Contents lists available at ScienceDirect

Neuropsychologia

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

Note

Inappropriate usage of the Brunner–Munzel test in recent voxel-based lesion-symptom mapping studies

Jared Medina*, Daniel Y. Kimberg, Anjan Chatterjee, H. Branch Coslett

Center for Cognitive Neuroscience and Department of Neurology, University of Pennsylvania, United States

ARTICLE INFO

Article history:
Received 22 June 2009
Received in revised form 4 September 2009
Accepted 13 September 2009
Available online 18 September 2009

Keywords:
Voxel-based lesion-symptom mapping (VLSM)
Brain-lesion mapping
Lesion
Voxel
Structure–function

ABSTRACT

Voxel-based lesion-symptom mapping (VLSM) techniques have been important in elucidating structure–function relationships in the human brain. Rorden, Karnath, and Bonilha (2007) introduced the non-parametric Brunner–Munzel rank order test as an alternative to parametric tests often used in VLSM analyses. However, the Brunner–Munzel statistic produces inflated *z* scores when used at any voxel where there are less than 10 subjects in either the lesion or no lesion groups. Unfortunately, a number of recently published VLSM studies using this statistic include relatively small patient populations, such that most (if not all) examined voxels do not meet the necessary criteria. We demonstrate the effects of inappropriate usage of the Brunner–Munzel test using a dataset included with MRIcron, and find large Type I errors. To correct for this we suggest that researchers use a permutation derived correction as implemented in current versions of MRIcron when using the Brunner–Munzel test.

Much of our understanding of brain function is based on observations of the consequences of brain injury. By examining the consequences of brain disruption, one can identify whether a brain region is required to perform a task, providing a stronger inference than afforded by measures of brain function such as functional imaging that identify regions involved in but not necessarily crucial to a task. Bates et al. (2003) introduced Voxel-based lesion-symptom mapping (VLSM), an update of a method for structure–function mapping that has been widely used for over a century. As its name suggests, the method considers the statistical relationship between behavior and the structural integrity of the brain, on a voxel-by-voxel basis. This technique can extend traditional lesion analysis by identifying novel brain areas (rather than being restricted to predefined regions of interest).

The original work of Bates et al. (2003) was based on the parametric *t*-test, which makes a number of assumptions regarding the distribution of the behavioral data. Rorden, Karnath, and Bonilha (2007) introduced analyses using non-parametric statistics that do not make such assumptions. As the distribution of behavioral scores from brain-damaged subjects are often non-normal, Rorden, Karnath, et al. (2007) proposed using the non-parametric rank order Brunner–Munzel test (Brunner & Munzel, 2000) as a complementary alternative to parametric tests for analyzing lesion–behavior relationships. They reported that the Brunner–Munzel test identified more areas significantly associated with a deficit than the *t*-test, without large differences in false alarm rates. The Brunner–Munzel test has been used in a number of VLSM studies since the publication of Rorden, Karnath, et al. (2007), likely due to its reported advantages for lesion data and the ease of usage of NPM (non-parametric mapping), a program included with MRIcron for VLSM analysis. However, we believe that inappropriate usage of the Brunner–Munzel test has resulted in a number of recent studies reporting potentially inaccurate relationships between lesion location and impairment.

The Brunner–Munzel rank order test was designed to detect differences between groups without making any assumptions regarding the shape or continuity of the underlying distribution. For large sample sizes, the Brunner–Munzel test statistic (tBM) behaves as a standard normal for generating *z* and *p* values. For moderate sized groups (>9), *p* values are generated using a *t*-distribution with a degrees of freedom correction. However, for small groups, accurate *p* values cannot be generated using the Brunner–Munzel test statistic approximation and a degrees of freedom correction. Brunner and Munzel (2000) stated that “for extremely small sample sizes ($n_i < 10$), simple and accurate approximations in a general nonparametric model cannot be expected.”

More specifically for VLSM analyses, it is not proper to use a Brunner–Munzel test statistic with a medium-sized group correction to generate *p* values at any voxels where there are less than 10 subjects in *either* the lesion or no lesion group. This was noted by

* Corresponding author at: 3 West Gates, 3400 Spruce Street, University of Pennsylvania, Philadelphia, PA 19104, United States.
Tel.: +1 215 614 0274; fax: +1 215 349 8260.
E-mail address: jared.medina@uphs.upenn.edu (J. Medina).

0028-3932/$ – see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.neuropsychologia.2009.09.016

Rorden, Karnath, et al. (2007) when discussing the Brunner–Munzel test:

> This test is relatively rapid to compute, and generates a statistic that is approximately normal for situations with at least 10 observations in each group. For smaller groups, one can either compute all possible more extreme permutations (to derive a precise *p* value) or use a permutation test to approximate the precise *p* value (Neubert & Brunner, 2007). Rorden, Bonilha, and Nichols (2007) have recently suggested that this test is suitable for voxel-based morphometry, albeit their implementation does not implement the permutation test for small groups. This small group correction is vital for lesion analysis, as the size of each group varies with lesion density (e.g., any voxel where only a few people have a lesion or almost all people have a lesion will require a small group permutation test).

Although not noted in the quoted article, a separate article in Neuroimage (Rorden, Bonilha, et al., 2007) stated that the degrees of freedom correction intended to be used for medium-sized groups only ($n > 9$ observations) was implemented in NPM for *all* sample sizes, including those when either group had less than 9 observations. Although this implementation is legitimate when the size of each group at a given voxel is >9, using this statistical test when either group contains less than 10 subjects can result in highly inflated Brunner–Munzel test scores.

To illustrate this point, we ran a Brunner–Munzel analysis of sample data included in the MRIcron software package, which includes 24 dummy left hemisphere lesions and a dummy behavioral score associated with each subject (..\example\lesions\continuous.val). Importantly, we ran this analysis using two different versions of NPM. In the earlier version (available when Rorden, Karnath, et al., 2007 was published), the Brunner–Munzel *z* score (*z*BM) was calculated using the medium-group size degrees of freedom (df) correction. In the current version of NPM, *z*BM was generated using the same method with greater than 15 subjects in each group, but was permutation derived when either group had less than 15 subjects (based on 20,000 permutations for the observed data). In this method, the precise *p* value is calculated by comparing the rank order of subjects in the lesion and no lesion groups at a voxel to the total number of possible permutations of rank orders that are more or less extreme at that voxel. In both analyses, we recorded the number of voxels that were significantly associated with poor performance on the dummy task, using false discovery rate (FDR) thresholds, a Bonferroni correction, and permutation thresholds set by taking the maximum Brunner–Munzel *z* score from 1000 permutations of the dataset, with significance of .05 set by the 50th greatest of the permutation generated maximum *z* scores (permFWE). Note that this dataset should not be analyzed using the Brunner–Munzel statistic, as only 626 out of 57,390 voxels (1.09%) had at least 10 subjects both with and without a lesion.

At all thresholds, a substantially greater number of voxels were significantly associated with a specific deficit when using the inappropriate medium-sized group corrected *z*BM than when using the permutation derived *z*BM scores (see Table 1). Furthermore, when using a Bonferroni correction of either .05 or .01, there are no significant voxels using the permutation derived *z*BM, whereas a large number of voxels were significantly associated with poor performance using the medium-sized group corrected *z* score. These differences are due in large part to the wildly skewed *z*BM scores generated using the medium-group size correction (see Table 2). In the voxels with the highest test statistic score both using the permutation-derived and medium-group size corrected *z*BM scores, (54, −2, −6), the 7 subjects with a lesion at those voxels were the seven poorest performers on the test (out of 24 subjects). Using the medium-group size correction, *z*BM is 55.09, whereas the permutation derived *z*BM is only 3.89. Since the *z*BM using the medium-group size correction does not reflect the actual probability of that pattern of performance occurring, both Bonferroni and FDR significance thresholds are inapplicable. Furthermore, permutation generated familywise error thresholds were not calculated properly in the earlier version of NPM, as large maximum *z*BM scores should have been generated from permutations far more frequently than observed using the software package.

Table 1
Using either the permutation derived correction (implemented in newer version of MRIcron/NPM) or medium-group size df correction (implemented in older versions), the number of voxels that are significantly correlated with impairment on the dummy tasks using the following types of thresholding: false discovery rate (FDR), permutation familywise error (permFWE), and Bonferroni correction.

	Permutation derived (new)	DF correction (old)
FDR .05	20,141	23,936
FDR .01	5,879	21,381
permFWE .05	5,128	12,682
permFWE .025	3,366	9,771
Bonferroni .05	0	12,619
Bonferroni .01	0	11,327

Table 2
Maximum Brunner–Munzel z score for each Brodmann Area (BA) in the left hemisphere, using the permutation derived OR the medium-group size df correction.

BA	Permutation derived (new)	DF correction (old)
2	2.40	4.44
3	2.24	3.81
4	3.09	6.14
6	3.72	11.33
20	3.72	32.87
21	3.72	32.87
22	3.89	55.09
37	0.89	0.94
38	3.72	32.87
40	0.08	0.08
41	1.07	1.08
42	2.15	2.82
43	3.72	14.68
44	3.19	7.01
45	1.98	3.00
47	0.98	1.05
48	3.89	55.09

Over the past few years, a number of papers have been published reporting VLSM analyses with the Brunner–Munzel test, as implemented in MRIcron. Unfortunately, some of these papers report results that are almost certainly due to the lack of appropriate procedures for infrequently lesioned voxels. Although these papers offer additional analyses to support their theoretical claims more broadly, these Brunner–Munzel VLSM analyses may be critically flawed. We therefore suggest that Brunner–Munzel VLSM analyses that include infrequently lesioned voxels be re-analyzed and, if necessary, revised using either the permutation generated test scores as implemented in the current version of NPM, or using other statistical tests (e.g. *t*-test). We also believe that anyone currently doing VLSM analyses using MRIcron and NPM should update to the newest version of the software.

Finally, we note that the apparent errors in analysis in the manuscripts cited here reflect the occasional but unfortunate consequence of using recently developed research methods. The error in computing Brunner–Munzel *z* scores came to our attention only after we initially made the mistake about which we hope to warn our peers.

Acknowledgments

The authors would like to thank an anonymous reviewer for helpful comments on the manuscript. This work was supported by NIH Grant R01: NS048130.

References

Bates, E., Wilson, S. M., Saygin, A. P., Dick, F., Sereno, M. I., Knight, R. T., et al. (2003). Voxel-based lesion-symptom mapping. *Nature Neuroscience*, *6*(5), 448–450.

Brunner, E., & Munzel, U. (2000). The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal*, *42*(1), 17–25.

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

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

Rorden, C., Karnath, H. O., & Bonilha, L. (2007). Improving lesion-symptom mapping. *Journal of Cognitive Neuroscience*, *19*(7), 1081–1088.