Published in final edited form as:
Neuroimage. 2009 February 15; 44(4): 1355–1362. doi:10.1016/j.neuroimage.2008.09.031.

An evaluation of traditional and novel tools for lesion behavior mapping

Chris Rorden, PhD[1], **Julius Fridriksson**[1], and **Hans-Otto Karnath, MD, PhD**[2]

1 *Department of Communication Sciences and Disorders, University of South Carolina, USA*

2 *Section Neuropsychology, Center of Neurology, Hertie-Institute for Clinical Brain Research, University of Tuebingen, Germany*

Abstract

Kinkingnéhun and colleagues (NeuroImage 37 [2007] 1237–1249) have recently described a novel approach for lesion-behavior mapping (LBM), referred to as Anatomo-Clinical Overlapping Maps (AnaCOM). Conventional voxelwise LBM tools apply statistics to contrast behavioral performance of patients with lesions that encompass given voxels to control patients where these voxels are spared. In contrast, AnaCOM contrasts performance of patients with injury involving given voxels to the performance of neurologically healthy participants. The authors correctly note that their procedure can offer substantially more statistical power than conventional LBM methods. We compared AnaCOM to conventional LBM techniques by examining hemiparesis (a common consequence of stroke) as the behavior of interest. We found that AnaCOM detected many regions of the middle cerebral artery territory not associated with the motor system. We suggest that conventional LBM techniques detect regions that are damaged in patients with a deficit while spared in those without a deficit, while AnaCOM detects regions that are associated with a deficit. Therefore, this new measure may offer poor specificity. Furthermore, on theoretical grounds we suggest that permutation-based thresholding will be a more sensitive method for controlling familywise error than the method of counting lesion-overlap clusters used by AnaCOM. Finally, we note that the within group variability tends to be smaller for neurologically healthy controls than in neurological patients, due to ceiling effects. Therefore, we suggest that nonparametric measures or the Welch's t-test are more appropriate than the conventional pooled variance t-test used by AnaCOM.

Keywords

lesion analysis; VLSM; power analysis; clustering

Introduction

Lesion behavior mapping (LBM) studies correlate the location of brain injury to neurological symptoms, revealing the critical anatomy for normal behavioral function. LBM studies have direct clinical implications for understanding neurological disorders and planning rehabilitation. LBM studies also allow a strong level of inference: determining whether a region

Address for correspondence: Chris Rorden, Department of Communication Sciences and Disorders, University of South Carolina, SC 29208, USA, E-mail: rorden@gwm.sc.edu.

is *critical* for a task. In contrast, brain activation studies aim to determine the brain regions *involved* with a task. Despite these strengths, LBM studies have their own set of limitations (Rorden and Karnath, 2004), and therefore a balanced approach is to combine brain activation techniques (like fMRI) with brain disruption techniques (such as LBM). However, current trends suggest that brain activation research is growing at a far faster rate than lesion studies (Fellows et al., 2005). One of the major limitations of LBM studies is that traditional statistical methods offer very low statistical power, and therefore LBM studies need large sample sizes, typically requiring years of data collection. Kinkingnéhun and colleagues (2007) have suggested a novel method (AnaCOM) for LBM that on the surface appears to offer substantially better sensitivity than traditional methods. Our aim was to explore this technique and compare it to existing methods commonly used in LBM.

There are several tools for lesion-behavior mapping. These can be broadly classified as region of interest based (ROI) LBM and voxelwise LBM. With ROI based methods, one examines whether the presence of damage (or amount of damage) to a small number of predefined anatomical regions predicts symptoms observed in neurological patients. For example, if both injury and deficit are scored as either intact or impaired, one can conduct a Fisher exact test, while if the extent of injury is a continuous measure one can employ the Mann-Whitney test (Herskovits, 1999). However, it is worth noting that ROI based methods can only identify patterns within predefined anatomical regions (e.g. if one divided the visual cortex between posterior and anterior regions, one could not accurately differentiate symptoms that correspond to ventral versus dorsal occipital damage). In contrast, with voxelwise LBM the entire brain is mapped as a volume of small 3D 'voxels' (typically, each voxel has a volume of $1mm^3$ to $27mm^3$), with an independent statistical test conducted for each voxel (i.e. for every voxel, one computes whether or not injury to that voxel predicts a deficit). Voxelwise LBM potentially offers better spatial precision than ROI-based LBM, and can reveal critical brain regions associated with a given deficit without *a priori* assumptions.

In this article, we focus on two major sources for the low statistical power of conventional voxelwise LBM studies. First, partial damage to a specific anatomical region can lead to deficits (the 'partial damage problem'). Second, voxelwise analysis must contend with the 'multiple comparisons problem'. AnaCOM attempts to address both of these problems. Therefore, we will investigate their solution to each problem in turn.

The Partial Injury Problem

Popular tools for voxelwise LBM include BrainVox (Frank et al., 1997), MRIcro (Rorden and Brett, 2000), VLSM (Bates et al., 2003), NPM (Rorden et al., 2007), the statistical modules of VoxBo (Kimberg et al., 2007), as well as the correlational methods described by Tyler et al. (2005). These tools differ in major respects including type of deficit (e.g. is the symptom a continuous measure with graded performance, or is the behavior binomial: either present or absent?), statistical test and statistical thresholding. However, all of these tools share a common assumption for computing statistics. Specifically, for every voxel a statistical test compares performance of individuals with injury to that voxel to the performance of individuals where that voxel is not injured. Therefore, these tests will detect brain regions that predict poor performance when injured and good performance when spared.

In stark contrast, Kinkingnéhun et al. (2007) propose a method they refer to as Anatomo-Clinical Overlapping Mapping (AnaCOM), which contrasts the behavioral performance of patients who have damage to a specific voxel to a group of neurologically healthy controls. Therefore, the difference between the conventional LBM and AnaCOM lies in the control population: conventional techniques examine the data from neurological patients while AnaCOM uses data from neurologically healthy individuals. While this distinction appears subtle, this change in the reference population dramatically influences the regions detected by

AnaCOM. Specifically, AnaCOM addresses a weakness with conventional LBM that we refer to as the 'partial injury problem'. This issue is illustrated in Figure 1 (see also Kinkingnéhun et al.'s Figure 11). Essentially, a functional module of the brain may be quite large, yet behavioral impairments can be observed when only a portion of this module is damaged. Therefore, two neurological patients may exhibit behavioral deficits due to damage to the same module even though there are no commonly injured voxels. With traditional LBM, the mutually exclusive nature of these injuries means that each individual is treated as the control for the other. On the other hand, AnaCOM contrasts each patient with a healthy control. As a consequence, AnaCOM should be able to detect lesion-behavior correlations with much smaller patient groups.

However, we argue that the assumptions made by this technique are often violated. One of the core problems with lesion mappings is that the locations of brain injury are not random, but rather reflect the brain's vascular architecture. As a concrete example of this problem, consider the common clinical condition of hemiparesis (weakness or paralysis moving the contralesional arm and/or leg). This symptom is due to damage to the motor cortex, the corticospinal tract, the basal ganglia, and is frequently associated with damage of the somatosensory cortices. However, all of these regions are supplied by blood from the middle cerebral artery, which also supplies blood to large portions of the lateral cortical convexity (Caviness et al., 2002). Therefore, patients with MCA territory injury will often show hemiparesis, but they will also often show damage to other portions of the cortex. We thus hypothesize that AnaCOM will detect injury to the entire MCA territory, while traditional LBM will be more specific for identifying regions associated with hemiparesis. This concern is illustrated in Figure 2. To test this hypothesis, we first conducted a Monte-Carlo simulation, based on a sample of 136 neurological stroke patients with right hemisphere lesions. We conducted a second analysis on a synthetic behavioral deficit – this analysis used the same dataset and parameters, with the sole exception being the replacement of the true behavioral measure (paresis) with a score that solely reflects the extent of damage to an arbitrary cortical region (roughly corresponding to Brodmann's Area 44). The logic behind this synthetic measure is that paresis can be the consequence of different anatomical injuries (e.g. motor cortex, the corticospinal tract, basal ganglia). If, as predicted, AnaCOM discovers a large territory associated with paresis, one could argue that this either reflects poor normalization or the discovery of new motor regions. In contrast, the sole factor influencing our synthetic behavior is the extent of damage to BA44, after normalization.

The multiple comparison problem

One of the primary reasons that any voxelwise analysis technique suffers from low statistical power is the 'multiple comparison problem' (MCP). As an extreme example, consider a study conducted with $1mm^3$ voxels where we test every single voxel in the gray and white matter. In this case, we will conduct around a million statistical tests, so with a conventional $p < 0.05$ alpha level (only detecting voxels that have a 5% or less chance of being due to random noise) one would expect around 50,000 false positives. A conventional solution to this problem is to apply Bonferroni Correction, where we adjust the statistical threshold to control for overall familywise error rate (FWE). Therefore, when conducting one million tests, we would adjust our threshold from $p < 0.05$ to $p < 0.00000005$, so that the chance of identifying false positives anywhere in the dataset is just 5%. Unfortunately, while this correction does control for multiple comparisons, it also necessarily leads to very low statistical power – very few real effects will be revealed. We discuss four methods for addressing the familywise error problem: overlap thresholding, cluster thresholding, permutation thresholding, and false discovery rate thresholding.

One method to tackle the MCP is to only compute statistics for voxels that are damaged in a reasonable proportion of the patients, a procedure we term 'overlap thresholding'. For example Kinkingnéhun et al. report data from 64 patients, and in their conventional LBM analysis they only compute statistics for voxels that are injured in at least three individuals, which reduces the number of statistical tests to 349,454. The logic for this is clear – the aim is to discover which brain regions are involved in common neurological syndromes. Therefore, by definition very rarely injured brain regions are unlikely to be the primary culprits in common disorders. Furthermore, regardless of whether one uses AnaCOM or traditional LBM methods, the very small number of observations found in rarely injured brain regions inherently results in low statistical power. While sensible, note that overlap thresholding does not dramatically reduce the number of statistical tests, so it is usually used in combination with the three other methods for addressing the MCP, which we describe next.

AnaCOM addresses the MCP by using a cluster thresholding method, which computes one statistical test for each spatial cluster, rather than for each voxel, with a cluster defined as a contiguous set of voxels that share a common lesion-overlap pattern (i.e. these voxels are all damaged in the same patients). Kinkingnéhun et al. report that their dataset could be defined as 1642 clusters, substantially reducing the number of comparisons relative to overlap thresholding alone (e.g. their average cluster spans 200 voxels). While cluster thresholding has been used by previous statistical LBM tools (Frank et al., 1997; Bates et al., 2003), Kinkingnéhun and colleagues are the first to have clearly described their implementation, including the issue of how to define contiguity (e.g. how do we define a neighboring voxel: must it share a face [6 neighbors per voxel], or do we also include edges [18 neighbors per voxel], or do we count corners as well [26 neighbors per voxel]?). The cluster thresholding method takes advantage of the redundancy in LBM datasets, where lesions are large contiguous regions.

While cluster thresholding is a principled solution to the MCP, permutation thresholding offers the optimal solution. Permutation thresholding has been previously described as a method of FWE control in LBM studies (Frank et al., 1997; Kimberg et al., 2007; Rorden et al., 2007). With LBM, our test statistic is based on contrasting the performance of patients with a lesion (consider the scores of three individuals: A = 3; B=2; C=2.5), to those without a lesion (consider four individuals D = 7; E = 6, F = 5; G = 9). With permutation testing (also known as 'randomization testing') we simply randomly scramble the order of the test scores between participants thousands of times to precisely compute how often we would observe a similar or more extreme set of test scores, offering a nonparametric measure of probability (this technique randomly samples the distribution revealed by 'exact testing', where one exhaustively tests all of the possible permutations). Permutation thresholding extends this concept to the MCP. Specifically, for LBM we create thousands of permutations of the participant's behavioral scores, and for each permutation we record the single most statistically significant voxel in the entire brain. Once completed, we rank-order these familywide maxima: if we have conducted 1000 permutations, the 50^{th} most extreme familywide maximum indicates the 5% threshold to control for familywise error. Just like cluster thresholding, permutation thresholding is sensitive to the inherent redundancy of lesion maps, and therefore the effective number of statistical tests will be fewer than voxelwise Bonferroni correction. However, we argue that permutation thresholding will offer more accurate FWE control than the cluster thresholding suggested by Kinkingnéhun and colleagues. Specifically, we note two facts that will tend to make cluster thresholding somewhat more conservative than the solution provided by permutation thresholding. First, the method described by Kinkingnéhun et al. suggests that the number of statistical tests is estimated by the number of contiguous clusters, while Permutation Thresholding reveals that the number of statistical comparisons is driven by the number of distinct voxel-lesion patterns (DLP). This is illustrated in Figure 3. Because the number of unique lesion patterns will generally be less (and never more) than the number of contiguous

clusters, this suggests that the cluster analysis tends to be somewhat conservative. The second reason that cluster analysis will tend to return a slightly more conservative threshold rests in the fact that in the real world, statistical tests such as the t-test tend to be slightly conservative. Specifically, these tests only deliver optimal results when all their inherent assumptions are met, while tending to be conservative in situations where the assumptions are violated (i.e. failing 'gracefully'). In the real world, statistical tests often include small violations, leading to slightly inaccurate p-values. Cluster analysis simply adjusts the statistical threshold based on the number of clusters, and therefore is not sensitive to this influence. On the other hand, permutation thresholds determine the actual statistical probabilities for the observed data and, in turn, return nominal values despite small violations of a test's assumptions. A clear illustration of this included in Rorden et al's (2007) Figure 3A which contrasts the conservative Fisher Exact test to the more accurate Liebermeister measure. This figure shows that the Liebermeister test is more sensitive than Fisher's test under FDR thresholding (described below), reflecting the more extreme values detected by the Liebermeister measure. However, the lower panel of this figure demonstrates that the two tests perform identically under permutation thresholding: the Fisher test is consistently conservative, so all the permutations deliver lower scores than the Liebermeister measure.

We suggest that counting the number of distinct voxel-lesion patterns (as first suggested by Kimberg et al., 2007) is slightly more accurate than counting the number of contiguous lesion clusters (as suggested by Kinkingnéhun et al.). Our reasoning is that this method accurately models the number of tests conducted during permutation thresholding (see Figure 3), and is computationally simpler (as one does not need to compute which neighbors are identical in order to define a contiguous cluster). Nevertheless, we acknowledge that cluster thresholding and counting distinct voxel-lesion patterns will tend to offer very similar solutions.

While permutation does offer the optimal solution to FWE correction, we note that there are clear reasons why people have sought alternatives such as cluster thresholds. First, traditional permutation thresholding is very computationally expensive, with the time required for an analysis scaling linearly with the number of permuations (e.g. a test that requires one minute will need over 16 hours if one wishes to generate one thousand permutations). Second, there are many situations where one wants to examine multiple factors (or remove variability described by nuisance covariates). In theory, it is possible to calculate permutation thresholds for some multifactorial designs (see Good, 2005), or remove nuisance variables by regressing nuisance effects from the data. In practice, an approximation of the permutation threshold (such as counting DLPs) offers an attractive alternative.

Our software (NPM; Rorden et al. 2007) implements a novel solution to the computational cost of permutation thresholding. Specifically, rather than computing permutations for each voxel, we compute permutations for each DLP. This method offers precisely the same result as full permutation thresholding, but is potentially much faster (due to the spatial contiguity observed with lesion maps). For example, consider an analysis where 51,000 voxels are tested, but there are only 9,100 DLPs – in this case the statistical computations for the DLP-based permutation threshold should be approximately five times quicker than the full permutation method. In practice, we predict that the benefit will be somewhat mitigated by the time required to generate and detect DLPs, so the benefit may be relatively small for computationally inexpensive tests (like the t-test), but more dramatic for demanding tasks such as the rank-order Brunner Munzel test. Our software also utilizes multithreading, a more conventional method for accelerating permutation thresholding. This should allow a computer with eight CPU cores to be much more rapid than a system with only a single core. Below we validate the efficiency of these techniques.

One final solution to the MCP is to control the False Discovery Rate (FDR) rather than the Familywise Error Rate. FWE controls the rate of making any false alarms: if we compute 1000 tests with a 5% FWE, there is only a 5% chance that there will be a single false alarm. On the other hand, FDR controls the ratio of false alarms to hits. For example, consider a test where 100 voxels survive a 5% FDR, in this case we expect that approximately 5 detected voxels are actually false alarms. FDR and Bonferroni FWE offer identical performance when only a tiny percentage of the voxels have a detectable signal. On the other hand, FDR is much more sensitive in situations where a large proportion of the sample shows signal. Therefore, FDR is dynamic, reflecting the signal in a given dataset. FDR provides a principled approach for offering reasonable statistical power when conducting many tests. Both VLSM (Bates et al., 2003) and NPM (Rorden et al., 2007) allow FDR thresholding for LBM studies. We suggest that this is often a useful threshold for smaller LBM studies, where one does not have the statistical power to achieve full FWE control.

In brief, we speculate that estimating the Bonferroni correction using the DLPs should closely approximate the FWE threshold identified through permutation thresholding. If this assertion is correct, DLPs would provide a computationally simple (and slightly less conservative) alternative to the cluster analysis method described by AnaCOM and would therefore provide a useful tool when permutation thresholding is not practical. To test this hypothesis, we conducted a Monte-Carlo simulation where we calculated both the DLPs and permutation thresholds for a sample of neurological patients. Furthermore, we predict that using DLPs to compute precise permutations is faster than computing permutations for each voxel. To test this prediction, we calculated the time required for these two methods.

Methods

In order to examine the performance of AnaCOM versus conventional LBM measures, we conducted an analysis including 136 consecutively admitted neurological patients with unilateral, right hemisphere strokes reported in Karnath et al. (2004) where hemiparesis scores were available. The degree of paresis of the upper and lower limbs was scored with the usual clinical ordinal scale, where ‘0’ stands for no trace of movement and ‘5’ for normal movement. For our analysis, we used the mean score of the upper and lower limb tests as our measure of hemiparesis. The mean score for these patients was 3.05 (with a range of 0.5), a standard deviation of 1.9 and a skew of −0.49. In contrast, it is exceptionally unusual for neurologically healthy controls to achieve a score other than 5. Therefore, for the AnaCOM test we contrasted patient data against a group of twenty controls scoring 5 on this task.

To contrast AnaCOM to conventional LBM we conducted a Monte-Carlo simulation. The simulation was repeated 25 times, with each simulation drawing 64 patients from the population of 136 stroke patients. For the AnaCOM simulations we contrasted these 64 selected patients to the scores of twenty healthy controls. Note that the sample size of patients and controls precisely matches the dataset described by Kinkingnéhun et al. in their validation of AnaCOM. Therefore, for each of the 25 simulations, the same sample of patients was analyzed using three tests: AnaCOM, traditional LBM using a t-test (Bates et al., 2003) and traditional LBM using the Brunner-Munzel test (Rorden et al., 2007). Statistics were only computed for voxels that were damaged in at least three individuals. The statistical maps for each test were thresholded using the DLPs to generate a Bonferroni 5% correction for familywise error. In other words, if there were 9100 distinct voxel-lesion patterns, only voxels which exceeded a Z-score of 4.3967 (Z-inverted for 0.05/9100) would be included as being detected. We then created a mean map for each test, revealing the percentage of simulations where each voxel exceeded the FWE threshold.

We conducted a second analysis using the same data and parameters, only changing the behavioral performance score. While the previous analysis used the actual paresis scores, in this second analysis the behavioral score was based linearly on the extent of Brodmann Area 44 injured: all controls and all patients with no injury to this region were given a score of 1000, patients with complete damage to this region received a score of 0, and an individual with 40% of this region damaged scored 600. The logic behind this synthetic score was to allow an objective measure for the test performance. One could argue that paresis might be due to large number of injuries, including white matter injury. In contrast, this simulated behavioral measure is purely driven by damage to a single region.

Implicit in our use of DLP thresholding is that this value closely approximates the threshold determined using the more computationally expensive Permutation Thresholding. To investigate this premise, during the Monte-Carlo simulation described above we simultaneously measured these two thresholding methods. For each simulation, we computed the 5% permutation threshold (estimated with 2000 permutations for both the traditional t-test and traditional Brunner-Munzel results), the 5% DLP threshold, and the lesion overlap threshold. For example, if a total of 51,000 voxels were lesioned in at least three individuals, the lesion overlap threshold would be 4.76. If there were 9100 distinct voxel-lesion patterns, the DLP threshold would be 4.3967. A unique lesion pattern is defined as the occurrence of a specific order of brain injury.

Our software (NPM; Rorden et al. 2007) implemented a novel form of permutation thresholding: conducting one statistical test for each DLP, rather than for each lesion. We hypothesized that this offers a computationally efficient method for estimating permutation thresholds. We conducted Monte-Carlo simulations to see if the DLP-based permutation threshold was significantly faster than the conventional permutation thresholding technique.

Our software also includes two techniques to accelerate analysis: DLP-based permutation threading and multithreading. We conducted twenty Monte-Carlo simulations to evaluate these techniques, with each simulation selecting 64 patients from our population and conducting a traditional t-test based LBM. Specifically, we compared the computation times for a single thread versus eight simultaneous threads. In addition, we computed the time required to estimate 1000 permutations using the DLP-based and voxelwise permutation thresholds. Note that this comparison is a stringent test of the DLP method: this method incurs a fixed cost for identifying the DLPs, but will see greater benefits for more permutations (e.g. if we had selected 4000 instead of 1000 permutations) and more computationally expensive tests (e.g. the t-test evaluated is much faster than rank-order tasks). The computation time excluded the time to load the dataset and save statistical maps (a single threaded process that required 10 seconds for each simulation, regardless of permutation method and thread count). All tests were run on an eight-core 3GHz Intel Xeon X5365 system with 5Gb of RAM running the Windows XP 64-bit operating system.

Results

Results for the paresis measure are shown in Figure 4, with the rows showing performance of AnaCOM (Fig. 4A and 4D), the traditional LBM t-test (Fig. 4B) and traditional LBM with the Brunner-Munzel test (Fig. 4C). Our planned repeated measures t-tests indicated that AnaCOM detected more voxels than traditional methods. Specifically, AnaCOM detected an average of 33,870 voxels per simulation (SD = 4208), while the conventional t-test detected 2657 voxels (SD = 1512) and the Brunner-Munzel test detected 1590 voxels (SD = 1428). Statistics revealed that AnaCOM detected statistically more voxels than the traditional t-test ($t(24)=37.12$, $p < 0.0001$) and the Brunner Munzel test ($t(24)=37.66$, $p < 0.0001$); further the traditional t-test detected more voxels than the traditional Brunner Munzel test ($t(24)=3.88$, $p < 0.0007$).

Furthermore, increasing the number of healthy controls used in the AnaCOM analysis from 20 to 64 individuals further boosted the number of voxels detected (mean 48,445; SD = 3508, compare Fig. 4A with 4D).

Note that in our simulation AnaCOM (Fig. 4A) always detected injury to areas associated with hemiparesis (e.g. motor cortex, corticospinal tract, putamen). On the other hand, AnaCOM also exhibits relatively low specificity – many simulations detected regions of the middle cerebral artery territory that are not classically associated with the motor system. In contrast, the t-test (Fig. 4B) sometimes fails to detect regions considered critical to movement, but also has a lower chance of false alarms (rarely detecting regions not associated with hemiparesis). Finally, the Brunner-Munzel test (Fig. 4C) appears less selective to regions associated with hemiparesis. We suggest the difference between the classical t-test and Brunner-Munzel test reflect that rank-order tests are not especially efficient with small sample sizes (see Rorden et al., 2007). In addition, increasing the number of controls included in the AnaCOM analysis from 20 to 64 individuals dramatically influenced the number of voxels detected by this test (Fig. 4D).

As noted, visual inspection of Figure 4 indicates that AnaCOM often detects regions not traditionally associated with the motor system. We argue that this reflects poor spatial specificity, but the previous description could in theory reflect improved sensitivity. In other words, one could interpret Figure 4A as revealing “previously undiscovered motor regions”. If this is the case, damage to these “novel motor areas” should predict paresis, even after one factors out deficits predicted by damage to the classic motor areas. However, analysis of the data suggests that many of the results probably reflect poor specificity. For example, consider the voxel at MNI coordinates X= 32, Y=−76 Z=0, which was detected in all 25 of the AnaCOM simulations using a control population of 64 individuals. The 15 individuals with injury to this voxel have a mean paresis score of 4.3 (SD = 1.1), while the 121 people where this voxel is spared actually perform better (2.9, SD = 1.9). To further investigate this effect, we conducted a multiple regression analysis that used paresis scores as the dependent measure with damage to this novel voxel as well as classic motor areas as independent variables. The classic areas of the motor system were defined as the precentral gyrus and putamen as localized by Tzourio-Mazoyer et al. (2002) and the corticospinal tract identified by Bürgel et al. (2006). For each individual we counted the number of voxels injured in this classic motor area (a continuous measure) as well as the presence or absence of injury to the novel location identified by AnaCOM. Damage to classic areas correlated with paresis severity ($r= 0.58$, $t[133]=7.923$, $p<0.0001$) while there was a numerical trend for damage to the novel area to predict better performance ($r=-0.232$, $t[133]=-1.54$, $p<0.126$). This suggests that this voxel does not impair motor performance.

Results from the synthetic behavioral data are shown in Figure 5. In this simulation, the extent of ‘impairment’ is defined as the percent damage to Brodmann’s Area 44 (this region is shown in panel 5D). This allows us to precisely measure the incidence of hits (accurately detecting the 1247 voxels that influence the behavior) and false alarms (reporting voxels that have absolutely no direct influence on the behavioral score). AnaCOM made an average of 1216 hits, but with an average of 38529 misses. In contrast, the t-test made an average of 1181 hits with 12259 misses. Finally, the Brunner-Munzel test made an average of just 132 hits (as noted earlier, this test can have low power with small sample sizes) with 4480 misses. Note that despite the conservative statistical threshold (5% corrected for multiple comparisons), all of the tests made a substantial number of false alarms. Presumably, this reflects the influence of vasculature in predicting lesion location, a concept we discuss in the discussion.

We also conducted repeated measures t-tests to determine if the DLP threshold accurately estimated the permutation threshold. For the t-test, the mean statistical threshold determined

using permutation thresholding was Z > 4.41 (SD 0.1242), while the unique lesion pattern method returned a mean threshold of 4.39 (St Dev 0.01454), finally the lesion overlap threshold yielded a mean threshold of 4.76 (St Dev 0.0116). Planned statistical comparisons using the repeated-measures t-test revealed no difference between the results derived from permutation and unique lesion patterns ($t(24)=0.77$, ns), while both were statistically different than the lesion overlap method (permutation method differed $t(24)=13.73$, $p < 0.0001$; unique pattern differed $t(24)=94.13$, $p < 0.0001$). Likewise, for the Brunner-Munzel test the mean threshold for the Permutation Test was Z= 4.42 (SD 0.1435), while the values for the DLP and lesion overlap method were identical to the values from the t-test thresholds. Planned t-tests revealed that the lesion overlap method was statistically different from the permutation ($t(24)=11.4$, $p <0.0001$) and the DLP measure ($t(24)=94.13$, $p <0.0001$), but that the permutation and DLP methods did not differ ($t(24)=0.98$ $P<0.3367$).

Full permutation thresholding (one computation per voxel for each permutation) was compared to our DLP-based permutation thresholding (one computation per DLP for each permutation). Further, we contrasted runs with one thread to runs using eight simultaneous threads. For one thread, the DLP method required 53.3 s (SD = 3.5) while the full method required 115.5 s (SD = 6.5). For eight simultaneous threads, the DLP method required 17.3 s (SD = 0.66) versus 24.8 s (SD = 1.15) for the full method. Therefore, increasing the number of threads lead to a nearly linear improvement (eight times the threads resulted in a factor of 5.9 to 7.1 increase in speed). The DLP-based method also resulted in a faster computation, but these effects decreased as the number of threads increased. This likely reflects our implementation where each thread generates its own cache of recently observed DLPs (we speculate that a multithreaded process that builds a single comprehensive list of all DLPs should scale better than our implementation).

Discussion

Our results suggest that AnaCOM offers poor specificity. We suggest that this test is prone to identify unrelated regions that are only associated with crucial regions in terms of vascular architecture.

A further concern regards a core assumption of AnaCOM – that stroke patients without lesions to a critical module will perform in the same range as normal controls. This hypothesis is probably violated when examining patients with acute stroke (which make up the large percent of LBM studies). Patients during acute care are often emotionally stressed, and adapting to an unfamiliar living condition. Many acute stroke patients exhibit generally poor performance on a wide range of tasks due to reasons that are not directly related to the location of their brain injury. Therefore, the traditional method of conducting statistical tests within a group of stroke patients with similar time since lesion onset seems more appropriate than AnaCOM's approach of comparing patients with acute brain injury to neurologically healthy controls.

We acknowledge that our investigation of hemiparesis, as well as our synthetic behavioral measure, offer an exceptionally difficult challenge for the AnaCOM method. In particular, the control population performs at ceiling on these tasks, showing no variability. Therefore, one could argue that the specificity of AnaCOM may fluctuate somewhat depending on the behavioral task applied. However, we argue that this performance is similar to many common neuropsychological tests.

Another unusual quirk of AnaCOM is the influence of the control population size. It is interesting that Kinkingnéhun et al. (2007) contrasted the performance of 64 neurological patients to a sample of just 20 neurologically healthy controls. It is worth noting that it is much easier to collect data from neurologically healthy controls (as one does not need to collect brain

scans, plot the extent of the injury, and normalize the data). By increasing the size of the control sample, one can dramatically increase the statistical significance, without actually modulating the effect size, as shown by contrasting Figures 4A and 4D. This increased statistical power also impairs the specificity of this measure.

To analyze their data, the authors of AnaCOM use a traditional t-test, which makes several assumptions regarding the data. Specifically, this test assumes that the data are normally distributed and that the two groups have similar variance. We suggest that both of these assumptions often may be violated if AnaCOM is applied to typical neuropsychological measures. Specifically, for many neuropsychological measures one might expect control patients to have ceiling effects, while stroke patients may show a graded range of performance (with individuals showing variable levels of impairment). The ceiling effects can lead to negatively skewed distributions (violating the normality assumption). Furthermore, if the neurologically healthy group exhibits ceiling effects, this group will have less variability than the patient group. On the other hand, more challenging tasks may lead to a graded range of performance in the healthy individuals, but floor effects in the brain damaged patients. This would lead to positively skewed data as well as variance differences between groups. We suggest that if the data are reasonably normal, Welch's t-test may be more appropriate than the classic Student's t-test. The Welch's t-test is sensitive to differences in the between group variance, while the traditional t-test pools variability. Further, in situations where the data is in fact normal, the Welch's t-test has similar statistical power to the conventional t-test (Ruxton, 2006), though it may offer slightly less power with unequal group sizes. Alternatively, a non-parametric test, such as the Brunner-Munzel measure could be used (Rorden et al., 2007).

In our synthetic behavioral analysis, the 'patient deficit' was actually a pure measure of the extent injury to BA44. Again, this analysis found that AnaCOM was particularly liberal –with this test typically identifying much of the middle cerebral artery territory. However, conventional LBM techniques also often detected regions outside BA44, even though by definition these voxels did not directly influence the behavioral measure. This analysis reveals a core weakness of the lesion mapping methods – damage to peripheral regions is often strongly predicted by injury closer to the root of the vascular branch. This can lead to LBM methods identifying central regions as being involved with functions (see Hillis et al., 2004; Husain & Nachev, 2007). This finding impacts the inference that can be drawn from contemporary lesion mapping studies. The tests accurately identify that regions near BA44 are predictive of the deficit variable (which might be clinically relevant). However, one cannot infer that all the regions identified using lesion mapping are necessary for the task (limiting the theoretical inference). This analysis demonstrates that this confound can pose a real challenge for lesion mapping, and emphasizes that this method must be complemented by other convergent methods that have different sets of assumptions.

The inherently low statistical power of conventional voxelwise LBM means that it is rarely suitable for the small sample sizes common when examining rare disorders. In these situations, researchers may be tempted to use AnaCOM. We argue that the inferences that can be drawn from the use of this tool are limited. Potential alternatives include region of interest analysis or conducting a Bayesian lesion-deficit analyses (BLDA) as described by Chen et al. (2008). This technique detects complex linear or nonlinear associations between brain-lesion locations and behavior. The multivariate BLDA complements the mass-univariate approach of conventional voxelwise LBM. Unlike conventional methods, BLDA can dissociate different regions that independently predict deficits. Furthermore, traditional methods ignore the fact that lesions are contiguous clusters (as they treat each voxel as an independent analysis). In contrast, BLDA can identify clusters that are strongly associated with a deficit (as BLDA explicitly models the spatial correlations among voxels), potentially offering better sensitivity and allowing the user to generate conditional probability tables that are useful for computing

the sensitivity and specificity of the cluster. These measures can also help plan power analyses for planning larger, objectively thresholded studies using traditional LBM techniques. Unfortunately, similar to traditional LBM methods, this technique is also sensitive to the influence of vasculature, this is demonstrated in red on Figure 5D. This figure shows the results of BLDA applied to the entire dataset of 136 individuals (as this software currently requires binomial behavioral data, we scored any individual with at least 10% injury to BA44 as impaired, with all other individuals counted as having normal performance). Note that the region identified by BLDA extends well beyond the border of BA44, though these results can not be directly compared to the other statistical analyses (as we used a binomial classifier for BLDA and a larger patient sample).

A clear result of our investigation is that counting the distinct lesion patterns provides a computationally simple approximation of permutation thresholding. This is useful in situations where permutation thresholding is not practical – for example, some forms of multifactorial analysis or estimating a power analysis. We note that the DLP threshold is very similar to the cluster threshold suggested by Kinkingnéhun et al. (2007), though it will tend to be slightly less conservative (as shown in Figure 3), and is computationally more efficient (as one does not need to consider a voxel's neighbors). Our simulations found that there was no statistically detectable difference between DLP and true permutation thresholding. In contrast, on theoretical grounds we predicted that DLP would tend to be slightly over-conservative. Indeed, this is precisely the pattern reported by Kimberg and colleagues (2007) where the DLP test reliably performed more conservatively than the permutation values. One important difference is that the lesions plotted by Kimberg and colleagues were plotted on each and every slice of their dataset, while our dataset used lesion maps that were plotted on one slice each 8mm (or 10mm for the most dorsal slices). As a result, our dataset had less spatial coherence than a complete voxelwise analysis. Therefore, we suggest that the results of Kimberg and colleagues probably more closely approximate typical usage, where lesions are drawn on all slices of a high-resolution image. With modern computers, the computational cost of permutation thresholding is negligible, and therefore we recommend this technique when applicable, but suggest that DLP provides an alternative for multi-factor designs and estimates of statistical power.

We also implemented and validated an accelerated permutation thresholding method. Our results suggest that this method is substantially quicker full permutation thresholding while returning identical results. In addition, our multithreaded implementation further reduces computational time for systems with multiple CPU cores. Where applicable, we suggest that permutation thresholding offers the optimal method for controlling familywise error.

In conclusion, we suggest that AnaCOM the inference one can draw based on AnaCOM are substantially different from traditional LBM. Therefore, results using this new technique need to be interpreted with caution. We acknowledge that traditional voxelwise LBM has low statistical power for small sample sizes, but suggest that region of interest or BLDA approaches avoid many of the limitations inherent with AnaCom.

Acknowledgements

NIH funding supports CR and JF (R01 NS054266 & R01 DC008255). HOK is supported by the Bundesministerium für Bildung und Forschung (BMBF-Verbundprojekt "Räumliche Orientierung" 01GW0641) and the Deutsche Forschungsgemeinschaft (SFB 550-A4). We wish to thank Rong Chen for helping with BLDA.

References

Bates E, Wilson SM, Saygin AP, Dick F, Sereno MI, Knight R, Dronkers N. Voxel-based lesion-symptom mapping. Nature Neurosci 2003;6:448–450. [PubMed: 12704393]

Bürgel U, Amunts K, Hoemke L, Mohlberg H, Gilsbach JM, Zilles K. White matter fiber tracts of the human brain: three-dimensional mapping at microscopic resolution, topography and intersubject variability. Neuroimage 2006;29:1092–1105. [PubMed: 16236527]

Caviness VS, Makris N, Montinaro E, Sahin NT, Bates JF, Schwamm L, Caplan D, Kennedy DN. Anatomy of Stroke, Part I. Stroke 2002;33(11):2549–2556. [PubMed: 12411641]

Chen R, Hillis AE, Pawlak M, Herskovits EH. Voxelwise Bayesian lesion-deficit analysis. Neuroimage 2008;40:1633–1642. [PubMed: 18328733]

Fellows LK, Heberlein AS, Morales DA, Shivde G, Waller S, Wu DH. Method matters: an empirical study of impact in cognitive neuroscience. J Cogn Neurosci 2005;17:850–858. [PubMed: 15969904]

Frank RJ, Damasio H, Grabowski TJ. Brainvox: an interactive, multimodal visualization and analysis system for neuroanatomical imaging. Neuroimage 1997;5:13–30. [PubMed: 9038281]

Herskovits EH, Megalooikonomou V, Davatzikos C, Chen A, Bryan RN, Gerring J. Is the spatial distribution of brain lesions associated with closed-head injury predictive of subsequent development of attention-deficit hyperactivity disorder? Analysis with brain image database. Radiology 1999;213:389–394. [PubMed: 10551217]

Hillis AE, Work M, Barker PB, Jacobs MA, Breese EL, Maurer K. Re-examining the brain regions crucial for orchestrating speech articulation. Brain 2004;127:1461–1462. [PubMed: 15197111]

Husain M, Nachev P. Space and the parietal cortex. Trends in Cognitive Sciences 2007;11:30–36. [PubMed: 17134935]

Karnath H-O, Fruhmann Berger M, Küker W, Rorden C. The anatomy of spatial neglect based on voxelwise statistical analysis - a study of 140 patients. Cerebral Cortex 2004;14:1164–1172. [PubMed: 15142954]

Kimberg DY, Coslett HB, Schwartz MF. Power in Voxel-based lesion-symptom mapping. J Cogn Neurosci 2007;19:1067–1080. [PubMed: 17583984]

Kinkingnéhun S, Volle E, Pélégrini-Issac M, Golmard J-L, Lehéricy S, du Boisguéheneuc F, Zhang-Nunes S, Sosson D, Duffau H, Samson Y, Levy R, Dubois B. A novel approach to clinical-radiological correlations: Anatomo-Clinical Overlapping Maps (AnaCOM): Method and validation. NeuroImage 2007;37:1237–1249. [PubMed: 17702605]

Rorden C, Brett M. Stereotaxic display of brain lesions. Behav Neurol 2000;12:191–200. [PubMed: 11568431]

Rorden C, Karnath HO. Using human brain lesions to infer function - a relic from a past era in the fMRI age? Nature Reviews Neuroscience 2004;5:813–819.

Rorden C, Karnath H-O, Bonilha L. Improving lesion-symptom mapping. J Cogn Neurosci 2007;19:1081–1088. [PubMed: 17583985]

Ruxton GD. The unequal variance t-test is an underused alternative to Student's t-test and the Mann–Whitney U test. Behavioral Ecology 2006;17:688–690.

Tyler LK, Marslen-Wilson W, Stamatakis EA. Dissociating neuro-cognitive component processes: voxel-based correlational methodology. Neuropsychologia 2005;43:771–778. [PubMed: 15721189]

Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M. Automated anatomical labelling of activations in spm using a macroscopic anatomical parcellation of the MNI MRI single subject brain. Neuroimage 2002;15:273–289. [PubMed: 11771995]

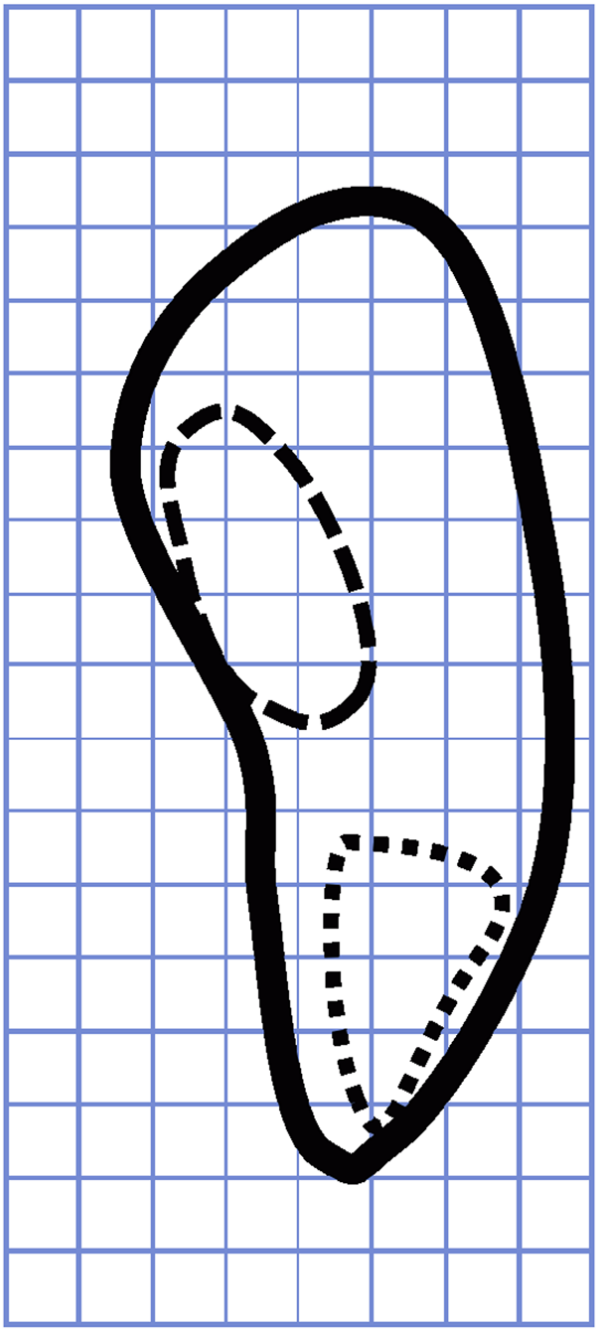

Figure 1.
The partial injury problem. Consider a module responsible for motor control (black shape). Two individuals have lesions that damage part of this module (dotted and dashed lines), impairing movement. Both patients have deficits, yet their injuries are mutually exclusive. This causes conventional lesion deficit mapping to have poor statistical power: voxels that are damaged in one patient with a deficit are spared in the other patient who has the same deficit. AnaCOM is resistant to this problem, as the performance of patients with damage to a particular voxel is contrasted with performance of neurologically healthy individuals.

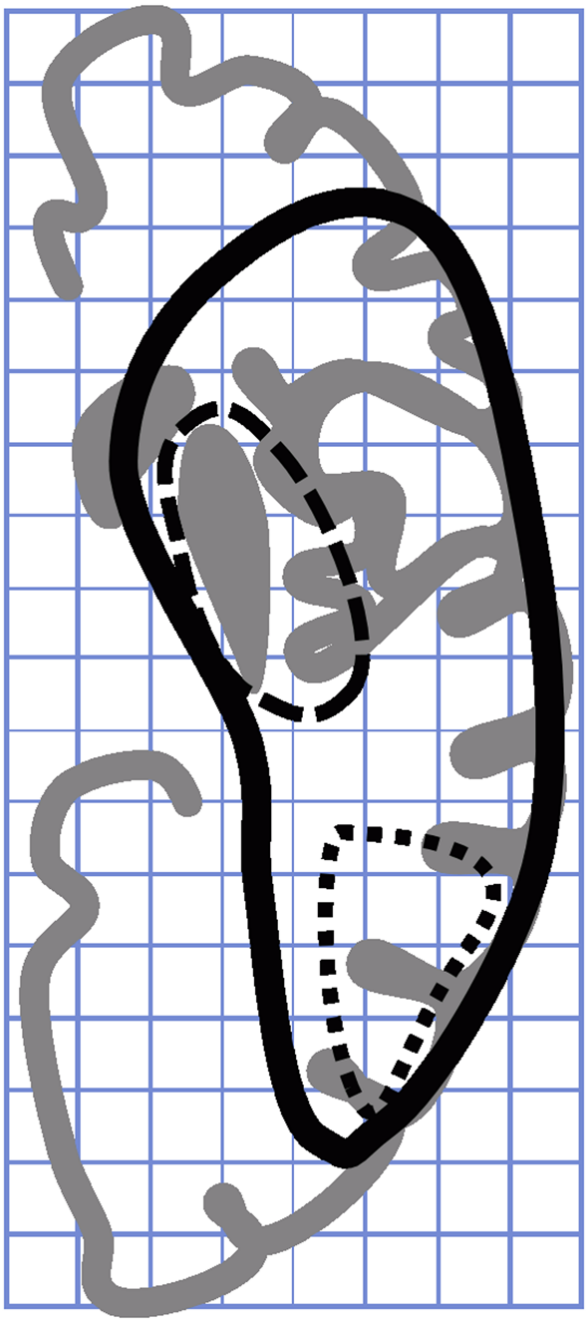

Figure 2.
Poor specificity of AnaCOM: strokes often damage portions of the middle cerebral artery territory (black shape). This territory includes regions that are critical for movement (dashed line) as well as regions that are not (dotted lines). However, damage to one region will often be accompanied by damage to another region. In this situation, AnaCOM has poorer specificity than conventional LBM techniques. Note that the critical features of this diagram are identical to Figure 1.

Patient	Observed	Perm 1	Perm 2	Perm 3
A	13	16	13	16
B	11	11	16	13
C	16	22	28	22
D	22	28	11	11
E	28	13	22	28

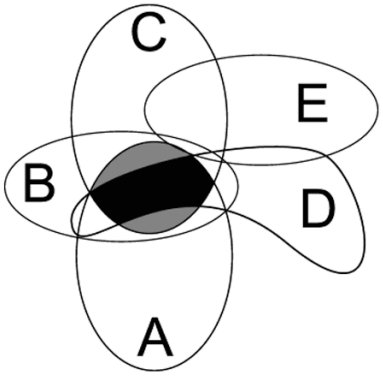

Figure 3.

A comparison of cluster thresholding as described by Kinkingnéhun and colleagues versus permutation thresholding. Consider the lesion maps of five patients (A..E). The region in gray is lesioned in patients A, B and C exclusively. Cluster thresholding counts the two gray areas as two independent tests (as they are not spatially contiguous). In contrast, these two locations will generate the same test statistic in every possible permutation. Therefore, Bonferroni correction based on the number of spatially contiguous clusters will tend to be more conservative than the value determined through permutation thresholding. We argue that counting the number of unique lesion overlap patterns best approximates the permutation thresholding.

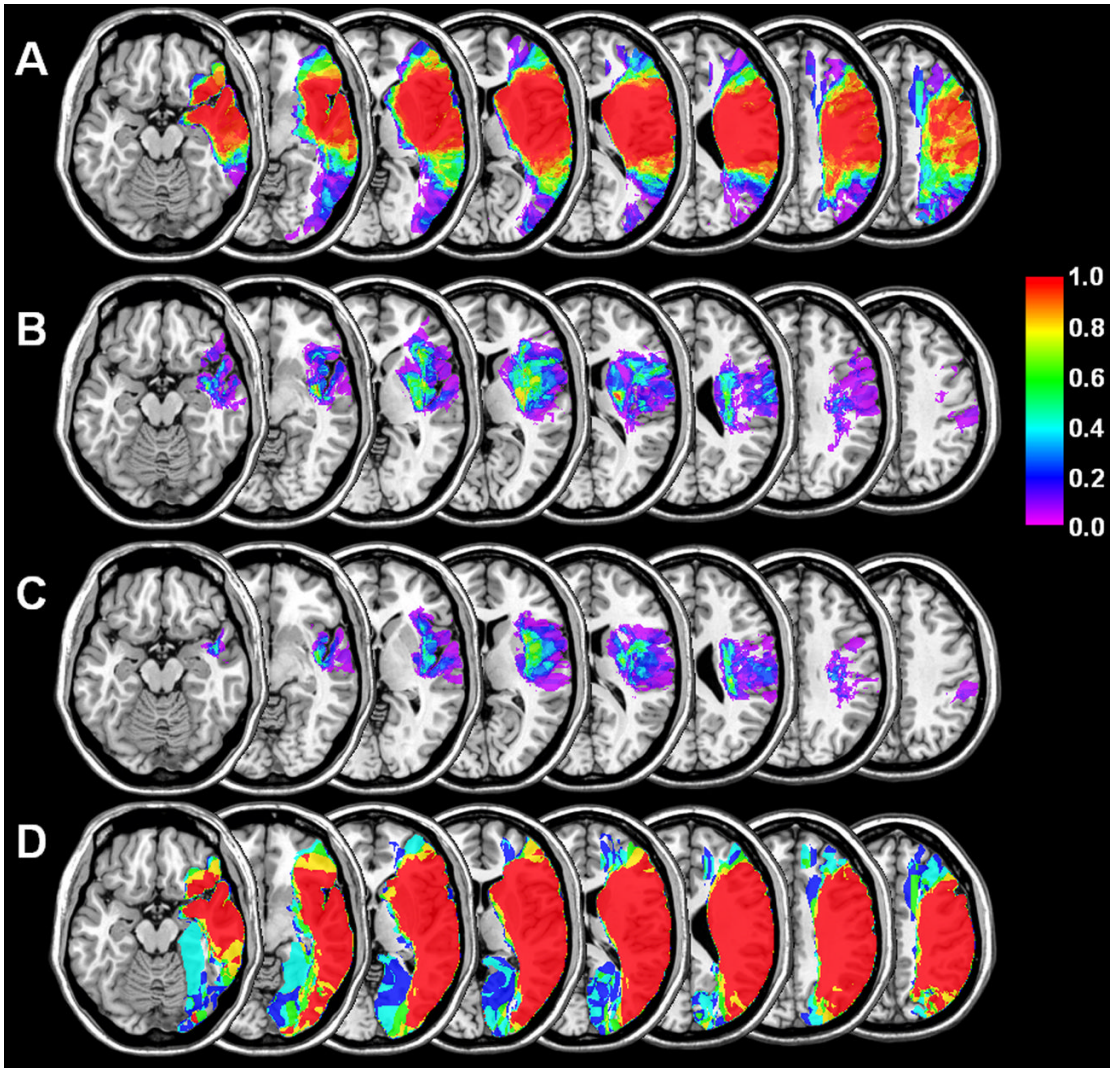

Figure 4.
Voxels that predict hemiparesis, as reported by AnaCOM (A, using 20 healthy controls), traditional LBM with the t-test (B), traditional LBM with the Brunner-Munzel test (C) and AnaCOM using 64 healthy controls (D). These maps show the results of twenty-five Monte-Carlo simulations, each selecting 64 stroke patients from a population of 136, and each thresholded at $p < 0.05$ corrected for multiple comparisons based on the number of unique lesion patterns. Therefore, regions that appear red were detected in all of the simulations, while regions in green were detected in 60% of the simulations. Axial slices correspond to −16, −8, 0, 8, 16, 24, 32 and 40mm in MNI space. Note that AnaCOM identifies large regions of middle cerebral artery territory (A).

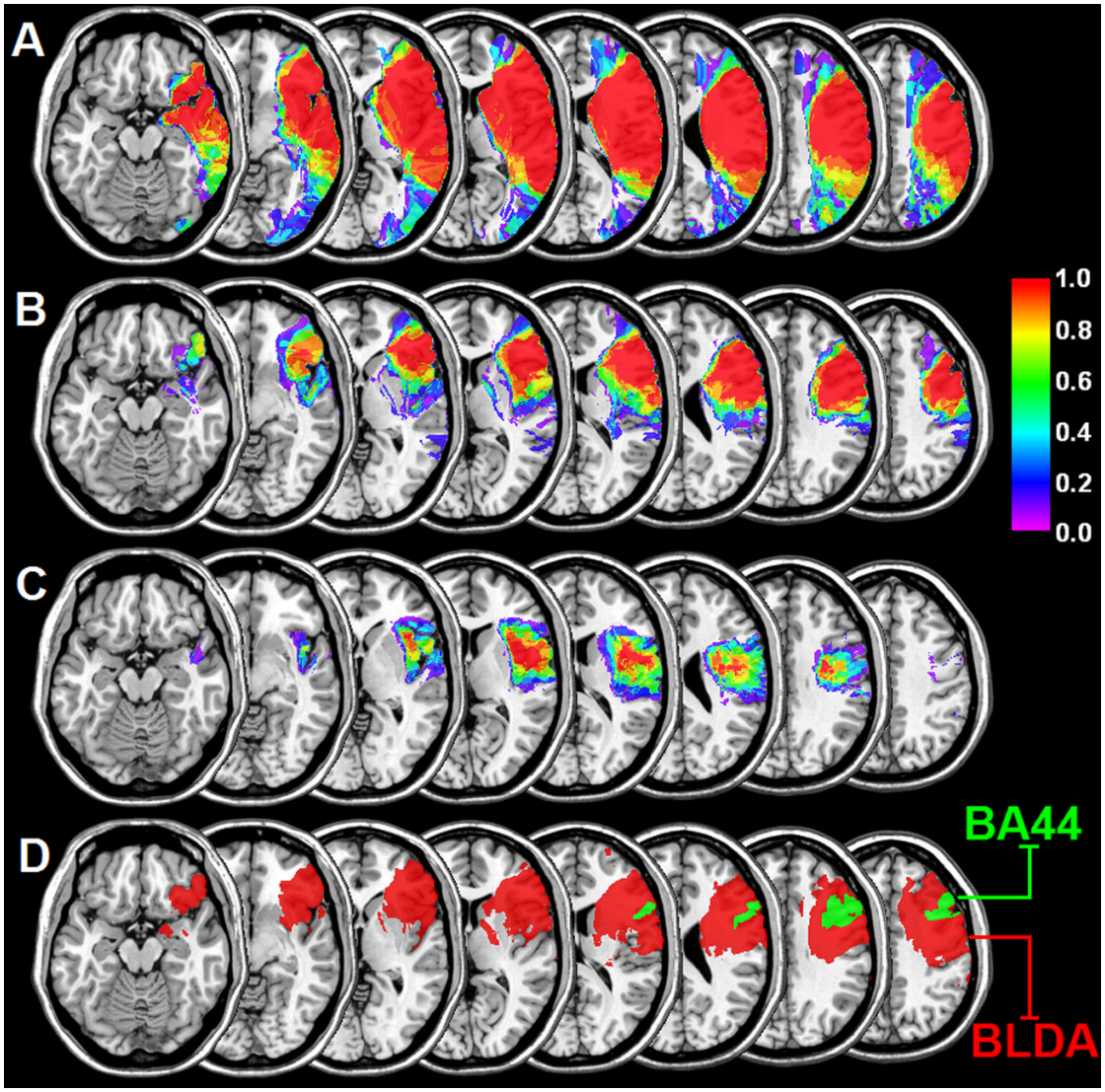

Figure 5.

The discrepancy displayed in Figure 4 between traditional LBM tests (Fig. 4B and C) and the AnaCOM method (Fig. 4A and D) could in theory be due to either enhanced sensitivity or poorer specificity of AnaCOM. Therefore, we conducted a new analysis where each patient's 'behavioral deficit' was actually calculated as the extent of injury to Brodmann's Area 44 (this region is shown in panel 5D). Therefore, by definition the 'deficit' can be perfectly predicted by damage to BA44, and no other region directly predicts this deficit. In all other respects, this analysis was identical to Figure 4. The results of 25 Monte-Carlo simulations are shown for AnaCOM (A, using 20 healthy controls), traditional LBM with the t-test (B), traditional LBM with the Brunner-Munzel test (C). Panel D shows the actual extent of our BA44 region in green, while in red we show the region detected by Bayesian lesion-deficit analyses [BLDA] (using the entire 136 patient dataset, as described in the text). Note that all tests appear biased by vasculature (e.g. damage to the rostral part of the MCA territory strongly predicts injury to BA44). However, this bias is especially pronounced for AnaCOM, which typically detects most of the middle cerebral artery territory.