Available online at www.sciencedirect.com

ScienceDirect

Magnetic Resonance Imaging 25 (2007) 1385–1396

MAGNETIC RESONANCE IMAGING

VBM lesion detection depends on the normalization template: a study using simulated atrophy

Shan Shen*, Andre J. Szameitat, Annette Sterr

Department of Psychology, University of Surrey, GU2 7XH Guildford, UK

Received 24 January 2007; revised 14 March 2007; accepted 14 March 2007

Abstract

Structural neuroimaging studies are of great interest for neuroscientists, which are reflected in the rising number of papers using voxel-based morphometry (VBM). One major step in VBM is the transformation of images to a standard template, a spatial normalization necessary to ensure that homologous regions are compared while interindividual characteristics are maintained. Templates can be created in different ways, and this may affect the likelihood that differences in gray/white matter density between groups are detected. However, studies investigating the interaction of normalization template and VBM accuracy are sparse. Existing work is based on patient–control group comparisons, and the emerging results are inconclusive. The present paper therefore used simulated atrophy in a simplified one-lesion model to systematically study template effects of VBM analyses implemented in SPM. This allowed us to characterize template-specific biases in reference to a set of prespecified parameters of anatomical difference. The data suggest that the likelihood of correctly detecting the prespecified lesion is modulated by the normalization template. Thereby, the relationship between template-related VBM accuracy and specific group/study characteristics is complex, and there does not appear to be one ‘best template.’ Our data show that template effects are critical and clearly suggest that the choice of template needs careful consideration in relation to the specific research question and study constraints.

Keywords: Voxel-based morphometry; Spatial normalization; Template; Simulated atrophy

1. Introduction

The investigation of structural brain characteristics among different study populations is an important issue in cognitive neuroscience research, which allows to link structural brain characteristics to cognition, behavioral phenomena and clinical symptoms. With the fast advancement of magnetic resonance imaging (MRI) with higher field strengths, structural imaging is becoming increasingly more feasible and popular. Voxel-based morphometry (VBM) is an automated algorithm that allows the identification of differences in tissue concentration. It enables the structural analysis of the whole brain without the necessity to define a prior region of interests [1–3]. Because of these advantages, this technique has been employed by many researchers to investigate gray matter (GM)/white matter (WM) changes in various clinical conditions, such as schizophrenia, Alzheimer’s disease and epilepsy [4–6], as well as the relationship between GM concentration and intelligence [7].

The SPM[1] (Statistical Parametric Mapping) software package (http://www.fil.ion.ucl.ac.uk/spm/) is currently the most popular VBM analysis tool [8–10]. In SPM, the standard procedure of VBM includes the following steps. First, MR images are spatially normalized to a standard template in order to coregister homologous brain regions across subjects. Then, the brain tissue is extracted from the normalized brain and smoothed to reduce the residual anatomical variability between subjects. Finally, a voxel-wise statistical analysis of the generalized linear model (GLM) is performed to identify the structural differences between groups. Among these steps, spatial normalization is crucial, as it is essentially required to ensure that homologous regions are compared across subjects and to

* Corresponding author. Tel.: +44 0 1483 682881; fax: +44 0 1483 689553.

E-mail address: shan.shen@surrey.ac.uk (S. Shen).

[1] Developed by the Wellcome Department of Imaging Neuroscience, University College London.

0730-725X/$ – see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.mri.2007.03.025

maintain interindividual anatomical differences at the same time [11]. More specifically, spatial normalization does not attempt to perfectly match every cortical feature but corrects for global brain shape differences [1]. Because of the complexity of brain structures, the normalization algorithm is always an approximation. We therefore hypothesized that the characteristics of the employed template may change normalized images in a way that modulates the likelihood of detecting GM/WM differences with VBM. We tested this idea in two stroke patients by applying two different templates, the MNI (Montreal Neurological Institute) and a so-called customer-specific template (details given below), and found clearly different results (see Fig. 1) for the two methods.

In the literature, one of the most prevalent templates for spatial normalization is the MNI template [3,8,10,12–15]. It is an average brain created from 152 T_1-weighted images, collected in a 1.5-T scanner. However, despite its popularity, the use of the MNI template is not undisputed and has been criticized for its contrast differences to most MR images [16]. The criticism includes the assumption that each scanner introduces specific nonuniformities in image intensity [17] and the differences in field strength of the scanner (e.g., 3-T scanner). A further critique relates to the demographic differences between the MNI template population and the cohort investigated in a particular experiment [16]. For instance, some recent experiments suggest GM density changes with aging [18] as well as

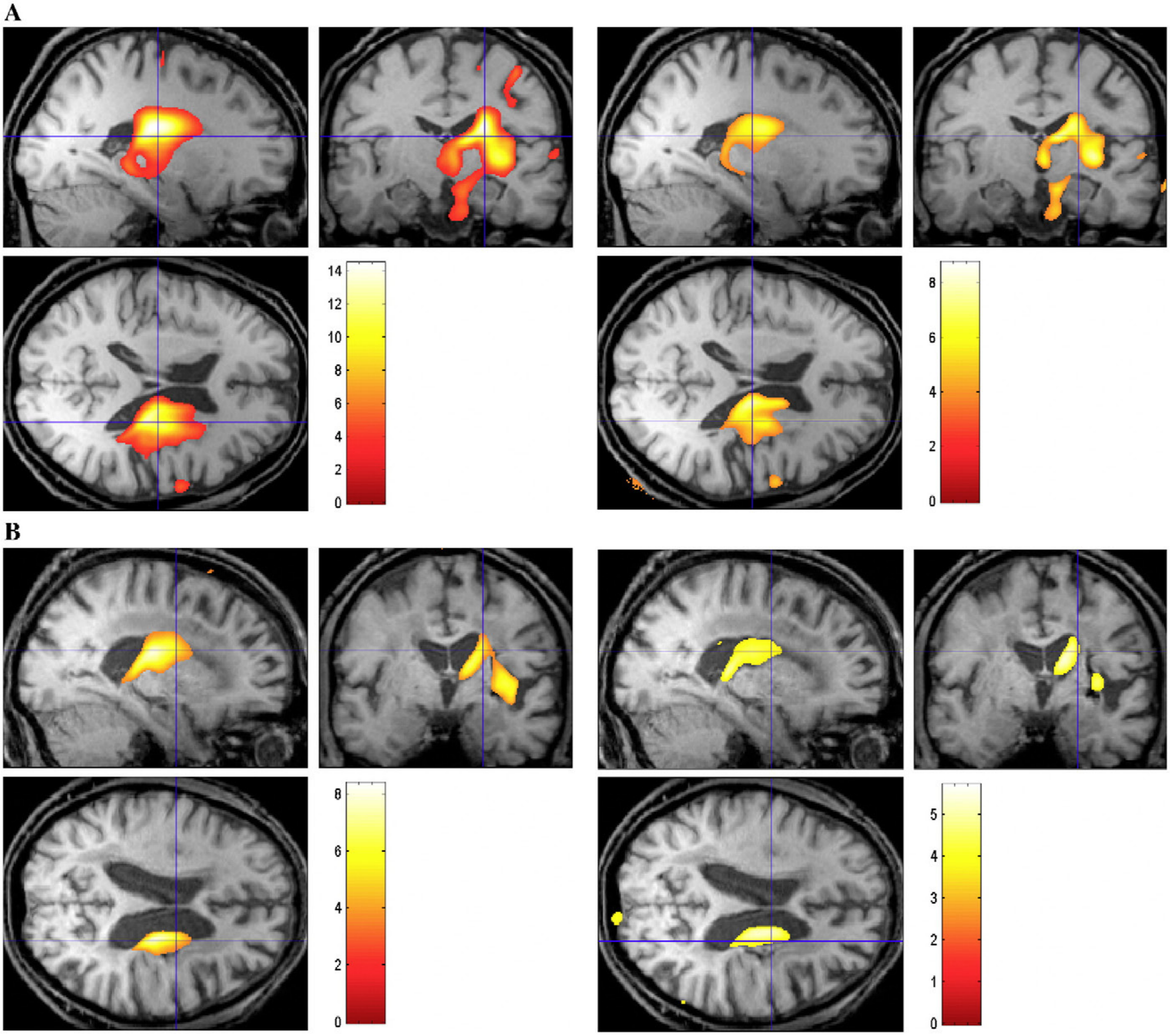

Fig. 1. Lesion detection in two stroke patients using VBM (single patient vs. 50 healthy control subjects; analysis details are the same as described in Section 2) with different templates. (A) Patient 1: left, MNI template (detected lesion size: 53,817 voxels); right, customer-specific template (detected lesion size: 28,075 voxels). (B) Patient 2: left, MNI template (detected lesion size: 19,163 voxels); right, customer-specific template (detected lesion size: 8744 voxels).

intelligence [7]. Therefore, if the MNI template is chosen, potentially influencing demographic factors are not taken into account.

In the light of these considerations, some researches have used customer-specific templates, that is, a template created specifically for one particular experiment [4,5,11]. The images used to create the template are typically acquired in the same scanner as images from the control group and the study group in this experiment, while major demographic variables are matched. The group of major interest, such as patients, is defined as the study group, while the compared group, composed of the healthy subjects, is called the control group. At first glance, customer-specific templates may appear to be the more appropriate method. However, there are several options for selecting images for template creation, and these different options may potentially create biases for VBM. In the literature, several image selections are presented: (a) Most popularly, images from the healthy control group in a particular study are chosen for template creation [6,19]. (b) Some research employed images from a subset of the control group [17]. The other options included (c) combined images from the control group and the study group [20–22] and (d) combined images from a subset of the control group plus a subset of a different healthy population [12]. (e) In the situation that a standard template may be required in a series of studies, images acquired from a different cohort rather than the specific control group were employed [11]. We call these types of templates 'customer-standard templates.' (f) Over and above these template-creation methods, Duchesne et al. [23] proposed the idea that normalizing the study group and the control group to their own templates (i.e., the study group has its own customer-specific template, so does the control group), rather than normalizing both to an identical template, would decrease the within-group anatomical variability and, therefore, increase the between-group separation. We call these templates 'group-specific templates.'

Given the crucial role of the normalization step in VBM, it is conceivable that the method of template creation, that is, the image information utilized, may interact with VBM sensitivity. Our data from two stroke patients (Fig. 1) clearly support this hypothesis. There are some studies present on the potential effects of different templates in VBM [11,20,21], but not one has investigated the effects of the type of image information selected for creating customer-specific templates. A systematic investigation of how this impacts on VBM appears, therefore, opportune.

In contrast to existing papers, which all conducted group comparisons, the present paper addressed this issue by using a simulated atrophy model that gave us a prespecified lesion, which allowed us to quantify the detection accuracy in terms of location and extent. In other words, through a simulation algorithm, we generated a study group with a known lesion from the images of normal control participants and subsequently applied different templates for normalization. By using the same group of participants for VBM analysis, our method excluded the potential group variation. More specifically, with a traditional approach using two different groups, the group variation may alter the size of detected atrophy and, hence, introduce a bias. Therefore, detected group difference in VBM=real difference (simulated atrophy)+template effect+group bias. Because the latter is most likely unknown, the template effects cannot be quantified. However, while excluding the group variation, our simplified design also excluded the possibility of the false-positive detection in the whole brain. This implied that the reported template effects are confined to the lesioned area. We hypothesized that the detection accuracy of the lesion would vary for the different templates and assumed that the smaller the disparity between detected and real (simulated) GM difference, the lower the template effect. It is worth noticing that using a simulated atrophy model instead of simulated stroke or other focal damage is to avoid the potential tissue misclassification, which may, in turn, affect registration results.

2. Methods

2.1. Participants and MR image acquisition

Fifty healthy subjects (24 males, 26 females; mean age=26.4 years, S.D.=9.2) were scanned on a 3-T Siemens Trio scanner (Erlangen, Germany). High-resolution 3D brain MR images were obtained using a T_1-weighted magnetization-prepared rapid acquisition gradient echo pulse sequence, with the following characteristics: TR=1830 ms, TE=4.43 ms, inversion time=1100 ms, 1 acquisition, flip angle=11°, FOV=256 mm, 176 slices, voxel size=1×1×1 mm^3, in-plane matrix=256×256. All participants were confirmed as right-handed using the Edinburgh Handedness Inventory [24]. Prior to scanning, all participants gave written informed consent according to the guidelines of the University of Surrey Ethical Review Board. Participants were paid for their participation.

2.2. Simulation of atrophy

The atrophy simulation method was developed by Karacali and Davatzikos [25,26] (University of Pennsylvania). It is

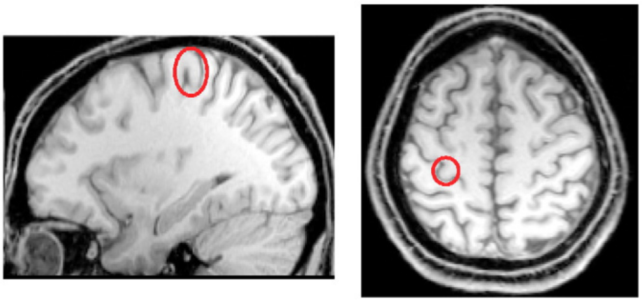

Fig. 2. Landmark for simulated atrophy: motor hand area on the left precentral gyrus.

based on an energy-minimization approach in which a warping transformation is optimized to produce a prescribed level of atrophy in a specified brain region. In other words, the corresponding level of volumetric changes, specifically, volumetric loss, is simulated using the acquired deformation array. In order to make the simulated atrophy more realistic, the following constrains are applied: Firstly, atrophy is obtained with no volumetric restrictions over the cerebrospinal fluid (CSF) as the CSF simply fills the void left by retreating tissue. Secondly, the deformation over the skull is zero because the atrophy is to brain tissues only. Finally, simulation is implemented within one class of brain tissues (i.e., either GM or WM) to preserve topology of the brain.

The simulation program runs in Linux. The landmark selected for atrophy simulation in our study is a knob-like structure on the left precentral gyrus. It is located in the motor hand area and is shaped like an omega or epsilon in the axial plane and like a hook in the sagittal plane [27]. This region was chosen because it represents a very robust anatomical landmark, easily identified in each individual (Fig. 2). In all 50 images acquired from the healthy participants, there was an average GM reduction of 5% (S.D.=1.37%) in the surrounding area of the landmark (a

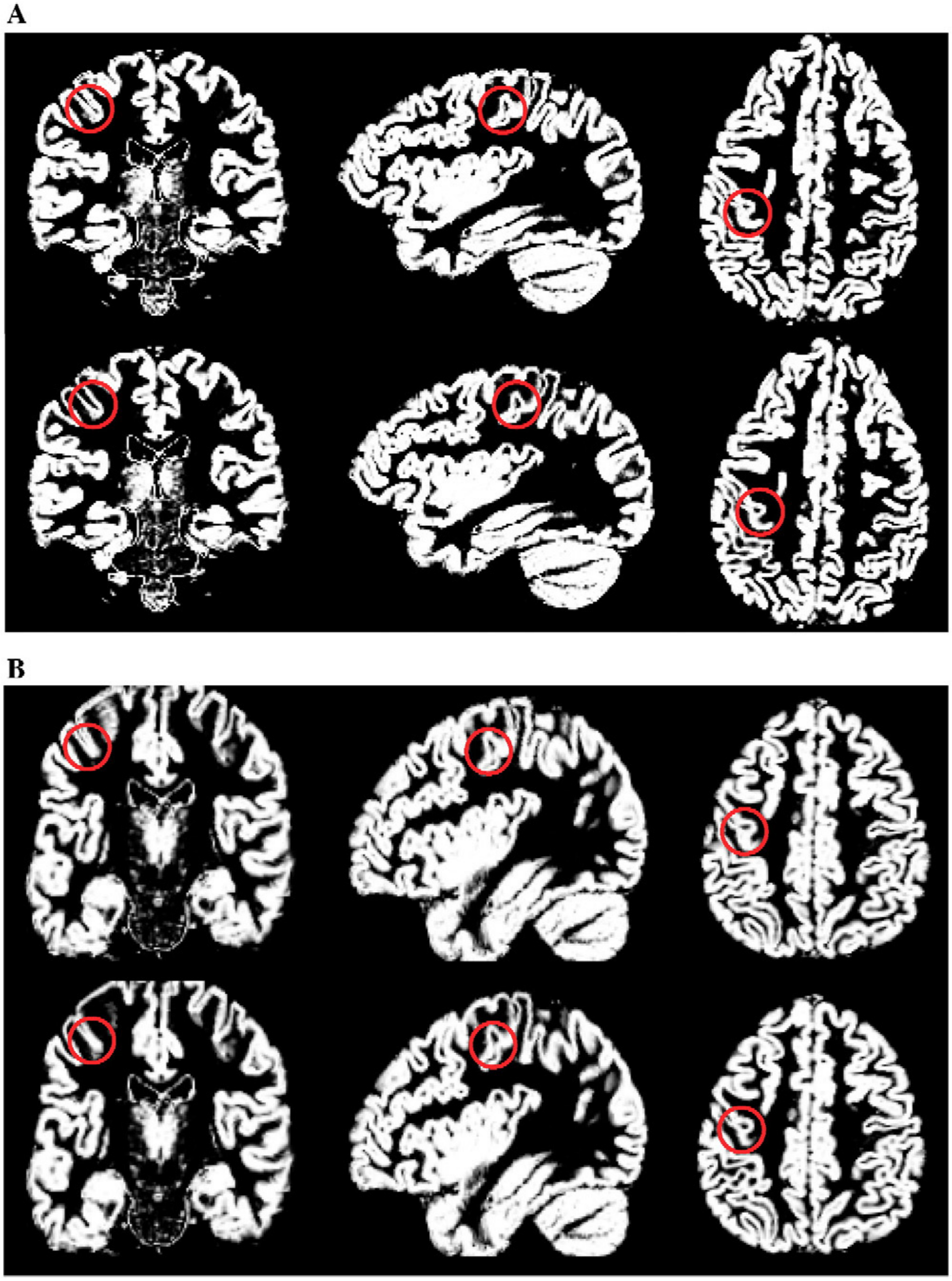

Fig. 3. An example of original and simulated GM images. (A) Original GM image (upper line) and GM image with simulated atrophy (lower line). (B) Normalized GM image (upper line) and normalized GM image with simulated atrophy (lower line).

Table 1
GM templates employed in the current study

Template		Description
General standard template	MNI[a]	1. Average of 152 images 2. Acquired in Montreal Neurological Institute 3. Participants: healthy subjects
Customer-specific template	T_CN	1. Average of N images (N=5, 15, 25, 50) 2. All images are included in the control group 3. Acquired in the 3-T scanner of the University of Surrey, UK (located in University of Royal Holloway, UK) 4. Participants: healthy students
	T_SN	1. Average of N images (N=15, 25, 50) 2. Each image contains simulated atrophy 3. All images are included in the study group 4. Acquired in the 3-T scanner of the University of Surrey, UK (located in University of Royal Holloway, UK) 5. The same participants as in T_CN
	$T_{CS}N$	1. Average of N images (N=100) 2. Fifty images are from the healthy control group and the other 50 are the simulated images, which compose the study group 3. Acquired in the 3-T scanner of the University of Surrey, UK (located in University of Royal Holloway, UK) 4. The same participants as in T_CN
Customer-standard template	T_C25*	1. Average of 25 images 2. All images are *not* included in the control group when employed 3. Acquired in the 3-T scanner of the University of Surrey, UK (located in University of Royal Holloway, UK) 4. Participants: healthy students

[a] The MNI template employed in this paper is not the whole-brain template but the GM template.

spherical region, radius=25 mm), while WM in this region was intact and CSF increased to equalize the shrinkage of GM [26]. Therefore, additional 50 images with localized GM atrophy were created. The original images were considered as the healthy control group and the corresponding images with simulated atrophy were the study group. With this approach, we solely focus on the detection of the atrophic region, which excludes the question of the false-positive detections in other regions. This will need to be addressed in a separate study. Fig. 3A shows an example of an original image and the same image with simulated atrophy.

2.3. VBM protocol and image preprocessing

The latest version of SPM (SPM5) was used for VBM analyses running in MATLAB 6.5. We chose an optimized VBM protocol, an approach designed to minimize errors while maximizing sensitivity, in which a brain-tissue-only template is employed instead of a whole-brain template [17]. The selected brain tissue (i.e., GM in our study) was extracted from the native brain and then normalized to the corresponding tissue template. All templates in this paper, therefore, refer to GM templates unless otherwise specified.

The non-brain-region images were removed from the original images using Brain Extraction Tool [28] integrated in MRIcro (http://www.mricro.com).

2.4. Customer-specific template

After removal of non-brain regions, the original and the simulated images were segmented into GM, WM and CSF, while the intensity inhomogeneities were corrected. SPM segmentation employed a mixture-model cluster analysis to identify voxel intensities matching particular tissue types [29,30]. It is combined with a priori knowledge of the spatial distribution of these tissues in normal subjects, derived from probability maps. The segmented GM images were then normalized to the GM prior map provided in SPM, and the normalized images were averaged. The final process of template creation was to smooth the average image using an isotropic Gaussian kernel with a full width at half maximum (FWHM) of 8 mm.

The cg_create_template[2] program developed by Gaser (http://dbm.neuro.uni-jena.de/vbm) was employed to generate the customer-specific templates. The prespecified numbers of GM images were randomly chosen from the control/study group for GM template creation.

2.5. Spatial normalization

After the templates were created, the GM images from both groups were spatially normalized to the GM templates. The default parameters in SPM5 were used (DCT cutoff=25 mm, nonlinear regularization=1, 16 iterations). Table 1 lists the templates employed in the current study. T_CN represents a template created by the control group, and T_SN represents a template created by the simulated atrophy group (i.e., study group). $T_{CS}N$ is created by both groups. N is the number of images, from which the template was created.

An example of normalization results using the MNI template is shown in Fig. 3B. The normalized GM images were then smoothed using a 12-mm FWHM kernel.

2.6. Statistical analysis

Group comparisons were performed using the random-effects analysis in SPM. The statistical model used was one-way ANOVA. Since the control group and the atrophy group included images from identical subjects, no confound correction, such as brain volume and age, was employed to remove the global differences. Thirteen comparisons were derived from the different templates applied for spatial normalization and the number of images in each group (Table 2). In Comparisons 1–10, the significance levels were set at P<.05, corrected for multiple comparison [31], with an extend threshold of 25 contiguous voxels. As the

[2] The algorithm was modified by us to be compatible with SPM5.

Table 2
List of comparisons with different templates and group sizes

Comparison no.	Template for control group	Template for atrophy group
Group size (n): 50 and 50		
1	MNI	MNI
2–5	$T_C N$ (N=5, 15, 25, 50)	$T_C N$ (N=5, 15, 25, 50)
6–8	$T_S N$ (N=15, 25, 50)	$T_S N$ (N=15, 25, 50)
9	$T_{CS} N$ (N=100)	$T_{CS} N$ (N=100)
10	$T_C 50$	$T_S 50$
Group size (n): 25 and 25		
11	MNI	MNI
12	$T_C 25$	$T_C 25$
13[a]	$T_C 25*$	$T_C 25*$

[a] In this comparison, $T_C 25*$ is regarded as a customer-standard template. The images created by the template are different from the images in the control group.

group sizes were smaller in Comparisons 11–13, a lower threshold was used (P<.000035, uncorrected for multiple comparison, cluster size>25).

3. Results

3.1. General results

The GM reduction was detected by finding the increased GM concentration using the contrast of control group versus study group. In order to quantify the template effects, the expected atrophy size was calculated using the average of the 50 simulated atrophy images (mean=1282 voxels, S.E.=54.7). The expected location of atrophy is around the left precentral gyrus.

The results showed that the location of atrophy was generally well detected in most comparisons, but the cluster size of atrophy differed for the various templates. Critically, Comparisons 6 to 9 revealed additional clusters located on the left postcentral gyrus. Table 3 lists the significant clusters in all comparisons. Fig. 4 shows the percentage difference between the detected atrophy size in the left precentral gyrus and the simulated atrophy size. The percentage was calculated as follows:

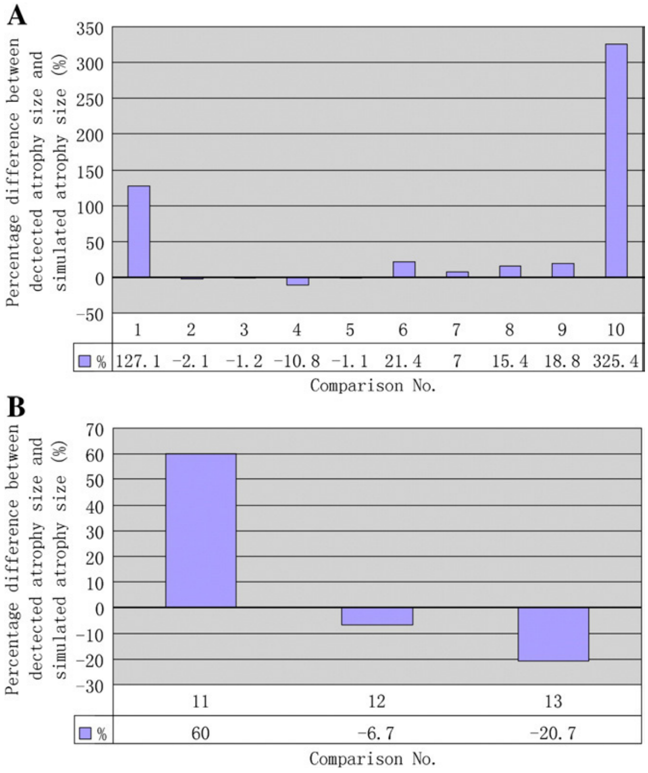

Fig. 4. Percentage difference between detected atrophy size on the left precentral gyrus and simulated atrophy size. (A) Comparisons 1–10: P<.05 corrected, cluster size>25 voxels. (B) Comparisons 11–13: P<.000035 uncorrected, cluster size>25 voxels.

$$\text{Percentage} = \frac{\text{number of detected significant voxels} - \text{number of simulated atrophy voxels}}{\text{number of simulated atrophy voxels}} \times 100\%.$$

A positive percentage represents overestimation of the atrophy size, and the negative percentage represents

Table 3
Significant clusters detected in Comparisons 1–13

Comparison no.	Voxel no. per resel	Left precentral gyrus			Left postcentral gyrus		
		Location (X Y Z)	T	Cluster size	Location (X Y Z)	T	Cluster size
P<.01 corrected, cluster size>25 voxels							
1	535.31	−39 −11 52	7.70	2911	–	–	–
2	504.30	−39 −9 53	7.0	1255	–	–	–
3	539.25	−39 −9 54	6.81	1267	–	–	–
4	517.17	−40 −9 53	6.69	1143	–	–	–
5	542.10	−39 −10 54	6.77	1268	–	–	–
6	538.61	−40 −8 52	7.53	1556	−35 −36 56	5.65	107
7	460.76	−40 −7 52	7.31	1372	−35 −34 57	5.64	28
					−42 −33 59	5.63	42
8	523.95	−40 −7 52	7.35	1480	−40 −34 50	5.74	200
9	541.35	−40 −8 52	7.27	1523	−35 −33 57	5.66	115
10	539.44	−40 −8 53	8.47	5454	–	–	–
P<.000035 uncorrected, cluster size>25 voxels							
11	637.63	−39 −10 51	6.72	2051	–	–	–
12	639.42	−40 −8 51	6.11	1196	–	–	–
13	639.80	−40 −7 51	6.04	1016	–	–	–

underestimation. The closer the percentage to zero, the higher the accuracy of atrophy detection. Selected T maps (Comparisons 1, 5, 8 and 9 represent the MNI, $T_{C}N$, $T_{S}N$ and $T_{CS}N$ templates, respectively; group size, $n=50$ and 50) are shown in standard SPM glass brain projections for illustration (Fig. 5).

3.2. MNI template (Comparisons 1 and 11) or customer-specific template (Comparisons 2–9, 12 and 13)

Overall, the MNI template obtained higher overestimation biases than customer-specific templates, except for Comparison 10, where group-specific templates were used. Although customer-specific templates generally overestimated the atrophy size in Comparisons 2–9 and underestimated it in Comparisons 12 and 13, they achieved higher detection accuracy than the MNI template (see Fig. 4 for percentages).

3.3. Customer-specific template created by the control group (Comparisons 2–5), the study group (Comparisons 6–8) or both (Comparison 9)

As shown in Fig. 4, the $T_{C}N$ template (Comparisons 2–5) achieved higher accuracies of detecting the atrophy than $T_{S}N$ (Comparisons 6–8) or $T_{CS}N$ (Comparison 9). For more detailed analyses, it indicated that a template containing images from the atrophy group ($T_{S}N$ and $T_{CS}N$) not only had lower detection accuracies than the template containing images only from the control group ($T_{C}N$) but also introduced additional clusters that were across the edges of the simulated region. It thus suggested that if customer-specific templates are chosen, they are best created from the healthy control group.

3.4. Effect of sample size (N) on the creation of customer-specific template (Comparisons 2–5 and 6–8)

Comparisons 2–5 tested the influence of sample size on template effects ($N=5$, 15, 25 and 50, i.e., number of images to create the template) from the control group. The results revealed that, within the tested range, the impact of sample size on detection accuracy is small. Comparisons 6–8, using the templates created by the study group, confirmed the finding that the number of images employed to create customer-specific templates affects VBM results only marginally. Our data provide no evidence that increasing the sample size would increase detection accuracy.

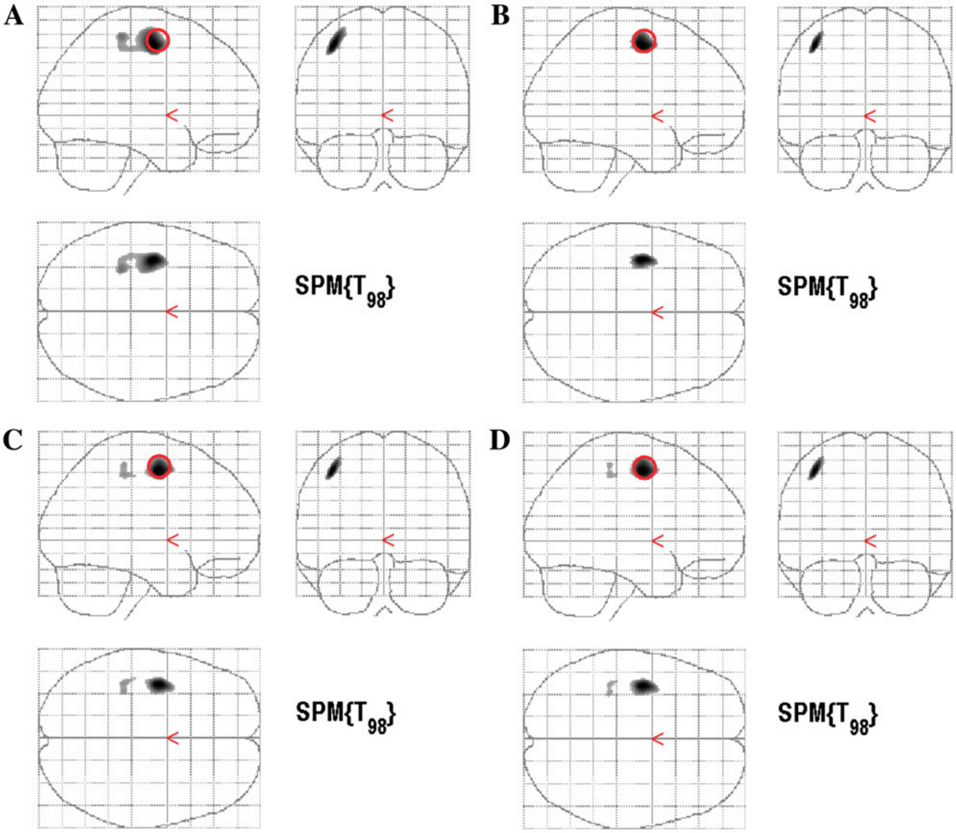

Fig. 5. T maps of significant GM atrophy in the comparisons with a group size of 50 ($P<.05$, corrected for multiple comparison). (A) Comparison 1: MNI (127.1% overestimation). (B) Comparison 5: $T_{C}50$ (1.1% underestimation). (C) Comparison 8: $T_{S}50$ (15.4% overestimation). (D) Comparison 9: $T_{CS}50$ (18.8% overestimation).

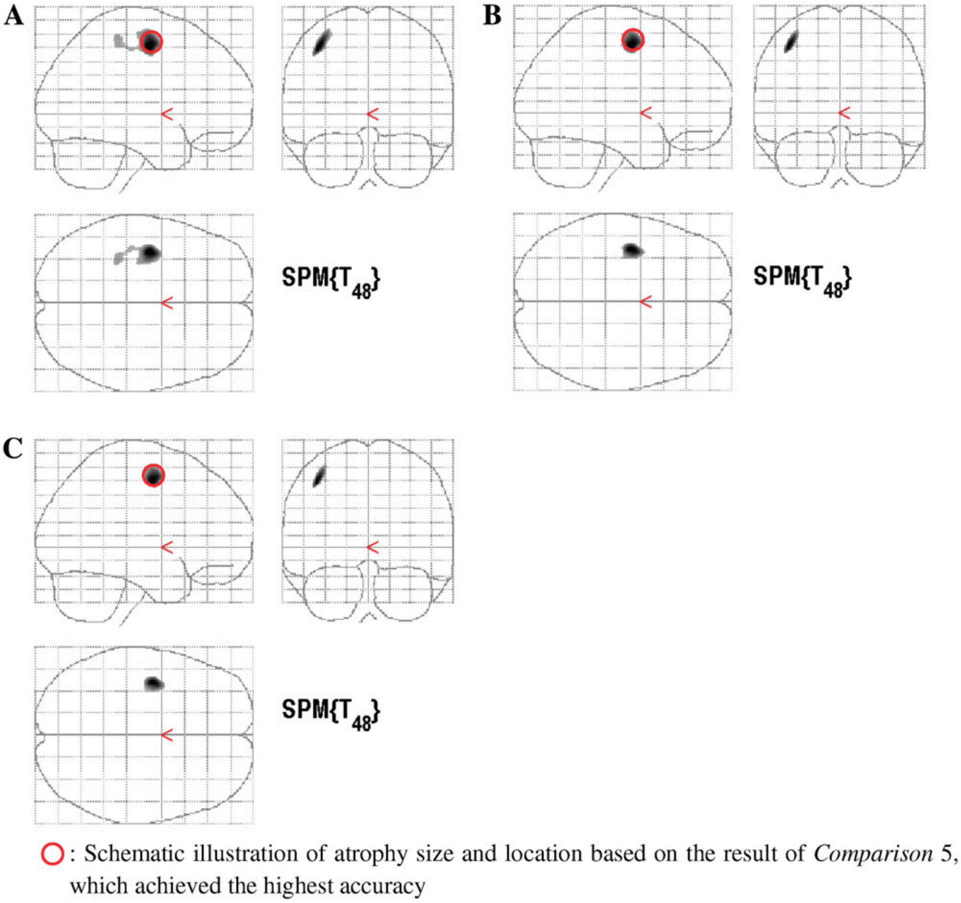
A
SPM{T48}
B
SPM{T48}
C
SPM{T48}
○ : Schematic illustration of atrophy size and location based on the result of *Comparison* 5, which achieved the highest accuracy

Fig. 6. T maps of significant GM atrophy in the comparisons with a group size of 25 ($P<.000035$, uncorrected). (A) Comparison 11: MNI (60% overestimation). (B) Comparison 12: $T_{\mathrm{C}}25$ (6.7% underestimation). (C) Comparison 13: $T_{\mathrm{C}}25^*$ (20.7% underestimation).

3.5. Comparison of MNI (Comparison 11), customer-specific template created by the control group (Comparison 12) and customer-standard template (Comparison 13)

A standard template is sometimes required across studies. However, considering the criticism the MNI template received, research groups may prefer to have an in-house template (i.e., customer-standard template), created from a number of normal images acquired in their own scanner. Typically, those images are not identical to the control images in each experiment, but they are matched for major demographic variables.

To investigate the performance of the customer-standard template, we randomly assigned the original 50 images to two sets of 25 images. One set of images composed the control group (the corresponding atrophy images were the study group) in all three comparisons (i.e., Comparisons 11–13). The $T_{\mathrm{C}}25$ template (i.e., the control-created customer-specific template) was formed by the same set of images. The other set of 25 images created a different template, $T_{\mathrm{C}}25^*$, which was considered as a customer-standard template. Comparisons 11–13 then employed the MNI, $T_{\mathrm{C}}25$ and $T_{\mathrm{C}}25^*$ template, respectively. Since these comparisons had smaller groups of images, no significance was found with corrected $P<.05$. Therefore, a reduced threshold, $P<.000035$, uncorrected with 25 contiguous voxels, was applied. The SPM glass brain projections

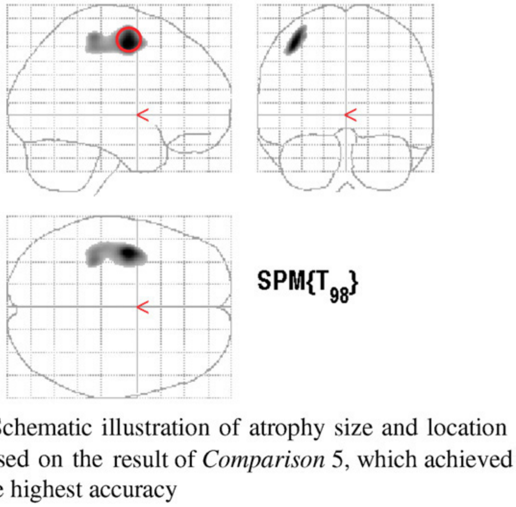
SPM{T98}
○ : Schematic illustration of atrophy size and location based on the result of *Comparison* 5, which achieved the highest accuracy

Fig. 7. T map of significant GM atrophy using group-specific templates in Comparison 10 ($P<.05$, corrected for multiple comparison, 325.5% overestimation).

(Fig. 6) confirmed the previous observation that the MNI template greatly overestimated the atrophy, while the control-created customer-specific template (6.7% underestimation) outperformed the customer-standard template (20.7% underestimation).

3.6. Identical or group-specific template for each group (Comparison 10)

Comparison 10 used group-specific templates for the control group and the study group, that is, T_C50 and T_S50. The detected atrophy region overestimated the atrophy size by 325.4%. This demonstrated that using group-specific templates introduced a huge amount of noise to the data (Fig. 7). An identical template is therefore suggested to be applied to both control and atrophy group.

4. Discussion

4.1. Template effects

To our own surprise, the data showed that the MNI template — the most commonly used normalization template at present — led to a considerable overestimation of the anatomical differences; that is, the estimated GM differences were much greater than the true lesion. We further found that the highest detection accuracy was achieved by those customer-specific templates that were created from images of the healthy control group. Most critically, the data suggest that templates generated from the atrophy group produced inaccurate VBM results, in which the atrophy was detected on the left postcentral gyrus besides the left precentral gyrus. This could be either true atrophy or false-positive detection as the location is at the edge of the simulated region. Group-specific templates, that is, different templates for the atrophy group and the control group, introduced the highest amount of overestimation biases among all templates tested. Translated into the clinical setting, this means that if patients with a lesion or the suspicion thereof are compared to healthy controls, group-specific templates may grossly inflate the structural differences between the two groups.

A further question addressed in this experiment concerned the role of the sample size (N) in creating customer-specific templates, that is, the number of images employed for the template. Our data showed that sample size had little impact on the normalization biases and suggested that even with a small number of images, customer-specific templates can be generated. This finding is particularly

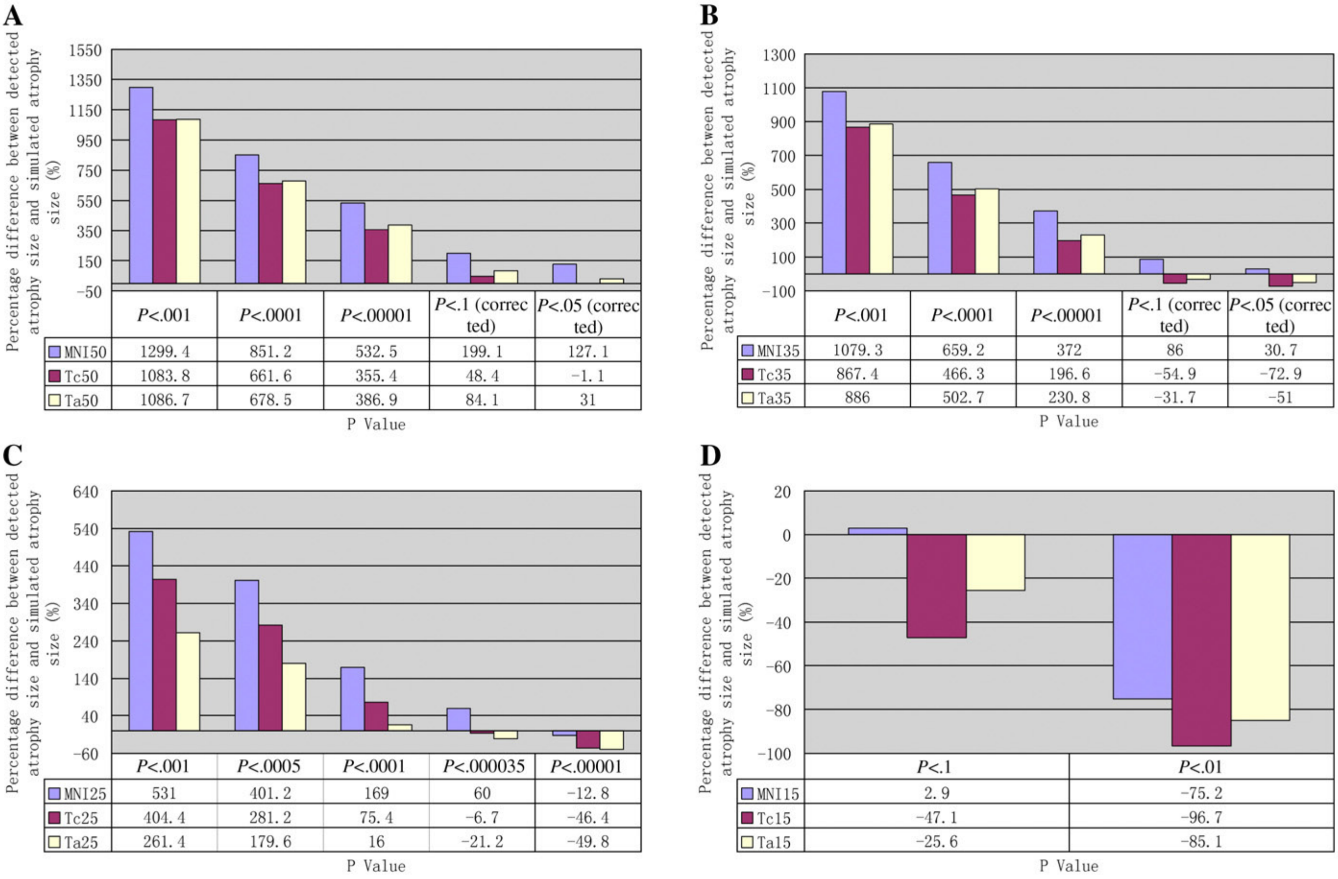

Fig. 8. Percentage difference between detected atrophy size and simulated atrophy size at a range of P values for MNI, T_CN and T_SN. (A) Group size=50, highest accuracy: −1.1% (T_C50, $P<.05$ corrected). (B) Group size=35, highest accuracy: 30.7% (MNI, $P<.05$ corrected). (C) Group size=25, highest accuracy: −6.7% (T_C25, $P<.000035$ uncorrected). (D) Group size=15, highest accuracy: 2.9% (MNI, $P<.1$ uncorrected).

relevant for the more clinically oriented areas of neuroimaging as they often work with small subject numbers.

Furthermore, our results indicated that the customer-standard template did not perform as well as the control-created template, but it performed better than the MNI.

A reasonable explanation for our findings is that the template effects depend on the degree of similarity between the template used and the normalization images. The detected group differences in VBM may be described as the real group difference multiplied by its modification because of the template (detailed discussion given below). If certain brain structures have similar shapes and volumes in the template (e.g., T_CN) and in the normalization images, these brain structures will not be greatly deformed as a result of spatial normalization. In this case, the VBM detection will be close to the real group difference and the template effect observed will be small. In contrast, if the templates (e.g., MNI) are rather different to the normalization images, the template effects may be evident in VBM results.

4.2. Effect of group size (n) and threshold on template effects

A critical issue relevant to all potential template effects is the group size[3] of the compared cohort (number of participants per group) and the threshold chosen for the GLM ANOVA. Both factors relate to test power and may, hence, lead to different conclusions in a test. An important question for the present study is how consistent the observed template effects are under different statistic power effects. We therefore tested a range of group sizes and thresholds on the template comparisons (MNI, T_CN and T_SN). Fig. 8 shows the detection accuracies acquired at a range of P values with four group sizes (n=15, 25, 35 and 50; images chosen randomly from the group). Our results revealed that the highest accuracy for each template was achieved with corrected P values (P<.05 or .1) and group size greater than 35. With group sizes smaller than 25, the lesion goes undetected if corrected P values are chosen. For all templates, the uncorrected P<.01 underestimated the size of the lesion when the group comprised 15 participants only but overestimated the lesion massively (i.e., by more than 1000%, n=50) with group sizes greater than 25. This interaction of VBM detection with group size/threshold indicates that selecting a standard or typically used statistical cutoff criterion might not necessarily produce the result that is closest to the actual structural difference.

The picture is further complicated if group size/threshold effects are considered for the different template scenarios. For MNI, T_CN and T_SN, the highest accuracy (1.1% underestimation) was achieved by T_C50 when the group size was 50 and a corrected P value of .05 was used. In the template comparisons with a group size of 35, the MNI template obtained an overestimation of 30.7% (P<.05, corrected), which was better than the customer-specific templates. For group sizes of 15 participants, the lesion was underestimated by all templates when an uncorrected P value of .01 was used.

These extended comparisons suggested that group size and threshold are relevant to the estimation of template effects. It also demonstrated that our initial finding, that the customer-specific template from the controls acquired greater detection accuracy than the MNI template, is not correct in all circumstances. When the group size is big or the threshold is low, MNI overestimates the size of the lesion, but when sample sizes are small and the threshold is high, MNI detection accuracy is better than T_CN. The comparisons also demonstrated that a VBM study with group sizes smaller than 25 may acquire unreliable detections regardless of the template used.

4.3. Potential impacts on template effects

An optimized VBM approach, which is suggested to produce superior VBM results than standard VBM [17], was chosen in our study. It involves spatial normalization of all images to a GM (or WM) template instead of a whole-brain template. This approach improves the match between the template and the GM images, preventing any confounding effect introduced by other tissues. However, the better registration may attenuate between-group differences. It therefore has a smaller likelihood of observing group differences compared to the standard VBM [20]. With regard to template effect, the improved match between the template and the GM images implies that the influence of the template on the images is enhanced. Therefore, we expect that the template effects would be more prominent in optimized VBM than in standard VBM.

Many VBM studies incorporated an additional modulation step after spatial normalization, which multiplies the partitioned images by the Jacobian determinants of the deformation field [1,22,32–34]. The reason for this is that, as a result of spatial normalization, the volumes of certain brain regions will grow, whereas others will shrink. It suggests that VBM is actually testing for regional differences in the concentration of a particular tissue (i.e., GM) [1]. In the current experiment, the template effects were tested based on the nonmodulation approach because of its popularity in VBM studies [3,6,11–15,19,35,36]. The simulated atrophy size was regarded as the ground truth to test the VBM detection accuracies with respect to the templates employed in spatial normalization. Arguably, without the modulation step, the detected atrophy size actually represents the loss of GM concentration while the simulated atrophy size refers to the GM volume reduction. However, as the template is created from the identical population to normalization images, the original volume of brain regions is most likely retained after normalization. On

[3] To prevent any confusion, we would like to point out that the group size (n) here refers to the number of images in each group, which were compared to detect the group differences, while the sample size stated above (Section 3.4) refers to the number of images for creating the templates.

the contrary, the volumes would be altered to match the volumes in the template. We assume that this impact is a major part of the template effects, whereas the image information included in the template is important. As stated above, spatial normalization is not required to match every cortical feature but corrects for global brain shape. A perfect normalization would remove GM differences if a modulation step was not applied because images are normalized to an exact match of the template [1]. This indicates that modulation is superior to nonmodulation in reserving between-group variability [20], and the influence of a template on normalized images will be reduced accordingly. It therefore could be hypothesized that the template effects in a modulation approach would be much less evident than in a nonmodulation approach.

Another questionable issue is whether, because of smoothing, the detected result in VBM does reveal the true differences between groups. The purpose of smoothing is to reduce the residual intersubject variances after normalization. Although smoothing may alter the atrophy size, without a smoothing step, it is more likely that no significant detection would be found using VBM. It is difficult to measure how much the true effect has been modified because of smoothing. However, it is reasonable to assume that choosing an optimal kernel size would maximize the signal-to-noise ratio. In other words, it can reduce the residual variances while retaining the effects to the most extent. This implies that VBM detections can be reasonably close to the real truth. In the circumstances that the residual variability within groups is minimal, a smaller kernel size would be used or the data would not be smoothed. The 12-mm kernel size employed in our study is generally applied in VBM studies [17,19,21,34]. We therefore consider that it is optimal to address the template effects.

The default parameters of spatial normalization in SPM5 were employed in the current experiment, except that the voxel size of normalized images was set to 1×1×1 mm^3. The purpose of this choice was to investigate the template effects in a VBM study with minimal parameter adjustment. In fact, different template effects may be observed when alternative parameters and methods in VBM are utilized. For example, Salmond et al. [11] demonstrated that using different numbers of nonlinear basis functions in spatial normalization would impact on the template effects. The study showed that with larger basis function sets (7×8×7), the choice of templates does affect the results of VBM. In contrast, with smaller sets (4×5×4), the VBM results are minimally dependent on the template used. It is worth noticing that SPM2/SPM5 does not provide the options for the number of basis function sets any longer but uses the selected nonlinear frequency cutoff (in mm) to determine an appropriate set of basis functions. By decreasing the cutoff value, the more basis functions will be used in spatial normalization; thus, it will increase the possibility of overfitting and enhance the template effects. Moreover, Senjem et al. [21] proposed an improved procedure for the use of the custom-specific template (i.e., using previous deformation estimates to initialize warping to the template). It resulted in a more reasonable VBM detection than the standard approach. This further confirmed that any methods involved in a VBM study would impact on the template effects.

Importantly, our data were analyzed using the standard normalization approach under normalization function [37] in SPM5, as opposed to the unified segmentation approach under segmentation. It is claimed that the template effects would be minimal if the unified approach is used [30]. However, it is too time-consuming for our experiments. For the same reason, the GM images were acquired using the tissue probability maps provided in SPM instead of the probability maps created by our data. It is likely that the use of the latter may increase segmentation accuracy.

Furthermore, the current study explored the template effects using the simulated atrophy located on the left precentral gyrus with 5% GM reduction. A question may be raised as to whether consistent results can be achieved when a different region or a different level of atrophy is employed. These issues will be investigated in future studies, in which various locations and atrophy sizes will be tested.

Many other factors such as the smoothing kernel size may also alter the template effects observed and will be addressed in the future.

5. Conclusion

The current experiment systematically investigated the impact of how images were selected for template creation, such as the source of images (e.g., healthy control group or study group) and the group size, on the likelihood of detecting structural differences between groups. Our data leave little doubt that it does matter for VBM which template is chosen for the normalization process, as the VBM methodology may have a higher or lower chance of detecting structural differences accurately. The choice of template therefore needs to be carefully considered. However, it is not a simple task to say which is the best template; more research is required to understand the respective interactions. It is worth noticing that our findings are specific for SPM5 or those softwares employing a similar spatial normalization approach. Different normalization implementations are likely to produce different template effects. Using the simplified scenario of simulated atrophy in one group plus an easily identifiable region to explore the role of template choice for VBM accuracy is a first step; further work will have to address critical issues such as atrophy size and different locations.

References

[1] Ashburner J, Friston KJ. Voxel-based morphometry — the methods. Neuroimage 2000;11(6):805–21.

[2] Ashburner J, et al. Computer-assisted imaging to assess brain structure in healthy and diseased brains. Lancet Neurol 2003;2(2):79–88.

[3] Bernasconi N, et al. Whole-brain voxel-based statistical analysis of gray matter and white matter in temporal lobe epilepsy. Neuroimage 2004;23(2):717–23.

[4] Keller SS, et al. Voxel based morphometry of grey matter abnormalities in patients with medically intractable temporal lobe epilepsy: effects of side of seizure onset and epilepsy duration. J Neurol Neurosurg Psychiatry 2002;73(6):648–55.

[5] Karas GB, et al. A comprehensive study of gray matter loss in patients with Alzheimers disease using optimized voxel-based morphometry. Neuroimage 2003;18(4):895–907.

[6] Job DE, et al. Voxel-based morphometry of grey matter densities in subjects at high risk of schizophrenia. Schizophr Res 2003;64(1):1–13.

[7] Gong Q-Y, et al. Voxel-based morphometry and stereology provide convergent evidence of the importance of medial prefrontal cortex for fluid intelligence in healthy adults. Neuroimage 2005;25(4):1175–86.

[8] Chan CHP, et al. Thalamic atrophy in childhood absence epilepsy. Epilepsia 2006;47(2):399–405.

[9] Draganski B, et al. Decrease of thalamic gray matter following limb amputation. Neuroimage 2006;31(3):951–7.

[10] Pannacciulli N, et al. Brain abnormalities in human obesity: a voxel-based morphometric study. Neuroimage 2006;31(4):1419–25.

[11] Salmond CH, et al. The precision of anatomical normalization in the medial temporal lobe using spatial basis functions. Neuroimage 2002;17(1):507–12.

[12] Maguire EA, et al. Navigation-related structural change in the hippocampi of taxi drivers. Proc Natl Acad Sci U S A 2000;97(8): 4398–403.

[13] Kubicki M, et al. Voxel-based morphometric analysis of gray matter in first episode schizophrenia. Neuroimage 2002;17(4):1711–9.

[14] Wilke M, et al. Gray matter-changes and correlates of disease severity in schizophrenia: a statistical parametric mapping study. Neuroimage 2001;13(5):814–24.

[15] Sowell ER, et al. Localizing age-related changes in brain structure between childhood and adolescence using statistical parametric mapping. Neuroimage 1999;9(6):587–97.

[16] Gaser C. VBM toolbox. http://dbm.neuro.uni-jena.de/vbm. 2004. [Accessed Nov 2005].

[17] Good CD, et al. A voxel-based morphometric study of ageing in 465 normal adult human brains. Neuroimage 2001;14(1):21–36.

[18] Tisserand DJ, et al. A voxel-based morphometric study to determine individual differences in gray matter density associated with age and cognitive change over time. Cereb Cortex 2004;14(9):966–73.

[19] Gaser C, Schlaug G. Brain structures differ between musicians and non-musicians. J Neurosci 2003;23(27):9240–5.

[20] Keller SS, et al. Comparison of standard and optimized voxel-based morphometry for analysis of brain changes associated with temporal lobe epilepsy. Neuroimage 2004;23(3):860–8.

[21] Senjem ML, et al. Comparison of different methodological implementations of voxel-based morphometry in neurodegenerative disease. Neuroimage 2005;26(2):600–8.

[22] Brenneis C, et al. Voxel-based morphometry in narcolepsy. Sleep Med 2005;6(6):531–6.

[23] Duchesne S, et al. Within-group nonlinear registration improves VBM results. Proc 11th Intl Soc Magn Reson Med 2006;913.

[24] Oldfield RC. The assessment and analysis of handedness: the Edinburgh Inventory. Neuropsychologia 1971;9(1):97–113.

[25] Karacali B, Davatzikos C. Estimating topology preserving and smooth displacement fields. IEEE Trans Med Imaging 2004;23(7):868–80.

[26] Karacali B, Davatzikos C. Simulation of tissue atrophy using a topology preserving transformation model. IEEE Trans Med Imaging 2006;25(5):649–52.

[27] Yousry T, et al. Localization of the motor hand area to a knob on the precentral gyrus. A new landmark. Brain 1997;120(1):141–57.

[28] Smith SM. Fast robust automated brain extraction. Hum Brain Mapp 2002;17(3):143–55.

[29] Ashburner J, Friston K. Multimodal image coregistration and partitioning — a unified framework. Neuroimage 1997;6(3):209–17.

[30] Ashburner J, Friston KJ. Unified segmentation. Neuroimage 2005;26(3):839–51.

[31] Worsley K, et al. A three-dimensional statistical analysis for CBF activation studies in human brain. J Cereb Blood Flow Metab 1992;12(6):900–18.

[32] Maguire E, et al. Navigation expertise and the human hippocampus: a structural brain imaging analysis. Hippocampus 2003;13(2):250–9.

[33] Haier R, et al. The neuroanatomy of general intelligence: sex matters. Neuroimage 2005;25(1):320–7.

[34] Apkarian AV, et al. Chronic back pain is associated with decreased prefrontal and thalamic gray matter density. J Neurosci 2004;24(46): 10410–5.

[35] Luders E, et al. A voxel-based approach to gray matter asymmetries. Neuroimage 2004;22(2):656–64.

[36] Davatzikos C, et al. Voxel-based morphometry using the RAVENS maps: methods and validation using simulated longitudinal atrophy. Neuroimage 2001;14(6):1361–9.

[37] John K, Ashburner JF. Nonlinear spatial normalization using basis functions. Hum Brain Mapp 1999;7(4):254–66.