels than the t test, $t(19) = 2.28$, $p < .034$. Figure 3C shows the mean results for these simulations.

DISCUSSION

We suggest that the Liebermeister test offers a suitable substitute for the Fisher or chi-square tests, whereas the BM test provides a truly nonparametric alternative for the t test. However, it should be emphasized that these improvements are incremental; although these tests may be slightly more sensitive at detecting lesion-deficit relationships, they do not provide a massive increase in sensitivity. Our sample data set suggests that both the t test and the BM test tend to identify the same regions, although the BM test may offer more power for skewed distributions (allowing one to detect regions with a smaller group of images).

For binomial lesion mapping, the Liebermeister measure always offers a more accurate estimation of probability than the Fisher exact test. However, one's choice of test may depend on the underlying data. Whereas the t test is a very robust test, the BM test may offer more power in situations where the assumptions of the t test have been violated. The t test will provide marginally more sensitivity when examining normally distributed data (Rorden et al., 2007). However, in many real-world situations the BM test offers similar or better power, without making any of the t test's assumptions (which remain untested in voxelwise analysis). A further benefit of this test is that it remains valid for ordinal data. Our implementation of the BM test computes permutations rather than relying on the test's normal approximation, so it remains accurate for small sizes (as revealed in our simulation, and demonstrated by Neubert & Brunner, in press).

Furthermore, it should be noted that there are certain situations where none of these tests are entirely appropriate. For example, all the tests we have described only account for a single symptom variable. Therefore, these tests are not appropriate for multiple factors or in situations when nuisance confounds are present (for instance, excluding the influence of overall lesion volume, or time between the stroke and testing). In these situations, analysis of covariance (Bates et al., 2003) or logistic regression (Karnath et al., 2004) are better options. However, we note that Brunner and colleagues (Brunner, Domhof, & Langer, 2002; Brunner, Munzel, & Puri, 1999, 2001) have recently described truly nonparametric solutions to these problems.

Our main aim was to validate the BM and Liebermeister test for voxelwise lesion mapping. However, it is worth noting that these tests are applicable throughout science. For example, the Liebermeister test offers an accurate measure for any binomial data where the marginals are not fixed. For example, many cognitive neuroscience studies use the Fisher exact test to examine behavioral responses. Therefore, we have created software that allows the user to compute the Liebermeister measure (as well as the Fisher exact test) probabilities for

any 2 × 2 binomial data (www.mricro.com/stats.html). The BM test is useful whenever the assumptions of the *t* test have been violated. For example, a participant's response time to a task is often positively skewed (as there is a minimum theoretical response time, with less limit on slow responses). In our experience, the BM test offers a slight benefit over the *t* test when the two conditions exhibit different amounts of skew, with slightly poorer power for normally distributed data.

It should be noted that voxelwise lesion mapping typically requires examining a relatively large group of patients. This is due to the inherent variability in lesion volume and extent as well as difficulties in computing the true functional extent of a lesion (e.g., disconnected tissue can appear intact). Because of these factors, a large number of observations are required to survive multiple comparison correction. Studies with relatively small sample sizes may be better served by pooling data across predefined anatomical regions, which can offer substantially more power than voxelwise analysis. Alternatively, subtraction analyses (cf. Rorden & Karnath, 2004) can help visualize the anatomical trends in these situations.

Although single-case studies have provided valuable insights into cognitive functions, the anatomical inference of these studies is necessarily very weak. Although single-case evidence can act as an important counterexample for anatomical models, they are difficult to generalize. For example, single case studies of individuals who exhibit Broca's aphasia following right-hemisphere injury illustrate that the left inferior gyrus is not always the crucial region for Broca's aphasia, but large group studies demonstrate that these individuals reflect the exceptions rather than the rule.

Our focus was to validate and optimize the statistical methods used for lesion mapping. However, it is crucial to stress that statistical power is also greatly influenced by the inherent variability of the data. A few major factors that can profoundly influence lesion–symptom correlations include differences in etiology, age at time of injury, time between injury and testing and residual perfusion of the spared brain (Fridriksson, Rorden, Morgan, Morrow, & Baylis, 2006), functional disconnection (Catani & ffytche, 2005), quality of the lesions, and the accuracy in normalizing lesion maps into a standard stereotaxic space (Brett, Leff, Rorden, & Ashburner, 2001). Although beyond the scope of this article, controlling for these factors can be crucial for detecting differences using lesion mapping.

We recognize that the acceptance of any neurotechnique depends not only on the theoretical strengths of the method but also on its accessibility to users. Therefore, we have created a free and open-source software package that allows users to easily convert medical images to a popular scientific format (the NIfTI format). It also allows the user to rapidly define regions of brain injury, includes a spreadsheet for entering behavioral performance, generates statistical maps, and computes FDRs and permutation-based statistical thresholds. This software provides a complete solution for lesion mapping when combined with software that "normalizes" brain images so that images from different people are registered to a common stereotaxic space. We note that there are many free automated implementations of normalization routines including SPM (Statistical Parametric Mapping; Ashburner & Friston, 1999), FSL (Jenkinson, Bannister, Brady, & Smith, 2002), and AIR (Automated Image Registration; Woods, Grafton, Watson, Sicotte, & Mazziotta, 1998). It is important to note that the presence of lesions will often disrupt automated normalization, and therefore it is typically worthwhile to employ a lesion-masked cost function (Brett et al., 2001). Once lesions have been normalized to a standard space, our software can compute *t*, Fisher exact, Liebermeister, BM, and logistic regression tests. The software and source code are available from www.mricro.com/mricron, and this package can run on most currently popular operating systems (with Windows, Linux, and OSX distributions available).

Acknowledgments

This research was supported by National Institutes of Health grant R01 NS054266 and German Federal Ministry of Education and Research (BMBF) grant 01GW0641.

Reprint requests should be sent to Chris Rorden, Department of Communication Sciences and Disorders, University of South Carolina, SC 29208, or via e-mail: chris@mricro.com.

Notes

1. For example, consider an experiment where we ask an observer to guess whether each of four coins have been placed face up or face down on the table. If we ran this experiment with fixed marginals, we would constrain both the placement of the coins and the responses. We could place two coins heads up, and two coins heads down, and tell the observer that they must guess heads (face up) precisely twice and tails (face down) precisely twice. In this situation, there are six possible permutations (HHTT, HTHT, HTTH, THHT, THTH, TTHH), and therefore the observer has a one in six chance of guessing the correct orientation of all the coins. This is the situation modeled by the Fisher exact test.
2. Consider a situation where we randomly flip each of the four coins and ask the observer to guess the resulting orientation. In this case, the outcome of each coin is unconstrained by the other coins, and one could observe zero (one permutation), one (four permutations), two (six permutations), three (four permutations), or four (one permutation) coins landing face up. In this instance, the observer has only a 1 in 16 chance of correctly guessing the orientation of the four coins. This unconstrained example is analogous to the situations observed with lesion mapping.

REFERENCES

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

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.

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.

Brunner, E., Domhof, S., & Langer, F. (2002). *Nonparametric analysis of longitudinal data in factorial designs.* New York: Wiley.

Brunner, E., & Munzel, U. (2000). The nonparametric Behrens–Fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal, 42,* 17–25.

Brunner, E., Munzel, U., & Puri, M. L. (1999). Rank-score tests in factorial designs with repeated measures. *Journal of Multivariate Analysis, 70,* 286–317.

Brunner, E., Munzel, U., & Puri, M. L. (2001). The multivariate nonparametric Behrens–Fisher-problem. *Journal of Statistical Planning and Inference, 108,* 37–53.

Catani, M., & ffytche, D. H. (2005). The rises and falls of disconnection syndromes. *Brain, 128,* 2224–2239.

Caviness, V. S., Makris, N., Montinaro, E., Sahin, N. T., Bates, J. F., Schwamm, L., et al. (2002). Anatomy of stroke, part I. *Stroke, 33,* 2549–2556.

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.

Feltovich, N. J. (2003). Nonparametric tests of differences in medians: Comparison of the Wilcoxon–Mann–Whitney and robust rank-order tests. *Experimental Economics, 6,* 273–297.

Fligner, M. A., & Policello, G. E. (1981). Robust rank procedure for the Behrens–Fisher problem. *Journal of the American Statistical Association, 76,* 162–168.

Frank, R. J., Damasio, H., & Grabowski, T. J. (1997). Brainvox: An interactive, multimodal visualization and analysis system for neuroanatomical imaging. *Neuroimage, 5,* 13–30.

Fridriksson, J., Rorden, C., Morgan, P. S., Morrow, K. L., & Baylis, G. C. (2006). Measuring the hemodynamic response in chronic hypoperfusion. *Neurocase, 12,* 146–150.

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.

Heinemann, A. W., Harvey, R. L., McGuire, J. R., Ingberman, D., Lovell, L., Semik, P., et al. (1997). Measurement properties of the NIH Stroke Scale during acute rehabilitation. *Stroke, 28,* 1174–1180.

Holmes, A. P., Blair, R. C., Watson, J. D. G., & Ford, I. (1996). Non-parametric analysis of statistic images from functional mapping experiments. *Journal of Cerebral Blood Flow and Metabolism, 16,* 7–22.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage, 17,* 825–841.

Karnath, H.-O., Fruhmann Berger, M., Küker, W., & Rorden, C. (2004). The anatomy of spatial neglect based on voxelwise statistical analysis—A study of 140 patients. *Cerebral Cortex, 14,* 1164–1172.

Kimberg, D. Y., Coslett, H. B., & Schwartz, M. F. (2007). Power in voxel-based lesion–symptom mapping. *Journal of Cognitive Neuroscience, 19,* 1067–1080.

Lancaster, H. O. (1961). Significance tests in discrete distributions. *Journal of the American Statistical Association, 58,* 223–234.

Liebermeister, C. (1877). Über Wahrscheinlichkeitsrechnung in Anwendung auf therapeutische Statistik. *Sammlung Klinischer Vorträge (Innere Medicin No. 31-64), 110,* 935–962.

Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of 2 random variables is stochastically larger than the other. *Annals of Mathematical Statistics, 18,* 50–60.

Neubert, K., & Brunner, E. (2007). A studentized permutation test for the nonparametric Behrens–Fisher problem. *Computational Statistics & Data Analysis, 51,* 5192–5204.

Pratt, J. W. (1964). Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association, 59,* 665–680.

Reiczigel, J., Zakariás, I., & Rózsa, L. (2005). Bootstrap test of stochastic equality of two populations. *The American Statistician, 59,* 156–161.

Rorden, C., Bonilha, L., & Nichols, T. E. (2007). Rank-order versus mean based statistics for neuroimaging. *Neuroimage, 35,* 1531–1537.

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.

Schlick, C. (1994). Fast alternatives to Perlin's bias and gain functions. In P. Heckbert (Ed.), *Graphic Gems IV* (pp. 401–403). Boston: Academic Press.

Seneta, E., & Phipps, M. C. (2001). On the comparison of two observed frequencies. *Biometrical Journal, 43,* 23–43.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for behavioural science.* London: McGraw-Hill.

Tocher, K. D. (1950). Extension of the Neyman–Pearson theory of tests to discontinuous variates. *Biometrika, 37,* 130–144.

Weintraub, S., & Mesulam, M.-M. (1985). Mental state assessment of young and elderly adults in behavioral neurology. In M.-M. Mesulam (Ed.), *Principles of behavioral neurology* (pp. 71–123). Philadelphia: F. A. Davis.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics, 1,* 80–83.

Woods, R. P., Grafton, S. T., Watson, J. D. G., 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.

Zumbo, B. D., & Coulombe, D. (1997). Investigation of the robust rank-order test for non-normal populations with unequal variances: The case of reaction time. *Canadian Journal of Experimental Psychology, 51,* 139–149.