Neuroscience and Biobehavioral Reviews 52 (2015) 49–55

Contents lists available at ScienceDirect

Neuroscience and Biobehavioral Reviews

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

Review

False positive rates in Voxel-based Morphometry studies of the human brain: Should we be worried?

Cristina Scarpazza[a,b,*,1], Stefania Tognin[a,1], Silvia Frisciata[a,c], Giuseppe Sartori[c], Andrea Mechelli[a]

[a] Department of Psychosis Studies, Institute of Psychiatry, Psychology & Neuroscience, King's College London, De Crespigny Park, London SE5 8AF, United Kingdom
[b] Center for Studies and Research in Cognitive Neuroscience (CSRNC), University of Bologna, Viale Europa 980, 47023 Cesena, Italy
[c] Department of Psychology, University of Padua, Via Venezia 12, 35131 Padova, Italy

ARTICLE INFO

Article history:
Received 16 September 2014
Received in revised form 10 February 2015
Accepted 11 February 2015
Available online 19 February 2015

Keywords:
Neuroimaging
Voxel-based Morphometry
False positive rate
Unbalanced design
Balanced design

ABSTRACT

Voxel-based Morphometry (VBM) is a widely used automated technique for the analysis of neuroanatomical images. Despite its popularity within the neuroimaging community, there are outstanding concerns about its potential susceptibility to false positive findings. Here we review the main methodological factors that are known to influence the results of VBM studies comparing two groups of subjects. We then use two large, open-access data sets to empirically estimate false positive rates and how these depend on sample size, degree of smoothing and modulation. Our review and investigation provide three main results: (i) when groups of equal size are compared false positive rate is not higher than expected, i.e. about 5%; (ii) the sample size, degree of smoothing and modulation do not appear to influence false positive rate; (iii) when they exist, false positive findings are randomly distributed across the brain. These results provide reassurance that VBM studies comparing groups are not vulnerable to the higher than expected false positive rates that are evident in single case VBM.

Contents

* Corresponding author at: Department of Psychosis Studies, King's College Health Partners, King's College London, De Crespigny Park, London SE5 8AF, United Kingdom. Tel.: +39 3896494919.
E-mail address: cristina.scarpazza@gmail.com (C. Scarpazza).
[1] These authors contributed to this work equally.

http://dx.doi.org/10.1016/j.neubiorev.2015.02.008

1. Introduction

Structural magnetic resonance imaging (MRI) allows the non-invasive and in vivo investigation of brain structure. Over the past two decades, the development of a number of automated techniques for the analysis of structural MRI data (Chung et al., 2003; Mechelli et al., 2005; Bandettini, 2009; Dell'Acqua and Catani, 2012) has led to a proliferation of studies on the neuroanatomical basis of neurological and psychiatric disorders. The popularity of these techniques can be explained by two critical advantages relative to traditional tracing methods: firstly, they allow detection of subtle morphometric group differences in brain structure that may not be discernible by visual inspection; secondly, they allow investigation of the entire brain, rather than a particular structure, in an automatic and objective manner.

The most widely used automated technique for the analysis of structural brain images is Voxel-based Morphometry (VBM) which involves a voxel-wise comparison of the local volume or concentration of gray and white matter between groups of subjects (Ashburner and Friston, 2000, 2001; Good et al., 2001; Mechelli et al., 2005). Over the past 15 years, VBM has been used successfully to investigate a wide range of neurological and psychiatric disorders including, but not limited to, Alzheimer's disease (Li et al., 2012), Parkinson's disease (Pan et al., 2013), multiple sclerosis (Lansley et al., 2013), unipolar (Lai, 2013) and bipolar (Selvaraj et al., 2012) depression, anxiety disorders (Radua et al., 2010) and psychosis (Honea et al., 2005; Bora et al., 2011 Mechelli et al., 2011). In addition, VBM has been used to compare groups of healthy subjects who differ with respect to biological or environmental variables of interest such as age (Kennedy et al., 2009; Takahashi et al., 2011), gender (Takahashi et al., 2011; Sacher et al., 2013), number of spoken languages (Mechelli et al., 2004), and exposure to stressful life events (Papagni et al., 2011).

1.1. Methodological factors influencing the results of a VBM study

Although overall VBM can be considered a user-friendly and practical tool, any user has to navigate a number of methodological options that are likely to influence the final results. These include, for example, the protocol for the acquisition of the MRI data, the type of pre-processing of the images and the statistical threshold used to identify significant effects.

Firstly, the accuracy and precision of the results are critically dependent on the quality of the input images including, for example, image resolution and acquisition sequence. Higher resolution is thought to result in more localized and more reliable results (Iwabuchi et al., 2013); this means that the results of identical comparisons performed at 1.5 T and 3 T respectively may differ for purely methodological reasons. The acquisition sequence is another source of variability that is often underestimated. Acquisition sequence includes different parameters such as image-to-noise ratio and uniformity, which are known to affect tissue classification leading to different results (Tardiff et al., 2009; Streitbürger et al., 2014).

Secondly, the results of a VBM study are dependent on the type of preprocessing. This may differ with respect to the segmentation procedure (Ashburner, 2012), the widely discussed normalization protocol (Crum et al., 2003; Ashburner and Friston, 2001) and the Gaussian smoothing kernel applied to the images (Salmond et al., 2002; Viviani et al., 2007; Smith and Nichols, 2009).

Thirdly, the results of a VBM study depend on the statistical analysis. For example, while nearly all studies use a correction for multiple comparisons based on random field theory, the user has the option of choosing the statistical threshold and the number of statistical tests (Smith and Nichols, 2009; Lieberman and Cunningham, 2009). In addition, some but not all studies use nuisance variables as covariates of no interest to reduced the amount of unexplained variance in the data (Hu et al., 2011).

From this brief overview, it appears that every step of a VBM study, from the acquisition of the data to the statistical analysis, involves a number of methodological choices that are likely to affect the final results.

1.2. Non-normality of the residuals

While the above methodological factors relate to how the data are acquired and the analyses are carried out, the validity of the final results are also dependent on the characteristics of the data. In particular, VBM assumes that the error terms in the statistical analysis are normally distributed; this is ensured through the Central Limit Theorem by applying a Gaussian smoothing kernel to the data at the preprocessing stage (Salmond et al., 2002). However, smoothing the data does not always ensure normal distribution of the error terms (Salmond et al., 2002; Silver et al., 2011; Scarpazza et al., 2013). For example a previous investigation found that, based on the Shapiro–Wilks test for normality, residuals in smoothed images were highly non-normal and, furthermore, deviation from normality was inversely related to the smoothing kernel (Silver et al., 2011). Moreover, in a recent investigation (Scarpazza et al., 2013), we estimated the likelihood of detecting significant differences in gray matter volume in individuals free from neurological or psychiatric diagnosis using two independent data sets (Scarpazza et al., 2013). This revealed that, when comparing a single subject against a group in VBM, the chance of detecting a significant difference which is not related to any psychiatric or neurological diagnosis is much higher than previously expected. As an example, using a standard voxel-wise threshold of $p < 0.05$ (corrected) and an extent threshold of 10 voxels, the likelihood of a single subject showing at least one significant difference is as high as 93.5% for increases and 71% for decreases. These results were unlikely to be due solely to the individual variability in neuroanatomy; this is because such variability would inflate the standard error estimated from the controls resulting in reduced rather than increased sensitivity. The most likely explanation for the very high false positive rate was that the data were not normally distributed; hence, the assumption of normality of the residuals required by the random field theory was violated. We concluded that interpretation of the results of single case VBM studies should be performed with caution, particularly in the case of significant differences in temporal and frontal lobes where false positive rates appear to be highest.

The above investigation raises the question of whether the surprisingly high false positive rate in single case VBM studies would also be evident in the context of balanced designs in which groups of equal size are compared. Although it is traditionally assumed that the use of smoothing is enough to ensure normality of the residuals when comparing groups of equal size (Mechelli et al., 2005), there is preliminary evidence that residuals in smoothed images

can still be non-normal and that this may result in high false positive rates even in the context of balanced designs (Salmond et al., 2002; Silver et al., 2011). A higher-than-expected false positive rate would have important implications for the validity of the hundreds of VBM studies comparing different experimental groups that are being published each year; conversely, a false positive rate of up to 5% (for a one-tailed test) or 10% (for a two-tailed test) would provide reassurance that any significant differences in group VBM studies is unlikely to result from the interaction between non-normality of the residuals and random field theory. So, the outstanding question which needs to be addressed is: should we be worried?

1.3. An experimental contribution to the existing literature

Since a revision of the existing literature is not sufficient to answer the above question, we decided to add an experimental contribution in which we examined false positive rates in group VBM studies by empirically estimating the likelihood of detecting significant differences in gray matter volume (GMV) between groups of the same size comprising of healthy individuals. In order to maximize the generalizability of our results, we used two independent data sets (Biswal et al., 2010) consistent with our previous investigation of false positive rates in single case VBM studies (Scarpazza et al., 2013). These two freely available data sets were acquired with the same images resolution (3 T) and acquisition sequence (MPRAGE) and comprised of a total of 396 subjects free from neurological or psychiatric diagnosis. A similar procedure to the one described in Scarpazza et al. (2013) was adopted, with the only difference being that in the present investigation we compared two groups of equal size rather than a single subject to a group. The impact of sample size (n = 8, 12, 16), smoothing (4 mm, 8 mm, 12 mm) and modulation (with and without modulation) was also investigated, as these factors have been found to influence false positive rates in previous studies (Salmond et al., 2002; Viviani et al., 2007; Silver et al., 2011; Scarpazza et al., 2013).

Our first hypothesis was that when VBM is used to compare groups of equal size, the rate of false positives would be about 5% (for one-tailed tests) or 10% (for two-tailed tests), in contrast with the very high false positive rates observed in the context of unbalanced designs (Scarpazza et al., 2013). Our second hypothesis was that false positive rate would vary as a function of sample size (with a higher number of differences detected for smaller sample size), degree of smoothing applied to the data (with a higher number of differences detected for smaller kernel smoothing), and modulation (with a higher number of differences detected for unmodulated data) as these variables have been reported to affect the number of significant effects in previous studies (Salmond et al., 2002; Viviani et al., 2007; Smith and Nichols, 2009; Scarpazza et al., 2013). Our third hypothesis was that, consistent with the results of our previous work (Scarpazza et al., 2013), significant differences would not be equally distributed across the whole brain but would be mainly located in the frontal and temporal lobes.

2. Methods

2.1. Subjects

Data from the Neuroimaging Informatics Tools and Resources Clearinghouse (NITRC) which are available at http://fcon_1000.projects.nitrc.org/fcpClassic/FcpTable.html were used (Biswal et al., 2010). The Cambridge (MA, USA) and Beijing (China) data sets were chosen because of their large sample size (n = 198) and their matched age range (18–28). All participants have never received a neurological or psychiatric diagnosis.

2.2. MRI data acquisition

All participants underwent the acquisition of a structural MRI scan using a 3 T MRI system. A T1-Weighted sagittal three-dimensional magnetization-prepared rapid gradient echo (MPRAGE) sequence was acquired, covering the entire brain. For the acquisition of the Cambridge data set, the following parameters were used: TR = 3; 144 slices, voxel resolution 1.2, 1.2, 1.2; matrix 192 × 192. For the acquisition of the Beijing data set, the following parameters were used: TR = 2; 128 slices, voxel resolution 1.0, 1.0, 1.3; matrix 181 × 175.

2.3. Data analysis

2.3.1. Preprocessing

Images were checked for scanner artifacts, and gross anatomical abnormalities, and then reoriented along the anterior–posterior commissure (AC–PC) line with the AC set as the origin of the spatial coordinates. The new segmentation procedure implemented in SPM8 (http://www.fil.ion.ucl.ac.uk/spm), running under Matlab 7.1 (Math Works, Natick, MA, USA) was used to segment all the images into gray matter (GM) and white matter (WM). A fast diffeomorphic image registration algorithm (DARTEL; Ashburner, 2007) was used to warp the GM partitions into a new study-specific reference space representing an average of all the subjects included in the analysis (Ashburner and Friston, 2009; Yassa and Stark, 2009). As an initial step, two different templates (one for each data set) and the corresponding deformation fields, required to warp the data from each subject to the new reference space, were created using the GM partitions (Ashburner and Friston, 2009). Each subject-specific deformation field was then used to warp the corresponding GM partition into the new reference space with the aim of maximizing accuracy and sensitivity (Yassa and Stark, 2009). Images were, finally, affine transformed into Montreal Neurological Institute (MNI) space and smoothed with a 4, 8 and 12-mm full-width at half-maximum (FWHM) Gaussian kernel. The above procedure was followed twice to create both unmodulated and modulated images, which were analyzed separately. The analysis on unmodulated data was performed on groups with sample size 16 and smoothing 8 mm only, consistent with our previous investigation (Scarpazza et al., 2013).

2.3.2. Group comparisons

Using SPM8, for each data set we performed 300 group comparisons including 100 comparisons between 2 groups of 16 subjects; 100 comparisons between 2 groups of 12 subjects; and 100 comparisons between 2 groups of 8 subjects. The groups used in all comparisons were created using randomization as implemented in Microsoft Excel software. A sample size of 8, 12 and 16 was chosen for three main reasons. Firstly, a typical neuroimaging study of regional differences includes 8–16 subjects per experimental group (Friston et al., 1999). Secondly, a recent analysis of the effect size in classical inference has suggested that, in order to optimize the sensitivity to large effects while minimizing the risk of detecting trivial effects, the sufficient sample size for a study is 16 (Friston, 2012); this investigation also highlighted the common misconception that smaller sample sizes lead to higher false positives rates. Thirdly, we wanted to examine the impact of decreasing sample size since parametric statistics appear to be more prone to deviation from normality for smaller sample sizes (Salmond et al., 2002; Scarpazza et al., 2013). In all comparisons, age and gender were entered into the design matrix as covariates of no interest. Voxels outside the brain were excluded by employing an implicit mask that removed all voxels whose intensity fell below 20% of the mean image intensity. The proportional scaling option was used to

identify regionally specific changes that were not confounded by global differences.

2.3.3. Statistical analysis

For each group comparison, two two-sample *t*-tests were used to identify increases and decreases in one group relative to the other respectively. Statistical inferences were made at voxel-level using a threshold of $p < 0.05$ with family-wise error (FWE) correction for multiple comparisons across the whole brain. No extent threshold was used since the main aim of the current investigation was to quantify the number of false positive results irrespective of cluster size. When significant between-group differences were detected, we refer to Group 1 > Group 2 to indicate increased GM volume in Group 1 compared to Group 2, while we refer to Group 1 < Group 2 to indicate decreased GM volume in Group 1 compared to Group 2.

For each data source (Beijing and Cambridge), we counted the number of comparisons yielding statistically significant differences (out of 100) over the three smoothing kernels (4, 8 and 12 mm), three sample sizes (16, 12 and 8 subjects per group), two preprocessing types (modulated, unmodulated) and two directions (Group 1 > Group 2; Group 1 < Group 2).

In order to investigate whether smoothing, sample size and direction had a significant impact on the number of false positive rates in the context of modulated data, we used the Statistical Package for the Social Sciences 22.0 (IBM SPSS Statistics 22.0, Chicago, IL, USA) to fit a logistic regression model from each data source, using the presence of a statistically significant difference in each comparison (yes or no) as dependent variable, and smoothing, sample size and direction as independent variables. For 8 mm smoothing and sample size of 16 subjects both modulated and unmodulated data were available, and therefore we also fit a further logistic regression model; here the dependent variable was the presence of a statistically significant difference in each comparison (yes or no), and the independent variables were modulation and direction (with only 8 mm smoothing and a sample size of 16 subjects, smoothing and sample size were not modeled). Both logistic regression models were assessed using the Hosmer–Lemeshow goodness-of-fit test, where a statistically significant *p*-value indicates lack-of-fit.

2.3.4. Brain areas individuation

From the SPM output, i.e. the list of MNI coordinates of the areas showing significant increases or decreases, we derived the corresponding areas using the Automated Anatomical Labeling (AAL) atlas as implemented in PickAtlas software (http://fmri.wfubmc.edu/software/PickAtlas).

3. Results

3.1. Number of comparisons yielding significant differences

When differences in each direction where considered separately (one-tailed), the number of comparisons yielding at least one false positive result was no more than 5% regardless of the sample size used and smoothing applied, consistent with our prediction for one-tailed tests. This was the case for both data sets, see Table 1 for details. When differences in the two directions were combined (two-tailed), the number of comparisons yielding at least one false positive result in either direction was no more than 10%, consistent with our prediction for two-tailed tests. Again, this was the case for both data sets, see Table 1 for details.

3.2. Impact of smoothing, sample size and direction of effect

The Hosmer–Lemeshow test was not significant ($p = 0.915$ and $p = 0.953$ for the Beijing and Cambridge data sets respectively), consistent with a null hypothesis of good model fit.

The impact of smoothing on the false positive rate was not significant, in either the Beijing ($p = 0.178$) or the Cambridge ($p = 0.162$) data set. Similarly, the impact of sample size on the false positive rate was not significant, in either the Beijing ($p = 0.847$) or the Cambridge ($p = 0.162$) data set. Finally, as one would expect given that all groups were created using randomization, the number of false positives did not vary depending on the direction of the effect under consideration (i.e. Group 1 > Group 2 or Group 1 < Group 2); this was the case both for the Beijing ($p = 0.636$) and the Cambridge ($p = 0.192$) data sets. Overall, these results indicate that smoothing, sample size and direction of the effect under investigation had no effect on the number of significant differences in the two data sets.

3.3. Impact of modulation

The Hosmer–Lemeshow test was not significant ($p = 0.153$ and $p = 0.669$ for the Beijing and Cambridge data sets, respectively), consistent with a null hypothesis of good model fit.

The impact of modulation on the false positive rate was not significant, in either the Beijing ($p = 1$) or the Cambridge ($p = 0.760$) data set.

3.4. Likelihood of detecting local maxima in a specific region

In addition to the number of comparisons yielding significant results, we also considered the location of the significant clusters (reported as absolute number in brackets in Table 1). With respect to comparisons performed on modulated images only, and pooling all the results obtained with different sample size and smoothing, 55 clusters were identified in the Beijing data set (2 of which out of the brain and then removed from the following statistics), and 50 clusters in the Cambridge data set (1 of which out of the brain and then removed from the following statistics). The significant differences were distributed throughout the cortex (44 clusters out of 53, 82.7% of the total findings in Beijing data set and 41 clusters out of 49, 83.6% of the total findings in Cambridge data set) with very few differences detected in subcortical regions (1 cluster in each data set, 1.8% and 2% in the Beijing and Cambridge data sets respectively). Additional differences were detected in the cingulate cortex (2 clusters out of 53, 3.8% of the total findings in the Beijing data set and 4 clusters out of 49, 8.1% of the total findings in the Cambridge data set), the insula (1 cluster only, 1.8% of the total findings, in the Cambridge data set) and the cerebellum (6 clusters out of 53, 11.3% of the total findings in the Beijing data set and 2 clusters out of 49, 4% of the total findings in the Cambridge data set). These results are summarized in Table 2 and represented graphically in Fig. 1; in addition, the location of each significant cluster can be found in the Supplementary Material.

Moreover, we observed that the significant differences were mainly located in the frontal lobe (21 clusters out of 53, 39.6% of the total findings in the Beijing data set and 16 clusters out of 49, 32.6% of the total findings in the Cambridge data set) compared to the other lobes (parietal: 10/53, 18.8% and 9/49, 18.3% in the Beijing and Cambridge data sets respectively; temporal: 4/53, 7.4% and 12/49, 24%; occipital: 9/53, 16.9% and 4/49, 8.1%).

In order to investigate whether the larger number of false positives in the frontal lobe relative to other regions of the brain could be explained by differences in size (Semendeferi et al., 1997), we estimated the volume (mm^3 and percentage) of each region of interest reported in Table 2 using PickAtlas. We then used the *z* test as implemented in SPSS (IBM SPSS Statistics 22.0, Chicago, IL, USA) to investigate whether the number of false positives in each region was proportional to the regional volume. The *z* test revealed that, in each region of interest, the number of false positives was proportional to the regional volume ($p > 0.05$). Output tables for

Table 1
Number of significant differences. Numbers of comparisons yielding statistically significant differences between groups as a function of smoothing (4 mm, 8 mm, 12 mm), sample size (n = 8, 12, 16) and modulation (modulated, unmodulated); as some comparisons yielded more than one significant difference, the total number of clusters across all comparisons is also reported in brackets. We report this information for increases and decreases separately (Group 1 > Group 2, Group 1 < Group 2) as well in combination (total). All differences were identified using a statistical threshold of $p < 0.05$ (FWE corrected).

		4 mm			8 mm			12 mm		
		Group 1 > Group 2	Group 1 < Group 2	Total	Group 1 > Group 2	Group 1 < Group 2	Total	Group 1 > Group 2	Group 1 < Group 2	Total
16 vs 16 Modulated	Beijing	4 (7)	2 (3)	6 (10)	3 (5)	1 (1)	4 (6)	3 (7)	3 (3)	6 (10)
	Cambridge	1 (1)	0 (0)	1 (1)	2 (2)	3 (3)	5 (5)	1 (1)	4 (5)	5 (6)
12 vs 12 Modulated	Beijing	4 (4)	3 (3)	7 (7)	0 (0)	2 (2)	2 (2)	1 (1)	0 (0)	1 (1)
	Cambridge	2 (3)	2 (2)	4 (5)	1 (1)	2 (2)	3 (3)	1 (1)	1 (1)	2 (2)
8 vs 8 Modulated	Beijing	3 (3)	3 (3)	6 (6)	2 (2)	3 (3)	5 (5)	2 (4)	3 (4)	5 (8)
	Cambridge	1 (2)	3 (3)	4 (5)	2 (6)	3 (3)	5 (9)	4 (9)	5 (5)	9 (14)
16 vs 16 Unmodulated	Beijing				1 (1)	3 (3)	4 (4)			
	Cambridge				4 (5)	2 (2)	6 (7)			

Table 2
The table reported the volume in mm^3 of each cerebral region. The percentage has been calculated on a total of 1583 mm^3 of total intracranial volume. Absolute number and proportion of statistically significant differences in different cortical and subcortical areas were reported for Beijing and Cambridge data sets, separately.

	Volume (mm^3)	%	Beijing (n = 53 clusters)		Cambridge (n = 49 clusters)	
			Raw number	%	Raw number	%
Frontal lobe	562.6	35.5	21	39.6	16	32.6
Parietal lobe	214.8	13.51	10	18.8	9	18.3
Temporal lobe	258.7	16.29	4	7.4	12	24
Occipital lobe	170.6	10.73	9	16.9	4	8.1
Insula	29	1.83	0	–	1	2
Cingulate	61.2	3.85	2	3.7	4	8.1
Subcorical structures	89.7	5.62	1	1.8	1	2
Cerebellum	196.9	12.38	6	11.3	2	4
Outside the brain			(2)		(1)	

Beijing and Cambridge respectively are reported in Supplementary Material (Tables S2 and S3).

Moreover, in order further explore the association between number of false positives and regional volume, we estimated Spearman's correlations for the two data sets separately. The correlations were significant both in the Beijing ($R = 0.80$, $p = 0.01$) and the Cambridge ($R = 0.84$, $p = 0.008$) data sets. These results are represented graphically in the Supplementary Material (Fig. S2).

4. Discussion

Previous investigations have used VBM to investigate brain abnormalities in a wide range of neurological and psychiatric disorders (Mechelli et al., 2005). However, previous simulations suggest that this technique may be susceptible to high false positive rates, particularly when the residuals are not normally distributed (Salmond et al., 2002; Scarpazza et al., 2013). The present study

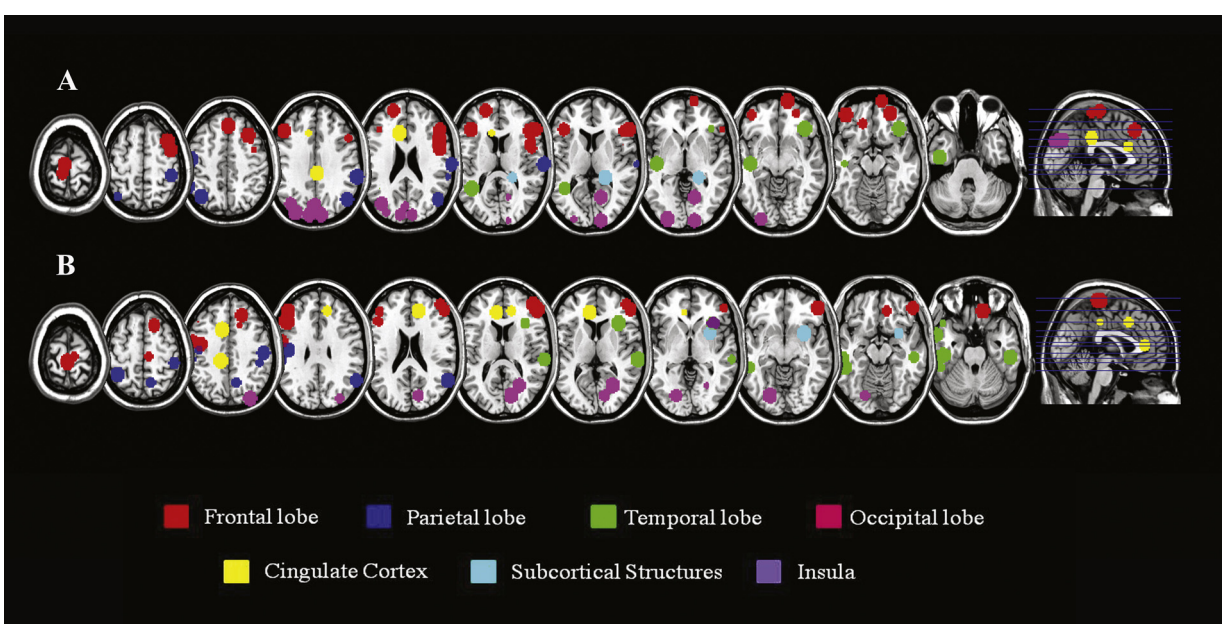

Fig. 1. Localization of statistically significant clusters in the Beijing (A) and Cambridge (B) data sets across all statistical analyses with modulated images. This image was created for illustration purposes using coordinate-based ROIs with 10 mm radius, with the center of each ROI located in the local maxima of the corresponding cluster. The 10 mm radius was chosen for display purposes in order to make each cluster clearly visible.

aimed to investigate whether the surprisingly high false positive rates found in single case VBM studies would also be evident in VBM studies in which groups of equal size are compared. This was achieved by empirically estimating the likelihood of detecting significant differences when comparing groups of healthy subjects in two independent, freely available data sets. Such empiricism was preferred to a simulation-based approach given recent evidence demonstrating a discrepancy in results between real and simulated neuroimaging data (Silver et al., 2011).

We tested three hypotheses based on the existing literature: firstly, we hypothesized that false positive rates would be about 5% (for one-tailed *t* test), in contrast with the very high false positive rates observed in the context of single case VBM; secondly, we expected that false positive rates would vary as a function of sample size (with a higher number of differences detected for smaller sample size), degree of smoothing applied to the data (with a higher number of differences detected for smaller kernel smoothing), and modulation (with and without modulation); thirdly, we hypothesized that significant differences would be mainly located in the frontal and temporal lobes.

Concerning the first hypothesis, when increases (i.e. Group 1 > Group 2) and decreases (i.e. Group 1 < Group 2) were considered separately, we detected a false positive rate of less than 5%. Critically, this result was replicated using two independent data sets acquired from subjects of different ethnicities, using different scanners, and different acquisition sequences. Therefore, our first hypothesis was confirmed: in VBM with balanced designs the likelihood of detecting a significant difference is not higher than expected. This provides reassurance that, when groups of equal size are compared, VBM is not susceptible to the violation of the assumption of normality that is responsible for high false positive rates in single case VBM (Scarpazza et al., 2013).

In contrast with our second hypothesis, we found that the number of false positives is not affected by the degree of smoothing, sample size or modulation. The null effect of smoothing replicates a previous investigation reporting that, in the context of balanced group comparisons, smoothing at 4 mm is sufficient to ensure that any non-normality has minimal impact on false positive rate (Salmond et al., 2002). In contrast, smoothing is not sufficient to prevent an escalation of false positive rate in the context of unbalanced comparisons (Salmond et al., 2002; Scarpazza et al., 2013). In addition the null effect of sample size suggests that, as long as a balanced design is employed, the number of subjects in each experimental group appears to have little or no impact on false positive rate. Again, this observation is in contrast with our previous finding that sample size moderates false positive rate in the context of single case VBM. Finally, the null effect of modulation suggests that false positive rates are comparable for modulated and unmodulated data, in contrast with our previous observation of higher false positive rates for unmodulated relative to modulated data in single case VBM (Scarpazza et al., 2013). Taken collectively, these results are consistent with the notion that VBM with balanced designs is robust against violation of the assumption of normality, regardless of the degree of smoothing, the sample size and the use of modulation. However, the non-significant effects of degree of smoothing, sample size and modulation might also be explained by the very small number of false positive effects in the present investigation relative to our previous study (Scarpazza et al., 2013), which may have resulted in reduced statistical sensitivity to these variables of interest.

In contrast with our third hypothesis, we found that significant differences were randomly distributed across the whole cortex; for example, the greater number of false positives in the frontal lobe relative to other lobes could be explained in terms of the former being larger than the latter. This is inconsistent with our previous report of a higher proportion of false positives in frontal and temporal regions in the context of single case VBM (Scarpazza et al., 2013). We speculate that greater individual variability in frontal and temporal cortices (Semendeferi et al., 1997) may result in greater violation of the assumption of normality in these regions, and that this is a concern in the context of single case VBM but not when groups of equal size are compared.

A limitation of the present study is that the statistical comparisons carried out within each data set were not completely independent, as the same subject could be present in more than one statistical comparison as a result of the repeated randomization process used to create each group. However, there is no reason to believe that this led to a systematic bias in our estimation of false positive rates. A second limitation is that we investigated false positive rates for a limited range of sample sizes ($n = 8, 12, 16$) and smoothing kernels (4 mm, 8 mm and 12 mm); however, these parameters were chosen based on the existing literature (Friston et al., 1999; Friston, 2012; Salmond et al., 2002; Scarpazza et al., 2013). The exploration of a larger range of parameters was outside the scope of the present investigation and would require greater much computational resources.

In conclusion, the present investigation provides empirical evidence that, in VBM studies employing a balanced design, the likelihood of detecting a significant difference is not higher than expected. This was replicated in two independent data sets, and did not appear to be influenced by the degree of smoothing, sample size or modulation. These results provide reassurance that VBM studies comparing groups of equal size are not vulnerable to the higher than expected false positive rates evident in single case VBM. It follows that non parametric statistics may be indicated in the context of single case VBM but are not required in VBM studies employing a balanced design. A final consideration is that the present investigation used two freely available data sets from the NITRC database; we believe that this well illustrates the potential of sharing large data sets for accelerating research about the human brain.

Acknowledgments

This research was supported by a grant (ID99859) from the Medical Research Council (MRC) to AM. The authors would like to thank Dr. Zang and Dr. Buckner for providing the data through the Neuroimaging Informatics Tools and Resources Clearinghouse. We are grateful to Dr. William Pettersson-Yeo for revising an initial draft of the manuscript.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.neubiorev.2015.02.008.

References

Ashburner, A., Friston, K., 2000. Voxel-based Morphometry – the methods. NeuroImage 11, 805–821.
Ashburner, A., Friston, K., 2001. Why Voxel-based Morphometry should be used. NeuroImage 14, 1238–1243.
Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. NeuroImage 38 (1), 95–113.
Ashburner, J., 2012. SPM: a history. NeuroImage 62 (2), 791–800.
Ashburner, J., Friston, K.J., 2009. Computing average shaped tissue probability templates. NeuroImage 45 (2), 333–341.
Bandettini, P.A., 2009. What's new in neuroimaging methods? Ann. N. Y. Acad. Sci. 1156, 260–293.
Biswal, B.B., Mennes, M., Zuo, X.N., Gohel, S., Kelly, C., Smith, S.M., Beckmann, C.F., Adelstein, J.S., Buckner, R.L., Colcombe, S., Dogonowski, A.M., Ernst, M., Fair, D., Hampson, M., Hoptman, M.J., Hyde, J.S., Kiviniemi, V.J., Kötter, R., Li, S.J., Lin, C.P., Lowe, M.J., Mackay, C., Madden, D.J., Madsen, K.H., Margulies, D.S., Mayberg, H.S., McMahon, K., Monk, C.S., Mostofsky, S.H., Nagel, B.J., Pekar, J.J., Peltier, S.J., Petersen, S.E., Riedl, V., Rombouts, S.A., Rypma, B., Schlaggar, B.L., Schmidt, S., Seidler, R.D., Siegle, G.J., Sorg, C., Teng, G.J., Veijola, J., Villringer, A., Walter, M., Wang,

L., Weng, X.C., Whitfield-Gabrieli, S., Williamson, P., Windischberger, C., Zang, Y.F., Zhang, H.Y., Castellanos, F.X., Milham, M.P., 2010. Toward discovery science of human brain function. Proc. Natl. Acad. Sci. U. S. A. 107 (10), 4734–4739.
Bora, E., Fornito, A., Radua, J., Walterfang, M., Seal, M., Wood, S.J., Yücel, M., Velakoulis, D., Pantelis, C., 2011. Neuroanatomical abnormalities in schizophrenia: a multimodal voxelwise meta-analysis and meta-regression analysis. Schizophr. Res. 127 (1–3), 46–57.
Chung, M.K., Worsley, K.J., Robbins, S., Paus, T., Taylor, J., Giedd, J.N., Rapoport, J.L., Evans, A.C., 2003. Deformation-based surface morphometry applied to gray matter deformation. NeuroImage 18 (2), 198–213.
Crum, W.R., Griffin, L.D., Hill, D.L.G., Hawkes, D.J., 2003. Zen and the art of medical image registration: correspondence, homology, and quality. NeuroImage 20, 1425–1437.
Dell'Acqua, F., Catani, M., 2012. Structural human brain networks: hot topics in diffusion tractography. Curr. Opin. Neurol. 25 (4), 375–383.
Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study. NeuroImage 10, 1–5.
Friston, K.J., 2012. Ten ironic rules for non-statistical reviewers. NeuroImage 61 (4), 1300–1301.
Good, C.D., Johnsrude, I.S., Ashburner, J., Henson, R.N.A., Friston, K.J., Frackowiak, S.J., 2001. A voxel based morphometric study of ageing in 456 normal adult human brains. NeuroImage 14, 21–36.
Honea, R., Crow, T.J., Passingham, D., Mackay, C.E., 2005. Regional deficits in brain volume in schizophrenia: a meta-analysis of voxel-based morphometry studies. Am. J. Psychiatry 162 (12), 2233–2245.
Hu, X., Erb, M., Ackermann, H., Martin, J.A., Grodd, W., Reiterer, S.M., 2011. Voxel based morphometry studies of personality: issue of statistical model specification – effect of nuisance covariates. NeuroImage 54 (3), 1994–2005.
Iwabuchi, A.J., Liddle, P.F., Palaniyappan, L., 2013. Clinical utility of machine learning approaches in schizophrenia: improving diagnostic confidence for translational neuroimaging. Front. Psychiatry 4, 95.
Kennedy, K.M., Erickson, K.I., Rodrigue, K.M., Voss, M.W., Colcombe, S.J., Kramer, A.F., Acker, J.D., Raz, N., 2009. Age-related differences in regional brain volumes: a comparison of optimized voxel-based morphometry to manual volumetry. Neurobiol. Aging 30 (10), 1657–1676.
Lai, C.H., 2013. Gray matter volume in major depressive disorder: a meta-analysis of voxel-based morphometry studies. Psychiatry Res. 211 (1), 37–46.
Lansley, J., Mataix-Cols, D., Grau, M., Radua, J., Sastre-Garriga, J., 2013. Localized grey matter atrophy in multiple sclerosis: a meta-analysis of voxel-based morphometry studies and associations with functional disability. Neurosci. Biobehav. Rev. 37 (5), 819–830.
Li, J., Pan, P., Huang, R., Shang, H., 2012. A meta-analysis of voxel-based morphometry studies of white matter volume alterations in Alzheimer's disease. Neurosci. Biobehav. Rev. 36 (2), 757–763.
Lieberman, M.D., Cunningham, W.A., 2009. Type I and Type II error concerns in fMRI research: re-balancing the scale. Soc. Cogn. Affect. Neurosci. 4 (4), 423–428.
Mechelli, A., Crinion, J.T., Noppeney, U., O'Doherty, J., Ashburner, J., Frackowiak, R.S., Price, C.J., 2004. Neurolinguistics: structural plasticity in the bilingual brain. Nature 431 (7010), 757.
Mechelli, A., Price, C.J., Friston, K.J., Ashburner, J., 2005. Voxel based morphometry of the human brain: methods and applications. Curr. Med. Imaging Rev. 1, 105–113.
Mechelli, A., Riecher-Rössler, A., Meisenzahl, E.M., Tognin, S., Wood, S.J., Borgwardt, S.J., Koutsouleris, N., Yung, A.R., Stone, J.M., Phillips, L.J., McGorry, P.D., Valli, I., Velakoulis, D., Woolley, J., Pantelis, C., McGuire, P., 2011. Neuroanatomical abnormalities that predate the onset of psychosis: a multicenter study. Arch. Gen. Psychiatry 68 (5), 489–495.
Pan, P.L., Shi, H.C., Zhong, J.G., Xiao, P.R., Shen, Y., Wu, L.J., Song, Y.Y., He, G.X., Li, H.L., 2013. Gray matter atrophy in Parkinson's disease with dementia: evidence from meta-analysis of voxel-based morphometry studies. Neurol. Sci. 34 (5), 613–619.
Papagni, S.A., Benetti, S., Arulanantham, S., McCrory, E., McGuire, P., Mechelli, A., 2011. Effects of stressful life events on human brain structure: a longitudinal voxel-based morphometry study. Stress 14 (2), 227–232.
Radua, J., van den Heuvel, O.A., Surguladze, S., Mataix-Cols, D., 2010. Meta-analytical comparison of voxel-based morphometry studies in obsessive–compulsive disorder vs other anxiety disorders. Arch. Gen. Psychiatry 67 (7), 701–711.
Sacher, J., Neumann, J., Okon-Singer, H., Gotowiec, S., Villringer, A., 2013. Sexual dimorphism in the human brain: evidence from neuroimaging. Magn. Reson. Imaging 31 (3), 366–375.
Salmond, C.H., Ashburner, J., Vargha-Khadem, F., Connelly, A., Gadian, D.G., Friston, K.J., 2002. Distributional assumptions in voxel-based morphometry. NeuroImage 17, 1027–1030.
Scarpazza, C., Sartori, G., De Simone, M.S., Mechelli, A., 2013. When the single matters more than group: very high false positive rates in single case Voxel Based Morphometry. NeuroImage 70, 175–188.
Selvaraj, S., Arnone, D., Job, D., Stanfield, A., Farrow, T.F., Nugent, A.C., Scherk, H., Gruber, O., Chen, X., Sachdev, P.S., Dickstein, D.P., Malhi, G.S., Ha, T.H., Ha, K., Phillips, M.L., McIntosh, A.M., 2012. Grey matter differences in bipolar disorder: a meta-analysis of voxel-based morphometry studies. Bipolar Disord. 14 (2), 135–145.
Semendeferi, K., Damasio, H., Frank, R., Van Hoesen, G.W., 1997. The evolution of the frontal lobes: a volumetric analysis based on three-dimensional reconstructions of magnetic resonance scans of human and ape brains. J. Hum. Evol. 32 (4), 375–388.
Silver, M., Montana, G., Nichols, T.E., The Alzheimer's Disease NeuroImage Initiative, 2011. False positives in neuroimaging genetics using voxel-based morphometry data. NeuroImage 54 (2), 992–1000.
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.
Streitbürger, D.P., Pampel, A., Krueger, G., Lepsien, J., Schroeter, M.L., Mueller, K., Möller, H.E., 2014. Impact of image acquisition on voxel-based-morphometry investigations of age-related structural brain changes. NeuroImage 87, 170–182.
Takahashi, R., Ishii, K., Kakigi, T., Yokoyama, K., 2011. Gender and age differences in normal adult human brain: voxel-based morphometric study. Hum. Brain Mapp. 32 (7), 1050–1058.
Tardiff, C.L., Collins, D.L., Pike, G.B., 2009. Sensitivity of voxel-based morphometry analysis to choice of imaging protocol at 3 T. NeuroImage 44, 827–838.
Viviani, R., Beschoner, P., Ehrhard, K., Schmitz, B., Thöne, J., 2007. Non-normality and transformations of random fields, with an application to voxel-based morphometry. NeuroImage 35 (1), 121–130.
Yassa, M.A., Stark, C.E., 2009. A quantitative evaluation of cross-participant registration techniques for MRI studies of the medial temporal lobe. NeuroImage 44 (2), 319–327.