Brain Imaging and Behavior (2019) 13:1538–1553
https://doi.org/10.1007/s11682-018-9986-x

ORIGINAL RESEARCH

fMRI data processing in MRTOOL: to what extent does anatomical registration affect the reliability of functional results?

Marco Ganzetti[1] · Gaia Amaranta Taberna[1] · Dante Mantini[1,2]

Published online: 22 November 2018

Abstract

Spatial registration is an essential step in the analysis of fMRI data because it enables between-subject analyses of brain activity, measured either during task performance or in the resting state. In this study, we investigated how anatomical registration with MRTOOL affects the reliability of task-related fMRI activity. We used as a benchmark the results from two other spatial registration methods implemented in SPM12: the Unified Segmentation algorithm and the DARTEL toolbox. Structural alignment accuracy and the impact on functional activation maps were assessed with high-resolution T1-weighted images and a set of task-related functional volumes acquired in 10 healthy volunteers. Our findings confirmed that anatomical registration is a crucial step in fMRI data processing, contributing significantly to the total inter-subject variance of the activation maps. MRTOOL and DARTEL provided greater registration accuracy than Unified Segmentation. Although DARTEL had superior gray matter and white matter tissue alignment than MRTOOL, there were no significant differences between DARTEL and MRTOOL in test–retest reliability. Likewise, we found only limited differences in BOLD activation morphology between MRTOOL and DARTEL. The test–retest reliability of task-related responses was comparable between MRTOOL and DARTEL, and both proved superior to Unified Segmentation. We conclude that MRTOOL, which is suitable for single-subject processing of structural and functional MR images, is a valid alternative to other SPM12-based approaches that are intended for group analysis. MRTOOL now includes a normalization module for fMRI data and is freely available to the scientific community.

Keywords Magnetic resonance imaging · Functional MRI · Image registration · Image normalization · SPM · MRTOOL

Introduction

Since its introduction in the early 1990s, functional magnetic resonance imaging (fMRI) based on blood oxygenation level-dependent (BOLD) signals has led to countless scientific and clinical breakthroughs on human brain function (Uludağ et al. 2015). The success of fMRI in the fields of basic and clinical neuroscience is mainly due to its high spatial resolution and ability to specifically identify changes in brain activity, which sets it apart from other brain imaging techniques. A typical fMRI study involves the acquisition of multiple, consecutive functional scans to track brain dynamics during task performance or rest, along with a structural scan to characterize the anatomy of the detected functional activity or connectivity. The structural scan is also typically used to make inferences about the anatomical correspondence of brain responses across individuals and studies. Accordingly, *spatial registration* (or *spatial normalization*) is today considered a fundamental step in fMRI data preprocessing. Spatial registration requires the non-linear transformation of an individual's structural image (typically a T1-weighted (T1-w) image) to a template image in MNI (Montreal Neurology Institute) coordinate space. Once this transformation is computed, it is applied to the original functional images (linearly aligned to the individual's structural image), thereby generating corresponding functional images in template space.

fMRI studies typically involve comparisons of brain activity and connectivity across several participants. Therefore, the detection of true effects may be hampered by both systematic and random between-subject misalignments during the spatial registration step. Random misalignments are likely to reduce

✉ Marco Ganzetti
marco.ganzetti@kuleuven.be

[1] Research Center for Movement Control and Neuroplasticity, KU Leuven, Tervuursevest 101, 3001 Leuven, Belgium

[2] Functional Neuroimaging Laboratory, IRCCS San Camillo Hospital, Venice, Italy

the statistical benefits of group-level analyses, leading to an increase in false negatives. Systematic misalignments may deceptively link neural activity to incorrect anatomical areas. Because anatomical labeling is template-based, classification of brain responses detected by fMRI may be inaccurate, especially when the functional responses are located close to the boundary between two anatomically defined brain regions. Establishing the validity and reliability of fMRI results is an area of concern for the neuroscientific community, and has attracted particular interest in recent years (Crinion et al. 2007; Crivello et al. 2002; Fischmeister et al. 2013).

To date, several solutions for spatial registration of structural magnetic resonance (MR) images have been proposed (Crinion et al. 2007; Klein et al. 2009). Many fMRI studies use the normalization/segmentation tools available in the SPM (Statistical Parametric Mapping) software package: the Unified Segmentation algorithm (Ashburner and Friston 2005; Pohl et al. 2005) and the DARTEL toolbox (Ashburner 2007). These yield significantly better registration quality than previous solutions and have been successfully used in numerous studies (Hoffman and Lambon Ralph 2018; Hope et al. 2015; Klöppel et al. 2008; Lorio et al. 2016; Michely et al. 2018; Ossenkoppele et al. 2015). DARTEL ranks among the top-performing registration methods currently available for group-level fMRI analyses (Klein et al. 2009).

Our recent work focused on improving the spatial registration of structural MR images in the presence of large deviations from standard brain morphology. To this end, we developed an SPM-based enhanced normalization approach, which we implemented in our MRI analysis toolbox, MRTOOL (Ganzetti et al. 2018). Our previous analyses demonstrated that MRTOOL yields more reliable anatomical co-localization across individuals with advanced levels of brain atrophy and ventricular enlargement than the standard Unified Segmentation algorithm. Our results also showed that the use of different spatial normalization approaches can lead to dissimilar results. Such differences can be more easily assessed in images where the brain anatomy deviates substantially from the standard template image being used.

The extent to which our new approach to spatial registration may affect functional imaging results has not yet been tested. Here, we aim to address this question by comparing the reliability of BOLD-fMRI results obtained via the normalization module of MRTOOL, the Unified Segmentation algorithm, and the DARTEL toolbox, all in SPM12. We apply these three registration methods to task-related BOLD data extracted from a publically available dataset and used for test–retest reliability analyses (Gorgolewski et al. 2013a). On the basis of our previous results obtained with structural MR data (Ganzetti et al. 2018), we hypothesize that MRTOOL will produce fewer registration errors than the Unified Segmentation algorithm, leading to greater statistical power and increased detection of functional activations, as well as increased test–retest reliability. Furthermore, we expect a similar level of performance from MRTOOL, which is suitable for single-subject data processing, to that attained by DARTEL, which is primarily used for group-level analyses.

Methods

Participants and procedures

MR imaging data included in this study were downloaded from the open access database OpenfMRI (Poldrack and Gorgolewski 2017). The dataset was originally acquired to validate fMRI tasks used in pre-surgical planning for tumor resection. The study was approved by the local South East Scotland Research Ethics Committee, and informed consent was obtained from all participants. MR data were collected from 10 normal healthy volunteers (four men and six women, aged 50–58 years). Three volunteers were left-handed and seven were right-handed, according to their own declaration. Each participant was scanned in two separate sessions, either 2 days (eight participants) or 3 days (two participants) apart; these are referred to as the 'Test' and 'Retest' sessions.

Tasks

Participants performed three behavioral tasks (Table 1): covert verb generation, overt word repetition, and motor movements. The first two tasks were aimed at mapping language areas of the brain either with (overt) or without (covert) actual speech production. The motor task consisted of finger tapping, foot flexion, and lip pursing interleaved with fixation on a cross. The behavioral paradigms were implemented in Presentation v. 17.0 (Neurobehavioral Systems, www.neurobs.com). Stimuli were synchronized and presented by NordicNeuroLab hardware (www.nordicneurolab.com). During the first scanning session, each participant familiarized themselves with the tasks by practicing them inside the scanner.

Motor task

Subjects had to move a body part corresponding to a picture presented on a screen. The following instructions were given: "You have to tap your index finger when you see a picture of a finger, flex your foot when you see a picture of a foot and purse your lips when you see a picture of lips". A block design with 15-s activation and 15-s rest period was employed. In every block, subjects moved the index finger of their dominant hand, or flipped their dominant foot or pouched their mouth. Movement was paced with a frequency of 0.4 Hz using visual stimuli. There were five repetitions of each activation/rest block, for a total scan time of 7 min and 40 s.

Table 1 Test–retest reliability analysis of segmented tissue maps

		ICC	Pairwise comparisons								
			MRTOOL - UNIF.SEG.			DARTEL - UNIF.SEG.			DARTEL - MRTOOL		
		Mean ± sd	*Effect size (d)*	*D*	*p*	*Effect size (d)*	*D*	*p*	*Effect size (d)*	*D*	*p*
GM	UNIF.SEG. MRTOOL DARTEL	0.59 ± 0.29 0.73 ± 0.30 0.72 ± 0.28	0.32	0.29	****	0.33	0.25	****	0.01	0.06	ns
WM	UNIF.SEG. MRTOOL DARTEL	0.63 ± 0.31 0.75 ± 0.31 0.71 ± 0.29	0.27	0.25	****	0.20	0.32	****	0.06	0.05	ns

Means and standard deviations of the Intraclass Correlation Coefficient (ICC) computed over the tissue probability maps (TPMs) for gray matter (GM) and white matter (WM). ICC values range from 0 to 1, where a value of 1 indicates complete correlation of the subject voxel intensities between the Test and Retest sessions. Cohen's effect size (d) is reported to allow a direct comparison between the mean ICC values across registration methods. The Kolmogorov–Smirnov (K-S) test statistic (D), measuring the largest distance between two empirical ICC distributions, and the significance level (p) are reported. Significance level: ****p < 0.0001; ns, not significant

Covert verb generation task

Subjects were asked to think of a verb complementing a noun presented to them visually. The following instructions were given: "When a word appears it will be a noun. Think of what you can do with it and then imagine saying 'With that I can ...' or 'That I can ...' ". A block design with 30-s task and 30-s rest blocks was employed. During the task blocks, ten nouns were presented for one second each followed by a fixation cross during which subject had to produce the response. The nouns were chosen at random from a set of 70 nouns. Rest blocks had an analogous structure but with each word replaced by scrambled visual patterns generated by scrambling the phase of the 'picture' of each word, i.e. the control patterns were matched in the amplitude spectrum. Seven activation/rest blocks were presented, for a total scan time of 7 min and 12.5 s.

Overt word repetition task

Subjects had to repeat aloud words that were heard via headphones. The following instructions were given: "When you hear a word, repeat it immediately". A block design with 30-s activation and 30-s rest blocks was employed in conjunction with a sparse sampling data acquisition technique to present and record stimuli during the silent periods. After 2.5 s of blank screen during which the fMRI data were acquired, subjects were presented with an auditory stimulus. This consisted of a pre-recorded native British English speaker reading a noun chosen at random from a set of 36 nouns. This was followed by a question mark prompting the subject to repeat the word. Subject responses were recorded using an MRI compatible microphone. The question mark disappeared after 1741 ms and the sequence was repeated 6 times in the same block. A blank screen was presented during rest periods. There were six activation/rest blocks, for a total scan time of 7 min and 40 s.

MRI acquisition protocol

Data were acquired on a GE Signa HDxt 1.5 T scanner with an 8-channel phased-array head coil at the Brain Research Imaging Center, University of Edinburgh. fMRI data were acquired with a single-shot gradient-echo echo-planar imaging pulse sequence with the following parameters: in-plane field of view (FOV) 256 × 256 mm, acquisition matrix 64 × 64, 30 slices per volume (interleaved slice order), slice thickness 4 mm, flip angle 90°, echo time (TE) 50 ms. The repetition time (TR) was 5 s for the overt word-repetition task and 2.5 s for all other tasks. In addition, in each session a high-resolution 3-D T1-w scan was acquired with a 3-D inversion recovery prepared (IRP) pulse sequence. The acquisition parameters for this additional sequence were: in-plane FOV 256 × 256 mm, acquisition matrix 256 × 256, 156 slices per volume, slice thickness 1.3 mm, flip angle 8°, TR 10 s, TE 4 s, and inversion time (TI) 500 ms.

Image processing

First, fMRI data were corrected for head motion with the realignment tool in SPM12. Then, the SPM12 coregistration tool was used to calculate an affine transformation aligning the mean fMRI volume to the structural T1-w image in original space. This affine transformation was then applied to the motion-corrected fMRI data. Spatial registration of the T1-w image to the standard SPM12 template in MNI space (ICBM152-MNI) was then performed with either the SPM12 Unified Segmentation algorithm (Ashburner and Friston 2005), the MRTOOL normalization module (Ganzetti et al. 2018), or the SPM12 DARTEL toolbox (Ashburner 2007).

The resulting non-linear deformation was subsequently applied to the fMRI data. Finally, spatial smoothing with a 6-mm full width at half maximum Gaussian kernel was applied to the resulting fMRI volumes in MNI space.

Subject-level analysis

Statistical analysis was also carried out in SPM12. fMRI time series were high-pass filtered with a cutoff of 128 Hz, and then subjected to general linear modeling (Friston et al. 1994). The design matrix consisted of task regressors and head-motion regressors, as well as their first derivatives. A single task regressor was employed for covert verb generation and overt word repetition. This regressor was generated by a boxcar function convolved with a canonical hemodynamic response function. For the motor tasks, each body part was modeled with a separate boxcar regressor. For the language tasks, we generated a simple contrast based on the task regressor. For the motor tasks, we generated three contrasts, one per condition, in which each body part was assessed against the other two. The finger, foot, and lips contrast images from left-handed subjects were flipped along the sagittal plane.

Group-level analysis

The average BOLD responses for the Test and Retest sessions were computed at the group level. For each subject, task-specific contrast images, as defined in the subject-level analysis, were averaged between sessions. The main effect of task for each contrast was then computed. The false discovery rate (FDR) method was employed at the cluster level to correct for multiple comparisons (Chumbley and Friston 2009; Woo et al. 2014). Specifically, statistical t-score maps were thresholded with a primary threshold level of $p < 0.001$, followed by a cluster-extent probability threshold set to 0.05.

Reliability analysis

Anatomical registration

The segmentation accuracy of the anatomical registration was assessed qualitatively and quantitatively. MRTOOL, the Unified Segmentation algorithm, and the DARTEL toolbox all use the same SPM12 processing framework to jointly execute spatial registration and segmentation. Individual tissue probability maps (TPM) of the gray matter (GM) and white matter (WM) in MNI space were thresholded ($p > 0.5$) and overlaid to generate consensus maps for each method and session separately. Likewise, the Dice Similarity Coefficient (DSC), which is a well-known metric for evaluating segmentation outputs (Fischmeister et al. 2013; Ganzetti et al. 2018; Shattuck et al. 2001; Van Leemput et al. 1999) was used to quantify the degree of overlap between and within subjects on the individual GM and WM tissue masks. The between-subject DSC was performed by computing the overlap between the thresholded map from a single subject and the thresholded maps of all the other subjects in the same session:

$$DSC = 2^{\cdot}\frac{|S_a \cap S_b|}{|S_a| \cup |S_b|} \tag{1}$$

where S_a and S_b are, respectively, the normalized and thresholded TPM for either GM or WM in two distinct subjects. The procedure was repeated for the Test and Retest sessions. Similarly, for within-subject overlap, the DSC was calculated as follows:

$$DSC = 2^{\cdot}\frac{|S_{test} \cap S_{retest}|}{|S_{test}| \cup |S_{retest}|} \tag{2}$$

where S_{test} and S_{retest} are the TPM for either GM or WM in a single subject for the Test and Retest sessions. DSC values range from 0 to 1, with higher values indicating greater overlap.

The test–retest reliability of the registration was assessed using the intraclass correlation coefficient (ICC). The ICC was computed voxel-wise over the probability values of GM and WM tissue maps following the approach described in Caceres et al. (2009). If registration (matching between brain tissues) is improved, the computed ICC should be a closer estimate of the true ICC for a specific voxel. This implementation is based on computing a single ICC value for each voxel, generating a distribution of ICC values. A two-way model (subject vs sessions) with no interaction (Shrout and Fleiss 1979) was used to compute the ICC as follows:

$$ICC(3,1) = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2} \tag{3}$$

where σ_r^2 is the between-subject variance and σ_e^2 is defined in the following manner:

$$\sigma_e^2 = \frac{\sum_{j=1}^{n}\left(\left(t_{1j} - \overline{t_1}\right) - \left(t_{2j} - \overline{t_2}\right)\right)^2}{n-1} \tag{4}$$

where n is the total number of subjects, t_{1j} and t_{2j} are the values of a single voxel for subject j for the Test and Retest sessions, and $\overline{t_1}$ and $\overline{t_2}$ are the between-subjects mean values for the Test and Retest sessions.

Effect on regional fMRI activations

We performed a reliability analysis on the BOLD responses for each task-specific region. As described above for the evaluation of anatomical registration, we measured the between- and within-subjects overlap of thresholded t-maps. Single-subject t-maps were thresholded according to an adaptive cluster-forming threshold (q = 0.05) criterion (Gorgolewski et al. 2012). This method combines Gamma–Gaussian mixture modeling with topological thresholding to improve cluster delineation at the subject level. Adaptive thresholding performs better than fixed thresholding in terms of over- or underestimation of the true activation border, especially in the presence of different signal-to-noise ratios. We repeated this analysis for both sessions and for the three registration methods.

Likewise, we performed a test–retest reliability analysis, based on the previously described voxel-wise ICC, for each task-specific contrast map generated from the subject-level analysis.

To ensure that the results were not biased toward low values because of a lack of reliability in the many regions that are not task-specific, we chose to report results only within specific anatomical regions (Gorgolewski et al. 2013a). For each of the five tasks, we used the probability maps available in the Anatomy toolbox to construct a region of interest (ROI) covering the area expected to show functional activity (Eickhoff et al. 2005, 2006, 2007). For the three motor conditions, the ROI was primary motor cortex, obtained by merging together Brodmann areas 4a and 4p (Geyer et al. 1996). For the covert verb-generation task, the ROI was Broca's area, obtained by combining Brodmann areas 44 and 45 (Amunts et al. 1999). For the overt word-repetition task, the ROI included the left primary auditory cortex (TE1.0, TE1.1, and TE1.2) and the posterior division of the left superior temporal gyrus (Wernicke's area) (Morosan et al. 2001, 2005).

To complement the group-level analysis of the average BOLD responses, we also extracted the activation-peak coordinates from the average contrast map between sessions for each subject and task.

Statistical analysis

We used modeling analysis to estimate whether the differences in performance across registration methods (assessed by the DSC) were significant. In particular, for the between-subjects analysis, we fitted the results to a linear mixed model with the registration method as a fixed effect and random intercepts for the subjects. The inclusion of the random intercepts significantly enhanced the fit of the model as it explained the variability introduced by the individuals. Before fitting the model, all mutually exclusive DSC values—generated by comparing an individual subject with the other subjects—were averaged together, creating a single measure for each subject. The resulting averaged DSC values were in turn averaged across sessions, for each subject within each method. For the within-subject analysis, the same model was fitted to the data. In both cases, assumption of normality was verified by inspecting Q-Q plots and computing the Shapiro–Wilk test on the residual distributions. This analysis was repeated for the evaluation of the anatomical alignment of GM and WM, for the volume overlap of the functional responses, and for the activation peak analysis. For the last two, the analysis was repeated for each separate task.

For the voxel-wise ICC reliability analysis, we computed the mean and standard deviation of the ICC distribution. To formally compare the reliability across registration methods, we reported the effect size using Cohen's d. Statistical differences between ICC distributions were assessed with the Kolmogorov–Smirnov (K-S) test (Stephens 1992). The K-S test statistic D measures the largest distance between two empirical ICC distribution functions.

Results

Structural registration

To qualitatively evaluate the registration performance of Unified Segmentation, MRTOOL, and DARTEL, we visually compared the spatial pattern of tissue overlap across subjects (Fig. 1). We observed inaccuracies in the spatial alignment of both GM and WM tissue maps with the Unified Segmentation algorithm. MRTOOL and DARTEL both showed substantial improvement over Unified Segmentation, with sharper transitions at the borders between tissues. There were no striking differences between the maps from the Test and Retest sessions.

To corroborate this qualitative assessment, we performed a statistical analysis to quantify the exact degree of overlap for both GM and WM thresholded maps (Fig. 2). The between-subjects DSC was significantly larger for DARTEL than for Unified Segmentation ($p < 0.0001$) or MRTOOL ($p < 0.0001$), for both GM and WM. MRTOOL had significantly larger DSC values than Unified Segmentation for both sessions and tissue types ($p < 0.0001$).

The within-subject overlap analysis confirmed the reliability of DARTEL registration (Fig. 3); however, the GM DSC values were only slightly higher for DARTEL than for MRTOOL ($p < 0.05$). For the WM maps, MRTOOL produced slightly larger DSC values than DARTEL but the difference was not significant. Of the three methods, Unified Segmentation clearly exhibited the lowest performance.

The test–retest reliability analysis on the voxel-wise ICC distributions (Table 1) revealed higher reliability for MRTOOL (ICC = 0.73 and ICC = 0.75 for GM and WM, respectively) and DARTEL (ICC = 0.72 and ICC = 0.71 for GM and WM, respectively) than for Unified Segmentation (ICC = 0.59 and ICC = 0.63 for GM and WM, respectively). Although MRTOOL produced higher ICC values than DARTEL for both GM and WM, the difference was not statistically significant.

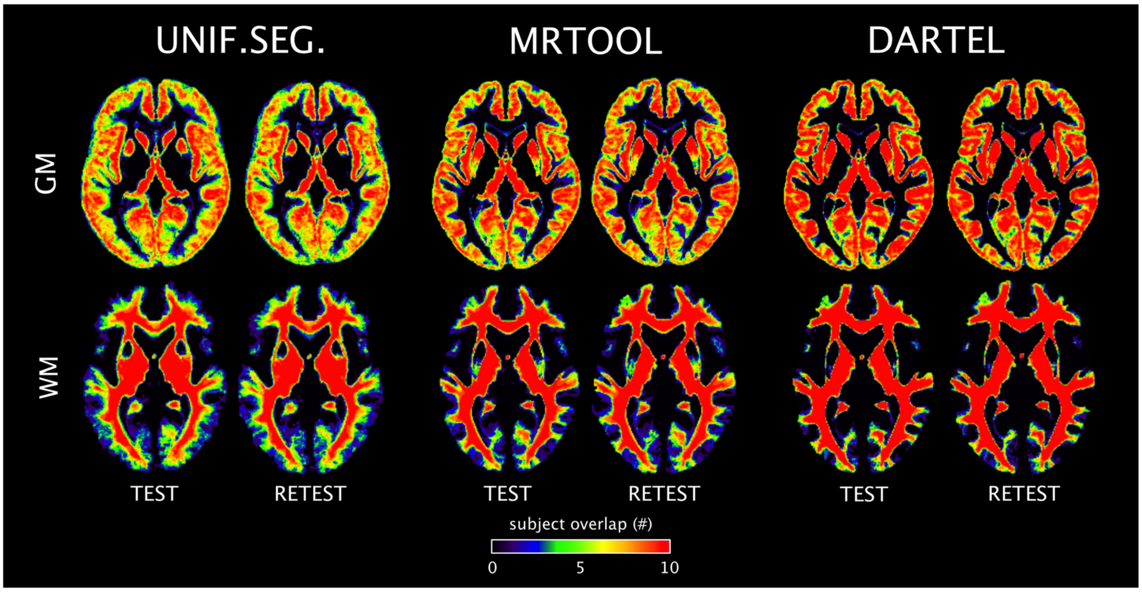


Fig. 1 **Consensus maps of segmented tissues.** Registration performance between methods was assessed by comparing the overlap of gray matter (GM) and white matter (WM) tissue probability maps (TPMs) across subjects. TPMs in the standard space of the SPM12 template were thresholded at 0.5. Values range from 0 to 10, where 10 indicates complete overlap. Consensus maps are reported for each method and session separately

Analysis of functional maps

Using the DSC as a performance metric, we next tested the impact of anatomical registration on task-related activations. Regional BOLD responses at the subject level were compared both within and between subjects. The within-subject reliability analysis did not reveal any significant differences between the three registration methods for the different tasks, suggesting a negligible effect of anatomical registration on the spatial overlap of functional activations across sessions (Table 2). However, differences between the methods became apparent in the between-subjects analysis of DSC values (Table 3). DARTEL and MRTOOL performed better than Unified Segmentation regardless of the task (Table 3). Despite the fact that DARTEL performed significantly better than MRTOOL in the analysis of tissue overlap (above), there were no significant differences between the two methods in the motor tasks or the verb-generation task. However, MRTOOL had lower DSC values than DARTEL in the word-repetition task ($p < 0.0001$). MRTOOL was relatively more reliable than Unified Segmentation, especially for the lip-pursing ($p < 0.001$) and verb-generation ($p < 0.01$) tasks.

The test–retest reliability analysis on the voxel-wise ICC distributions showed that DARTEL and MRTOOL performed better than Unified Segmentation regardless of the task (Table 4). There were no tangible differences in effect size between MRTOOL and DARTEL ($d < 0.2$). Except for the foot-flexion ($p < 0.05$) and word-repetition ($p < 0.0001$) tasks, there were no statistically significant differences between the ICC distributions for MRTOOL and DARTEL.

To extend these reliability analyses, we investigated the impact of anatomical registration on group-level BOLD responses for both motor (Fig. 4) and language (Fig. 5) conditions. The locations of movement-related activity, as detected by the three registration methods, were largely in line with those reported in the literature (Gorgolewski et al. 2013a). The activations were more spatially confined with MRTOOL and DARTEL than with Unified Segmentation, with regional differences in the t values and the spatial extent of the clusters, as well as in the activation-peak coordinates (Table 5).

With the Unified Segmentation algorithm, the finger-tapping response was mainly located in the postcentral gyrus, specifically in the primary somatosensory cortex. With MRTOOL and DARTEL, the finger-tapping response was not limited to sensory areas, but spanned anteriorly toward the middle contralateral precentral gyrus. The foot-flexion response was also confined to the hemisphere contralateral to the movement. With MRTOOL and DARTEL, the cluster, corresponding to the precentral and postcentral gyri, was well-defined and characterized by relatively high t values. In contrast, the activation cluster obtained with Unified Segmentation was located more medially, straddling the mid-sagittal plane. In the lip-pursing condition, all three methods yielded consistent bilateral activations in the inferior part of the precentral gyrus, although the Unified Segmentation approach resulted in the lowest t values, indicating lower reliability. Activation was also observed in the cerebellum, ipsilaterally in both the finger and foot tasks, and bilaterally in the lips task.

The results from the language tasks largely confirmed those from the motor tasks (Fig. 5). The verb-generation condition yielded consistent activation in the language-specific areas of the brain. For all registration methods, significant clusters were found in the Broca's area (BA 44 and 45), the anterior superior part of the left putamen, the left posterior superior temporal gyrus, and the supplementary motor area. However, the left thalamus was significantly activated only when MRTOOL was used for anatomical registration. The activation cluster in the Broca's area showed higher reliability with MRTOOL and DARTEL than with Unified Segmentation. In the word-repetition task, we found significant clusters over the bilateral superior temporal gyrus, at the level of the primary auditory cortex (Morosan et al. 2001), and the Wernicke's area (Caspers et al. 2006). A major difference between methods was the detection of a significant response in the left anterior putamen with MRTOOL and DARTEL, but not with Unified Segmentation. Furthermore, the activation patterns in the auditory areas TE 1.1, TE 1.2, and TE 1.3 had lower t values with Unified Segmentation, indicating reduced reliability with this method. Overall, MRTOOL and DARTEL yielded similar activation maps and consistent localization of activation peaks (Table 5).

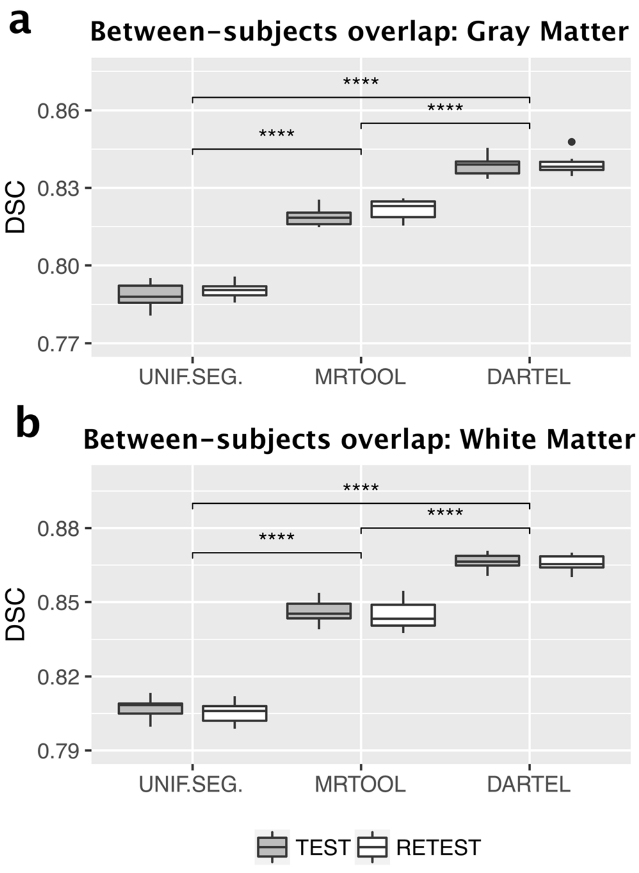


Fig. 2 **Anatomical registration: between-subjects tissue overlap.** Box and whisker plots showing the Dice Similarity Coefficient (DSC) between the tissue probability map (TPM) for a single subject and the respective maps for all other subjects. The procedure was iterated for each subject and repeated separately for the Test and Retest sessions. The DSC was computed on the normalized and thresholded ($p > 0.5$) TPMs for gray matter and white matter. DSC values range from 0 to 1, where 1 indicates complete overlap. Significance level: ****$p < 0.0001$

Discussion

Over the past few decades, fMRI has become a key technique for measuring brain activity noninvasively (Bullmore 2012). However, no consensus on the reliability of BOLD responses has yet been reached (Bennett and Miller 2010; Giussani et al. 2010). The inherent complexity of fMRI data processing is one of the major determinants of this impasse (Savoy 2005). Given the multilevel structure of fMRI data processing, numerous factors can influence the final outcome. In the present study, we investigated how anatomical registration—which

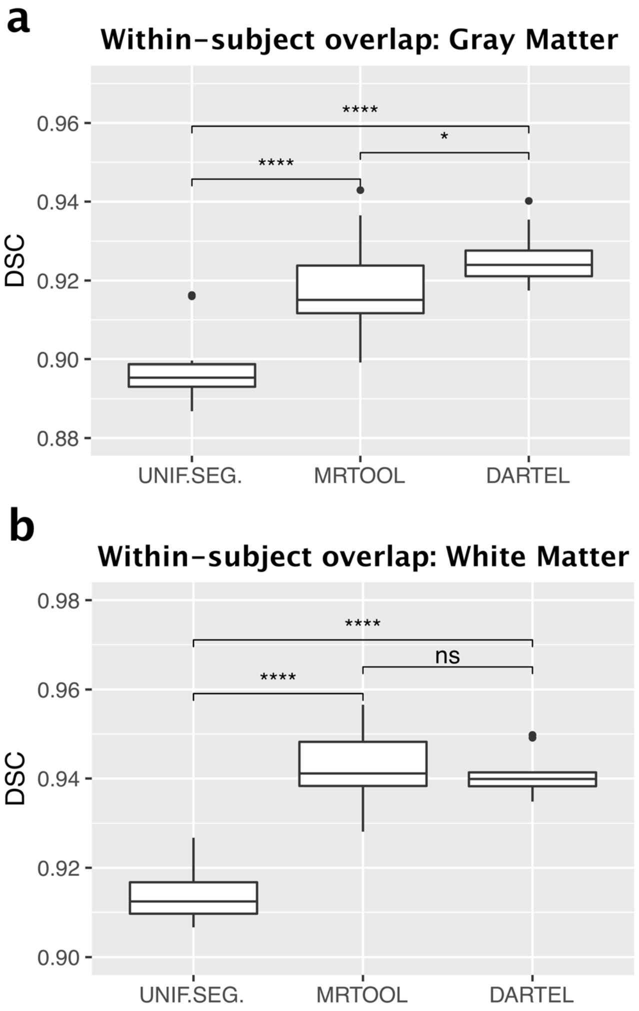


Fig. 3 **Anatomical registration: within-subject tissue overlap.** Box and whisker plots show the Dice Similarity Coefficient (DSC) between the tissue probability map (TPM) for a single subject in the Test session and the respective map for the same subject in the Retest session. The procedure was iterated for each subject. The DSC was computed on the normalized and thresholded (p > 0.5) TPMs for gray matter and white matter. DSC values range from 0 to 1, where 1 indicates complete overlap. Significance levels: *$p < 0.05$, ****p < 0.0001; ns, not significant

Table 2 Volume overlap of thresholded task-related responses: within-subject analysis

			Pairwise comparisons								
		DSC	MRTOOL - UNIF.SEG.			DARTEL - UNIF.SEG.			DARTEL - MRTOOL		
		Mean ± sd	*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*
Finger	UNIF.SEG.	0.644 ± 0.183	0.002 [−0.040,0.045]	0.08	ns	−0.029 [−0.072,0.014]	−0.41	ns	−0.031 [−0.074,0.011]	−0.48	ns
	MRTOOL	0.646 ± 0.184									
	DARTEL	0.615 ± 0.200									
Foot	UNIF.SEG.	0.546 ± 0.205	0.026 [−0.006,0.058]	0.71	ns	0.008 [−0.024,0.040]	0.18	ns	−0.018 [−0.050,0.014]	−0.38	ns
	MRTOOL	0.572 ± 0.208									
	DARTEL	0.554 ± 0.212									
Lips	UNIF.SEG.	0.614 ± 0.262	0.079 [−0.024,0.183]	0.54	ns	0.048 [−0.056,0.152]	0.28	ns	−0.031 [−0.135,0.072]	−0.4	ns
	MRTOOL	0.694 ± 0.142									
	DARTEL	0.663 ± 0.138									
Verb Gen	UNIF.SEG.	0.584 ± 0.180	0.024 [−0.013,0.062]	0.52	ns	0.049 [0.011,0.086]	1.08	**	0.024 [−0.014,0.062]	0.46	ns
	MRTOOL	0.609 ± 0.190									
	DARTEL	0.633 ± 0.171									
Word rep	UNIF.SEG.	0.524 ± 0.194	−0.028 [−0.085,0.030]	−0.48	ns	0.024 [−0.034,0.081]	0.25	ns	0.051 [−0.006,0.109]	0.87	ns
	MRTOOL	0.496 ± 0.219									
	DARTEL	0.547 ± 0.207									

Means and standard deviations of the Dice Similarity Coefficient (DSC) between the suprathreshold maps of single subjects in the Test and Retest sessions. For each task-related response, the DSC was computed over suprathreshold voxels restricted to a specific region of interest. Average DSC values across subjects are reported for each separate method along with the respective pairwise comparisons. Significance level: **p < 0.01; ns, not significant

Table 3 Volume overlap of thresholded task-related responses: between-subjects analysis

			Pairwise comparisons								
		DSC	MRTOOL - UNIF.SEG.			DARTEL - UNIF.SEG.			DARTEL - MRTOOL		
		Mean ± sd	*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*
Finger	UNIF.SEG.	0.452 ± 0.113	0.032 [−0.011,0.075]	0.44	ns	0.039 [−0.004,0.081]	0.9	ns	0.007 [−0.036,0.050]	0.13	ns
	MRTOOL	0.484 ± 0.151									
	DARTEL	0.490 ± 0.120									
Foot	UNIF.SEG.	0.386 ± 0.107	0.020 [−0.029,0.070]	0.69	ns	0.055 [0.005,0.104]	0.63	*	0.034 [−0.015,0.084]	0.45	ns
	MRTOOL	0.406 ± 0.087									
	DARTEL	0.441 ± 0.072									
Lips	UNIF.SEG.	0.457 ± 0.093	0.078 [0.031,0.125]	1.37	***	0.099 [0.052,0.146]	1.37	****	0.021 [−0.026,0.068]	0.35	ns
	MRTOOL	0.535 ± 0.070									
	DARTEL	0.556 ± 0.050									
Verb gen	UNIF.SEG.	0.400 ± 0.089	0.030 [0.008,0.053]	1.1	**	0.045 [0.022,0.068]	1.17	****	0.014 [−0.007,0.037]	0.98	ns
	MRTOOL	0.430 ± 0.109									
	DARTEL	0.445 ± 0.118									
Word rep	UNIF.SEG.	0.358 ± 0.093	0.021 [−0.011,0.054]	1.1	ns	0.084 [0.051,0.116]	1.62	****	0.062 [0.030,0.095]	1.36	****
	MRTOOL	0.379 ± 0.107									
	DARTEL	0.441 ± 0.110									

Means and standard deviations of the Dice Similarity Coefficient (DSC) between the suprathreshold map for a single subject and the respective maps for all other subjects. The procedure was iterated for each subject and session. For each task-related response, DSC was computed over suprathreshold voxels restricted to a specific region of interest. DSC values for each pair of subjects were averaged together. The resulting mean values averaged across sessions are reported for each separate method together with the respective pairwise comparisons. Significance levels: *p < 0.05, **p < 0.01, ***p < 0.001, ****p < 0.0001; ns, not significant

Table 4 Test–retest reliability analysis of task-related responses

			Pairwise comparisons								
		ICC	MRTOOL - UNIF.SEG.			DARTEL - UNIF.SEG.			DARTEL - MRTOOL		
		Mean ± sd	*Effect size (d)*	*D*	*p*	*Effect size (d)*	*D*	*p*	*Effect size (d)*	*D*	*p*
Finger	UNIF.SEG.	0.43 ± 0.21	0.20	0.12	****	0.22	0.14	****	0.04	0.05	ns
	MRTOOL	0.49 ± 0.22									
	DARTEL	0.50 ± 0.24									
Foot	UNIF.SEG.	0.39 ± 0.21	0.15	0.12	****	0.21	0.15	****	0.06	0.06	*
	MRTOOL	0.44 ± 0.22									
	DARTEL	0.46 ± 0.24									
Lips	UNIF.SEG.	0.33 ± 0.22	0.20	0.13	****	0.22	0.15	****	0.02	0.05	ns
	MRTOOL	0.41 ± 0.24									
	DARTEL	0.41 ± 0.23									
Verb gen	UNIF.SEG.	0.50 ± 0.24	0.09	0.08	***	0.03	0.09	***	0.05	0.05	ns
	MRTOOL	0.53 ± 0.24									
	DARTEL	0.51 ± 0.27									
Word rep	UNIF.SEG.	0.52 ± 0.21	0.22	0.15	****	0.33	0.26	****	0.13	0.15	****
	MRTOOL	0.58 ± 0.19									
	DARTEL	0.62 ± 0.21									

Means and standard deviations of the Intraclass Correlation Coefficient (ICC) computed over each task-specific contrast map. ICC values range from 0 to 1, where 1 indicates complete correlation of the subject voxel intensities between the Test and Retest sessions. Cohen's effect size (d) is reported to allow a direct comparison between the mean ICC values across registration methods. The Kolmogorov–Smirnov (K-S) test statistic (D), measuring the largest distance between two empirical ICC distributions, and the significance level (p) are reported. Significance levels: *$p < 0.05$, ***$p < 0.001$, ****$p < 0.0001$; ns, not significant

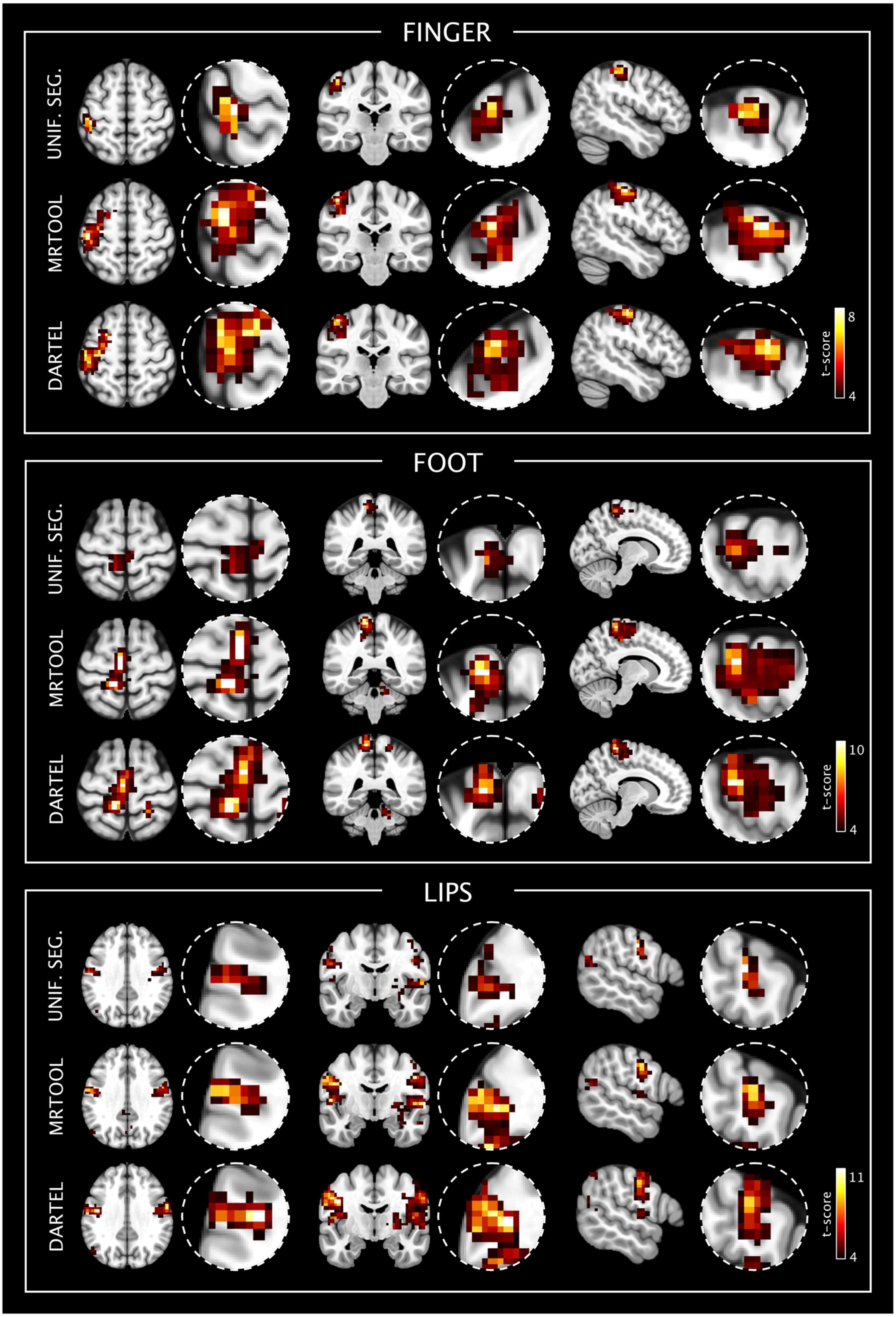
transforms an individual's structural image to a standard-space template—can impact the reliability of task-related BOLD activity.

Between-subjects reliability analysis

Our findings suggest that anatomical registration is a crucial step in the data-processing routine that contributes significantly to the total inter-subject variance of the resulting activation maps (Figs. 4 and 5). During normalization of the anatomical T1-w image to the standard SPM12 MNI template, a decrease in registration performance (Fig. 2) is directly linked to a reduction in the Dice overlap (DSC) between thresholded fMRI maps (Table 3). According to the present results, and in line with our previous study (Ganzetti et al. 2018), a key factor in the superior performance of MRTOOL compared with Unified Segmentation and DARTEL may be the explicit skull-stripping of the structural image. In Unified Segmentation and DARTEL, implicit brain extraction is executed according to a probabilistic weighting of non-brain structures. That is, in the Unified Segmentation registration framework, the GM, WM, and cerebrospinal fluid TPMs are used to segment the brain, and three supplementary non-brain tissue maps (i.e. bone, fat and air) are used to improve the implicit skull-stripping. In line with other studies, we found a clear benefit of skull-stripped images on the quality of spatial registration (Acosta-Cabronero et al. 2008; Fein et al. 2006; Fischmeister et al. 2013).

Our results demonstrate that the use of different spatial normalization approaches can lead to dissimilar results. However, it is worth noting that the difference between MRTOOL and the other two methods can be appreciated more easily in cases where the brain anatomy deviates markedly from the standard template image. In contrast to our previous validation of MRTOOL, MR data in the present study were collected from healthy volunteers characterized by normal anatomy and limited age range (50–58 years).

DARTEL uses the TPM generated by the Unified Segmentation module as input, yet produces the best registration output. We believe that this is because of the multilevel registration process in DARTEL. Instead of producing a non-linear transformation that directly matches the individual brain to a standard template in MNI space, DARTEL estimates the deformations that best align the subjects by iteratively registering the individual brains with the group average. After several iterations, a population-specific template is generated that is then registered to the standard template, permitting the transformations to be combined so that all the individual spatially normalized brains can be transformed to MNI space. In this way, the residual inter-subject variability not accounted for in Unified Segmentation is reduced. In line with the works of Crinion et al. (2007) and Fischmeister et al. (2013), who



also investigated language tasks, our results show that the temporal area is a substantial source of structural variability not completely accounted for during the normalization procedure. Indeed, the superior performance of DARTEL in this

◀ **Fig. 4 Group-level cortical activations: motor tasks.** Statistical maps of the BOLD response averaged across sessions for three motor conditions: finger tapping (top panel), foot flexion (middle panel), and lip pursing (bottom panel). The effect of anatomical registration on group-level activations is illustrated for the Unified Segmentation, MRTOOL, and DARTEL algorithms. We report thresholded values calculated using a cluster-based significance criterion to control for multiple comparisons (false discovery rate: primary threshold $p < 0.001$, cluster-extent threshold $p < 0.05$)

region compared with the other methods suggests that there may be an increased level of inter-subject variability in the temporal lobe that is not fully compensated in the other methods.

According to our group-level analysis (Fig. 5), MRTOOL produced superior results at the subcortical level. Compared with Unified Segmentation, MRTOOL produced significant activation clusters in the left thalamus in the verb-generation task and in the anterior portion of the left putamen in the word-repetition task. We believe that the superiority of MRTOOL in this case is mainly due to the improved registration in the region of the lateral ventricles. Indeed, one of the advantages of MRTOOL over the other methods is the use of a specific registration step to address the anatomical variability in the periventricular WM. As shown in our previous work (Ganzetti et al. 2018), spatial alignment around the ventricles is considerably improved with MRTOOL.

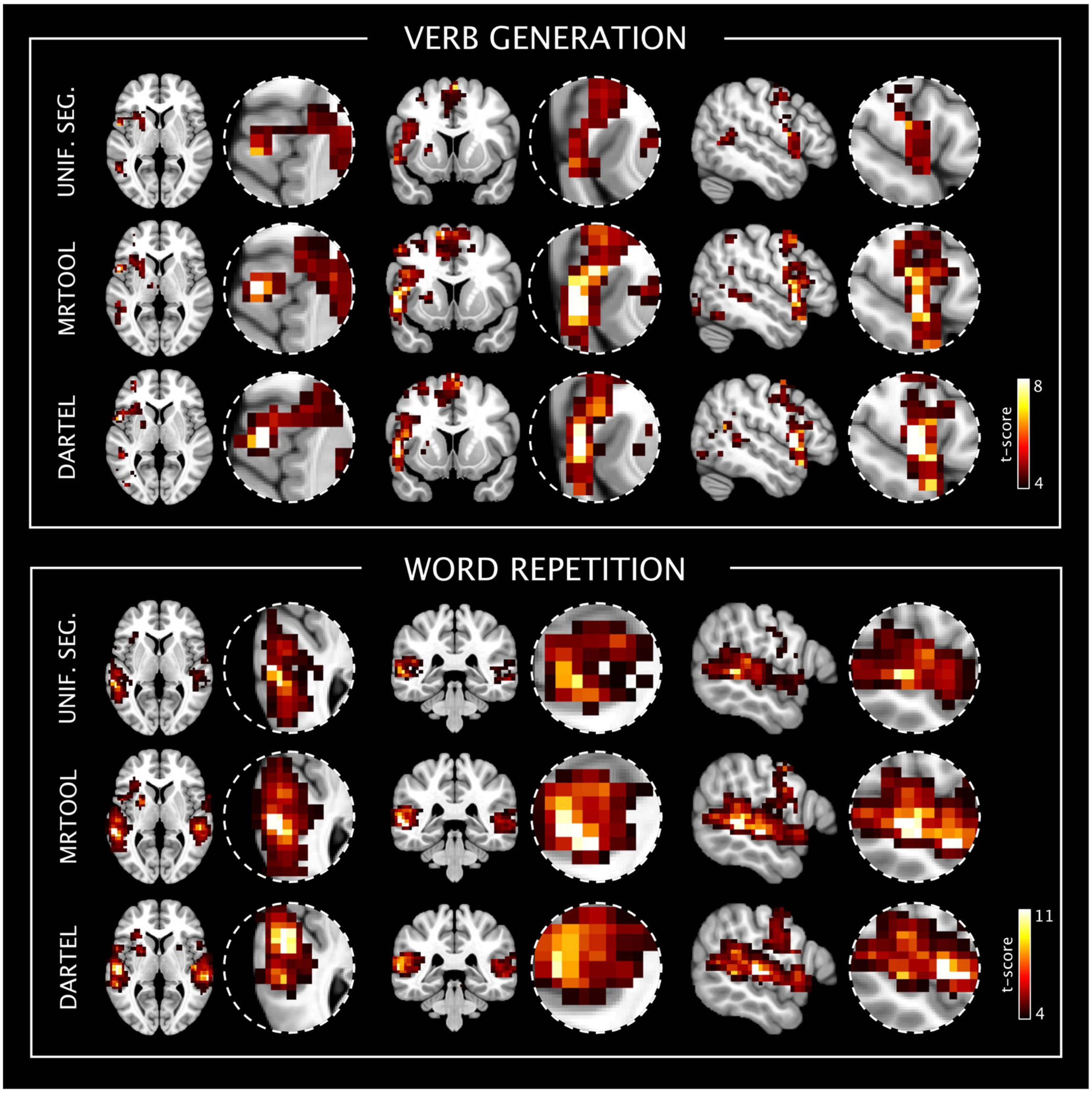


Fig. 5 Group-level cortical activations: language tasks. Statistical maps of the BOLD response averaged across sessions for two language conditions: covert verb generation (top panel) and overt word repetition (bottom panel). The effect of anatomical registration on group-level activations is illustrated for the Unified Segmentation, MRTOOL, and DARTEL algorithms. We report thresholded values calculated using a cluster-based significance criterion to control for multiple comparisons (false discovery rate: primary threshold $p < 0.001$, cluster-extent threshold $p < 0.05$)

Table 5 Group-level cortical activations: peak analysis

				Pairwise comparisons											
		Activation peak			MRTOOL-UNIF.SEG.				DARTEL-UNIF.SEG.				DARTEL-MRTOOL		
			Mean ± sd (mm)		*Estimate [C.I.]*	*Effect size (d)*	*p*		*Estimate [C.I.]*	*Effect size (d)*	*p*		*Estimate [C.I.]*	*Effect size (d)*	*p*
Finger	UNIF.SEG.	*x* *y* *z*	−36.2 ± 4.1 −17.4 ± 5.9 50.2 ± 5.5	*x*	−4.6 [−9.18, −0.02]	0.92	*	*x*	−3.5 [−5.72, −1.28]	0.79	**	*x*	1.1 [−3.17, 5.37]	0.22	ns
	MRTOOL	*x* *y* *z*	−40.8 ± 5.2 −21.8 ± 3.6 52.3 ± 2.5	*y*	−4.4 [−7.69, −1.11]	0.86	*	*y*	−3 [−7.59, 1.59]	0.52	ns	*y*	1.4 [−2.04, 4.84]	0.3	ns
	DARTEL	*x* *y* *z*	−39.7 ± 4.2 −20.4 ± 5.1 53.6 ± 5.1	*z*	2.1 [−0.43, 6.44]	0.66	ns	*z*	3.4 [−1.28, 8.08]	0.61	ns	*z*	1.3 [−3.10, 3.91]	0.09	ns
Foot	UNIF.SEG.	*x* *y* *z*	−6 ± 4.7 −34.6 ± 9.7 66.2 ± 3.16	*x*	−4.02 [−7.86, −0.54]	1.06	*	*x*	−4 [−7.93, −0.07]	0.96	*	*x*	0 [−1.37, 1.77]	0.06	ns
	MRTOOL	*x* *y* *z*	−10 ± 2.4 −31.9 ± 4.9 70.4 ± 1.7	*y*	2.7 [−5.7, 11.11]	0.33	ns	*y*	5.6 [1.86, 9.34]	0.62	**	*y*	2.9 [−3.53, 9.33]	0.43	ns
	DARTEL	*x* *y* *z*	−10 ± 2.9 −29 ± 7.3 73.2 ± 4.1	*z*	4.2 [1.38, 7.02]	1.56	**	*z*	7 [4.19, 9.81]	1.81	**	*z*	2.8 [0.01, 5.60]	0.84	*
Lips	UNIF.SEG.	*x* *y* *z*	−47.3 ± 8.6 −12.3 ± 2.9 31.9 ± 5.1	*x*	−5.1 [−11.24, 1.14]	0.68	ns	*x*	−4.5 [−10.66, 1.76]	0.63	ns	*x*	0.6 [−0.37, 1.57]	0.12	ns
	MRTOOL	*x* *y* *z*	−52.4 ± 4.9 −9.8 ± 4.1 34.6 ± 1.8	*y*	2.5 [0.86, 4.24]	0.68	**	*y*	2.9 [0.81, 5.08]	1.04	*	*y*	0.4 [−2.19, 2.99]	0.11	ns
	DARTEL	*x* *y* *z*	−51.8 ± 3.8 −9.4 ± 2.4 35.4 ± 2.4	*z*	2.7 [−1.37, 6.67]	0.66	ns	*z*	3.5 [−0.26, 7.17]	0.83	ns	*z*	0.8 [−1.74, 3.34]	0.36	ns
Verb gen	UNIF.SEG.	*x* *y* *z*	−44.7 ± 5.06 3.2 ± 7.5 13 ± 8.5	*x*	−6.7 [−11.9, −1.47]	1.34	*	*x*	−7.7 [−12.7, −2.65]	1.4	**	*x*	−1 [−4.77, 2.77]	0.19	ns
	MRTOOL	*x* *y* *z*	−51.4 ± 4.1 7.2 ± 2.9 18.4 ± 8.1	*y*	4 [−0.27, 8.17]	0.65	ns	*y*	3.6 [−1.26, 8.36]	0.58	ns	*y*	−0.4 [−3.62, 2.82]	0.12	ns
	DARTEL	*x* *y* *z*	−52.4 ± 5.3 6.8 ± 3.1 20 ± 9.9	*z*	5.4 [−0.02, 10.8]	0.62	*	*z*	7 [0.38, 13.61]	0.72	*	*z*	2.6 [−0.99, 4.19]	0.17	ns

Table 5 (continued)

			Activation peak	Pairwise comparisons									
					MRTOOL-UNIF.SEG.			DARTEL-UNIF.SEG.			DARTEL-MRTOOL		
			Mean ± sd (mm)		*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*	*Estimate [C.I.]*	*Effect size (d)*	*p*
Word rep	UNIF.SEG.	*x*	−53.4 ± 5.2	*x*	−3 [−6.94, 0.94]	0.67	ns	−4 [−6.52, −1.48]	0.84	**	−1 [−3.26, 1.26]	0.28	ns
		y	−13.2 ± 7.2										
		z	0.6 ± 2.8										
	MRTOOL	*x*	−56.4 ± 2.9	*y*	−2.8 [−8.32, 2.72]	0.44	ns	−2.2 [−7.17, 2.77]	0.35	ns	0.6 [−2.56, 3.76]	0.12	ns
		y	−16 ± 4.3										
		z	4.4 ± 3.8										
	DARTEL	*x*	−57.4 ± 3.6	*z*	3.8 [0.46, 7.13]	1.06	*	4.8 [2.01, 7.59]	1.69	**	1 [−1.81, 3.80]	0.29	ns
		y	−15.4 ± 4.5										
		z	5.4 ± 2.5										

Means and standard deviations for the group-level coordinate values of the activation peaks for each task. For each subject, 3-D coordinates were extracted from the mean contrast map computed by averaging across the Test and Retest sessions. Average coordinate values across subjects are reported for each separate method together with the respective pairwise comparisons. Significance levels: *p < 0.05, **p < 0.01; ns, not significant

Within-subject reliability analysis

The within-subject overlap of thresholded activation maps (Table 2) was greater than the between-subjects overlap (Table 3) for each task. This was the main motivation behind our decision to average across sessions prior to the group-level analysis. Given the lower within-subject tissue overlap with Unified Segmentation than with MRTOOL or DARTEL (Fig. 3), it is reasonable to expect a similar trend for the corresponding volume overlap in the task-related analysis. However, we did not find any statistical differences between the methods, except between DARTEL and Unified Segmentation during the verb-generation task. A possible explanation is that the spatial resolution of the fMRI acquisition may not have been high enough to capture the small within-subject variations introduced by different registration methods. It is worth noting that, despite the significant differences between methods, the Dice overlap computed for the assessment of GM and WM segmentations (Fig. 3) yielded relatively high DSC values (range 0.89–0.94 and 0.91–0.96 for GM and WM, respectively). Therefore, in the within-subject case, it is conceivable that thresholded task-related activations, which focus on local activation patterns, may not be affected by the registration process.

However, with the test–retest reliability analysis of the functional maps, we confirmed that anatomical registration does have a significant impact on the reliability of task-related responses in our study (Table 4). Since the ICC metric combines both between-subjects and between-sessions variance, the same ICC value can be due to either high between-subjects variance and low between-sessions variance, or low between-subjects variance and high between-sessions variance. However, Gorgolewski et al. (2013b) and Wei et al. (2004) suggested that the ICC is more heavily influenced by between-subjects variance than between-sessions variance. Indeed, the results of our test–retest reliability analysis of task-related responses (Table 4) are more in line with the between-subjects DSC results than the within-subject DSC results (Tables 3 and 4).

Methodological considerations

Our decision to use Dice overlap as a reliability metric was mostly driven by its utility for analysis of thresholded maps. Thresholded maps are generally considered the reference output in a standard fMRI analysis. Not only are the conclusions from group studies drawn from such maps, but neurosurgeons use single-subject thresholded maps to design surgical procedures (Wengenroth et al. 2011). We acknowledge the limitation raised by Smith et al. (2005) that thresholded single-session maps do not provide accurate information about inter-session variability. However, we emphasize that our study focused on the differences between registration methods. Therefore, given that we processed and evaluated the same group of individuals equivalently with all methods, the between-session variability can be considered a constant term in our analysis. Hence, we can conclude that this constraint does not affect the interpretation of our results.

Furthermore, to minimize the effect of between-session variability in the fMRI time series at the single-subject level, we implemented an adaptive thresholding method instead of the classic fixed-threshold strategy. We used the cluster-forming threshold approach described in Gorgolewski et al. (2012). In fMRI, global effects and nuisance variation are common sources of noise in the BOLD signal (Murphy et al. 2009), causing shifts of the overall t-distribution around zero. Adaptive thresholding can correct for null-distribution imperfections by shifting the mean of the Gaussian component in the mixture model. The ability to adapt to different levels of noise translates to superior between-session reliability, and thus greater between-session overlap of thresholded maps.

It could be argued that the low spatial resolution of our fMRI readout may be not sufficient to capture small differences introduced by differences in spatial registration, leading to a limitation in the interpretation of our results. However, the relative performance of the three methods on high-resolution structural MR data (i.e. the anatomical registration; Figs 2 and 3, Table 1) is expected to translate to comparable performance at the functional level; therefore, we would expect to observe similar results even with higher-resolution fMRI acquisition.

Conclusion

At present, BOLD-fMRI is one of the most useful methods available for studying human brain function (Rosen and Savoy 2012). Many fMRI studies rely on the use of large cohorts of participants to make robust inferences about brain structure–function relationships (Boisgontier et al. 2018; Pauwels et al. 2015; Solesio-Jofre et al. 2018). However, such large cohorts can include participants with a wide range of ages, and thus different structural properties of the brain. In this context, the availability of accurate tools for spatial registration of fMRI data is important to ensure the reliability of the reported findings. Recently, the use of fMRI for clinical purposes has increased (Detre 2006). Although neuroscience studies often rely on group analyses, clinical applications focus on a single patient (de Bertoldi et al. 2015; Dubois and Adolphs 2016). Indeed, being able to accurately associate a specific functional activation to the underlying anatomical structure is crucial for diagnosis (Demirci and Calhoun 2009; Nahab and Hallett 2010) and rehabilitation of neurological disorders (Dong et al. 2006; Laatsch et al. 2004). Our results suggest that MRTOOL permits the mapping of brain activity with greater accuracy and reliability than Unified Segmentation. Furthermore, MRTOOL was developed for single-subject applications (in which an individual brain is directly registered to a reference template), yet its performance is comparable to that achieved by DARTEL (a multilevel registration approach that iteratively registers individual brains with the group average).

Given its effectiveness in the presence of brain atrophy (Ganzetti et al. 2018), which occurs in normal aging as well as in a number of pathologies, MRTOOL may be particularly suitable for single-subject clinical purposes.

Information sharing statement

All the subjects included in this study were extracted from the open access database OpenfMRI (Poldrack and Gorgolewski 2017). The dataset was originally acquired to validate fMRI tasks used in pre-surgical planning for tumor resection. This study was approved by the local South East Scotland Research Ethics Committee, and informed consent was obtained from all participants. Both the Unified Segmentation algorithm and the DARTEL toolbox are implemented within the SPM12 software package (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/). Statistical analyses were conducted in R (https://www.r-project.org/). The MRTOOL normalization module is freely available and integrated in our software toolbox MRTOOL (http://www.nitrc.org/projects/mrtool/). We hope that our work will contribute to a more widespread use of robust and automated approaches for large-scale analyses of MR data.

Funding The study received funding from the FWO and European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie action (grant 665501) and the KU Leuven Special Research Fund (grant C16/15/070).

Compliance with ethical standards

Conflict of interest Author Marco Ganzetti declares that he has no conflict of interest. Author Gaia Amaranta Taberna declares that she has no conflict of interest. Author Dante Mantini declares that he has no conflict of interest.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This article does not contain any studies with animals performed by any of the authors.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

Acosta-Cabronero, J., Williams, G. B., Pereira, J. M. S., Pengas, G., & Nestor, P. J. (2008). The impact of skull-stripping and radio-frequency bins correction on grey-matter segmentation for voxel-based morphometry. *Neuroimage, 39*(4), 1654–1665. https://doi.org/10.1016/j.neuroimage.2007.10.051.

Amunts, K., Schleicher, A., Bürgel, U., Mohlberg, H., Uylings, H. B. M., & Zilles, K. (1999). Broca's region revisited: Cytoarchitecture and intersubject variability. *Journal of Comparative Neurology, 412*(2), 319–341. https://doi.org/10.1002/(SICI)1096-9861(19990920)412:2<319::AID-CNE10>3.0.CO;2-7.

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage, 38*(1), 95–113. https://doi.org/10.1016/j.neuroimage.2007.07.007.

Ashburner, J., & Friston, K. J. (2005). Unified segmentation. *Neuroimage, 26*(3), 839–851. https://doi.org/10.1016/j.neuroimage.2005.02.018.

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Annals of the New York Academy of Sciences, 1191*, 133–155. https://doi.org/10.1111/j.1749-6632.2010.05446.x.

Boisgontier, M. P., Cheval, B., van Ruitenbeek, P., Cuypers, K., Leunissen, I., Sunaert, S., Meesen, R., Zivari Adab, H., Renaud, O., & Swinnen, S. P. (2018). Cerebellar gray matter explains bimanual coordination performance in children and older adults. *Neurobiology of Aging, 65*, 109–120. https://doi.org/10.1016/j.neurobiolaging.2018.01.016.

Bullmore, E. (2012). The future of functional MRI in clinical medicine. *NeuroImage, 62*, 1267–1271. https://doi.org/10.1016/j.neuroimage.2012.01.026.

Caceres, A., Hall, D. L., Zelaya, F. O., Williams, S. C. R., & Mehta, M. A. (2009). Measuring fMRI reliability with the intra-class correlation coefficient. *NeuroImage, 45*, 758–768. https://doi.org/10.1016/j.neuroimage.2008.12.035.

Caspers, S., Geyer, S., Schleicher, A., Mohlberg, H., Amunts, K., & Zilles, K. (2006). The human inferior parietal cortex: Cytoarchitectonic parcellation and interindividual variability. *NeuroImage, 33*(2), 430–448. https://doi.org/10.1016/j.neuroimage.2006.06.054.

Chumbley, J. R., & Friston, K. J. (2009). False discovery rate revisited: FDR and topological inference using Gaussian random fields. *Neuroimage, 44*(1), 62–70. https://doi.org/10.1016/j.neuroimage.2008.05.021.

Crinion, J., Ashbumer, J., Leff, A., Brett, M., Price, C., & Friston, K. J. (2007). Spatial normalization of lesioned brains: Performance evaluation and impact on fMRI analyses. *Neuroimage, 37*(3), 866–875. https://doi.org/10.1016/j.neuroimage.2007.04.065.

Crivello, F., Schormann, T., Tzourio-Mazoyer, N., Roland, P. E., Zilles, K., & Mazoyer, B. M. (2002). Comparison of spatial normalization procedures and their impact on functional maps. *Human Brain Mapping, 16*(4), 228–250. https://doi.org/10.1002/hbm.10047.

de Bertoldi, F., Finos, L., Maieron, M., Weis, L., Campanella, M., Ius, T., & Fadiga, L. (2015). Improving the reliability of single-subject fMRI by weighting intra-run variability. *NeuroImage, 114*, 287–293. https://doi.org/10.1016/j.neuroimage.2015.03.076.

Demirci, O., & Calhoun, V. D. (2009). Functional magnetic resonance imaging - implications for detection of schizophrenia. *European Neurological Review, 4*(2), 103–106. https://doi.org/10.17925/ENR.2009.04.02.103.

Detre, J. A. (2006). Clinical applicability of functional MRI. *Journal of Magnetic Resonance Imaging, 23*, 808–815. https://doi.org/10.1002/jmri.20585.

Dong, Y., Dobkin, B. H., Cen, S. Y., Wu, A. D., & Winstein, C. J. (2006). Motor cortex activation during treatment may predict therapeutic gains in paretic hand function after stroke. *Stroke, 37*(6), 1552–1555. https://doi.org/10.1161/01.STR.0000221281.69373.4e.

Dubois, J., & Adolphs, R. (2016). Building a science of individual differences from fMRI. *Trends in Cognitive Sciences, 20*, 425–443. https://doi.org/10.1016/j.tics.2016.03.014.

Eickhoff, S. B., Stephan, K. E., Mohlberg, H., Grefkes, C., Fink, G. R., Amunts, K., & Zilles, K. (2005). A new SPM toolbox for combining probabilistic cytoarchitectonic maps and functional imaging data. *NeuroImage, 25*(4), 1325–1335. https://doi.org/10.1016/j.neuroimage.2004.12.034.

Eickhoff, S. B., Heim, S., Zilles, K., & Amunts, K. (2006). Testing anatomically specified hypotheses in functional imaging using cytoarchitectonic maps. *NeuroImage, 32*(2), 570–582. https://doi.org/10.1016/j.neuroimage.2006.04.204.

Eickhoff, S. B., Paus, T., Caspers, S., Grosbras, M. H., Evans, A. C., Zilles, K., & Amunts, K. (2007). Assignment of functional activations to probabilistic cytoarchitectonic areas revisited. *NeuroImage, 36*(3), 511–521. https://doi.org/10.1016/j.neuroimage.2007.03.060.

Fein, G., Landman, B., Tran, H., Barakos, J., Moon, K., Di Sclafani, V., & Shumway, R. (2006). Statistical parametric mapping of brain morphology: Sensitivity is dramatically increased by using brain-extracted images as inputs. *Neuroimage, 30*(4), 1187–1195. https://doi.org/10.1016/j.neuroimage.2005.10.054.

Fischmeister, F. P. S., Hollinger, I., Klinger, N., Geissler, A., Wurnig, M. C., Matt, E., et al. (2013). The benefits of skull stripping in the normalization of clinical fMRI data. *Neuroimage-Clinical, 3*, 369–380. https://doi.org/10.1016/j.nicl.2013.09.007.

Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. P., Frith, C. D., & Frackowiak, R. S. J. (1994). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping, 2*(4), 193–1097. https://doi.org/10.1002/hbm.460020402.

Ganzetti, M., Liu, Q., & Mantini, D. (2018). A spatial registration toolbox for structural MR imaging of the aging brain. *Neuroinformatics, pp., 16*, 1–13. https://doi.org/10.1007/s12021-018-9355-3.

Geyer, S., Ledberg, A., Schleicher, A., Kinomura, S., Schormann, T., Burgel, U., et al. (1996). Two different areas within the primary motor cortex of man. *Nature, 382*(6594), 805–807. https://doi.org/10.1038/382805a0.

Giussani, C., Roux, F. E., Ojemann, J., Sganzerla, E. P., & Pirillo, D. (2010). Is preoperative functional magnetic resonance imaging reliable for language areas mapping in brain tumor surgery? Review of language functional magnetic resonance imaging and direct cortical stimulation correlation studies. Neurosurgery. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=358341166.

Gorgolewski, K., Storkey, A. J., Bastin, M. E., & Pernet, C. (2012). Adaptive thresholding for reliable topological inference in single subject fMRI analysis. *Frontiers in Human Neuroscience, 6*. https://doi.org/10.3389/fnhum.2012.00245.

Gorgolewski, K., Storkey, A. J., Bastin, M. E., Whittle, I., & Pernet, C. (2013a). Single subject fMRI test-retest reliability metrics and confounding factors. *Neuroimage, 69*, 231–243. https://doi.org/10.1016/j.neuroimage.2012.10.085.

Gorgolewski, K., Storkey, A., Bastin, M. E., Whittle, I. R., Wardlaw, J. M., & Pernet, C. R. (2013b). A test-retest fMRI dataset for motor, language and spatial attention functions. *Gigascience, 2*, Artn 6. https://doi.org/10.1186/2047-217x-2-6.

Hoffman, P., & Lambon Ralph, M. A. (2018). From percept to concept in the ventral temporal lobes: Graded hemispheric specialisation based on stimulus and task. *Cortex, 101*, 107–118. https://doi.org/10.1016/j.cortex.2018.01.015.

Hope, T. M. H., Jones, O. P., Grogan, A., Crinion, J., Rae, J., Ruffle, L., et al. (2015). Comparing language outcomes in monolingual and bilingual stroke patients. *Brain, 138*(4), 1070–1083. https://doi.org/10.1093/brain/awv020.

Klein, A., Andersson, J., Ardekani, B. A., Ashburner, J., Avants, B., Chiang, M. C., Christensen, G. E., Collins, D. L., Gee, J., Hellier, P., Song, J. H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R. P., Mann, J. J., & Parsey, R. V. (2009). Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *Neuroimage, 46*(3), 786–802. https://doi.org/10.1016/j.neuroimage.2008.12.037.

Klöppel, S., Stonnington, C. M., Chu, C., Draganski, B., Scahill, R. I., Rohrer, J. D., et al. (2008). Automatic classification of MR scans in

Alzheimer's disease. *Brain, 131*(3), 681–689. https://doi.org/10.1093/brain/awm319.

Laatsch, L. K., Thulborn, K. R., Krisky, C. M., Shobat, D. M., & Sweeney, J. A. (2004). Investigating the neurobiological basis of cognitive rehabilitation therapy with fMRI. *Brain Injury, 18*(10), 957–974. https://doi.org/10.1080/02699050410001672369.

Lorio, S., Kherif, F., Ruef, A., Melie-Garcia, L., Frackowiak, R., Ashburner, J., Helms, G., Lutti, A., & Draganski, B. (2016). Neurobiological origin of spurious brain morphological changes: A quantitative MRI study. *Human Brain Mapping, 37*(5), 1801–1815. https://doi.org/10.1002/hbm.23137.

Michely, J., Volz, L. J., Hoffstaedter, F., Tittgemeyer, M., Eickhoff, S. B., Fink, G. R., & Grefkes, C. (2018). Network connectivity of motor control in the ageing brain. *NeuroImage: Clinical, 18*, 443–455. https://doi.org/10.1016/j.nicl.2018.02.001.

Morosan, P., Rademacher, J., Schleicher, A., Amunts, K., Schormann, T., & Zilles, K. (2001). Human primary auditory cortex: Cytoarchitectonic subdivisions and mapping into a spatial reference system. *NeuroImage, 13*(4), 684–701. https://doi.org/10.1006/nimg.2000.0715.

Morosan, P., Schleicher, A., Amunts, K., & Zilles, K. (2005). Multimodal architectonic mapping of human superior temporal gyrus. In *Anatomy and Embryology* (Vol. 210, pp. 401–406). https://doi.org/10.1007/s00429-005-0029-1.

Murphy, K., Birn, R. M., Handwerker, D. A., Jones, T. B., & Bandettini, P. A. (2009). The impact of global signal regression on resting state correlations: Are anti-correlated networks introduced? *NeuroImage, 44*(3), 893–905. https://doi.org/10.1016/j.neuroimage.2008.09.036.

Nahab, F. B., & Hallett, M. (2010). Current role of functional MRI in the diagnosis of movement disorders. *Neuroimaging Clinics of North America, 20*, 103–110. https://doi.org/10.1016/j.nic.2009.08.001.

Ossenkoppele, R., Mattsson, N., Teunissen, C. E., Barkhof, F., Pijnenburg, Y., Scheltens, P., van der Flier, W. M., & Rabinovici, G. D. (2015). Cerebrospinal fluid biomarkers and cerebral atrophy in distinct clinical variants of probable Alzheimer's disease. *Neurobiology of Aging, 36*(8), 2340–2347. https://doi.org/10.1016/j.neurobiolaging.2015.04.011.

Pauwels, L., Vancleef, K., Swinnen, S. P., & Beets, I. A. M. (2015). Challenge to promote change: Both young and older adults benefit from contextual interference. *Frontiers in Aging Neuroscience, 7*(JUL). https://doi.org/10.3389/fnagi.2015.00157.

Pohl, K. M., Fisher, J., Levitt, J. J., Shenton, M. E., Kikinis, R., Grimson, W. E. L., & Wells, W. M. (2005). A unifying approach to registration, segmentation, and intensity correction. *Medical Image Computing and Computer-Assisted Intervention - Miccai 2005, Pt 1, 3749*, 310–318.

Poldrack, R. A., & Gorgolewski, K. J. (2017). OpenfMRI: Open sharing of task fMRI data. *Neuroimage, 144*, 259–261. https://doi.org/10.1016/j.neuroimage.2015.05.073.

Rosen, B. R., & Savoy, R. L. (2012). FMRI at 20: Has it changed the world? *NeuroImage, 62*, 1316–1324. https://doi.org/10.1016/j.neuroimage.2012.03.004.

Savoy, R. L. (2005). Experimental design in brain activation MRI: Cautionary tales. In *Brain Research Bulletin* (Vol. 67, pp. 361–367). https://doi.org/10.1016/j.brainresbull.2005.06.008.

Shattuck, D. W., Sandor-Leahy, S. R., Schaper, K. A., Rottenberg, D. A., & Leahy, R. M. (2001). Magnetic resonance image tissue classification using a partial volume model. *Neuroimage, 13*(5), 856–876. https://doi.org/10.1006/nimg.2000.0730.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428. https://doi.org/10.1037/0033-2909.86.2.420.

Smith, S. M., Beckmann, C. F., Ramnani, N., Woolrich, M. W., Bannister, P. R., Jenkinson, M., Matthews, P. M., & McGonigle, D. J. (2005). Variability in fMRI: A re-examination of inter-session differences. *Human Brain Mapping, 24*(3), 248–257. https://doi.org/10.1002/hbm.20080.

Solesio-Jofre, E., Beets, I. A. M., Woolley, D. G., Pauwels, L., Chalavi, S., Mantini, D., & Swinnen, S. P. (2018). Age-dependent modulations of resting state connectivity following motor practice. *Frontiers in Aging Neuroscience, 10*(FEB). https://doi.org/10.3389/fnagi.2018.00025.

Stephens, M. A. (1992). Introduction to Kolmogorov (1933) on the empirical determination of a distribution. In *Breakthroughs in Statistics: Methodology and Distribution*. https://doi.org/10.1007/978-1-4612-4380-9_9.

Uludağ, K., Uğurbil, K., & Berliner, L. (2015). fMRI: From nuclear spins to brain functions. fMRI: From nuclear spins to brain functions. https://doi.org/10.1007/978-1-4899-7591-1.

Van Leemput, K., Maes, F., Vandermeulen, D., & Suetens, P. (1999). Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging, 18*(10), 897–908. https://doi.org/10.1109/42.811270.

Wei, X., Yoo, S. S., Dickey, C. C., Zou, K. H., Guttmann, C. R., & Panych, L. P. (2004). Functional MRI of auditory verbal working memory: Long-term reproducibility analysis. *NeuroImage, 21*, 1000–1008. https://doi.org/10.1016/j.neuroimage.2003.10.039.

Wengenroth, M., Blatow, M., Guenther, J., Akbar, M., Tronnier, V. M., & Stippich, C. (2011). Diagnostic benefits of presurgical fMRI in patients with brain tumours in the primary sensorimotor cortex. *European Radiology, 21*(7), 1517–1525. https://doi.org/10.1007/s00330-011-2067-9.

Woo, C. W., Krishnan, A., & Wager, T. D. (2014). Cluster-extent based thresholding in fMRI analyses: Pitfalls and recommendations. *Neuroimage, 91*, 412–419. https://doi.org/10.1016/j.neuroimage.2013.12.058.