



Published in final edited form as:

Neuropsychologia. 2010 January ; 48(1): 341–343. doi:10.1016/j.neuropsychologia.2009.09.016.

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

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

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

Abstract

Voxel-based lesion symptom mapping (VLSM) techniques have been important in elucidating structure-function relationships in the human brain. Rorden, Karnath & Bonilha (2007b) introduced the nonparametric 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 ten 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.

Keywords

voxel-based lesion symptom mapping; VLSM; brain-lesion mapping; lesion; voxel; structure-function

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 and colleagues (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 and colleagues was based on the parametric t-test, which makes a number of assumptions regarding the distribution of the behavioral data. Rorden, Karnath & Bonilha (2007b) introduced analyses using non-parametric statistics that do not make such assumptions. As the distribution of behavioral scores from brain-damaged subjects are often

Corresponding address: Jared Medina, 3 West Gates, 3400 Spruce Street, University of Pennsylvania, Philadelphia, PA 19104, Phone: 215-614-0274, Fax: 215-349-8260, jared.medina@uphs.upenn.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

non-normal, Rorden and colleagues 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 et al. (2007b), 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 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 Rorden and colleagues when discussing the Brunner-Munzel test (2007b):

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 *et al.*, 2007a) 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 et al. 2007b was published), the Brunner-Munzel Z score (zBM) was calculated using the medium group size degrees of freedom (df) correction. In the current version of NPM, zBM 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 57390 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 zBM than when using the permutation derived zBM scores (see Table 1). Furthermore, when using a Bonferroni correction of either .05 or .01, there are no significant voxels using the permutation derived zBM, 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 zBM 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 zBM scores, (54,-2,-6), the seven 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, zBM is 55.09, whereas the permutation derived zBM is only 3.89. Since the zBM 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 zBM scores should have been generated from permutations far more frequently than observed using the software package.

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 (e.g. Coulthard *et al.*, 2008a; Coulthard *et al.*, 2008b; Malhotra *et al.*, 2009; Molenberghs *et al.*, 2008; Moro *et al.*, 2008; Pazzaglia *et al.*, 2008a; Pazzaglia *et al.*, 2008b; Rossit *et al.*, 2009; Sarri *et al.*, 2009; Zihl *et al.*, 2009) report results that are almost certainly due to the lack of appropriate procedures for infrequently lesioned voxels. These results may be critically flawed, and we therefore suggest that these papers (and any others that have been affected) be re-analyzed and 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 SM, Saygin AP, Dick F, Sereno MI, Knight RT, et al. Voxel-based lesion-symptom mapping. *Nature Neuroscience* 2003;6(5):448–450.

- Brunner E, Munzel U. The nonparametric behrens-fisher problem: Asymptotic theory and a small-sample approximation. *Biometrical Journal* 2000;42(1):17–25.
- Coulthard E, Rudd A, Husain M. Motor neglect associated with loss of action inhibition. *Journal of Neurology Neurosurgery and Psychiatry* 2008a;79(12):1401–1404.
- Coulthard EJ, Nachev P, Husain M. Control over conflict during movement preparation: Role of posterior parietal cortex. *Neuron* 2008b;58(1):144–157. [PubMed: 18400170]
- Malhotra P, Coulthard EJ, Husain M. Role of right posterior parietal cortex in maintaining attention to spatial locations over time. *Brain* 2009;132:645–660. [PubMed: 19158107]
- Molenberghs P, Gillebert CR, Peeters R, Vandenberghe R. Convergence between lesion-symptom mapping and functional magnetic resonance imaging of spatially selective attention in the intact brain. *Journal of Neuroscience* 2008;28(13):3359–3373. [PubMed: 18367603]
- Moro V, Urgesi C, Pernigo S, Lanteri P, Pazzaglia M, Aglioti SM. The neural basis of body form and body action agnosia. *Neuron* 2008;60(2):235–246. [PubMed: 18957216]
- Pazzaglia M, Pizzamiglio L, Pes E, Aglioti SM. The sound of actions in apraxia. *Current Biology* 2008a;18(22):1766–1772. [PubMed: 19013068]
- Pazzaglia M, Smania N, Corato E, Aglioti SM. Neural underpinnings of gesture discrimination in patients with limb apraxia. *Journal of Neuroscience* 2008b;28(12):3030–3041. [PubMed: 18354006]
- Rorden C, Bonilha L, Nichols TE. Rank-order versus mean based statistics for neuroimaging. *Neuroimage* 2007a;35:1531–1537. [PubMed: 17391987]
- Rorden C, Karnath HO, Bonilha L. Improving lesion-symptom mapping. *Journal of Cognitive Neuroscience* 2007b;19(7):1081–1088. [PubMed: 17583985]
- Rossit S, Malhotra P, Muir K, Reeves I, Duncan G, Birschel P, et al. The neural basis of visuomotor deficits in hemispatial neglect. *Neuropsychologia* 2009;47(10):2149–2153. [PubMed: 19394349]
- Sarri M, Greenwood R, Kalra L, Driver J. Task-related modulation of visual neglect in cancellation tasks. *Neuropsychologia* 2009;47(1):91–103. [PubMed: 18790703]
- Zihl J, Samann P, Schenk T, Schuett S, Dauner R. On the origin of line bisection error in hemianopia. *Neuropsychologia* 2009;47(12):2417–2426. [PubMed: 19383505]

Table 1

	Perm Derived (new)	DF Correction (old)
FDR .05	20141	23936
FDR .01	5879	21381
permFWE .05	5128	12682
permFWE .025	3366	9771
Bonferroni .05	0	12619
Bonferroni .01	0	11327

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.

Table 2

BA	Perm 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

Maximum Brunner-Munzel Z score for each Brodmann Area (BA) in the left hemisphere, using the permutation derived and the medium-group size df correction.