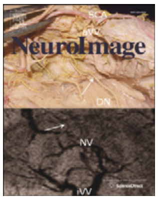
NeuroImage 56 (2011) 1164–1170

Contents lists available at ScienceDirect

NeuroImage

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



Multi-site voxel-based morphometry — Not quite there yet

N.K. Focke [a,*], G. Helms [b], S. Kaspar [a], C. Diederich [a], V. Tóth [a], P. Dechent [b], A. Mohr [c], W. Paulus [a]

[a] *Dept. of Clinical Neurophysiology, University Medical Center, Georg-August University Göttingen, Germany*
[b] *MR-Research in Neurology and Psychiatry, University Medical Center, Georg-August University Göttingen, Germany*
[c] *Dept. of Neuroradiology, University Medical Center, Georg-August University Göttingen, Germany*

ARTICLE INFO

Article history:
Received 30 November 2010
Revised 15 January 2011
Accepted 9 February 2011
Available online 13 February 2011

Keywords:
Voxel-based morphometry
Multi-site
Validation
Bias
Segmentation accuracy

ABSTRACT

Voxel-based morphometry (VBM) is a widely applied method in computational neurosciences but it is currently recommended to compare only data collected at a single MRI scanner. Multi-site VBM would be a desirable approach to increase group size and, thus, statistical power. We aimed to assess if multi-site VBM is feasible on similar hardware and compare the magnitude of inter- and intra-scanner differences.

18 healthy subjects were scanned in two identical 3 T MRI scanners using different head coil designs, twice in scanner A and once in scanner B. 3D T1-weighted images were processed with SPM8 and FSL4.1 and compared as paired *t*-test (scan versus re-scan) on a voxel basis by means of a general linear model (GLM). Additionally, coefficient-of-difference (coeffD) maps were calculated for respective pairs of gray matter segmentations.

We found considerable inter-scanner differences clearly exceeding a commonly used GLM significance threshold of $p<0.05$ (FWE corrected). The spatial pattern of detected differences was dependent on whether SPM8 or FSL4.1 was used. The inclusion of global correcting factors either aggravated (SPM8) or reduced the GLM detected differences (FSL4.1). The coeffD analysis revealed markedly higher variability within the FSL4.1 stream both for the inter- and the intra-scanner comparison. A lowered bias cutoff (30 mm FWHM) in SPM8 improved the comparability for cortical areas. Intra-scanner scan/re-scan differences were generally weaker and did not exceed a $p<0.05$ (FWE corrected) threshold in the GLM analysis.

At 3 T profound inter-scanner differences are to be expected that could severely confound an unbalanced VBM analysis. These are like related to the receive bias of the radio-frequency hardware.

Introduction

MRI offers the unique opportunity to non-invasively access the morphology of the human brain at high resolution. Thus, in vivo MR brain scans can be performed repeatedly or on large cohorts. The anatomical complexity, however, makes purely visual interpretation of the obtained data challenging. To achieve an objective and operator-independent comparison, computational methods have been devised; one of the most commonly used is voxel-based morphometry (VBM) (Wright et al., 1995; Ashburner and Friston, 2000). VBM is usually based on high-resolution, T1-weighted MRI scans that are segmented into tissue classes, spatially normalized, smoothed and analyzed voxel-by-voxel in general linear or non-parametric models. The method has been widely applied in group studies (Focke et al., 2008a; Keller and Roberts, 2008) and also in comparing single subjects against a group (Focke et al., 2008b; Woermann et al., 1999; Focke et al., 2009). Currently it is a consensus that a single study should be conducted on a single site, i.e. a single MRI scanner only. This, however, may impose practical limitations when the condition in question is rare or a large cohort is needed, like in genetic studies. Exchange of hardware components or upgrades of software are almost inevitable within the life cycle of an MRI scanner. Thus, especially in a longitudinal study, it is very demanding to keep every scanner parameter constant over a prolonged period of time. Therefore even a single-site study can truly become multi-site when the scanner undergoes major replacement works, be it scheduled or (even worse) unscheduled due to a technical fault. Recently researchers have tried to establish multi-site VBM analysis and have shown that, pooling of data is possible and that the pattern of interest is not affected when the site and/or upgrade level are included as a regressor into the analysis and a balanced design is used including an adequate number of controls per site (Stonnington et al., 2008; Pardoe et al., 2008; Segall et al., 2009). Another recent study suggested using VBM-preprocessed data in combination with atlas-based regions of interest for multi-site volumetry (Huppertz et al., 2009). One of the most challenging technical problems for the segmentation software is to correct for the spatial variation of image intensities. This so-called bias is pivotal to multi-site VBM, since it can be caused by the specific MRI equipment (radio-frequency (RF) hardware, static magnetic field inhomogeneity, non-linearity of imaging gradients) and subject specific factors (Jovicich et al., 2006). Within the preprocessing of VBM data, correction of image bias is an important step that is

[*] Corresponding author at: Dept. of Clinical Neurophysiology, University Medical Center, Georg-August University Göttingen, 37099 Göttingen, Germany.
E-mail address: nfocke@uni-goettingen.de (N.K. Focke).

1053-8119/$ – see front matter © 2011 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2011.02.029

iteratively improved together with the segmentation, e.g. in FAST4 (Zhang et al., 2001) or in the unified segmentation algorithm of the recent SPM packages (Ashburner and Friston, 2005). Whether this is sufficient to reduce the between site differences and allow for valid multi-site VBM is, however, unclear. In the current study we aimed to assess if multi-site VBM with current software packages is feasible and quantify differences introduced by two technically similar MRI scanners available on site. We also evaluated whether including global scaling regressors or a different bias cutoff could improve the comparability.

Methods

We enrolled 18 healthy subjects into the study after written informed consent. All subjects were scanned three times on different days in two different MR systems. Both scanners A and B were 3 T Siemens Magnetom TIM Trio (Siemens Healthcare, Erlangen, Germany). Scanner A (MR-Research in Neurology and Psychiatry) was equipped with an 8-channel head coil for signal reception (Invivo, Gainesville, FL), scanner B (Dept. of Neuroradiology) with the vendor's 12-channel head coil. The body coil was used for transmission. We each acquired two repetitions of a T1-weighted, 3D magnetization-prepared rapid gradient-echo (Mugler and Brookeman, 1990) (3D-MPRAGE, TI = 900 ms, $\alpha = 9°$, TE = 3.0–3.2 ms, TR = 2250 ms) as recommended by the ADNI initiative (Jack et al., 2008). Like in a "real-life" situation, no specific calibrations except the standard pre-scan procedures were performed. Every subject was scanned twice in scanner A (intra-scanner test–retest) and once in scanner B (inter-scanner test–retest). The mean interval between reference scan and re-scan was 151 days for the inter-scanner comparison (range 12–456), and 191 days for the intra-scanner comparison (range 23–465). The study was approved by the Ethics Committee of the University Medical Center, Göttingen.

The original DICOM images were transferred to a Linux-based image server and converted to 3D NIFTI format using mriconvert (http://www.lcni.uoregon.edu/~jolinda/MRIConvert). First, the two repetitions of the T1-weighted scans were coregistered (2nd scan rigidly coregistered to 1st scan) and averaged with the FMRIB Software Library toolkit version 4.1 (FSL4.1) (http://www.fmrib.ox.ac.uk/fsl). The VBM preprocessing was done in parallel with Statistical Parametric Mapping version 2008 (SPM8, http://www.fil.ion.ucl.ac.uk/spm) running on a Matlab 7.7 platform (the MathWorks Inc., Natick, MA) and FSL4.1. In the SPM8 stream images were segmented with the unified segmentation and default settings (bias regularization 0.0001, bias cutoff FWHM 60 mm) in native space. Next images were normalized and modulated with the DARTEL toolbox to a study specific template (generated iteratively) in MNI space with 1.5 mm cubic resolution (as recommended for the DARTEL procedure) and smoothed with an 8 mm FWHM Gaussian kernel. In the FSL-VBM stream images were brain-extracted with BET (Smith, 2002), segmented with FAST4 (Zhang et al., 2001) with default settings (bias field smoothing extent 20 mm FWHM) and normalized to a study specific template with FNIRT in 2 mm cubic resolution (as recommended by the FSL developers) including modulation with the respective Jacobian determinant map. The resulting maps of gray matter partial volume were then smoothed with an 8 mm FWHM Gaussian kernel in SPM8. The final analysis was then done as a paired *t*-test in SPM8 (for all pre-processing streams) with a significance threshold of $p<0.05$ corrected for multiple comparisons with family-wise error rate (FWE) and with a threshold of $p<0.0001$ (uncorrected). To define the final analysis space and to exclude remote non-gray matter areas, an absolute gray matter probability threshold of 0.05 was applied. Both the intra- and inter-scanner pairs were compared in this fashion. We additionally included the total intracranial volume (TIV) or the total gray matter volume (TGV) as regressors of no interest into the analysis. TIV and TGV were estimated by summing up the voxel-values of all tissue classes (TIV) or the gray matter class (TGV) in Matlab. Resulting statistical maps were overlaid on the averaged T1 weighted image of all subjects with MRIcron (http://www.cabiatl.com/mricro/mricron/index.html) for visualization in each processing stream respectively. Following a recent publication suggesting that at 3 T a reduced bias cutoff of 30–50 mm FWHM is preferable (Zheng et al., 2009) we repeated the SPM8 analysis with a bias cutoff of 30 mm FWHM (bias regularization was left unchanged at 0.0001) within the unified segmentation algorithm.

To allow for a concise comparison of the detected inter- and intra-scanner differences we calculated the overall number of suprathreshold voxels and the maximum t-score in each comparison. Lower values (both for t-score and voxel count) indicate a better comparability of the respective scans.

As an additional analysis step we calculated the voxel-based coefficient of difference (coeffD) maps. This was done in Matlab following the formula coeffD = mean(2 * abs(gray_matter_image_A − gray_matter_image_B)/(gray_matter_image_A + gray_matter_ image_B)). CoeffD, thus, is an absolute percentage value expressing the mean scan to re-scan difference over the whole group. Higher values indicate increased variability between the image pairs. To remove artifacts in non-brain areas the resulting maps were masked with an overall probability threshold of 0.05 (all voxels with at least one gray matter probability of <0.05 were set to 0). This approach is identical to the absolute threshold masking used in the SPM8 analysis. Unthresholded coeffD maps were calculated for all image pairs (scan–re-scan) and all processing streams (compare Fig. 4).

Results

The inter-scanner comparison of T1-based VBM showed strong, spatially distributed gray matter volume differences that survived FWE correction and could, thus, severely confound an unbalanced multi-site analysis. Intra-scanner differences, in contrast, were mild, non-systematic, and did not survive FWE correction.

Detailed results of the inter-scanner comparison are shown in Table 1, intra-scanner results are summarized in Table 2.

T1-based VBM in SPM8

In the inter-scanner comparison of SPM8-processed T1 images, VBM showed a relative gray matter volume increase (reference < re-scan) in left temporal, inferior parietal and occipital lobe and the left cerebellum. A relative gray matter volume decrease (reference > re-scan) was detected in the anterior frontal lobe with emphasis on the left side (see Fig. 1). Without including a global scaling regressor the relative increase was predominant. With TIV included as regressor the relative decrease was more prominent, when TGV was included both effects were nearly balanced with a slight emphasis on the relative decrease. Including global regressors (TIV or TGV) aggravated the gross inter-scanner differences evident in this comparison both for suprathreshold voxel count and maximum t-score (see Table 1).

When using a modified bias cutoff of 30 mm FWHM within the unified segmentation of SPM8 the observed differences were reduced but did still exceed a $p<0.05$ (FWE corrected) threshold. The inclusion of TIV or TGV, as in the processing stream with standard settings, increased the scan/re-scan differences (see Fig. 2).

When comparing the intra-scanner SPM8-processed T1 images (for standard settings as well as for bias cutoff of 30 mm) the detected differences were clearly less (see Table 2). No cluster surpassed the $p<0.05$ (FWE corrected) threshold. With the $p<0.0001$ (uncorrected) threshold minor differences were seen.

Table 1
Overview of inter-scanner VBM results.

Regressor	Threshold	Increase (ref.<re-scan)		Decrease (ref.>re-scan)	
		# of voxels	Max t-score	# of voxels	Max t-score
		SPM8 T1-based			
None	$p<0.05$ (FWE)	9107	10.65	0	N/A
	$p<0.0001$ (uncorr.)	55,165	10.65	1010	6.23
TIV	$p<0.05$ (FWE)	91	9.51	21,282	16.83
	$p<0.0001$ (uncorr.)	5823	9.51	81,260	16.83
TGV	$p<0.05$ (FWE)	4468	13.12	19,718	15.74
	$p<0.0001$ (uncorr.)	51,440	13.12	65,196	15.74
		FSL4.1 T1-based			
None	$p<0.05$ (FWE)	3828	17.21	10,691	15.88
	$p<0.0001$ (uncorr.)	17,112	17.21	40,921	15.88
TIV	$p<0.05$ (FWE)	749	13.43	6067	15.36
	$p<0.0001$ (uncorr.)	8741	13.43	33,505	15.36
TGV	$p<0.05$ (FWE)	154	12.84	8	9.06
	$p<0.0001$ (uncorr.)	2861	12.84	1894	9.06
		SPM8 T1-based (30 mm FWHM bias cutoff)			
None	$p<0.05$ (FWE)	435	8.84	12	7.06
	$p<0.0001$ (uncorr.)	9565	8.84	621	7.06
TIV	$p<0.05$ (FWE)	89	11.47	8882	15.33
	$p<0.0001$ (uncorr.)	2740	11.47	72,996	15.33
TGV	$p<0.05$ (FWE)	2727	19.31	9,277	16.25
	$p<0.0001$ (uncorr.)	37,143	19.31	47,709	16.25

The summarized VBM results of the inter-scanner comparison (i.e. re-scan in a different scanner) for each processing stream are shown. The overall (=summed up) number of supra-threshold voxels for each comparison (# of voxel) and the maximum t-score is listed. Applied thresholds were family-wise error rate (FWE) corrected $p<0.05$ or uncorrected $p<0.0001$. Total intracranial volume (TIV) or total gray matter volume (TGV) were also included as regressors of no interest into the GLM analysis respectively. Lower maximum t-scores and fewer suprathreshold voxels indicate a better comparability of scan and inter-scanner re-scan.

Table 2
Overview of intra-scanner VBM results.

Regressor	Threshold	Increase (ref.<re-scan)		Decrease (ref.>re-scan)	
		# of voxels	Max t-score	# of voxels	Max t-score
		SPM8 T1-based			
None	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	90	6.07	0	N/A
TIV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	0	N/A	41	6.14
TGV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	94	6.06	405	6.37
		FSL4.1 T1-based			
None	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	0	N/A	43	5.87
TIV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	0	N/A	59	6.08
TGV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	0	N/A	57	5.65
		SPM8 T1-based (30 mm FWHM bias cutoff)			
None	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	6	4.96	0	N/A
TIV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	16	5.87	573	6.28
TGV	$p<0.05$ (FWE)	0	N/A	0	N/A
	$p<0.0001$ (uncorr.)	64	6.81	350	7.48

The summarized VBM results of the intra-scanner comparison (i.e. re-scan in the same scanner) for each processing stream are shown. The overall (=summed up) number of supra-threshold voxels for each comparison (# of voxel) and the maximum t-score is listed. Applied thresholds were family-wise error rate (FWE) corrected $p<0.05$ or uncorrected $p<0.0001$. Total intracranial volume (TIV) or total gray matter volume (TGV) were also included as regressors of no interest into the GLM analysis respectively. Lower maximum t-scores and fewer suprathreshold voxels indicate a better comparability of scan and intra-scanner re-scan.

T1-based VBM with FSL4.1 pre-processing

When using the FSL4.1 pre-processing, we observed the same global picture of stronger inter-scanner than intra-scanner differences. The spatial localization of GM volume changes, however, was quite different compared to the SPM8 processing stream (see Figs. 1 and 3). In the inter-scanner comparison we found predominantly gray matter volume decreases scattered over the brain with a frontal/anterior emphasis. Relative increases were detected in the right temporal and occipital lobe. When including TIV differences were slight, when including TGV they were clearly reduced. However, even with TGV included a relevant number of voxels still surpassed the $p<0.05$ (FWE corrected) threshold.

In the intra-scanner comparison, again, no differences exceeding the $p<0.05$ (FWE corrected) threshold were found. With the $p<0.0001$ (uncorrected) threshold some relative gray matter volume decreases were evident.

Coefficient of difference maps

Voxel-based coeffD maps are shown in Fig. 4. In general the results of the general linear model analysis were confirmed: areas with significant differences in the conventional VBM analysis also showed higher coeffD values. However, in comparison of the different analysis streams the FSL processed images showed higher coeffD values both for the inter-scanner and the intra-scanner comparison. When comparing the SPM8 analysis with default settings and with a reduced bias cutoff the latter, as in the VBM analysis, showed reduced variability in the cortical convexity but higher coeffD values in the putamen/pallidum region both for the intra- and inter-scanner comparison.

Discussion

We could show that VBM of pooled multi-site/multi-scanner data still is challenging even when using similar scanners and imaging protocols. The observed volume differences in the T1-based methods (both SPM8 and FSL4.1) would pose severe problems in VBM analysis of unbalanced sub-cohorts and did clearly exceed even a strict FWE corrected significance threshold of $p<0.05$. With a lowered bias cutoff (30 mm FWHM) in SPM8 differences could be reduced but would still be relevant. Our findings can aid in estimating the potential impact of inter-scanner differences for planning and interpreting VBM studies at 3 T.

Multi-site VBM

Making multi-site VBM possible is an important goal for MR imaging of rare diseases, large patient groups or longitudinal studies. Image bias caused by scanner- and subject specific factors is a major obstacle for this endeavor and is generally more severe when using magnetic field strengths of 3 T (and above) and phased-array head coils (Zheng et al., 2009; Bernstein et al., 2006). This experimental setup, however, is becoming increasingly common. Higher magnetic field strengths (7 T already on the horizon for clinical usage) and multi-element phased-array coils (32 channel coils commonly available) will improve SNR and anatomical detail but may also exacerbate bias-related problems. At 1.5 T, disease specific differences (Alzheimer disease) were markedly stronger than scanner specific effects (maximum t-score 5.94) in a T1-based VBM study conducted across multiple scanner upgrades, but with transmit-receive head coil of the same design (Stonnington et al., 2008). The authors of this study concluded that "data can be pooled from different scanners without corroding the integrity of results is reassuring for large multi-site studies". Another study combining three different sites (2×1.5 T, 1 x 3 T, all with transmit-receive head coils) have reported stronger scanner specific than disease specific (childhood absence epilepsy) differences, exceeding a $p<0.05$ (FWE corrected) threshold (Pardoe et al., 2008). A large multi-site VBM study in schizophrenia (including 1.5, 3 and 4 T

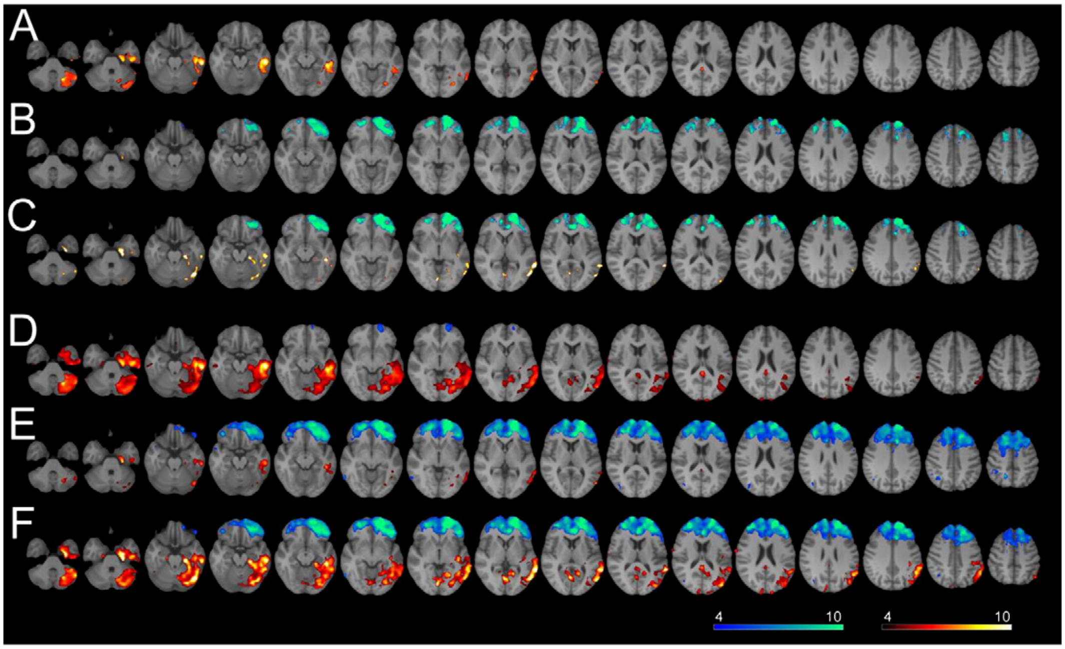


Fig. 1. SPM8 inter-scanner results. Result images for the inter-scanner comparison (paired *t*-test) are shown for the T1-based, SPM8 analysis. Colors indicate t-scores and are in red-yellow for relative increases (reference<re-scan), colors in blue-green are for relative decreases (reference>re-scan). Rows are for different significance thresholds and including no regressors, total intracranial volume or total gray matter volume into the GLM model. Thresholded maps are overlaid on the averaged, normalized (SPM8) T1 image. Images are in radiological convention (left in the image is right in the subject). A: p<0.05 (FWE corrected), no regressors. B: p<0.05 (FWE corrected), total intracranial volume included. C: p<0.05 (FWE corrected), total gray matter volume included. D: p<0.0001 (uncorrected), no regressors. E: p<0.0001 (uncorrected), total intracranial volume included. F: p<0.0001 (uncorrected), total gray matter volume included.

scanners, no information provided about the head coils used) also reported relevant inter-scanner differences but found that these did not affect the interpretation of their data and conclude "that it is feasible to combine data from different sites [...] for a VBM analysis" (Segall et al., 2009). Another multi-site VBM study in schizophrenia using identical 1.5 T scanners (no information about the used head coils) at four different sites found between sites "differences in a small number of brain regions" but with "little impact on the latter findings" (Meda et al., 2008). Our study, using technical identical 3 T scanners and an 8-channel vs. 12-channel phased-array of surface coils, showed severe inter-scanner differences (maximum t-scores from 10.65–17.21). This emphasizes the importance of receiver B1 bias, i.e.

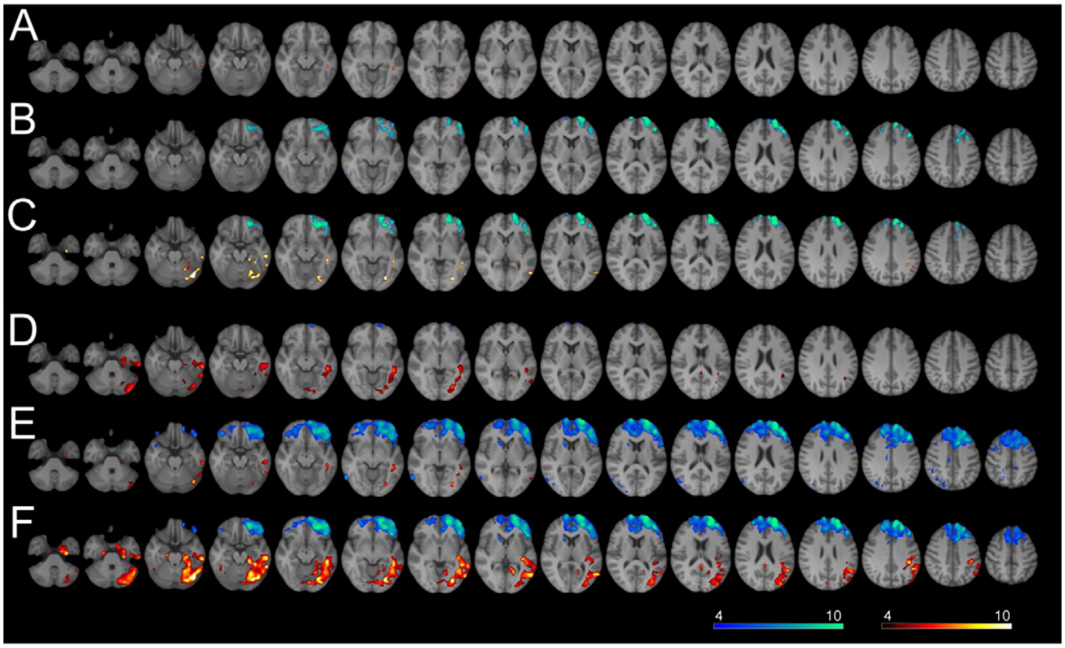


Fig. 2. SPM8 inter-scanner results with 30 mm bias cutoff. Result images for the inter-scanner comparison (paired *t*-test) are shown for the T1-based, SPM8 analysis with reduced bias cutoff (30 mm FWHM). Colors indicate t-scores and are in red-yellow for relative increases (reference<re-scan), colors in blue-green are for relative decreases (reference>re-scan). Rows are for different significance thresholds and including no regressors, total intracranial volume or total gray matter volume into the GLM model. Thresholded maps are overlaid on the averaged, normalized (SPM8) T1 image. Images are in radiological convention (left in the image is right in the subject). A: p<0.05 (FWE corrected), no regressors. B: p<0.05 (FWE corrected), total intracranial volume included. C: p<0.05 (FWE corrected), total gray matter volume included. D: p<0.0001 (uncorrected), no regressors. E: p<0.0001 (uncorrected), total intracranial volume included. F: p<0.0001 (uncorrected), total gray matter volume included.

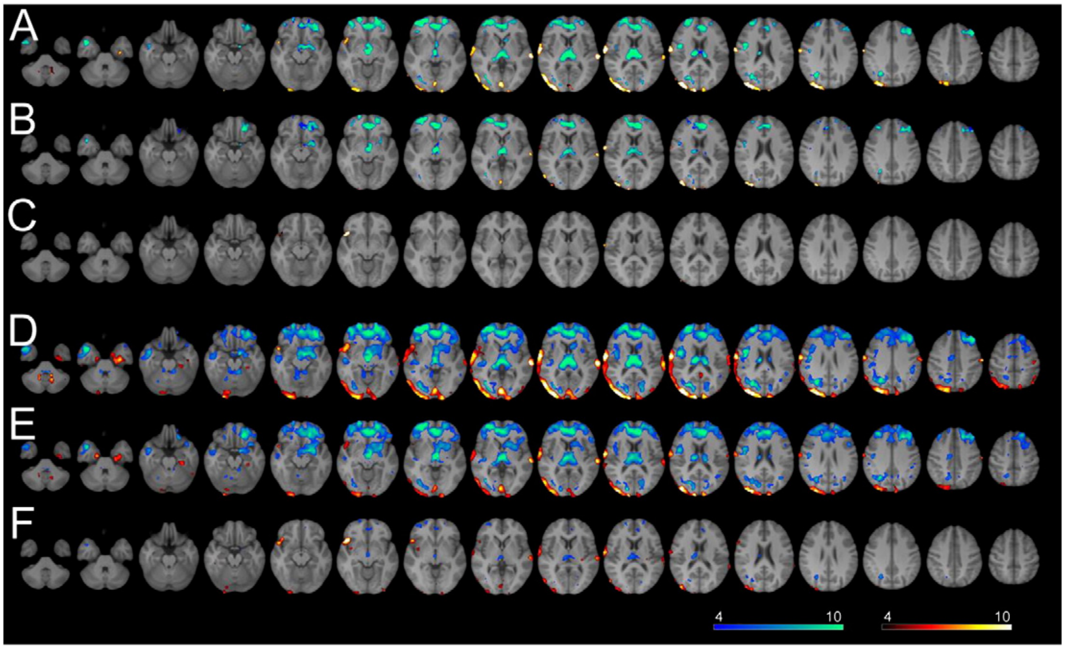


Fig. 3. FSL inter-scanner results. Result images for the inter-scanner comparison (paired *t*-test) are shown for the T1-based, FSL analysis. Colors indicate t-scores and are in red-yellow for relative increases (reference<re-scan), colors in blue-green are for relative decreases (reference>re-scan). Rows are for different significance threshold and including no regressors, total intracranial volume or total gray matter volume into the GLM model. Thresholded maps are overlaid on the averaged, normalized (FSL) T1 image. Images are in radiological convention (left in the image is right in the subject). A: p<0.05 (FWE corrected), no regressors. B: p<0.05 (FWE corrected), total intracranial volume included. C: p<0.05 (FWE corrected), total gray matter volume included. D: p<0.0001 (uncorrected), no regressors. E: p<0.0001 (uncorrected), total intracranial volume included. F: p<0.0001 (uncorrected), total gray matter volume included.

systematic effects of the head coils, that may have been underexplored in previous studies. The relatively favorable assumptions about multi-site VBM drawn from the 1.5 T/transmit-receive head coil scenario seem, thus, not to hold true in our more contemporary setup. Careful balancing of patient and control groups and including the scanner/site as a regressor into the analysis has to be recommended for 3 T/phased-array coil multi-site VBM studies especially when different head-coil configurations are used. Adequate balancing of patients and controls may help to control site-specific effects but this can be difficult to achieve especially in a longitudinal scenario with inevitable upgrades or hardware replacement needed for maintenance or with MRI scanners going out of service.

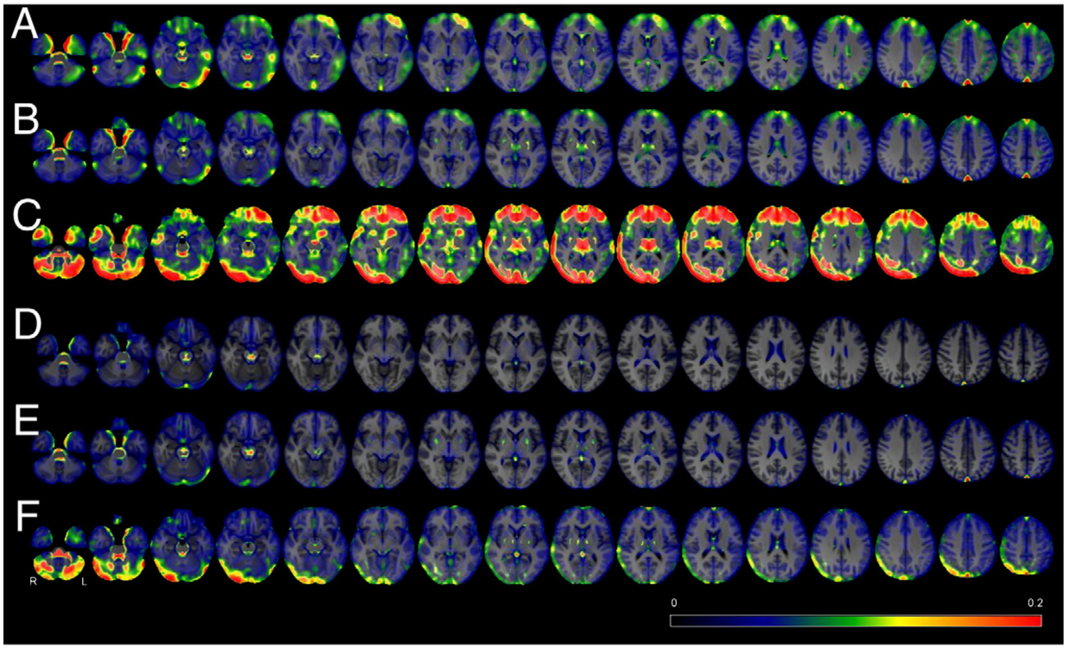


Fig. 4. Coefficient of difference maps. Unthresholded coefficient of difference (coeffD) maps for the different analysis streams are shown overlaid on the averaged, normalized T1 images. The color scale represents coeffD values (corresponding range 0–0.2, identical for all rows). Note the markedly higher coeffD values for the FSL processing stream (C and F). Images are in radiological convention (left in the image is right in the subject). A: Inter-scanner comparison, SPM8 processing stream, default settings. B: Inter-scanner comparison, SPM8 processing stream, 30 mm FWHM bias cutoff. C: Inter-scanner comparison, FSL4.1 processing stream. D: Intra-scanner comparison, SPM8 processing stream, default settings. E: Intra-scanner comparison, SPM8 processing stream, 30 mm FWHM bias cutoff. F: Intra-scanner comparison, FSL4.1 processing stream.

Correcting image bias

Since spatial signal bias is one of the main obstacles for comparing and processing MR images several (post-)processing methods for bias correction have been proposed, implemented on MR-systems or distributed in various software packages (Vovk et al., 2007). In principle there are retrospective methods, e.g. included in image segmentation (Zhang et al., 2001; Ashburner and Friston, 2005), that employ objective criteria (like minimal entropy) and do not require extra information (Sled et al., 1998) and parametric correction methods requiring additional reference scans in conjunction with image acquisition (Sled and Pike, 2000; Samson et al., 2006). In our study we applied the more common retrospective approach, using either SPM8 (unified segmentation) or FSL4.1 (FAST4). Although the pattern of the eventually detected inter-scanner differences was specific to the applied method, both failed to sufficiently remove the scanner-specific variations. GLM-detected differences in the FSL4.1 stream could be improved by including TGV as regressor of no interest, relevant inter-scanner discrepancies, however, persisted. Also the coeffD maps showed strong scan–rescan differences especially in the FSL4.1 stream. To simulate a "real-life" scenario we applied a widely used imaging protocol (Jack et al., 2008) and the default bias correction settings for both SPM8 and FSL4.1. A recent publication suggested that for 3 T studies a more liberal bias FWHM cutoff of (30–50 mm) is preferable (Zheng et al., 2009). A lowered bias cutoff (30 mm FWHM) within the unified segmentation algorithm in SPM8, in keeping with Zheng et al., reduced the detected differences in our study as well both for the GLM analysis and the coeffD maps. These did, however, still exceed a FWE corrected $p<0.05$ threshold; the reduction was also weaker when including the TIV or TGV regressors. The coeffD maps additionally revealed higher variability in basal ganglia regions (putamen/pallidum) in the analysis with a lowered bias cutoff.

The two MRI scanners used in our study were technically identical (both 3T Siemens TIM Trio) but did use a different, phased-array head coil (8-channel versus 12-channel). In view of Stonnington et al., our results suggest that the RF coil design may have a crucial influence on comparability. It appears likely that this is a potential cause of the severe differences observed but our technical setup did not allow for a structured comparison (e.g. switching the head-coils). Nevertheless our findings indicate that, for multi-site VBM studies, head-coil selection is a significant factor and that within a given study any (substantial) change of the receiver equipment should be avoided especially when using bias-susceptible sequences.

Quantitative MRI methods like DTI do effectively remove the receive bias (Vollmar et al., 2010). The use of high-resolution maps of R1 (Weiskopf et al., 2011) or magnetization transfer (Helms et al., 2009) for segmentation may be an alternative solution to improve inter-scanner comparability. The latter are, however, not yet generally available.

Intra-scanner differences

When using surface coils, the bias may be influenced by the positioning of individual subjects. Such variation, however, is not expected to be systematic and accordingly, we generally found no differences for the intra-scanner comparison (re-scan in the same scanner) exceeding a conservative $p<0.05$ (FWE corrected) threshold. With lower thresholds (we applied an uncorrected $p<0.0001$) chance findings are inevitable with any voxel-based method; the exact nature of the detected intra-scanner differences is, thus, not clear. A biological cause (i.e. aging) is, although theoretically possible, rather unlikely given the mean interval between the two scans of 191 days (range 23–465). In keeping with the GLM results lower coeffD values were found in the intra-scanner comparison as compared to the inter-scanner analysis.

Effect of TIV and TGV regressors

It is common practice to include a correcting factor e.g. TIV as regressor into a VBM study when modulated images are used that preserve the overall volume of gray matter. Otherwise global differences i.e. head size (correlated to body height and sex) could confound the comparison if not carefully controlled by matching. In our study (scan/re-scan comparison) the overall group characteristics should have been matched perfectly between the compared groups (identical subjects). Therefore, in theory, including any global regressors should not be needed. We did, however, observe a strong influence of TIV or TGV. For the SPM8-based comparisons, including these as regressors did exacerbate the GLM detected inter-scanner differences, for the FSL4.1-based pre-processing GLM detected inter-scanner differences were reduced. For the intra-scanner comparison, with generally mild differences, no clear pattern was evident. The most likely cause of this observation is a global scaling issue, in our case not caused by different subject specific factors but globally different image intensities, global differences of the applied segmentation algorithms, distortions or differing calibration between the two sites.

Comparison of the GLM and coeffD results

One potential weakness of a GLM analysis is that statistical significance (t-scores) is not only influenced by the signal in question but also by the observed standard deviation. In particular, increased noise introduced by the segmentation algorithm could lead to lower t-scores and, thus, to the incorrect assumption that the method is superior. To account for this effect we calculated voxel-based coeffD maps that should be more representative of the total difference. Also this approach eliminates the need to specify an (arbitrary) threshold (i.e. significance cutoff for the GLM analysis) and allows for a direct comparison of the different processing streams. In general coeffD maps confirmed the GLM results with strong variability focused on the cortical convexity that is close to the RF coil. High variability was also found in regions prone to susceptibility like the temporal pole and non-gray matter areas like large vessels, ventricles or the brainstem. In comparison of the different segmentation algorithms the FSL4.1 processed gray matter maps showed markedly higher coeffD values distributed over the entire brain for the inter-scanner and, although weaker, also for the intra-scanner comparison. This probably explains why the inclusion of global correcting regressors (TIV/TGV) in the GLM analysis was particularly effective within the FSL4.1 stream whereas in the SPM8 stream a reduction of global differences (=noise) resulted in higher significance scores. Thus, the applied version of the SPM8 segmentation algorithm seems to be more robust than the FSL4.1/FAST4 method. A lowered bias cutoff in the unified segmentation of SPM8 reduced cortical differences but resulted in higher coeffD values in basal ganglia regions (putamen/pallidum).

Limitations

Our study was done with healthy controls only; we, thus, could not directly compare the observed, "false positive" volumetric differences to "true positive" changes i.e. due to a disease condition. Further studies are needed to compare these effects.

Conclusion

We could show that, at present, multi-site VBM still is problematic when using 3 T and multi-channel head coils. None of the applied software packages was capable of sufficiently removing relevant inter-scanner differences post-hoc. Considering the coeffD results, however, SPM8 outperforms FSL4.1 processing. Our findings also indicate that a lowered bias cutoff threshold (30 mm FWHM) may be preferable for cortical areas within the SPM8 unified segmentation.

Supplementary materials related to this article can be found online at doi:10.1016/j.neuroimage.2011.02.029.

Acknowledgment

This study was supported by the University Medical Center Göttingen(“Rückkehrer-Startförderung” to NKF).

References

Ashburner, J., Friston, K.J., 2000. Voxel-based morphometry—the methods. Neuroimage 11 (6), 805–821.

Ashburner, J., Friston, K.J., 2005. Unified segmentation. Neuroimage 26 (3), 839–851.

Bernstein, M.A., Huston, J., Ward, H.A., 2006. Imaging artifacts at 3.0 T. J. Magn. Reson. Imaging 24 (4), 735–746.

Focke, N.K., Thompson, P.J., Duncan, J.S., 2008a. Correlation of cognitive functions with voxel-based morphometry in patients with hippocampal sclerosis. Epilepsy Behav. 12 (3), 472–476.

Focke, N.K., Symms, M.R., Burdett, J.L., Duncan, J.S., 2008b. Voxel-based analysis of whole brain FLAIR at 3T detects focal cortical dysplasia. Epilepsia 49 (5), 786–793.

Focke, N.K., Bonelli, S.B., Yogarajah, M., Scott, C., Symms, M.R., Duncan, J.S., 2009. Automated normalized FLAIR imaging in MRI-negative patients with refractory focal epilepsy. Epilepsia 50 (6), 1484–1490 Jun.

Helms, G., Draganski, B., Frackowiak, R., Ashburner, J., Weiskopf, N., 2009. Improved segmentation of deep brain grey matter structures using magnetization transfer (MT) parameter maps. Neuroimage 47 (1), 194–198 Mar 31.

Huppertz, H.-J., Kröll-Seger, J., Klöppel, S., Ganz, R.E., Kassubek, J., 2009. Intra- and interscanner variability of automated voxel-based volumetry based on a 3D probabilistic atlas of human cerebral structures. Neuroimage 49 (3), 2216–2224.

Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., et al., 2008. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. J. Magn. Reson. Imaging 27 (4), 685–691.

Jovicich, J., Czanner, S., Greve, D., Haley, E., van der Kouwe, A., Gollub, R., et al., 2006. Reliability in multi-site structural MRI studies: effects of gradient non-linearity correction on phantom and human data. Neuroimage 30 (2), 436–443.

Keller, S.S., Roberts, N., 2008. Voxel-based morphometry of temporal lobe epilepsy: an introduction and review of the literature. Epilepsia 49 (5), 741–757 May.

Meda, S.A., Giuliani, N.R., Calhoun, V.D., Jagannathan, K., Schretlen, D.J., Pulver, A., et al., 2008. A large scale (N = 400) investigation of gray matter differences in schizophrenia using optimized voxel-based morphometry. Schizophr. Res. 101 (1–3), 95–105 Apr.

Mugler, J.P., Brookeman, J.R., 1990. 3-Dimensional magnetization-prepared rapid gradient-echo imaging (3D MP-Rage). Magn. Reson. Med. 15 (1), 152–157 Jul.

Pardoe, H., Pell, G.S., Abbott, D.F., Berg, A.T., Jackson, G.D., 2008. Multi-site voxel-based morphometry: methods and a feasibility demonstration with childhood absence epilepsy. Neuroimage 42 (2), 611–616.

Samson, R.S., Wheeler-Kingshott, C.A., Symms, M.R., Tozer, D.J., Tofts, P.S., 2006. A simple correction for B1 field errors in magnetization transfer ratio measurements. Magn. Reson. Imaging 24 (3), 255–263.

Segall, J.M., Turner, J.A., van Erp, T.G., White, T., Bockholt, H.J., Gollub, R.L., et al., 2009. Voxel-based morphometric multisite collaborative study on schizophrenia. Schizophr. Bull. 35 (1), 82–95 Jan.

Sled, J.G., Pike, G.B., 2000. Correction for B1 and B0 variations in quantitative T2 measurements using MRI. Magn. Reson. Med. 43 (4), 589–593.

Sled, J.G., Zijdenbos, A.P., Evans, A.C., 1998. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imaging 17 (1), 87–97 Feb.

Smith, S.M., 2002. Fast robust automated brain extraction. Hum. Brain Mapp. 17 (3), 143–155 Nov.

Stonnington, C.M., Tan, G., Klöppel, S., Chu, C., Draganski, B., Jack Jr., C.R., et al., 2008. Interpreting scan data acquired from multiple scanners: a study with Alzheimer's disease. Neuroimage 39 (3), 1180–1185.

Vollmar, C., O'Muircheartaigh, J., Barker, G.J., Symms, M.R., Thompson, P., Kumari, V., et al., 2010. Identical, but not the same: intra-site and inter-site reproducibility of fractional anisotropy measures on two 3.0 T scanners. Neuroimage 51 (4), 1384–1394.

Vovk, U., Pernus, F., Likar, B., 2007. A review of methods for correction of intensity inhomogeneity in MRI. IEEE Trans. Med. Imaging 26 (3), 405–421 Mar.

Weiskopf, N., Lutt, A., Helms, G., Novak, M., Ashburner, J., Hutton, C., 2011. Unified segmentation based correction of R1 brain maps for RF transmit field inhomogeneities (UNICORT). Neuroimage 54 (3), 2116–2124. Epub 2010 Oct 18.

Woermann, F.G., Free, S.L., Koepp, M.J., Ashburner, J., Duncan, J.S., 1999. Voxel-by-voxel comparison of automatically segmented cerebral gray matter—a rater-independent comparison of structural MRI in patients with epilepsy. Neuroimage 10 (4), 373–384.

Wright, I.C., McGuire, P.K., Poline, J.B., Travere, J.M., Murray, R.M., Frith, C.D., et al., 1995. A voxel-based method for the statistical analysis of gray and white matter density applied to schizophrenia. Neuroimage 2 (4), 244–252 Dec.

Zhang, Y., Brady, M., Smith, S., 2001. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. IEEE Trans. Med. Imaging 20 (1), 45–57.

Zheng, W.L., Chee, M.W.L., Zagorodnov, V., 2009. Improvement of brain segmentation accuracy by optimizing non-uniformity correction using N3. Neuroimage 48 (1), 73–83 Oct 15.