Neuropsychologia 115 (2018) 112–123

Contents lists available at ScienceDirect

Neuropsychologia

journal homepage: www.elsevier.com/locate/neuropsychologia

Corrections for multiple comparisons in voxel-based lesion-symptom mapping

Daniel Mirman[a,*], Jon-Frederick Landrigan[b], Spiro Kokolis[b], Sean Verillo[b], Casey Ferrara[c], Dorian Pustina[d]

[a] University of Alabama at Birmingham, Birmingham, AL, USA
[b] Drexel University, Philadelphia, PA, USA
[c] Moss Rehabilitation Research Institute, Elkins Park, PA, USA
[d] University of Pennsylvania, Philadelphia, PA, USA

ARTICLE INFO

Keywords:
Voxel-based lesion-symptom mapping
VLSM
Multiple comparisons
Permutation tests
Cluster size correction
Family-wise error correction

ABSTRACT

Voxel-based lesion-symptom mapping (VLSM) is an important method for basic and translational human neuroscience research. VLSM leverages modern neuroimaging analysis techniques to build on the classic approach of examining the relationship between location of brain damage and cognitive deficits. Testing an association between deficit severity and lesion status in each voxel involves very many individual tests and requires statistical correction for multiple comparisons. Several strategies have been adapted from analysis of functional neuroimaging data, though VLSM faces a more difficult trade-off between avoiding false positives and statistical power (missing true effects). We used simulated and real deficit scores from a sample of approximately 100 individuals with left hemisphere stroke to evaluate two such permutation-based approaches. Using permutation to set a minimum cluster size identified a region that systematically extended well beyond the true region, making it ill-suited to identifying brain-behavior relationships. In contrast, generalizing the standard permutation-based family-wise error correction approach provided a principled way to balance false positives and false negatives. Comparison with the widely-used parametric false discovery rate (FDR) correction showed that FDR produces anti-conservative results at smaller sample sizes (N = 30–60). An implementation of the continuous permutation-based FWER correction method described here is included in the lesymap package for lesion-symptom mapping (https://dorianps.github.io/LESYMAP/).

1. Introduction

Identifying relationships between location of brain damage and cognitive deficits is a foundational method in cognitive neuroscience, tracing its history at least to the behavioral neurologists of the mid-19th century (e.g., Lichtheim, 1885). Those early studies were based on individual case studies and, as data accumulated, researchers used lesion overlays to identify the locations where damage consistently produced deficits of interest. Recent advances in neuroimaging technology have allowed much finer-grained analyses at the level of individual voxels (Bates et al., 2003; Rorden and Karnath, 2004). In voxel-based lesion-symptom mapping (VLSM), an association between deficit severity and lesion status (lesioned vs. not lesioned) is tested in each voxel, producing a statistical map of the strength of relationship between lesion status and deficit. However, this map is the result of individual tests across tens or even hundreds of thousands of voxels.

The large number of tests involved in analysis of neuroimaging data requires some kind of statistical correction for multiple comparisons. Several strategies have been proposed, often by adaptation from analysis of functional neuroimaging (e.g., fMRI). One standard strategy of correction for multiple comparisons is to control voxel-level family-wise error rate (FWER), which is the probability of making one or more false positive (Type 1) errors among the entire set of tests. The Bonferroni correction is a classic FWER correction method, though it is generally considered overly conservative for neuroimaging data. An alternative, non-parametric, approach to FWER correction is to use permutations of the observed data to build a null distribution of test statistics and compare the observed test statistic against that null distribution to determine the likelihood of observing the result if the null hypothesis were true. Because it is based on permuting the real data, this approach has the important advantage of not making assumptions about the distributions of scores or test statistics – assumptions that are likely to be violated by skewed distributions of behavioral deficit (symptom) scores and by the spatial contiguity inherent to stroke lesions. Building

* Corresponding author.
E-mail address: dan@danmirman.org (D. Mirman).

http://dx.doi.org/10.1016/j.neuropsychologia.2017.08.025
Received 1 March 2017; Received in revised form 18 August 2017; Accepted 22 August 2017
Available online 26 August 2017
0028-3932/

a null distribution based on permutations of real data offers a rather literal way to compute *p*-values: the *p*-value is literally the probability of observing a particular outcome if there were no relationship between the behavioral scores and lesion patterns (i.e., random permutations). A null distribution of the test statistic can be built based on permutations of real data and used to reject voxels where the true analysis does not sufficiently differ (e.g., $p > 0.01$) from the permutation-based null distribution to warrant rejecting the null hypothesis (e.g., Kimberg et al., 2007; Rorden et al., 2007).

Controlling the probability of making one or more false positive errors is based on the idea that each test is critically related to the researcher's interpretation or inference, thus, a single false positive could potentially undermine the inference and needs to be controlled. Voxel-level FWER does not align with VLSM interpretation, which never depends on a single voxel. The misalignment between standard FWER correction and VLSM interpretation makes standard voxel-level FWER correction unnecessarily conservative: if no inferences are made based on a single voxel, then a single false positive voxel cannot be responsible for an invalid inference about lesion-symptom relations.

The general approach of permutation-based correction for multiple comparisons can be implemented in many different ways, depending on what aspect of the results is to be controlled. Instead of controlling the rate of a single false positive voxel, permutation-based FWER can be used to set a minimum cluster size, thus controlling the rate of a single false positive *cluster* of voxels. Setting a minimum cluster size is a common "clean-up" step in neuroimaging data analysis; the addition of a principled strategy for selecting the minimum cluster size is the critical component that turns this into a statistical correction method. The permutation-based strategy is to set a voxel-wise cluster-forming threshold (e.g., $p < 0.0001$), then use permutations to determine the null distribution of cluster sizes that pass this threshold, and use that distribution to set a minimum cluster size (e.g., Nichols and Holmes, 2002). This approach is appealing because stroke lesions are inherently contiguous and VLSM interpretation tends to focus on clusters.

Another technique for controlling the rate of false positives is False Discovery Rate (FDR), which quantifies the proportion of above-threshold results that can be expected to be false positives (Genovese et al., 2002). That is, at FDR threshold $q = 0.05$, 5% of above-threshold voxels are expected to be false positives, which is likely to be substantially more than one voxel but not likely to affect interpretation of the overall patten (for a clear description see Bennett et al. 2009). FDR is widely used for analysis of functional neuroimaging data and VLSM, however, we have encountered informal criticism that FDR is inappropriate for VLSM. FDR is certainly less conservative than FWER (see also Rorden et al., 2007), but that is true by design – FDR is designed to allow a small percentage of false positive voxels, whereas FWER aims to make it unlikely that there is even a single false positive voxel – and we are not aware of any published analysis showing that FDR incorrectly quantifies the rate of false positive voxels in VLSM.

Correction for multiple comparisons is an attempt to manage variability, but it cannot remove all of the noise and leave all of the signal. Either some noise will get left behind or some of the signal will be removed. That is, there is an inherent trade-off between false positives and false negatives; incorrectly generalizing a result and overlooking a generalization that is warranted. By convention, data analysis requires setting a threshold to identify results that warrant rejection of the null hypothesis. There is a substantial price associated with adopting the conservative position that the probability of even a single false positive voxel needs to be controlled: VLSM analysis is based on a single data point per participant (each participant only has one lesion and only one deficit profile) and sample sizes are often limited by the practical challenges of recruiting and testing large numbers of participants with the targeted neurogenic deficits. This price is further exacerbated by publication bias: studies that meet the statistical threshold may be published, but studies that fall short are relegated to the "file drawer", leaving a biased scientific literature. Publication bias also encourages various forms of "p-hacking" or "researcher degrees of freedom", in which researchers try alternative analysis strategies (excluding certain "outlier" participants, transforming scores, etc.) until they find one that surpasses the statistical threshold. The result is a report that appears to use rigorous statistical methods, but the actual rate of false positives far exceeds the nominal *p*-value (e.g., Simmons et al., 2011; Nosek et al., 2012; Gelman and Loken, 2014). In addition to statistical soundness, the analytical strategy should allow researchers to transparently report their observations and the strength of the evidence that supports their conclusions.

The present study investigated two permutation-based methods of correcting for multiple comparisons in VLSM. The next section describes our investigation of using permutations to determine a minimum cluster size. Our analyses found that this approach produces consistent spill-over into neighboring regions (i.e., the identified region extends well beyong the boundaries of the true lesion-symptom relation), making it not well-suited to identifying brain-behavior relationships. Although some spill-over is a necessary consequence of high spatial correlations among neighboring voxels inherent to stroke lesion data, cluster-based correction was substantially more susceptible to this problem than standard voxel-level FWER correction. The subsequent section describes a generalization of the permutation-based FWER correction approach that captures some of the inferential advantages of FDR and cluster-based correction without making parameteric assumptions about the data. This approach makes it possible to balance control of false positives against risk of false negatives, and to transparently report results in a way that allows others to evaluate the evidence. We also compare this approach with the parametric FDR method and describe conditions under which FDR may produce misleading results. The final section of this report summarizes our findings and conclusions, and discusses future directions.

2. Minimum cluster size

Using permutations to determine a minimum cluster size proceeds as follows: (1) permute behavioral data and conduct VLSM analysis, (2) apply a pre-set cluster-forming threshold for each voxel (e.g., $p < 0.0001$), (3) compute size of largest supra-threshold voxel cluster, (4) repeat steps 1–3 many times to build up a null distribution of supra-threshold cluster sizes (e.g., Nichols and Holmes, 2002). This null distribution is the distribution of largest cluster sizes that are observed when there is no relationship between deficit scores and lesion location. Clusters from the original (true) VLSM analysis that are larger than 95% of the null distribution of cluster sizes are taken to reflect true lesion-symptom associations (for examples of application to VLSM see Pillay et al. (2014, 2017), Binder et al. (2016), Mirman et al. (2015a)). This strategy involves two separate thresholds: the first is a pre-set voxel-level *p*-threshold; the second is a permutation-based cluster size threshold. This strategy has two intuitively appealing properties. First, strokes and other neurological disorders tend to produce spatially contiguous lesions, resulting in high spatial correlations between the lesion status of neighboring voxels. Using permutation to determine a null distribution of cluster sizes intuitively controls for this spatial correlation and produces a minimum cluster size threshold that should not be observed by chance. Second, it is typical for interpretation of VLSM (and other neuroimaging) results to focus on clusters, so correcting at the cluster level (rather than the voxel level) makes this statistical strategy more closely aligned with the interpretation strategy. In the following analyses we examine this permutation-based cluster-size correction strategy for detecting true lesion-symptom relations.

2.1. Data

The lesion maps were from 124 participants with aphasia following left hemisphere stroke confirmed by computed tomography (CT) or magnetic resonance imaging (MRI) and collected as part of a larger,

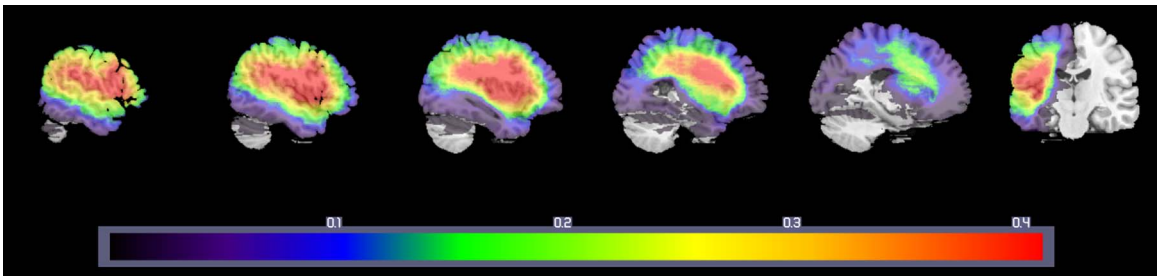

Fig. 1. Lesion overlap map for 124 left hemisphere stroke cases included in the present analyses. Hotter colors indicate that a larger proportion of the participant sample had lesions in that area.

ongoing project investigating the anatomical basis of psycholinguistic deficits in post-acute aphasia.[1] The structural data were based on 108 research scans (65 MRI and 43 CT) and 16 clinical scans (5 MRI and 11 CT). Lesions imaged with MRI were manually segmented on the structural image by a trained technician and reviewed by an experienced neurologist, then registered first to a custom template constructed from images acquired on the same scanner, and then from this intermediate template to the Montreal Neurological Institute space "Colin27" volume. Lesions imaged with CT were drawn by the experienced neurologist directly onto the Colin27 volume, after rotating (pitch only) the template to approximate the slice plane of the patient's scan. Fig. 1 shows the lesion overlap map for these 124 lesion maps, which have been used in VLSM analyses reported elsewhere (Mirman et al., 2015a, 2015b). Following best practices in VLSM analysis, only voxels where more than 10% of participants had lesions were included (338,831 voxels) and the analyses controlled for overall lesion volume (Sperber and Karnath, 2017; Zhang et al., 2014).

In order to have a deficit score with a known neural correlate, we calculated the percent damage in two brain regions that are widely-studied and frequently damaged in middle cerebral artery stroke aphasia: BA 45 and BA 39. An effective statistical correction strategy should approximately identify these areas; that is, damage in BA 45 should be the "neural correlate" of percent damage in BA 45. Effects in these areas may be vulnerable to mis-localization in VLSM (Mah et al., 2014), though this issue is substantially reduced in more realistic analyses that control for etiology and overall lesion size, and only test voxels that are affected in a reasonable number of participants (Sperber and Karnath, 2017), as we do here.

2.2. Analysis strategy

For each of the simulated deficit (percent damage) scores, we conducted a basic VLSM, applied a pre-set threshold, then calculated the size of the largest supra-threshold voxel clusters. We then repeated this analysis 1000 times, permuting the deficit scores for each repetition to create a random association between the scores and lesion profiles. The cluster sizes from the permutations were used to set a 95% threshold (i.e., larger than 95% of permutation-based clusters) for the original VLSM data. These analyses were carried out in R version 3.3.1 (R Core Team, 2016) using the ANTsR package version 0.3.3 (Avants et al., 2016) and the LESYMAP package (Pustina et al., 2017).

Four different pre-set thresholds were tested within the same set of 1000 permutations: 0.05, 0.01, 0.001, 0.0001. This covers the range from the most permissive threshold (0.05) to a reasonably conservative threshold (0.0001) for initially identifying voxels for subsequent cluster size correction. The more permissive thresholds will allow more voxels into the cluster size calculation, which should produce larger clusters. Therefore, there should be a positive correlation between the pre-set *p*-threshold and the permutation-based cluster size threshold. This positive correlation is an inverse strictness relationship: more permissive *p*-thresholds produce more conservative cluster size thresholds. One motivation for this study was to examine how one might balance these inversely related factors for optimal VLSM interpretation and inference.

2.3. Results

As expected, there was a positive relationship between cluster size threshold (95th percentile of maximum cluster sizes across 1000 permutations) and *p*-threshold (Fig. 2): more permissive *p*-thresholds allow more voxels into the cluster analysis, thus producing larger clusters. Indeed, the relationship is almost perfectly linear in the log-log plot in Fig. 2. The next stage was evaluating how well this method recovers the true neural correlates for each deficit score. The two less conservative *p*-thresholds produced extremely large cluster thresholds: more than 20,000 voxels at $p < 0.05$ and more than 6500 voxels at $p < 0.01$. Any clusters of that size or larger would not be neuroanatomically specific enough to provide useful insights into lesion-symptom relationships.

The top row in Fig. 3 shows the results of permutation-based cluster size correction (at voxel-wise $p < 0.001$ and $p < 0.0001$, and family-wise cluster size $p < 0.05$) for simulated deficit scores of percent damage in BA 45 and BA 39. The identified region expands beyond the bounds of the true region, covering an area that is approximately twice the size of the Brodmann Area where percent damage was used as the behavioral score. For comparison, the middle row in Fig. 3 shows the same VLSM analyses thresholded using permutation-based FWER correction ($p < 0.05$) based on the maximal test statistic in each permutation, and the bottom row shows the actual regions as defined in the Brodmann Area atlas. The FWER correction did a substantially better job of identifying the critical regions, with about 25–50% fewer voxels surviving correction.

2.4. Discussion

We explored the use of a permutation-based approach to determine a minimum cluster size threshold for statistical correction of VLSM. This approach is adapted from analysis of functional neuroimaging data (Nichols and Holmes, 2002) and has been previously used in VLSM (Pillay et al., 2014, 2017; Binder et al., 2016; Mirman et al., 2015a). Using structural lesion data from 124 participants with left hemisphere stroke, we constructed deficit scores using percent damage in BA 45 and BA 39. As expected, there was a positive relationship between pre-set *p*-threshold and the resulting cluster thresholds: more permissive *p*-thresholds allow more voxels into the cluster analysis, thus producing larger clusters. As a result, a less conservative cluster-forming *p*-threshold will produce larger - and less anatomically precise - cluster(s) in the final results.

This cluster-based correction correctly identified the critical BA regions, but the supra-threshold clusters extended well beyond the boundaries of the correct BA regions. This pattern suggests that the permutation-based cluster size approach can correctly reject cases in

[1] That project was funded by National Institutes of Health grant R01DC000191 to Myrna F. Schwartz and we are grateful to Dr. Schwartz and her team for sharing these data with us to make these analyses possible.

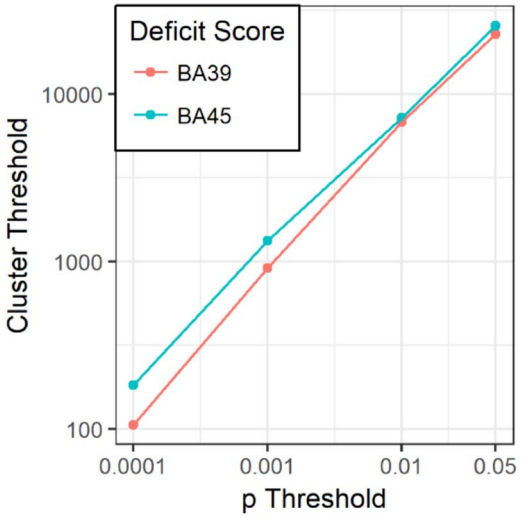

Fig. 2. Cluster size thresholds based on largest cluster from each permutation at each *p*-threshold. Note that both axes are logarithmically scaled.

based method, adjacent voxels have a much weaker threshold than non-adjacent voxels, thus exacerbating the spill-over effect.[2] That is, the cluster-based method is least able to reject false positives in the area where they are most likely to occur – adjacent to the critical (true positive) region.

Although permutation-based FWER out-performed cluster-based correction in these analyses, percent damage is a very strong relationship and, as discussed in the Introduction, this FWER method is very conservative because it controls the possibility of a single false-positive voxel. This single-voxel standard does not align with how VLSM results are interpreted and this conservatism carries real costs for scientific progress. In the next section we explore a generalization of this approach that allows balancing false positives against false negatives and transparently reporting the evidence. In the process, we also evaluate the FDR correction method against permutation-based FWER methods.

3. Continuous permutation-based FWER

The standard permutation-based FWER correction method proceeds as follows: (1) permute behavioral data and conduct VLSM analysis, (2)

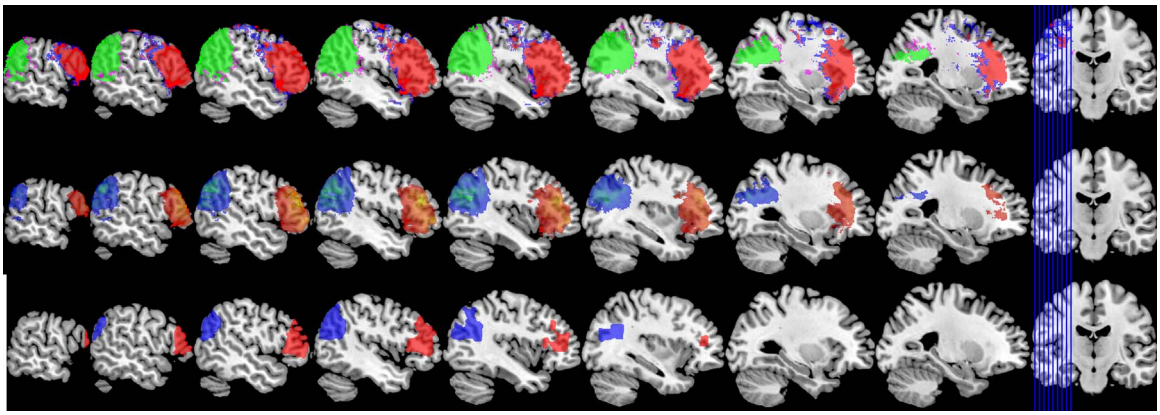

Fig. 3. Results of VLSM analysis of percent damage to BA 45 and BA 39. Top row shows results thresholded using permutation-based cluster size correction ($p < 0.05$). The primary clusters (red for BA45, green for BA39) are based on voxel-wise $p < 0.0001$, the additional voxels (blue for BA45, purple for BA39) are included when the voxel-wise threshold is $p < 0.001$. Middle row shows results thresholded using voxel-wise permutation-based FWER at $p < 0.05$ (BA45 in red-yellow, BA39 in blue-green). Bottom row shows the true regions: BA45 and BA39 as defined in the Brodmann Area atlas (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

which there is no consistent relationship between behavioral score and lesion location, but it appears to be insufficiently spatially specific when a true relationship exists. If a lesion-symptom relationship does exist, this method will detect the correct region, but spatially contiguous regions will also be included in the critical cluster. To some degree, this spill-over is a necessary consequence of the high spatial correlation inherent in stroke lesion data, and the quite strong relationship between our simulated deficit scores (percent damage) and their corresponding brain regions. Even the more conservative permutation-based FWER method had some spill-over into adjacent regions, although it was substantially less than the cluster-based method. The cluster-based method may be particularly vulnerable to spill-over resulting from spatial correlations precisely because it looks for clusters of voxels. Consider a voxel that is just outside the critical region, but that has a moderately high spatial correlation with voxels inside the critical region; that is, when voxels inside the critical region are lesioned, it is also often lesioned, and when they are not, it is also usually not lesioned. This adjacent voxel will have a moderate lesion-symptom association due to its correlation with the critical voxels, but will it survive correction? Under the FWER method, that association must be stronger than 95% of the maximal associations in the permutations – the same as any other voxel in the analysis. Under the cluster-based method, it only has to be stronger than the pre-set cluster-forming *p*-threshold. In other words, unlike the FWER method, under the cluster-identify the maximal test statistic (typically, the most extreme *t*-value), (3) repeat steps 1 and 2 many times to build a null distribution of maximal *t*-values, (4) compute the *n*-th percentile of that null distribution to determine a threshold for the test statistic, which corresponds to *n*% of the permutations having 0 voxels that exceed this threshold (Rorden et al., 2007). A typical value of *n* is 95, which produces a FWER-corrected $p < 0.05$: less than 5% of the permutations had even a single voxel that exceeded this *t*-threshold.

This approach controls the rate of single-voxel false positives, but it can be generalized to multi-voxel false positives by focusing on the *v*-th most extreme test statistic.[3] The standard strategy is the special case when $v = 1$, thus using the most extreme voxel-wise test statistic from each permutation, and controlling the rate of 1 false positive voxel. If, for example, $v = 10$, one would similarly use the 10th most extreme voxel-wise test statistic from each permutation, and control the rate of up to 10 false positive voxels. An example is shown in Fig. 4 where the left panel shows the sorted *t*-values from the first 10 permutations and the right panel shows permutation-based *t*-value distributions and 95%

[2] An exception is the special case where the pre-set cluster-forming *p*-threshold is equivalent to the FWER threshold, but that would render the cluster-based correction unnecessary.

[3] In the statistical literature, controlling for some number *k* of false positives is also known as *k*-FWER control (e.g., Romano and Wolf, 2007).

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
1	4.81	3.80	3.93	4.23	3.87	4.23	4.68	3.45	4.83	4.21
2	4.70	3.57	3.93	4.23	3.87	4.21	4.51	3.44	4.78	4.06
3	4.55	3.57	3.93	4.23	3.60	4.19	4.50	3.40	4.61	3.97
4	4.52	3.54	3.93	4.23	3.60	4.19	4.41	3.35	4.59	3.91
5	4.51	3.49	3.89	4.23	3.57	4.19	4.37	3.34	4.49	3.91
6	4.50	3.49	3.87	4.23	3.57	4.19	4.17	3.31	4.49	3.88
7	4.50	3.42	3.81	4.23	3.57	4.19	3.97	3.31	4.47	3.85
8	4.50	3.42	3.79	3.98	3.57	4.19	3.92	3.29	4.46	3.85
9	4.49	3.42	3.77	3.98	3.57	4.19	3.92	3.29	4.45	3.85
10	4.48	3.40	3.77	3.98	3.55	4.19	3.88	3.29	4.42	3.82
11	4.41	3.29	3.76	3.93	3.49	4.19	3.87	3.27	4.42	3.79
12	4.38	3.26	3.73	3.91	3.48	4.16	3.81	3.27	4.41	3.78
13	4.36	3.26	3.73	3.91	3.47	4.14	3.75	3.24	4.41	3.78
14	4.33	3.25	3.73	3.89	3.47	4.11	3.70	3.23	4.37	3.77
15	4.33	3.24	3.71	3.74	3.44	4.11	3.65	3.23	4.36	3.77
16	4.33	3.23	3.69	3.59	3.44	4.11	3.56	3.22	4.31	3.71
17	4.32	3.22	3.67	3.59	3.42	4.11	3.56	3.22	4.30	3.71
18	4.32	3.21	3.62	3.51	3.40	4.11	3.53	3.21	4.28	3.70
19	4.27	3.21	3.61	3.51	3.40	4.08	3.52	3.21	4.27	3.70
20	4.27	3.20	3.61	3.51	3.39	3.94	3.52	3.21	4.26	3.70
...										
100	2.90	2.84	2.82	2.90	3.32	2.18	3.74	3.17	2.97	2.41
...										
1000	2.45	2.39	2.42	2.51	2.75	1.80	2.98	2.81	2.60	2.04

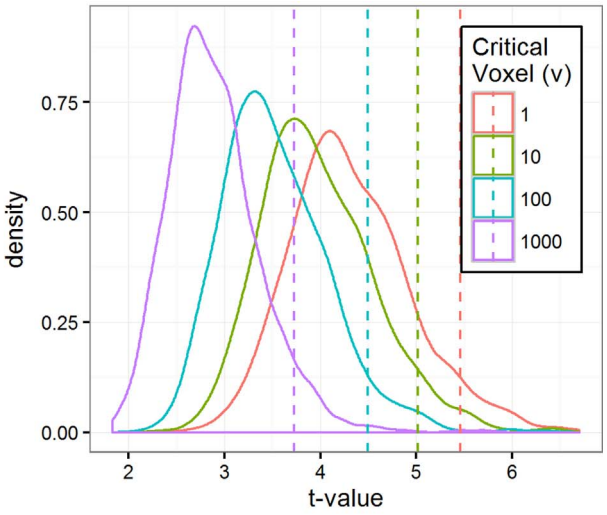

Fig. 4. Example of continuous FWER threshold calculation. Left panel shows the top sorted *t*-values from the first 10 permutations. The red box highlights the standard $v = 1$ permutation *t*-values, which produce the red distribution in the right panel and the 95% threshold indicated by the red dashed line. The green box highlights the $v = 10$ permutation *t*-values, which produce the green distribution in the right panel and the 95% threshold indicated by the green dashed line. Analogous *t*-values, distributions, and 95% thresholds are also shown for $v = 100$ (blue) and $v = 1000$ (purple) (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

thresholds at $v = 1$ (red), $v = 10$ (green), $v = 100$ (blue), and $v = 1000$ (purple). A single set of permutations produced the set of possible 95% thresholds at different critical voxel (v) values, reflecting the expected number of false positive voxels at the corresponding t threshold. Not surprisingly, that t threshold decreases as the v value increases, but the relationship remains the same as the standard FWER case: the v value specifies the maximum number of voxels that exceeded the corresponding t-threshold in 95% of the permutations. That is, a reasonable upper bound on number of false positive voxels at that t threshold. We refer to this extension of the standard permutation-based FWER as *continuous* permutation-based FWER because it uses the same permutation-based FWER strategy but allows values of $v > 1$.

Since interpretation of VLSM results typically relies on a large set of voxels, that interpretation is unlikely to be affected by a small (but >1) number of false positive voxels. This is not to say that, for example, $v = 10$ should be adopted instead of $v = 1$; rather, this generalization of FWER allows investigators to assess the strength of their evidence in a flexible way and to report that assessment in a way that allows readers (and reviewers) to evaluate the claims.

Continuous permutation-based FWER is somewhat similar to the false discovery rate (FDR) approach in that both allow multiple false positive voxels and quantify that rate of false positives. However, the two methods differ in two important ways. First, FDR is designed to control the *proportion* of supra-threshold voxels that are expected to be false positives (this proportion is usually reported as the q-value), whereas the FWER approach quantifies the *number* of possible false positive voxels, which may be a high or low proportion of supra-threshold voxels. Second, continuous FWER is permutation-based, which means that (unlike FDR) it makes no assumptions about distributions of data or test statistics. The latter property is important because lesion data may violate the assumptions of FDR severely enough to make FDR unreliable for VLSM. Here we report results from application of this continuous permutation-based FWER approach in several contexts. In addition, we used the permutation data to evaluate whether the nominal FDR q-value correctly quantifies the proportion of supra-threshold voxels that are expected to be false positives.

3.1. Analysis strategy

We conducted analogous analyses on three sets of data. The first was the same full dataset used in the cluster size threshold analyses above: 124 left hemisphere stroke cases with two simulated behavioral scores, percent damage in BA 45 and BA 39. This provides a relatively large data set with a known correct outcome. The second was randomly sampled sub-sets of these data to examine how continuous FWER and FDR perform for smaller data sets. We used 50 random half-samples (N = 62) and 50 random quarter-samples (N = 31). The third data set was speech recognition deficit data from 99 left hemisphere stroke cases reported in a recent article (Mirman et al., 2015a). This data set provided an opportunity to test continuous FWER and FDR in the context of real behavioral data where the outcome was relatively uncontroversial: deficits in speech perception and spoken word recognition should be associated with lesions in left superior temporal lobe regions. Although not quite as certain as using simulated behavioral scores, this outcome is very strongly expected and using real behavioral data allowed us to test these statistical methods in the context real-world variability.

For each analysis, we conducted standard VLSM analysis and computed continuous permutation-based FWER 95th percentile thresholds at $v = 1, 10, 100, 1000$ based on 1000 permutations. That is, t-value thresholds where 95% of the permutations had fewer 1, 10, 100, or 1000 supra-threshold voxels. We then computed the number of voxels in the original VLSM that had t values greater than the t threshold at each v threshold, which is the number of FWER-corrected ($p < 0.05$) voxels at each v threshold. The v threshold and the number of supra-threshold voxels were then used to compute an *effective* q value: the proportion of supra-threshold voxels that can be expected to be false positives based on the v value. For example, if 500 voxels survived the correction at $v = 10$, that would correspond to $q = 10/500 = 0.02$. To evaluate the standard FDR method, we computed a FDR-corrected t-threshold using the effective q value from the continuous FWER analysis. If FDR correctly estimates the nominal proportion of false positives (q), then the t threshold produced by FDR should (approximately) match the permutation-based t-threshold with the corresponding effective q value. That is, the effective q from FWER was aligned with the q for FDR, so the t-threshold computed by FDR should match the t-threshold computed by FWER. Any consistent discrepancies in critical t values will provide insight into whether FDR is anti-conservative or overly conservative. These analyses were carried out in R version 3.2.4 (R Core Team, 2016) using the ANTsR package version 0.3.3 (Avants et al., 2016) and the FDR implementation in the AnalyzeFMRI package version 1.1–16 (Bordier et al., 2011). A basic implementation of the continuous permutation-based FWER correction method is available at https://gist.github.com/dmirman/

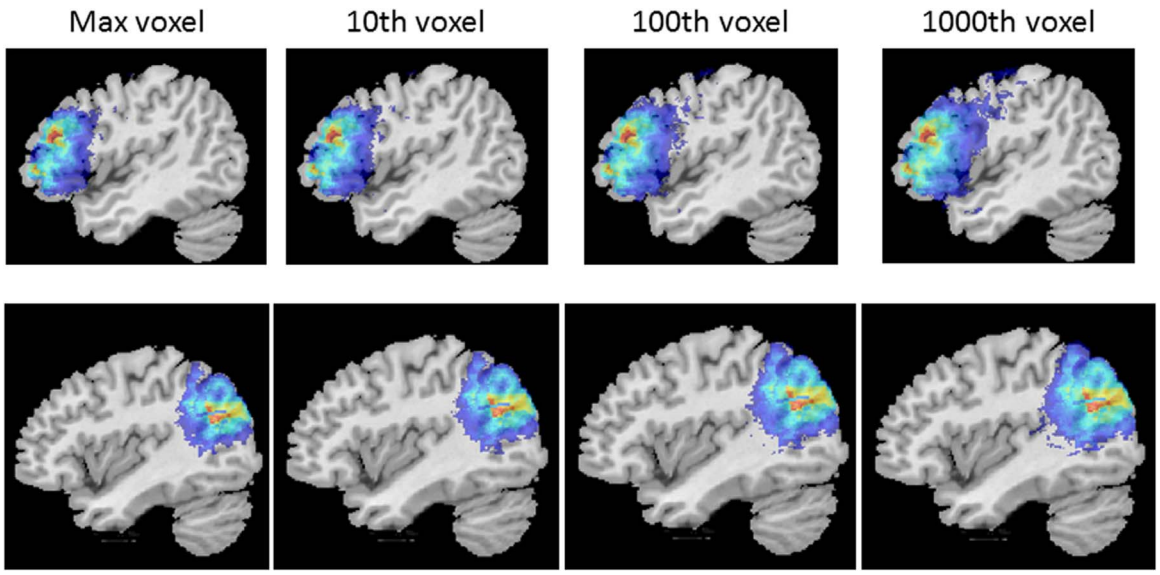

Fig. 5. Results of VLSM analysis of percent damage to BA 45 (top row, at $x = 45$) and BA 39 (bottom row, at $x = 50$). Thresholded using permutation-based continuous FWER at $p < 0.05$.

05a92e0e9e0027f6fe6e528c648143d7 and is included in the lesymap package (https://dorianps.github.io/LESYMAP/), which provides access to a wider variety of lesion-symptom mapping methods.

3.2. Results

3.2.1. Simulated scores, full sample

Fig. 5 shows the VLSM results corrected at $p < 0.05$ using continuous permutation-based FWER with $v = 1, 10, 100, 1000$ voxels. The first column in Fig. 5 corresponds to the standard FWER correction, which is also shown in the right column of Fig. 3. The other columns show that (unsurprisingly) the supra-threshold region increases as the number of allowed false positive voxels increases.

The left panel of Fig. 6 shows the relationship between the voxel number threshold (v) in the continuous FWER correction and the resulting effective q value. The points in the bottom left corner correspond to the standard, $v = 1$, FWER correction. As v was increased, there was a corresponding increase in effective q, the proportion of supra-threshold voxels that can be expected to be false positives. This relationship between v and effective q was essentially linear (on log-log scale) and virtually identical for the BA45 and BA39 test cases. Note that even at the most lenient threshold, $v = 1000$, the effective q value was still quite low (0.011 for BA45; 0.016 for BA39), presumably because of the very strong relationship between percent damage in a BA (the simulated deficit score) and lesion in that region.

The right panel of Fig. 6 shows the relationship between the critical t-value (i.e., the corrected t-threshold) as computed by continuous FWER and by FDR. The FDR-corrected t-value was computed using the effective q from the left panel of Fig. 6. Since effective q is a non-parametric estimate of the true rate of false positive voxels at the corresponding t-threshold, if FDR works as intended, then it should produce t-thresholds that are very similar to those computed by the non-parametric continuous FWER method. This is the pattern shown in the right panel of Fig. 6: virtually identical critical t-thresholds computed by FDR correction and by continuous FWER correction, for both BA45 and BA39 scores. That is, the permutations confirm that, at $q = 0.01$, FDR correction accurately produced a critical t-value such that up to 1% of supra-threshold voxels could be expected to be false positives.

This is an encouraging result for application of FDR to VLSM data because it shows the q-value correctly quantifies the proportion of false positive voxels. However, this is an ideal scenario in at least two ways: (1) a very strong relationship between simulated score (percent damage in BA 45 or BA 39) and lesion location, and (2) a relatively large sample size (N = 124). To evaluate the contribution of this second factor we conducted further analyses of these same data but using smaller subsamples of the data.

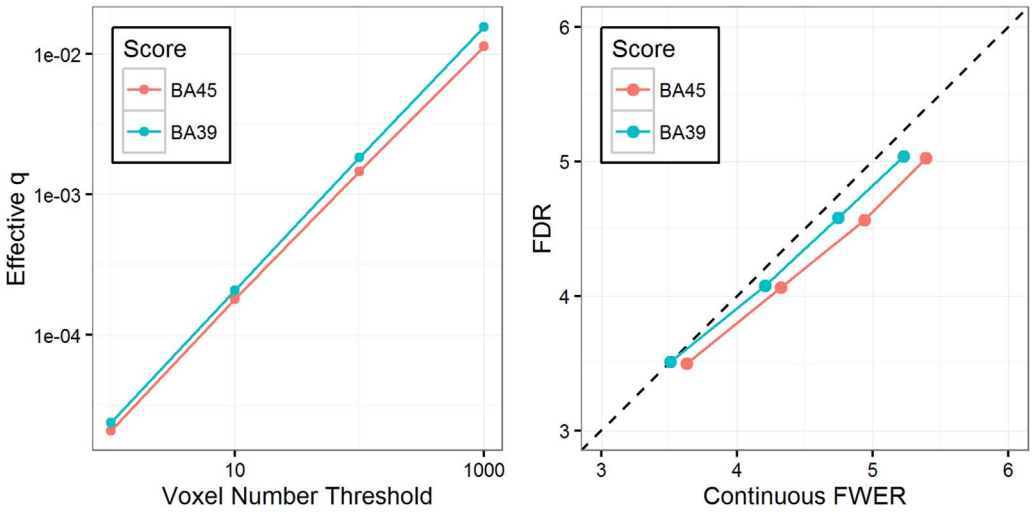

Fig. 6. Left: Relationship between voxel number threshold (v) and the proportion of supra-threshold voxels that can be expected to be false positives (effective q). Right: Relationship between critical t-values (t-thresholds) determined by continuous FWER correction and FDR correction.

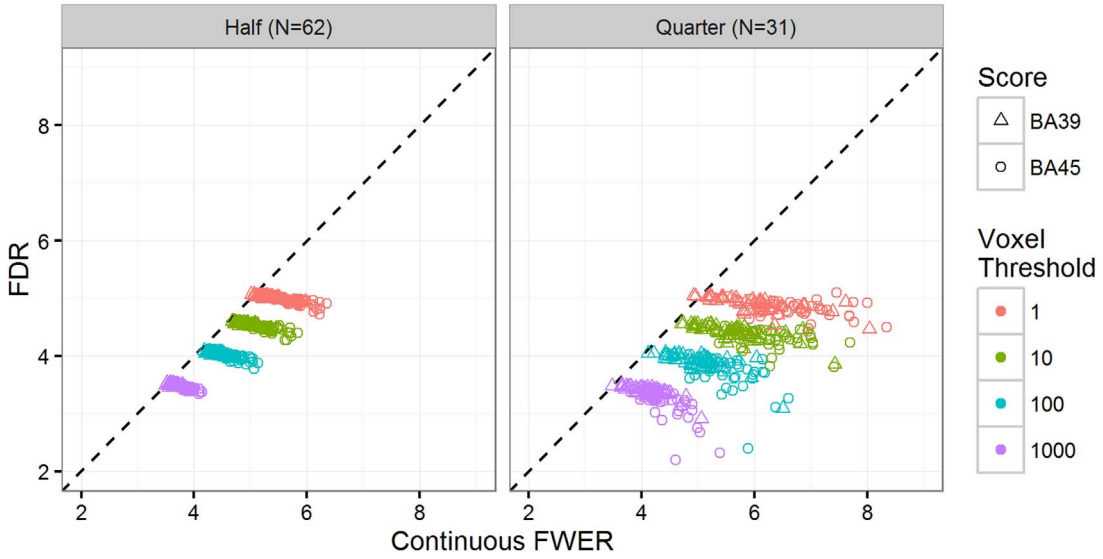

Fig. 7. Relationship between critical *t*-values (*t*-thresholds) determined by continuous FWER correction and FDR correction for 50 randomly selected half-samples (left panel) and 50 quarter-samples (right panel). Deficit scores are percent damage in BA39 (triangles) or BA45 (circles). Each point represents one of the random sub-samples.

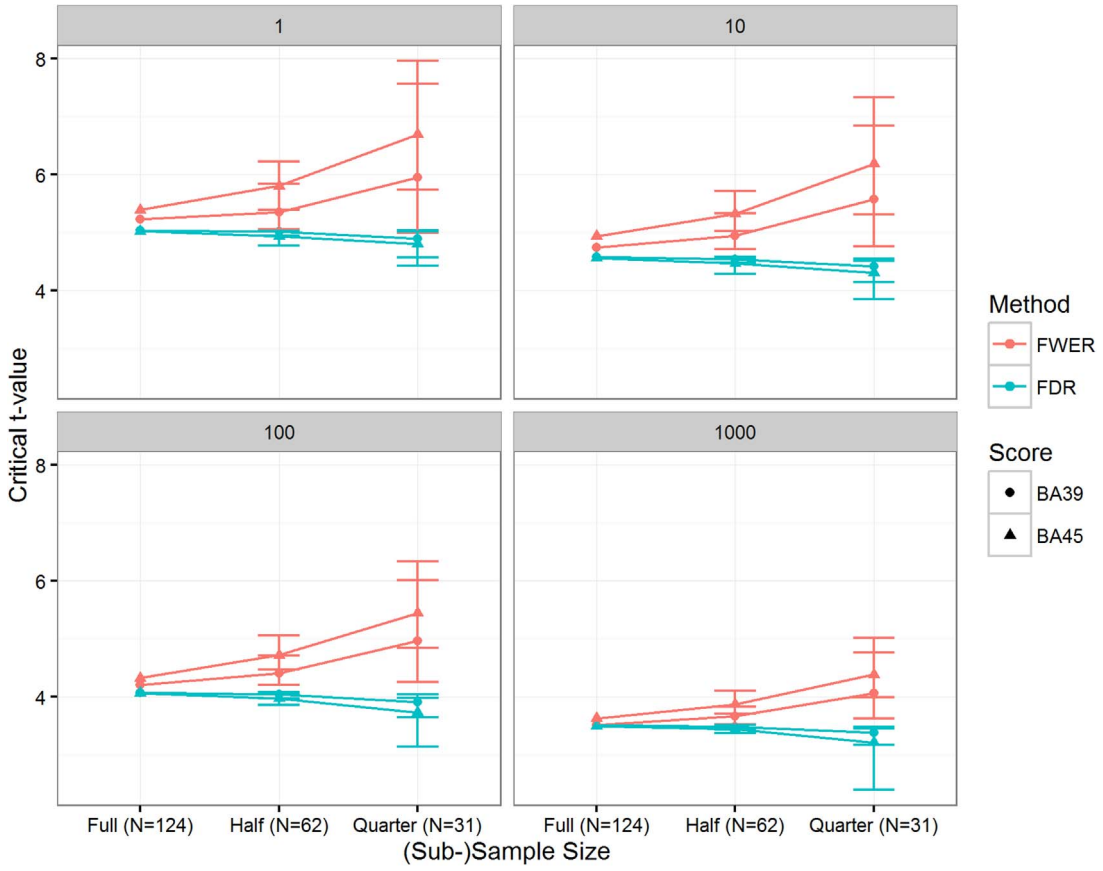

Fig. 8. Critical *t*-values (*t*-thresholds) determined by continuous FWER correction and FDR correction as a function of sample size. Deficit scores are percent damage in BA39 (circles) or BA45 (triangles). Error bars represent 95% confidence intervals.

3.2.2. Simulated scores, sub-samples

The initial analysis of 124 participants constitutes a fairly large sample size by VLSM standards. More modest sample sizes (e.g., 40–60) are far more common and many studies report even smaller samples (e.g., 20–40). Smaller sample sizes are more likely to (more severely) violate assumptions of FDR, so, although FDR worked as intended for the full N = 124 sample, it may not be robust at smaller sample sizes. In particular, the spatial coherence of lesions means that the voxel-wise tests violate the test independence assumption and symptom scores are often non-normal – both of these problems will tend to be more severe for smaller sample sizes. However, FDR is robust to some degree of assumption violation (Groppe et al., 2011a, 2011b), so it may produce approximately correct results even under these conditions. To evaluate this, we repeated the comparison of continuous FWER and FDR using 50 half (N = 62) and 50 quarter (N = 31) random sub-samples of the full data set. Fig. 7 shows scatterplots of the critical *t*-values based on

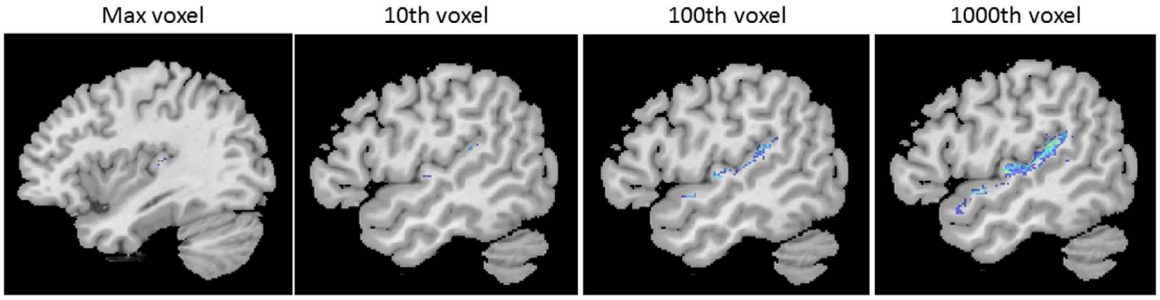

Fig. 9. Results of VLSM analysis of Speech Recognition scores. Thresholded using permutation-based continuous FWER at $p < 0.05$, panels show results at different v thresholds: $v = 1, 10, 100, 1000$. Note: at $v = 1$, the small set of supra-threshold voxels was located more medially than the main group of voxels in the other panels, so the left panel shows the results at $x = 53$ while the other panels are at $x = 40$.

continuous FWER (at each of the four v thresholds) and the corresponding FDR critical t-value computed using the effective q value. The dashed line represents exact equivalence between continuous FWER and FDR, which was approximately true for the full sample. For these smaller samples, the FDR method tends to produce a substantially less conservative critical t-value threshold for each effective q value. There was substantial variability in how far the FDR threshold deviated from the continuous FWER threshold, with some sub-samples showing fairly close correspondence (as was observed for the full sample), but most falling short of that. Comparing the half-sample and quarter-sample data (left vs. right panels in Fig. 7) shows that both the departure of FDR from continuous FWER and the variability of their relation became more extreme for smaller samples.

The interpretation of continuous FWER is transparent, so this discrepancy between the methods represents a problem for FDR. For example, for the first quarter-sample in the BA 45 analysis, the (traditional) $v = 1$ FWER threshold produced a critical t-value of 6.0 and 5904 voxels had t-values above that threshold. That is, permutation analysis indicates that only 1 out of those 5904 can be expected to be a false positive, which is an effective $q = 1/5904 = 0.00017$. Applying FDR to these data with $q = 0.00017$ produced a critical t-value of 4.9 and 11,429 voxels passed that t-threshold. According to FDR, only 0.017% of those 11,429 voxels are expected to be false positives, which is approximately 2 voxels. However, the permutation data reveal that, if there were no relationship (i.e., if the null hypothesis were true), then approximately 100 voxels could be expected to exceed a critical t-value of 4.9; about 50 times more than the nominal rate implied by the q-value.

Sample size appears to have different effects on FWER correction and FDR correction (see Fig. 8). FDR-corrected thresholds are relatively constant across sample sizes, but FWER-corrected thresholds increase as sample size becomes smaller. This pattern may arise because violations of assumptions have a bigger effect in smaller samples. In particular, because stroke lesions are spatially contiguous, in smaller samples, many neighboring voxels become indistinguishable – their spatial correlation is 1.0. For example, voxels were nearly unique in the full (N = 124) sample – the average size of patches of equivalent voxels was only 1.7 voxels. In the half-samples (N = 62), it was 4.5–5 voxels, and in the quarter-samples (N = 31) it was approximately 25 voxels (for a more thorough evaluation using largely the same dataset see Pustina et al. 2017). The FDR correction is influenced by the degree of skew toward small p-values (or large test statistics) in the observed *voxel-level* test results. In the presence of a true signal (as in the simulations here), large patches of equivalent voxels may enhance this skew, thus overestimating the signal strength and producing a somewhat anti-conservative FDR-corrected threshold.

The continuous FWER approach offers a permutation-based alternative that incorprates the greater statistical power of FDR (i.e., quantifying the expected upper bound of false positives) while using the distributional properties of both the behavioral data and the lesion data to naturally account for the spatial correlation and other distributional properties of the data. This combination of increased power to detect true effects while accurately quantifying false positives is particularly important for the many VLSM studies that have sample sizes in the 30–60 range: FDR may produce anti-conservative results for these smaller sample sizes, but the conservative standard ($v = 1$) FWER correction may relegate these studies to the file drawer. We return to this issue further after examining a real deficit example.

3.2.3. Speech recognition scores

All of the preceding analyses used simulated deficit scores that had rather strong lesion-symptom relations. A strong signal is easy to detect and may obscure weaknesses of a statistical method, so it is important to test statistical methods with more realistic data. Adding noise to simulated lesion-symptom relations would effectively weaken them, but real lesion-symptom relations are not simply randomly noisy, so there is no guarantee that adding random noise would capture the ways that real lesion-symptom relations differ from simulated ones. However, using real deficit data is somewhat risky because the true lesion-symptom relation is not known. To mitigate this concern, we chose a relatively uncontroversial case: composite speech recognition deficit scores determined by a factor analysis of data from 99 individuals with left hemisphere stroke (Mirman et al., 2015a). These scores primarily reflect phoneme discrimination and auditory lexical decision performance (for details see Mirman et al. (2015a, 2015b)) and it is quite well-established that these tasks primarily engage left superior temporal lobe structures (e.g., Hickok and Poeppel, 2015; DeWitt and Rauschecker, 2012). As a result, this dataset allows us to investigate how continuous FWER and FDR would work in a real VLSM context while being fairly confident about what the correct result should be.

Fig. 9 shows the VLSM results after continuous FWER correction at $v = 1, 10, 100, 1000$. As expected, the identified region is in the superior temporal lobe and, as in the simulated scores analyses, a more relaxed v threshold produces a larger supra-threshold region. At the standard, $v = 1$, threshold, the FWER corrected critical t-value was 5.45 and 57 voxels passed this threshold. On one hand, this is a positive result: it is unlikely ($p < 0.05$) that even one of those 57 voxels is a false positive, so we should feel confident about interpreting those 57 voxels as being critically important for speech recognition. On the other hand, those 57 voxels are virtually invisible in the figure (left-most panel in Fig. 9; even the 261 voxels that passed the $v = 10$ threshold of $t = 5.04$ are hard to see) and it seems unlikely that editors, reviewers, and readers would be convinced by a 57-voxel result (or even a 261-voxel result). Such a small cluster might even be within the margin of error of the lesion segmentation algorithms and the warping algorithm used to align individual lesion maps to a common template for analysis. Relaxing the v threshold reveals an easier to interpret result. For example, at $v = 100$, the critical t-value was 4.42 and 1527 voxels passed this threshold. Up to 100 of those 1527 voxels (6.5%) can be reasonably expected to be false positives, but that probably would not affect how one would interpret the result in the $v = 100$ panel of Fig. 9.

This is not to say that the threshold should be moved from $v = 1$ to

$v = 100$ – there may be circumstances where 100 voxels (or 6.5% of the supra-threshold voxels) *would* affect the interpretation of VLSM results. A flexible v threshold gives the continuous FWER approach two important advantages. First, researchers can select the v threshold that is most appropriate for testing their hypothesis and can report their results at multiple v thresholds. If the evidence is strong, they can draw strong conclusions; if the evidence is not so strong, they can draw tentative conclusions. For example, the small anterior-most cluster of voxels that passed the $v = 100$ threshold may be smaller than 100 voxels and disappears at the $v = 10$ threshold. This is relatively weak evidence that anterior superior temporal regions are critical for speech recognition, especially compared to the much stronger evidence that posterior superior temporal regions are critical for speech recognition. This is importantly different from the standard (and increasingly criticized) dichotomous logic that an effect is either "significant" or non-existent (see also Amrhein et al., 2017; Chen et al., 2016). Second, the likely upper limit of false positive voxels is transparently available to the reviewers and readers, who can then evaluate how the v threshold influences the conclusions; for example, whether the possibility of 100 false positive voxels undermines the conclusions or not. Transparently reporting the strength of the evidence allows the science to accumulate – multiple studies that weakly or tentatively show the same pattern can be aggregated to strongly support a conclusion, and contradictory results can be evaluated on the strength of their evidence.

Fig. 10 shows the relationship between t-value thresholds based on continuous FWER correction and FDR correction for the speech recognition data. As in the sub-sample analyses, continuous FWER is consistently more conservative than FDR, indicating that FDR produced incorrect results. For example, at $q = 0.018$, the FDR-corrected critical t-value was 4.37 and 1703 voxels passed that threshold. The nominal expectation is that up to 1.8% of those voxels may be false positives (about 30 voxels), but the permutation data indicate that more than 100 voxels can be expected to be false positives, or more than 3 times higher than expected. As in the sub-sample analyses, this result suggests that researchers should be wary of using FDR correction with VLSM data.

3.3. Discussion

Permutation-based FWER correction uses permutations of the observed data to build a null distribution of voxel-wise test statistics, then uses this null distribution to set thresholds for evaluating the test statistics in the original (true) analysis. The standard version of this approach uses only the most extreme voxel-wise test statistic from each permutation and the resulting threshold makes it unlikely that even a single false positive voxel will be observed. We examined a generalization of this approach in which the threshold is not based on only the most extreme ($v = 1$) test statistic. Using the v-th most extreme test statistic (where $v > 1$) provides a way to quantify the possible rate of false positive voxels: up to v voxels can be reasonably expected to be false positives. We refer to this approach as *continuous permutation-based FWER correction*. Since single voxels rarely (if ever) affect interpretation of VLSM results, this extension aligns the correction method with how VLSM results are interpreted, and allows for transparently reporting the strength of the evidence. Analyses of speech recognition deficit scores provided a particularly clear demonstration of the value of quantifying rates of false positive voxels within a flexible framework. At the standard $v = 1$ threshold, only 57 voxels survived the correction – a statistically "significant" result that is hard to interpret. Examining the data at $v = 10$, 100, and 1000 revealed a clear (and unsurprising) relationship between posterior superior temporal lobe damage and speech recognition deficits. Researchers can calibrate the v threshold to their hypotheses or regions of interest. For example, a hypothesis about a very specific region (e.g., role of area Spt in speech processing, Rogalsky et al., 2015) might require a small v threshold, whereas a hypothesis about a broader region (e.g., role of the inferior frontal gyrus in lexical selection, Harvey and Schnur, 2015; Mirman and Graziano, 2013) might allow a larger v threshold. In addition to providing researchers with more flexibility in evaluation of their data, this approach provides a simple and transparent way to report the expected upper limit of false positive voxels, which allows readers to evaluate the conclusions as well.

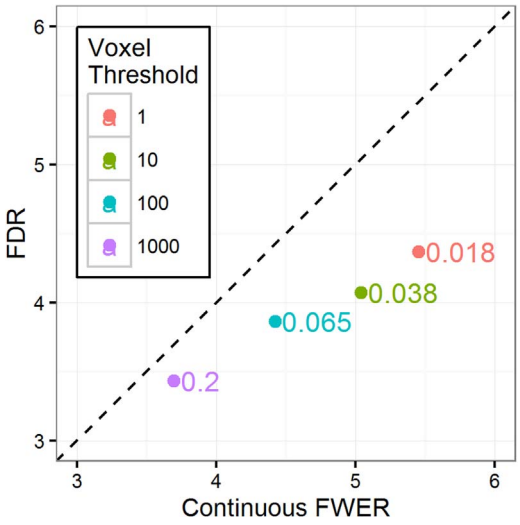

Fig. 10. Critical t-values (t-thresholds) determined by continuous FWER correction (at $v = 1$, 10, 100, 1000) and FDR correction for analysis of speech recognition scores. Points indicate the critical t-values, numbers next to the points indicate corresponding q-values.

As can be seen in the left panel of Fig. 4, there are likely to be some ties among the sorted voxel-wise test statistics, especially in smaller samples, where patches of equivalent voxels will necessarily have the same test results. As discussed above, this may contribute to making FDR correction anti-conservative for smaller samples and it is important to consider how it affects calculation and interpretation of the continuous FWER threshold. As an example, consider a threshold calculated for $v = 100$. When there are ties, the 100th voxel is equivalent to some $v < 100$ voxel as well as some $v > 100$ voxel, so the threshold calculated at $v = 100$ corresponds to a range from more strict to less strict than the nominal $v = 100$ level. If that threshold is applied as an open interval (i.e., only voxels that exceed the threshold), then the functional threshold is somewhat more conservative (i.e., corresponding to the $v < 100$ end of the patch) and the v value correctly specifies the upper limit of the number of false positives that can be expected. However, if the threshold were applied as a closed interval (i.e., voxels that are equal to or exceed the threshold), then the functional threshold is somewhat less conservative (i.e., corresponding to the $v > 100$ end of the patch). Therefore, it is important that the threshold be applied as an open interval.

One might consider calculating and applying correction thresholds using unique patches rather than voxels, but this strategy is problematic because patches are not geometrically or spatially equivalent. We tested patch-based corrections and noticed that the error correction became impossible to interpret because the number and configuration of voxels under each patch varies (e.g., areas in the periphery of MCA have larger patches) and depends on the sample size (smaller groups produce larger patches). Because voxels are geometrically equivalent, it is fairly straight-forward to interpret a voxel-based threshold that allows, for example, up to 10 false positive voxels (continuous FWER) or 1% false positive voxels (FDR) out of some reasonably large number of supra-threshold voxels. In contrast, up to 10 false positive patches or 1% false positive patches could cover any part of the results if those 10 or 1% happened to be large patches. As a result, even a small number or proportion of patches could undermine any inferences about lesion-symptom relationships. Therefore, both continuous FWER and FDR corrections need to be calculated at the voxel level rather than the

patch level.

The v threshold can be regarded as a minimum number of voxels that must exceed the threshold, which makes the continuous FWER approach somewhat similar to cluster-based correction methods. The two methods are also similar in that their thresholds are derived through permutation. However, there are two important differences. First, under cluster-based correction, the cluster size threshold is a true "critical" threshold: if the computed cluster size threshold is 100 voxels, then a cluster of 101 voxels is "significant" whereas a cluster of 99 voxels is not. In contrast, under continuous FWER, the v threshold is an estimate of the upper bound of the number of false positive voxels that can be expected: if $v = 100$, observing 101 supra-threshold voxels provides very weak evidence and observing 99 supra-threshold voxels provides only slightly weaker evidence. This is useful because it provides a graded way to quantify the strength (or weakness) of the evidence in a way that is straightforward to report and interpret. Second, the cluster size threshold (by definition) requires the supra-threshold voxels to be contiguous and (as discussed above) appears to be particularly vulnerable to spill-over effects arising from the inherent spatial autocorrelation in stroke lesion data. In contrast, continuous FWER defines the minimum test statistic that must be exceeded by any voxel, regardless of its location. This makes it less vulnerable to such spill-over effects, though it is impossible to be completely immune to them.

Importantly, continuous permutation-based FWER maintains the advantages of the standard non-parameteric permutation-based FWER correction strategy. These advantages became apparent in the comparison between continuous FWER correction and FDR correction. Like continuous FWER, FDR (nominally) quantifies the expected rate of false positive voxels. For a relatively large sample (N = 124) with a very strong simulated lesion-symptom relation, FDR quite accurately quantified the rate of false positive voxels. However, at smaller sample sizes (N = 62, N = 31) and with real deficit data (presumably a less strong lesion-symptom relation), FDR consistently under-estimated the rate of false positives. To our knowledge, this is the first concrete evidence that FDR correction may not be appropriate for VLSM analysis.

4. General discussion and conclusions

Permutation-based FWER correction is the current "gold standard" correction for multiple comparisons in VLSM. The standard permutation-based FWER strategy is to build a null distribution using only most extreme voxel-wise test statistic from each permutation. The main weakness of this approach is that it aims to control the occurrence of even a single false positive voxel, which is not the scale at which VLSM results are interpreted. We described an extension - continuous permutation-based FWER - which better aligns with VLSM interpretation and allows researchers a more flexible balance between false positives and false negatives. Continuous FWER uses the v-th most extreme voxel-wise test statistic, so the standard approach is the special case where $v = 1$, but other values $v > 1$ may be used as appropriate for a particular data set and hypothesis. This provides a principled way for researchers to flexibility set the upper limit of how many false positive voxels are allowed and to transparently report this limit along with their results, so readers can also evaluate the evidence. Since single voxels rarely (if ever) affect the interpretation of VLSM results, this flexibility lets researchers align their statistical method with their interpretations of the results.

In addition to continuous FWER, we examined two other methods of correction, but those results were not encouraging. Using permutations to set a minimum cluster size tended to produce clusters that extended well beyond the correct region. This was partly due to a true correlation between damage in adjacent regions and damage in the correct region (i.e., spatial coherence); however, since the other correction methods seemed less susceptible to this spill-over problem, spatial auto-correlation does not appear to be the full explanation. Instead, we suspect this spill-over occurred because weak or noisy effects in adjacent voxels were incorporated into true clusters, with the unfortunate consequence of blurring the boundary of the true symptom-related region. Cluster-based correction appeared to be effective at controlling the occurrence of false positive clusters, but the spill-over effect poses a problem for identifying brain-behavior relationships. The spill-over effect may have been further exacerbated by the strong relationship between our simulated deficit scores (percent damage) and their corresponding brain regions. The false discovery rate (FDR) approach is inferentially similar to continuous FWER: it aims to quantify the rate of false positive voxels. FDR performed quite well for larger samples with strong lesion-symptom relations, but consistently underestimated the rate of false positive voxels when the sample sizes were smaller and in a real data case (where the lesion-symptom relation is likely to be weaker). This suggests that researchers should be wary of using FDR in conventional VLSM analyses.

There is an inherent trade-off between false positives and false negatives: striving to eliminate false positives will necessarily result in missing many true effects, but generalizing from every observation will necessarily produce some incorrect inferences. Setting arbitrary thresholds of statistical significance makes evidence appear more dichotomous than it really is; statistical thresholds encourage binary thinking in which an effect is either significant or non-existent. This dichotomy is further exacerbated by publication bias because weaker, not statistically significant results are simply not published. This reifies the sense that effects that do not pass the significance threshold are non-existent, leading to a biased scientific literature and undermining evidence accumulation. Balancing false positives and false negatives is particularly challenging in VLSM, where participant recruitment and testing is difficult and relatively expensive, and samples are generally large relative to other research in neuropsychology and cognitive neuroscience. A typical VLSM research project might require a long period of expensive data collection to reach a reasonable sample size of, say, N = 50. If analysis of that data set produced results just short of the standard FWER $v = 1$ statistical threshold, the researchers would be left with an unpublishable result. Substantially increasing the sample size is likely to be impractical (and perhaps impossible) as it would require another long, labor-intensive, and expensive data collection effort. Addressing this challenge requires a statistical correction method that allows researchers to flexibly balance false positives and false negatives and to report how they struck that balance in a transparent fashion so that readers can interpret the evidence. The v threshold plays this role in the continuous permutation-based FWER correction method: v is the expected upper limit of false positive voxels, which can be adjusted to suit the researchers' hypotheses and reported for readers to use in their evaluation. Making the full statistical maps easily available to other researchers through a repository would further support evidence accumulation through re-analysis and meta-analysis.

We have deliberately avoided recommending a specific v threshold to be used in continuous FWER correction, or how many supra-threshold voxels (effective q) should be considered "significant". Such thresholds are fundamentally arbitrary – there is nothing qualitatively different between $v = 10$ and $v = 11$, just as there is nothing qualitatively different between $p = 0.04$ and $p = 0.06$. Setting significance thresholds contributes to mis-interpretation of these continuous statistics and results that do not pass the threshold tend not to be published, which creates an incentive for researchers to use (undisclosed) variations in data pre-processing and analysis methods to achieve a "significant" result (for more discussion see Amrhein et al. 2017). Thus, in lieu of recommending specific choices of hard thresholds, we offer guidelines for effective use of continuous FWER. First, the v threshold should match the neural specificity of the hypothesis under investigation. Neurally precise hypotheses (e.g., anterior vs. posterior portion of the inferior frontal gyrus) may require a small v threshold; in contrast, a high v threshold may be sufficient for evaluating a neuroanatomically broader hypothesis (e.g., role of anterior temporal lobe vs. the temporo-parietal cortex). It may also be useful to consider the overall number of

voxels that are entered into the analysis, especially for VLSM analyses that are constrained to specific regions of interest. Second, the number of supra-threshold voxels should be substantially larger than v (the effective q, which should be fairly small, may serve as a useful summary statistic). For example, observing 101 supra-threshold voxels at $v = 100$ would not license strong conclusions because almost that many could be false positives. That is, the conclusions drawn by investigators should reflect both the number of supra-threshold voxels relative to the v threshold (effective q) and the neural precision of their hypothesis. These general guidelines may not be sufficient for selection of a single appropriate v threshold, so it may be useful to test and report results at multiple v thresholds as a way to assess the robustness of the effects. Continuous FWER is an extension of standard permutation-based FWER that enables the researcher to explore and report the strength of the effects by testing different v thresholds. Further investigation by independent groups will help to refine and set community standards for how v thresholds should be set and explored. We have made available a basic implementation[4] of continuous FWER and one that is integrated into a user-friendly lesion-symptom mapping package,[5] both of which make it easy to test multiple different v thresholds in order to provide a more complete understanding of the strength of the evidence.

It may be possible to further improve the correction methods described here by considering unique patches (where lesion status is perfectly correlated across participants) rather than individual voxels (Kimberg et al., 2007). Our (unsystematic) comparisons suggest that, in a typical data set, the number of unique patches may be an order of magnitude smaller than the number of voxels. Such a vast reduction in the number of (redundant) tests would have substantial consequences for these methods, as well as for the processing time required to compute them. More generally, including voxel neighborhood information as part of voxel-level corrections may lead to even more effective algorithms. In addition, it may be possible to develop algorithms to automatically establish an optimal v-threshold by using spatial correlations in the data (e.g., number of patches, neighborhood correlations) as an indication of the amount of continuous FWER correction required for the specific dataset under investigation. Finally, recent development of multivariate lesion-symptom mapping methods (Zhang et al., 2014; Pustina et al., 2017), which evaluate lesion-symptom relationships across all voxels simultaneously, provide a better method for studying brain-behavior relationships and mitigate the need for multiple comparisons correction. Such methods are not in wide use yet, but they offer a promising alternative approach.

In addition to optimizing corrections for multiple comparisons, getting the most out of lesion-symptom studies requires optimizing the test statistics themselves. Recent analysis show that correcting for overall lesion volume and only testing voxels with sufficient lesion involvement is important for avoiding mis-localization (Sperber and Karnath, 2017). For functional neuroimaging data, threshold-free cluster enhancement (TFCE; Smith and Nichols, 2009) has been shown to improve voxel-wise signal-to-noise ratio by incorporating local support information from nearby voxels. This method considers all possible cluster-forming thresholds, so it can capture support from small strong clusters as well as weaker diffuse clusters. Once the voxel-wise TFCE statistic has been calculated, it still requires correction for multiple comparisons. Because they operate at different steps in the analysis process, calculating voxel-wise test statistics using TFCE, then correcting those test statistics for multiple comparisons using continuous FWER correction may jointly enhance power in VLSM analysis.[6]

VLSM is important for basic and translational human neuroscience. Analysis of lesion-symptom relations has been at the core of cognitive neuroscience research since the mid-19th century and remains critical to the field. Lesions that produce chronic deficits in a particular domain or task provide the strongest evidence that the damaged neural structures were critical for that domain or task. This method is an important complement to functional neuroimaging in neurologically-intact populations, but the differences in data collection challenges create somewhat different statistical demands. Cognitive neuroscience has tremendous potential for stimulating development of new and improved diagnosis, treatment, rehabilitation, and education strategies. That potential cannot be realized without testing the affected populations. VLSM offers a unique opportunity for research that answers fundamental questions about the neural basis of cognition while addressing the real-world problem of understanding neurogenic cognitive deficits. Robust, flexible, and transparent statistical methods play an important role in maximizing the impact of VLSM research.

[4] https://gist.github.com/dmirman/05a92e0e9e0027f6fe6e528c648143d7.

[5] https://dorianps.github.io/LESYMAP/.

[6] To our knowledge, TFCE has not been evaluated for lesion-symptom mapping analyses. The binary lesion/non-lesion status of individual voxels and inherently high spatial correlations of lesion data may affect local support (as it is used in TFCE), or may require some adjustment of TFCE implementation parameters.

Acknowledgements

We are very grateful to Dr. Myrna F. Schwartz and her research team for sharing the anatomical data that made these analyses possible and for her comments on an early draft of this report. This manuscript benefitted from helpful discussions with Branch Coslett and the Laboratory for Cognition and Neural Stimulation, the Moss Rehabilitation Research Institute's Cognitive Area group, and constructive criticism from anonymous reviewers. We thank Yongsheng Zhang for sharing his Matlab implementation of the cluster size thresholding, which formed the basis of our re-implementation. This research was funded in part by Drexel University, University of Alabama at Birmingham, and National Institutes of Health grant R01DC010805 to DM.

References

Amrhein, V., Korner-Nievergelt, F., Roth, T., 2017. The earth is flat ($p > 0.05$): significance thresholds and the crisis of unreplicable research. PeerJ 5 (e3544). http://dx.doi.org/10.7717/peerj.3544.

Avants, B.B., Kandel, B.M., Duda, J.T., Cook, P.A., Tustison, N.J., Pustina, D., 2016. ANTsR: ANTs in R: Quantification Tools for Biomedical Images. R Package Version 0.3.3.

Bates, E., Wilson, S.M., Saygin, A.P., Dick, F., Sereno, M.I., Knight, R.T., Dronkers, N.F., 2003. Voxel-based lesion-symptom mapping. Nat. Neurosci. 6 (5), 448–450.

Bennett, C.M., Wolford, G.L., Miller, M.B., 2009. The principled control of false positives in neuroimaging. Soc. Cogn. Affect. Neurosci. 4 (4), 417–422.

Binder, J.R., Pillay, S.B., Humphries, C.J., Gross, W.L., Graves, W.W., Book, D.S., 2016. Surface errors without semantic impairment in acquired dyslexia: a voxel-based lesion-symptom mapping study. Brain 139 (5), 1517–1526.

Bordier, C., Dojat, M., de Micheaux, P.L., 2011. Temporal and spatial independent component analysis for fmri data sets embedded in the AnalyzeFMRI R package. J. Stat. Softw. 44 (9), 1–24.

Chen, G., Taylor, P.A., Cox, R.W., 2016. Is the statistic value all we should care about in neuroimaging? NeuroImage 147, 952–959.

DeWitt, I., Rauschecker, J.P., 2012. Phoneme and word recognition in the auditory ventral stream. Proc. Natl. Acad. Sci. USA 109 (8), E505–E514.

Gelman, A., Loken, E., 2014. The statistical crisis in science. Am. Sci. 102 (6), 460–465.

Genovese, C.R., Lazar, N.A., Nichols, T.E., 2002. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. NeuroImage 15 (4), 870–878.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011a. Mass univariate analysis of event-related brain potentials/fields i: a critical tutorial review. Psychophysiology 48 (12), 1711–1725.

Groppe, D.M., Urbach, T.P., Kutas, M., 2011b. Mass univariate analysis of event-related brain potentials/fields ii: simulation studies. Psychophysiology 48 (12), 1726–1737.

Harvey, D.Y., Schnur, T.T., 2015. Distinct loci of lexical and semantic access deficits in aphasia: evidence from voxel-based lesion-symptom mapping and diffusion tensor imaging. Cortex 67, 37–58.

Hickok, G., Poeppel, D., 2015. Neural basis of speech perception. In: In: Celesia, G., Hickok, G. (Eds.), The Human Auditory System: Fundamental Organization and Clinical Disorders 129. pp. 149–160.

Kimberg, D.Y., Coslett, H.B., Schwartz, M.F., 2007. Power in voxel-based lesion-symptom mapping. J. Cogn. Neurosci. 19 (7), 1067–1080.

Lichtheim, L., 1885. On aphasia. Brain 7, 433–484.

Mah, Y.-H., Husain, M., Rees, G., Nachev, P., 2014. Human brain lesion-deficit inference remapped. Brain 137 (9), 2522–2531.

Mirman, D., Chen, Q., Zhang, Y., Wang, Z., Faseyitan, O.K., Coslett, H.B., Schwartz, M.F., 2015a. Neural organization of spoken language revealed by lesion-symptom

mapping. Nat. Commun. 6 (6762), 1–9.
Mirman, D., Graziano, K.M., 2013. The neural basis of inhibitory effects of semantic and phonological neighbors in spoken word production. J. Cogn. Neurosci. 25 (9), 1504–1516.
Mirman, D., Zhang, Y., Wang, Z., Coslett, H.B., Schwartz, M.F., 2015b. The ins and outs of meaning: bhavioral and neuroanatomical dissociation of semantically-driven word retrieval and multimodal semantic recognition in aphasia. Neuropsychologia 76, 208–219.
Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.
Nosek, B.A., Spies, J.R., Motyl, M., 2012. Scientific Utopia: II. Restructuring incentives and practices to promote truth over publishability. Perspect. Psychol. Sci. 7 (6), 615–631.
Pillay, S.B., Binder, J.R., Humphries, C., Gross, W.L., Book, D.S., 2017. Lesion localization of speech comprehension deficits in chronic aphasia. Neurology 88 (10), 970–975.
Pillay, S.B., Stengel, B.C., Humphries, C., Book, D.S., Binder, J.R., 2014. Cerebral localization of impaired phonological retrieval during rhyme judgment. Ann. Neurol. 76 (5), 738–746.
Pustina, D., Avants, B., Faseyitan, O., Medaglia, J., and Coslett, H.B., 2017. Improved accuracy of lesion to symptom mapping with multivariate sparse canonical correlations, pages 1–50, (Manuscript under review).
R Core Team, 2016. A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.
Rogalsky, C., Poppa, T., Chen, K.-H., Anderson, S.W., Damasio, H., Love, T., Hickok, G., 2015. Speech repetition as a window on the neurobiology of auditory-motor integration for speech: a voxel-based lesion symptom mapping study. Neuropsychologia 71, 18–27.
Romano, J.P., Wolf, M., 2007. Control of generalized error rates in multiple testing. Ann. Stat. 35 (4), 1378–1408.
Rorden, C., Karnath, H.-O., 2004. Using human brain lesions to infer function: a relic from a past era in the fMRI age? Nat. Rev. Neurosci. 5, 813–819.
Rorden, C., Karnath, H.-O., Bonilha, L., 2007. Improving lesion-symptom mapping. J. Cogn. Neurosci. 19 (7), 1081–1088.
Simmons, J.P., Nelson, L.D., Simonsohn, U., 2011. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. Psychol. Sci. 22 (11), 1359–1366.
Smith, S.M., Nichols, T.E., 2009. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. NeuroImage 44 (1), 83–98.
Sperber, C., Karnath, H.-O., 2017. Impact of correction factors in human brain lesion-behavior inference. Hum. Brain Mapp. 38 (3), 1692–1701.
Zhang, Y., Kimberg, D.Y., Coslett, H.B., Schwartz, M.F., Wang, Z., 2014. Multivariate lesion-symptom mapping using support vector regression. Hum. Brain Mapp. 35 (12), 5861–5876.