Optimization and comparative evaluation of nonlinear deformation algorithms for atlas-based segmentation of DBS target nuclei

Siobhan Ewert a,b, Andreas Horn a, Francisca Finkel b,c, Ningfei Li a,d, Andrea A. Kühna, Todd M. Herrington b,*

a Charité – University Medicine Berlin, Department of Neurology, Movement Disorders and Neuromodulation Unit, Berlin, Germany
b Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston, MA, USA
c Program in Behavioral Neuroscience, Northeastern University, Boston, MA, USA
d Institute of Software Engineering and Theoretical Computer Science, Neural Information Processing Group, Technische Universität Berlin, Germany

ARTICLE INFO

Keywords:
Manual segmentation
Automated segmentation
Atlas-based segmentation
Subcortical normalization
Spatial image deformation
DBS

ABSTRACT

Nonlinear registration of individual brain MRI scans to standard brain templates is common practice in neuroimaging and multiple registration algorithms have been developed and refined over the last 20 years. However, little has been done to quantitatively compare the available algorithms and much of that work has exclusively focused on cortical structures given their importance in the fMRI literature. In contrast, for clinical applications such as functional neurosurgery and deep brain stimulation (DBS), proper alignment of subcortical structures between template and individual space is important. This allows for atlas-based segmentations of anatomical DBS targets such as the subthalamic nucleus (STN) and internal pallidum (GPI).

Here, we systematically evaluated the performance of six modern and established algorithms on subcortical normalization and segmentation results by calculating over 11,000 nonlinear warps in over 100 subjects. For each algorithm, we evaluated its performance using T1- or T2-weighted acquisitions alone or a combination of T1-, T2- and PD-weighted acquisitions in parallel. Furthermore, we present optimized parameters for the best performing algorithms. We tested each algorithm on two datasets, a state-of-the-art MRI cohort of young subjects and a cohort of subjects age- and MR-quality-matched to a typical DBS Parkinson’s Disease cohort. Our final pipeline is able to segment DBS targets with precision comparable to manual expert segmentations in both cohorts. Although the present study focuses on the two prominent DBS targets, STN and GPI, these methods may extend to other small subcortical structures like thalamic nuclei or the nucleus accumbens.

1. Introduction

Translating single-subject imaging into common space such as the MNI space is a core concept in the field of brain mapping (Evans et al., 2012). It allows for comparison of structural and functional MRI findings across brains, cohorts and research centers. Increasingly these techniques are being applied in clinical contexts including stroke (Fox et al., 2014) and deep brain stimulation (DBS) (Horn et al., 2017a, 2017b). In the latter, subcortical atlases allow for delineation of DBS targets such as the subthalamic nucleus (STN) and internal part of the globus pallidus (GPI) that are imperfectly resolved on standard clinical MRI (Ewert et al., 2017). By transforming single-subject data into a common reference frame such as MNI space, one can leverage the considerable brain imaging data available in public repositories (Horn et al., 2017a; Klein et al., 2009; Crivello et al., 2002). Example subcortical atlases include the probabilistic ATAG atlas (Keuken et al., 2014), the 7T based “ultra-high field atlas for DBS planning” (Wang et al., 2016) and the DISTAL DBS atlas (Ewert et al., 2017). Similarly, connectomic resources such as normative structural and functional connectomes in MNI space are now available (Horn and Blankenburg, 2016; Horn et al., 2014) and have been used to explore the functional and structural networks targeted by DBS (Horn et al., 2017b, 2017a). Critical to this approach is the accurate transformation of single-subject imaging into the atlas space (or vice versa).

In the context of DBS, we often want to determine the spatial relationship between the implanted electrode and its specific target. This is also important for studies that assess electrode localizations across patients to identify spatial correlates of treatment outcomes (Horn et al., 2017a, 2017b, 2017c). The optimization of such transformations is vital so that clinicians can plan surgery and adjust the positions of the implanted electrodes to achieve optimal clinical outcomes.

* Corresponding author.
E-mail address: therrington@mgh.harvard.edu (T.M. Herrington).

https://doi.org/10.1016/j.neuroimage.2018.09.061
Received 18 April 2018; Received in revised form 16 August 2018; Accepted 21 September 2018
Available online 26 September 2018
1053-8119/© 2018 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
2017b; Neumann et al., 2017; Israel and Bergman, 2016; Welter et al., 2014; Tisch et al., 2007). Another application for transforming atlas information to single-subject imaging is the segmentation of target nuclei on preoperative MRI acquisitions for DBS surgical planning.

Such patient-to-template or template-to-patient registrations are termed “atlas-based segmentation” and are important because manual segmentations of DBS related structures is highly time consuming, requires expert anatomical knowledge (Forstmann et al., 2017a, b; Zwirner et al., 2016; Visser et al., 2016b; Chakravarty et al., 2013) and may not be straightforward on clinical MRI data given insufficient signal-to-noise ratio or resolution. Thus, the automated segmentation of STN and GP has been a field of vital and ongoing work with innovative contributions by multiple research groups (Garzón et al., 2017; Visser et al., 2016b, 2016a; Chakravarty et al., 2013; Haegelen et al., 2012; Helms et al., 2009; Lim et al., 2013; Villegas et al., 2009). In addition, many more general methods facilitating imaging co-registration have been developed over the last 20 years, primarily within the field of brain imaging (Ashburner, 2007; Ashburner and Friston, 2011; Andersson et al., 2010; Avants et al., 2010; Schönkecker et al., 2009).

Despite the many developments in this field, comparative studies evaluating the alignment of subcortical structures in image registration are lacking, perhaps due to the focus on cortical structures in support of the MRI literature (Glasser et al., 2016). In 2009, Klein and colleagues published an influential study comparing algorithm performance for cortical structures (Klein et al., 2009). However, only a few, coarsely defined subcortical structures were assessed. Moreover, since 2009 the algorithms studied have been improved substantially.

Here, we evaluated and optimized the preset parameters of six commonly used deformation algorithms. Multiple parameters were evaluated for most algorithms. Each algorithm, except for one, was evaluated for results of both mono- and multi-spectral datasets. Specifically, we assessed the performance of each algorithm using T1-or T2-weighted imaging alone versus a combination of T1-, T2-and Proton Density weighted MRI. In total, we estimated over 11,000 deformations from 103 manually labeled subject brains to template space. Results were compared to manual expert segmentations. Accuracy of normalizations was estimated in a high-quality dataset consisting of healthy subjects acquired in the Human Connectome Project. These MR scans exhibit a high signal-to-noise ratio and were acquired on specialized hardware, representing an as-good-as-it-gets example dataset for 3T data. Second, we estimated results on data from the IXI project which resembles MR image quality commonly acquired on 1.5 or 3T magnets. These MRIs represent a more typical example of what is acquired in routine clinical practice or when the focus is not on specialized structural imaging. In addition, the IXI data set includes T1, T2 and PD acquisitions, allowing us to assess the performance of multi-vs. monospectral normalizations.

The present study was done with a particular focus on the two most common DBS targets, STN and GP, though a translation to other small subcortical structures such as thalamic nuclei or the nucleus accumbens may potentially be inferred. To facilitate the broadest utility of these results to the scientific community, we have distributed all parameters and Tissue Probability Maps (TPM) and include practical information on the required computational processing times required for each algorithm presented.

2. Methods

2.1. Subjects and data

Two datasets were analyzed, a high quality 3T data set in young subjects (“Young Cohort”; YC from here on; image resolution 0.7 mm3 isometric; age range 22–35 yrs, two subjects 36 - yrs; mean age 28.7 yrs, std = 3.42 yrs; F:M = 49:24) and a data set resembling clinically acquired MRIs for DBS patients (“Pseudo-Clinical”; PC; image resolution 0.94 × 0.94 × 1 mm3; age range 55–70 yrs, mean age 62.9 yrs, std = 4.2 yrs; F:M = 17:13; for an exact list of YC- and PC-subjects please see section “Lists of segmented subjects” in supplementary material). 3T data was obtained from The Human Connectome Project (HCP) (https://db.humanconnectome.org/; WU-Minn HCP Data - 900 subjects; Fischl, 2012; Jenkinson et al., 2002; Marcus, 2018; Van Essen et al., 2012). 73 unrelated subjects with sufficient image contrast for manual segmentation were selected. For each subject, both structural image modalities, T1-weighted (T1w) and T2-weighted (T2w) images were selected. Minimal preprocessing steps had been done within the HCP standard preprocessing workflow and included correction of gradient nonlinearity and field map distortion, removal of spatial artifacts, within-subject cross-modal registration and aligning the subject’s native volume header to MNI space using a rigid body transform estimated by the FLIRT tool (FSL 5.0, Jenkinson et al., 2012). For details on the HCP preprocessing pipeline please see (Glasser et al., 2013).

To address how accurate automated segmentation algorithms would perform on clinical data (sometimes acquired on 1.5T with low signal-to-noise ratio and including older patients), a “typical clinical” dataset was obtained from the IXI project (http://brain-development.org/ixi-dataset/). A sub-cohort of 30 subjects aged between 55 and 70 years (age-matching a typical age of Parkinson’s disease (PD) patients undergoing DBS surgery), whose imaging was free of gross motion artifact were randomly selected for this study. These subjects were scanned on either 1.5T or 3T MR hardware and provided T1w, T2w and Proton-Density-weighted images. In contrast to the YC, these images were neither coregistered cross-modally within-subject nor rigidly aligned to MNI space. Cross-modal, within-subject registration was performed using SPM12.

Additional details on the quality of the data used and a list of specific subjects is available in the supplementary materials.

2.2. Manual segmentation

In total, left and right GPi and left and right STN of 103 brains were manually segmented (73 YC and 30 PC). Manual segmentations were performed following a specified and published protocol (Ewert et al., 2017) in 3DSlicer (https://www.slicer.org/) (see also “Detailed Segmentation Protocol” in supplementary material). Manual segmentations were carried out by either one of two raters (S.E. and F.F.) with overall 41% of HCP and 33% of IXI images segmented by both raters to estimate inter-rater reliability (see supplementary materials for details). The anatomical outlines of the segmented structures were defined with reference to neuroanatomical atlases and in-house acquired high resolution post-mortem 7T MRI scans that showed the target structures in detail (Eldow et al., 2018; Ding et al., 2016; Mai et al., 2015; Massey et al., 2012; Naidich et al., 2009).
2.3. Measures of inter-rater agreement

Inter-rater agreement was determined for each dataset by calculating both Dice coefficient (Dice, 1945) and the mean absolute surface distance (SD) (Wang et al., 2016) which measures the mean euclidean distance between two surfaces. All analyses were performed using MATLAB 2015b (The MathWorks, Inc., Natick, Massachusetts, United States).

2.4. Normalization and automatic segmentation

The MNI ICBM 152 LIN 2009b template was used (Fonov et al., 2011). Native subject volumes were nonlinearly deformed into template space (normalization) using established algorithms included in the Lead-DBS v2.1.0 software package (www.lead-dbs.org; (Horn and Kühn, 2015)). Using these algorithms, deformation fields that project patient volumes into template space were estimated, inverted and applied to a precise definition of target structures (STN and GPI) in the DISTAL atlas in template space (Ewert et al., 2017). This process will subsequently be referred to as “atlas-based segmentation”.

The DISTAL atlas consists of a precise manual segmentation of the subthalamic nucleus and internal pallidum. This segmentation was performed on a high definition MNI template (ICBM 2009b lin asym, Fonov et al., 2011) that is available in multiple modalities. These were used to generate an automatic pre-segmentation that was used alongside each individual spectra (T1, T2, etc.) to segment the structures of interest.

To assess the accuracy of each normalization, the atlas-based (automatic) segmentations were compared to manual segmentations of the corresponding nucleus in native space (Fig. 1). Agreement between atlas-based and manual segmentations was again assessed by calculating the Dice coefficient and mean surface distance. Nucleus volumes (in mm$^3$) for automatic and manual segmentations were correlated by correlating the sum of voxel volumes of the respective nuclei.

We practically optimized parameters and compared subcortical alignment for the following nonlinear image registration tools: the SyN (Avants et al., 2010) and BSplineSyN (Tustison, 2013) symmetric diffeomorphic image registration methods implemented in Advanced Normalization Tools 2.2.0 (http://stnava.github.io/ANTS/), three methods implemented in SPM12 (7219) software (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) (“New Segment” (Ashburner and Friston, 2005), “DARTEL” (Ashburner, 2007) and “SHOOT”, (Ashburner and Friston, 2011), FNIRT as implemented in FSL 5.0.10 (Andersson et al., 2010), and a three-step linear method that had been specifically developed to register deep brain stimulation electrodes to MNI space (Schönecker et al., 2009). Table 1 gives a summary of each method, and in the following sections we describe in more detail how we optimized the input parameters of our tested algorithms.

For ANTs we evaluated ANTs SyN (Avants et al., 2010) and ANTs BSplineSyN (Tustison, 2013), each with multiple different presets and with uni-versus multimodal datasets (T1 vs. T1 & T2 vs. T2-weighted MRI). We evaluated different degrees of gradient steps, shrinking factors, penalties for high frequency image deformations, convergence parameters, winsorizing and smoothing sigmas to produce three levels of allowable deformation field strength — termed low, mid and high variance. For the winning method (ANTS SyN with the preset “Low Variance”), we additionally evaluated another step of subcortical refine similar to the one described in (Schönecker et al., 2015). Here, as a last step after the global non-linear alignment of the whole brain, a subsequent warp was applied guided by a subcortical mask that focused on the basal ganglia. The final optimized parameters for registration of subcortical structures evaluated in the present study is the ANTs-based

Table 1

| Software | ID   | Full name | Citation | URL                |
|----------|------|-----------|----------|--------------------|
| SPM12    | 7219 | New Segment | Ashburner and Friston, 2005 | http://www.fil.ion.ucl.ac.uk/spm/software/spm12/ |
| DARTEL   |      | A fast diffeomorphic image registration algorithm. | Ashburner, 2007 |                     |
| SHOOT    |      | Diffeomorphic registration using geodesic shooting and Gauss-Newton optimization | Ashburner and Friston, 2011 |                     |
| FSL 5.0.10 | FNIRT | FSL Non-linear registration | Andersson et al., 2010 | https://fsl.fmrib.ox.ac.uk/hb/fulwiki/FSL/ANTS/ |
| ANTs 2.2.0 | SyN | Symmetric diffeomorphic image registration | Avants et al., 2010 | http://stnava.github.io/ANTS/ |
| BSplineSyN |      | Explicit B-spline regularization in diffeomorphic image registration | Tustison, 2013 | |
| Lead-DBS 2.1.0 | Linear Threestep | Linear Threestep Registration with Focus on Basal Ganglia and Brainstem | Schönecker et al., 2009 | http://www.lead-dbs.org/ |

![Fig. 1. Study workflow. A) First, native subject images for the Young (upper row) and Pseudo-Clinical (lower row) Cohort were normalized into MNI template space in which the DBS target regions STN and GPI had been precisely defined (Ewert et al., 2017). B) The deformation fields estimated during normalization were inverted and applied to the template STN and GPI. This results in atlas-based (automatic) segmentations (orange labels) in patient native space which can then be compared to manual segmentations, considered ground truth. C) Shows the overlap for one subject of each cohort (YC and PC) between atlas-based (orange labels) and manual (green labels) segmentation. D) These results were then compared to inter-rater overlap. Green and blue labels indicate manual labels from different raters in the same subject.](image-url)
SyN approach with subcortical refinement termed “Effective (Low Variance)” which is now the default normalization method of Lead-DBS software v.2.1.6 (Horn et al., 2018). Specific parameters of this method and the principal competing presets are published online (https://github.com/leaddbs/leaddbs/blob/master/ext_libs/ANTs/presets/). The mask used for the final subcortical refinement step is also supplied with Lead-DBS software and is shown in Fig. S1 relative to the T1 ICBM 2009b nlin asym template.

For the SPM methods we evaluated the accuracy of subcortical image registration using different tissue priors. The SPM methods use Tissue Probability Maps (TPMs) as priors for their image segmentation and subsequent registration with the segmented template image. Critical to this method is a proper definition of subcortical target structures in both subject and template TPM. We tested the results of two other TPMs against the original SPM TPM which was derived from the same IXI dataset (http://brain-development.org/ixi-dataset/) that formed our clinically age- and SNR-matched dataset, the PC (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/SPM12_Release_Notes.pdf). The second TPM tested was the one created by (Lorio et al., 2016). To create the third TPM, we adopted the approach described in (Lorio et al., 2016) and created a specialized template TPM with a refined definition of our subcortical target structures. Using the Lorio-template as a starting point, we used SPM New Segment to segment our target template. Subsequently, we used these segmentation as priors for SPM New Segment and our evaluated two cohorts. This TPM is currently implemented in Lead-DBS and can be downloaded at www.lead-dbs.org.

For the final, linear image registration algorithm, we used ANTs linear as registration method with the default input parameters. To improve the accuracy of this method we did, however, adapt the two-step approach described in (Schönecker et al., 2009). The first step computes a global alignment between the individual and the target brain. The second step focuses on the subcortex using a mask that focuses on the basal ganglia and brainstem (Fig. S1).

ANTS and SPM methods were the most accurate registration methods in a previous comparative study that focused on cortical structures (Klein et al., 2009). These two pipelines also allow for multispectral warps, which in theory may help compensate for deficiencies of individual MRI sequences. For example, the STN is not visible on T1, which is often the volume with best spatial resolution, so a combined warp that uses information from both T1- and T2-weighted volumes appears sensible. The best performing ANTs preset with an additional step of subcortical refine similar to the one described in (Schönecker et al., 2015) and the best performing TPM (for the SPM methods) were then subjected to detailed comparison and assessment.

### 2.5. Comparison of results

Similarity measures between manual and automatic segmentations for each method were analyzed using one-way ANOVA analyses. Pairwise multiple comparison post-hoc tests between each pair of approaches were estimated using Tukey’s Honest Significant Difference procedure (as is default in Matlab’s anova1 and multcompare functions).

![DICE COEFFICIENT ANTs COMPARISON](image)

**Fig. 2. Evaluation of different ANTs presets and datasets.** Violin plots of Dice coefficients with the black bar indicating the median Dice coefficient. The upper row shows results for the YC, while the lower row shows results for PC. For all presets three datasets with different MRI weightings were evaluated in the YC (T1; T1 & T2; T2) while four different MRI weightings were evaluated in the PC (T1; T1 & T2; T2, with T1,T2 & PD not displayed here). For ANTs presets, the variance indicates the level of image distortions allowed. ANTs presets evaluated from left to right were: 1. BSplineSyN high variance; 2. BSplineSyN mid variance; 3. BSplineSyN low variance; 4. SyN high variance; 5. SyN mid variance; 6. SyN low variance; 7. SyN low variance with the additional step of subcortical refinement. The last preset yielded the best overall outcome. Of all evaluated datasets a multi-modal non-linear deformation (T1 & T2) yielded the best outcome, especially for the winning preset no. 7 (SyN low variance with the additional step of subcortical refinement).
3. Results

For each algorithm tested, we compared the accuracy of the atlas-based automated segmentation compared to the manually segmented target by means of a Dice coefficient. We first assessed the performance of two ANTs nonlinear warping algorithms (Syn and BSplineSyN) each with different degrees of gradient steps, shrinking factors, penalties for high frequency image deformations, convergence parameters, warping and smoothing sigmas to produce three levels of allowable deformation field strength — termed low, mid and high variance. The performance of each algorithm is shown in Fig. 2 by means of violin plots showing the distribution of Dice coefficients. The best performing algorithm was ANTs Syn Low Variance with subcortical refinement, using a T1 + T2-weighted multimodal imaging data set. This ANTs Syn preset penalizes high-frequency image deformations and includes an additional final normalization step using a subcortical mask.

We turned next to the SPM methods, New Segment, SHOOT and DARTEL. The SPM methods use Tissue Probability Maps (TPMs) as priors for the image segmentation and subsequent registration with the segmented template image. Critical to this method is a proper definition of subcortical target structures in both subject and template TPM. We evaluated the performance of the original SPM TPM (which was derived from the IXI dataset), against the TPM described by Lorio (Lorio et al., 2016) and a novel TPM implemented within Lead-DBS. Fig. 3 shows the automated segmentation performance of SPM New Segment Results using each of the three TPMs. For the YC, we see a slight improvement of image registration accuracy for the Lead-DBS TPM. In contrast, in the PC the original SPM TPM performed slightly better than the Lead-DBS template, perhaps owing to it having been developed on the IXI dataset from which the PC cohort was selected. The Lead-DBS TPM was used in the subsequent main analysis.

We then compared these optimized algorithms to a subcortically focused linear coregistration and FNIRT as implemented in FSL. The comparative performance of all methods is summarized in Table 2a+b and Fig. 4a+b. Results are shown separately for YC (Fig. 4a and Table 2a) and PC (Fig. 4b and Table 2b). Of all the methods compared, SPM Segment and ANTs SyN performed best yielding an accurate segmentation with Dice coefficients typical of inter-rater reliability. Overall results for YC (median Dice) from lowest to highest are 0.56 (Linear); 0.67 (SPM SHOOT); 0.68 (SPM DARTEL); 0.70 (FNIRT); 0.73 (SPM Segment); 0.74 (ANTS Syn); 0.78 (Inter-Rater). Overall results for PC (median Dice) from lowest to highest are 0.57 (Linear); 0.57 (SPM DARTEL); 0.68 (FNIRT); 0.68 (SPM SHOOT); 0.68 (SPM Segment); 0.68 (ANTS Syn). Representative examples of the two best and two worst performing subjects for each algorithm and each data set are shown in Fig. 5a and Fig. 5b.

Fig. 5a+b and Table 3a+b show a more detailed analysis for the two best performing methods, SPM Segment and ANTs Syn. Here, in addition to the Dice coefficient, the mean surface distance (SD) as well as volume correlations between atlas-based and manual segmentation for each subject are shown. Results are displayed for STN and GPI separately. Fig. 5a and Table 3a show the results for the YC, while Fig. 5b and Table 3b show the same results for the PC. Notably, the Dice coefficients and SD achieved with both methods, SPM Segment and ANTs Syn, almost reached inter-rater level accuracy for the YC. For the PC there was no significant difference in accuracy between inter-rater and atlas-based segmentations.

Specifically, in the YC for the STN, inter-rater agreement by means of SD was not significantly better than ANTs Syn (median SD 0.41 mm vs. 0.38 mm) indicating a very high normalization accuracy for this method in the subcortical region of the STN (Fig. 5a and Table 3a). For the GPI,

Table 2

| Young Cohort (YC) | Both Nuclei | STN | GPI |
|------------------|------------|-----|-----|
| Method           | Median     | Std. | Range | Median | Std. | Range | Median | Std. | Range |
| Linear           | 0.56       | 0.21 | 0.6-0.86 | 0.48   | 0.21 | 0.6-0.81 | 0.64   | 0.18 | 0.61-0.86 |
| SPM SHOOT        | 0.67       | 0.10 | 0.26-0.83 | 0.65   | 0.11 | 0.26-0.82 | 0.67   | 0.09 | 0.34-0.83 |
| SPM DARTEL       | 0.68       | 0.08 | 0.39-0.83 | 0.68   | 0.07 | 0.39-0.83 | 0.68   | 0.06 | 0.50-0.82 |
| FNIRT            | 0.70       | 0.10 | 0.21-0.87 | 0.67   | 0.09 | 0.21-0.80 | 0.76   | 0.09 | 0.30-0.87 |
| SPM Segment      | 0.73       | 0.10 | 0.35-0.88 | 0.70   | 0.10 | 0.35-0.87 | 0.78   | 0.08 | 0.46-0.88 |
| ANTs             | 0.74       | 0.07 | 0.47-0.88 | 0.72   | 0.06 | 0.52-0.83 | 0.75   | 0.07 | 0.47-0.88 |
| Inter-Rater      | 0.78       | 0.08 | 0.54-0.88 | 0.76   | 0.09 | 0.54-0.88 | 0.80   | 0.05 | 0.70-0.88 |
SPM Segment performed slightly better than ANTs SyN showing no significant difference between inter-rater agreement by means of Dice or SD (\(p > 0.1\)). In fact, for the mean SD of GPi segmentations no method proved significantly different than inter-rater mean SD (\(p > 0.8\)). To further compare the results of atlas-based and manual segmentation in the YC, we compared the volumes of the segmented nuclei (right panels in Fig. 5a + b). For both methods and nuclei we see a small but significant correlation between manual and automated segmentation volumes.

### Table 2

| Pseudo-Clinical Cohort (PC) | Both Nuclei | STN | GPI |
|----------------------------|------------|-----|-----|
| Method                     | Median     | Std. | Range| Median     | Std. | Range| Median     | Std. | Range|
| Linear                     | 0.57       | 0.19 | 0.06-0.82| 0.48       | 0.20 | 0.06-0.75| 0.62       | 0.15 | 0.26-0.82|
| SPM DARTEL                 | 0.57       | 0.14 | 0.13-0.79| 0.55       | 0.14 | 0.13-0.79| 0.61       | 0.14 | 0.21-0.78|
| FNIRT                      | 0.58       | 0.18 | 0.06-0.82| 0.58       | 0.17 | 0.12-0.76| 0.58       | 0.19 | 0.06-0.82|
| SPM SHOOT                  | 0.59       | 0.16 | 0.02-0.81| 0.58       | 0.18 | 0.02-0.78| 0.62       | 0.13 | 0.27-0.81|
| SPM Segment                | 0.64       | 0.13 | 0.17-0.85| 0.62       | 0.12 | 0.16-0.79| 0.68       | 0.14 | 0.23-0.85|
| Inter-Rater                | 0.66       | 0.08 | 0.41-0.81| 0.66       | 0.10 | 0.41-0.81| 0.65       | 0.06 | 0.51-0.79|
| ANTs                       | 0.68       | 0.10 | 0.41-0.86| 0.67       | 0.09 | 0.41-0.82| 0.71       | 0.09 | 0.44-0.86|

### Fig. 4. a) Normalization accuracy for the Young Cohort. Dice coefficient as measurement of agreement between atlas-based segmentation and manual segmentation is shown for each method. Only the best performing dataset and preset is shown for each method. A) Results for STN and GPi combined. B) and C) show results for STN and GPi separately. For the exact Dice values please see Table 2a. Linear performed best with MRI dataset T1 & T2; SHOOT with T1; DARTEL with T1; FNIRT with T2; SPM Segment with T1 & T2 and the Lead-DBS TPM; ANTs with T1 & T2 and the protocol “Low Variance with Subcortical Refine”. b) shows the normalization results for the Pseudo-Clinical Cohort in the same format as Fig. 4a. For the exact Dice values please see Table 2b.
Atlas-based segmentations by SPM Segment and ANTs SyN are reasonable in size (mean volume STN and GPi by ANTs SyN: 148 mm³ and 406 mm³; SPM Segment: 119 mm³ and 377 mm³). For additional details of volumes resulting from atlas-based segmentations compared to volumes of manual segmentations in the YC and PC please see lower sections of Table 3a and b.

The median Dice coefficient for inter-rater agreement in the YC for subjects segmented by both raters (S.E. and F. F.) was 0.76 ± 0.09 (standard deviation) for STN and 0.80 ± 0.05 for GPi. We found no significant effect of hemisphere on inter-rater Dice coefficients (p > 0.1). Median surface distance (SD) was 0.41 mm ± 0.20 for STN and 0.45 mm ± 0.14 for GPi, corresponding to less than the width of one voxel, again without significant effect of hemisphere (p > 0.1). Inter-rater agreement was lower in the PC. Dice coefficients for the PC were 0.66 ± 0.10 STN and 0.65 ± 0.06 for GPi. Median SD was 0.64 mm ± 0.25 for STN and 0.82 mm ± 0.16 for GPi. Again, no significant effect of hemisphere was found in any of these metrics (p > 0.2).

Tables 3a and 3b presents inter-rater results in detail.

In the YC, segmentation mean volumes of GPi were 359 mm³ ± 50 (standard deviation), range 231–743 mm³) and of STN 120 mm³ ± 20 (range 68–168 mm³). For volumes of the GPi, there was no significant difference between left and right hemisphere (p = 0.49) and volumes correlated significantly between hemispheres (R = 0.605, p < 0.001). As previously reported in the literature (Shen and Gao, 2009; Nowinski et al., 2005), a significant difference between volumes of left and right STN was found (p = 0.031; mean volume for left STN = 122 mm³, std = 22, mean volume for right STN = 118 mm³, std = 18). Segmented volumes of GPi and STN within the same subject also correlated significantly (R = 0.326, p < 0.001). In the PC, both STN and GPi volumes were significantly smaller compared to the YC (p < 0.05) which is likely attributable to a combination of lower data resolution and lower subcortical contrast (mean volumes GPi: 324 mm³ ± 62, range 157–434 mm³; STN: 112 mm³ ± 19, range 74–154 mm³). Again, volumes correlated significantly across hemispheres (STN: R = 0.525, p < 0.005; GPi: R = 0.376, p < 0.05). However, no significant difference in volume was found between hemispheres (p = 0.07 and p = 0.67 respectively). Furthermore, no significant correlation could be found for segmented volumes of GPi and STN within the same subject (R = 0.132, p = 0.315).

The manual segmentation of the PC dataset was more challenging due to the lower subcortical contrast and image resolution. This is reflected in the lower Dice coefficient between the manual segmentations of YC and PC (median Dice for both nuclei 0.78 and 0.66 respectively). Nevertheless, automated segmentation performed well (Fig. 3b and Table 3b). For the PC, the performance of both methods (SPM Segment and ANTs SyN) was statistically indistinguishable from inter-rater agreement for Dice coefficient and SD for both nuclei (inter-rater median Dice STN and GPi: 0.66 and 0.65; median SD STN and GPi in mm: 0.64 and 0.82). SPM Segment performed slightly better for SD than ANTs SyN (median SD STN and GPi Segment vs. ANTs SyN in mm: 0.52 and 0.73 vs. 0.61 and 0.75), whereas ANTs SyN performed slightly but significantly better as measured by Dice coefficient (median Dice STN and GPi for SPM Segment vs. ANTs SyN: 0.62 and 0.68 vs. 0.67 and 0.71). For additional details please see Table 3b. Due to a lower number of segmented brains in the PC versus the YC cohorts (30 vs. 73 brains), we have lower statistical power. Only SPM Segment for the GPi showed a statistically significant correlation between manual and automated segmented volumes. This may be due to the reduced SNR of the PC-dataset and increased variability of manually segmented volumes. Still, atlas-based segmentation volumes in the PC cohort remain reasonable (mean volume STN and GPi by ANTs SyN: 118 and 380 mm³; SPM Segment: 112 and 357 mm³).

3.1. Computational time per method

Computational time and memory requirements are important practical considerations when implementing these methods in a research or clinical workflow. Table 4 presents an estimate of required computation time for each method based on commonly available computational resources. The MacBook Pro (left row) had the following specifications: Memory 16 GB 1867 MHz DDR3; Processor 2.9 GHz Intel® Core™ i5. The desktop PC (right row) was better equipped: Memory 64 GB; Processor Intel® Core™ i7-7700K CPU @ 4.20 GHz × 8. The computations are based on a test subjects with multimodal input data consisting of a T1- and T2-weighted MRI. SPM Segment proved to be the fastest evaluated algorithm. ANTs SyN with the additional refinement of subcortical alignment was substantially slower. The processing times were primarily dependent on CPU throughput and were not limited by RAM in either of these configurations.

4. Discussion

We draw three main conclusions from this study. First, we evaluated six commonly used nonlinear deformation algorithms for human brain MRI and identified two methods that perform superiorly. We optimized their parameters and underlying data for alignment of subcortical structures with high precision. Second, we report that the accuracy of these optimized atlas-based segmentations are similar to inter-rater accuracy of expert manual raters, especially for data with lower signal to noise ratio that is common in clinical practice. These results also allow us to quantify the amount of error introduced when comparing subcortical results—such as DBS electrode placements—across patients in standard space. Third, we demonstrate that multimodal image registration is more accurate than unimodal image registration for subcortical structures. This motivates the preoperative acquisition of multispectral data in patients undergoing functional neurosurgery.

We compared six established, openly available nonlinear deformation algorithms from four software suites with a focus on subcortical structures. Similar to prior studies focused on cortical segmentations, this was done by computing the overlap of automated segmentations with manual segmentations (Klein et al., 2009) and comparing accuracy of this overlap to the inter-rater accuracy of two independent manual raters. Dice coefficients of inter-rater YC-STN segmentations (mean 0.75) indicate a good inter-rater agreement and are comparable with inter- and intra-rater STN segmentations reported by other groups (Garzón et al., 2017; Lorio et al., 2016; Wang et al., 2016; de Hollander et al., 2014; Keuken et al., 2014). For the GPi the mean Dice coefficient was 0.79 which is also comparable to previously reported results (Makowski et al., 2017; Lorio et al., 2016; Wang et al., 2016; Keuken et al., 2014). Both results indicate sufficient overlap to establish a valid ground truth from which to assess automated segmentation.

Similar to prior results in the cortical domain, subcortically optimized ANTs and SPM implementations performed the best. However, there were some notable differences. First, within SPM, the oldest of the evaluated methods, the Unified Segmentation algorithm (Ashburner and Friston, 2005) yielded slightly superior results in comparison to newer implementations DARTEL and SHOOT (Ashburner, 2007; Ashburner and Friston, 2011) although these outperformed the Unified Segmentation approach in previous studies with a cortical focus (Klein et al., 2009). There are three potential explanations. First, DARTEL and SHOOT were created and optimized for precision in the cortex (Ashburner and Friston, 2011) and have primarily been utilized for fMRI. Second, the algorithms build upon subject-specific Tissue Probability Maps (TPMs) that result from the preceding Unified Segmentation step. They may only improve results if a proper definition of target structures in both subject and template TPM is given. In contrast to cortical gyri, for small subcortical structures this may not always be the case. To account for this, we did adopt (Lorio et al., 2016), as well as create a specialized template TPM with refined definition of these structures which did improve results (Fig. 3). Third, DARTEL and SHOOT were designed to perform group-wise registrations. Here, they were evaluated for pair-wise registrations since i) this represents a more suitable use-case for clinical neuroimaging and ii) in a prior study, group-wise and pair-wise results of DARTEL were reported not to differ (Klein et al., 2009). A second
Fig. 5. a): Normalization accuracy for the two best performing methods SPM Segment and ANTs SyN for the Young Cohort. The two best performing methods SPM Segment and ANTs SyN are displayed and compared to inter-rater agreement for the YC. The left half (red) shows results for the STN, the right half (blue) for the GPI. Measurement of agreement are shown as Dice coefficient and surface distance (SD) (left half of both panels). Correlation of volumes between atlas-based and manual segmentations are shown on the right hand side of each panel. For detailed numerical results please see Table 3a. b) Normalization accuracy for the two best performing methods SPM Segment and ANTs SyN for the Pseudo-Clinical Cohort. The two best performing methods SPM Segment and ANTs SyN are displayed and compared to inter-rater agreement for the PC. Both methods showed statistically equivalent or better agreement between atlas-based and manual segmentations compared to the inter-rater comparison. Results for the STN are shown in red (left half) and for GPI in blue (right half). For both STN and GPI the left upper left graph shows the overlap between atlas-based and manual segmentations or inter-rater manual segmentations by means of Dice coefficient; the lower left graph shows the mean surface distance between those segmentations (for numeric values please see Table 3b); the upper and lower right graphs show a correlation of volumes for each subject between atlas-based and manual segmentations.

difference was that the ANTs based SyN approach outperformed the BSplineSyN approach. Again, BSplineSyN is the newer algorithm that had outperformed SyN for cortical normalization when it was introduced (Tustison, 2013). In summary, models with low variance in general outperformed models with high variance. One reason for this may be that small structures in the subcortex do not exhibit high image contrast and best results may be achieved by aligning surrounding anatomical structures instead of the nuclei themselves.

4.1. The choice of normalization technique is important

To date, there is no accepted standard for performing normalization of small subcortical nuclei. Fig. 6 demonstrates how variable normalization results can be depending on the method chosen. Here, we show the result of three example methods (ANTS SyN, SPM SHOOT and ANTs BSplineSyN) that were applied to multi-modally warp a single patient volume to MNI space. The patient had bilateral electrode implanted in the GPI. The ground truth of the electrode position can be seen on the native patient image (Fig. 6A). Three-dimensional reconstructions of DBS.

Table 3

| YOUNG COHORT (YC) | STN | Median | Mean | Std. | Range | GPI | Median | Mean | Std. | Range |
|-------------------|-----|--------|------|------|-------|-----|--------|------|------|-------|
| Dice Coeff.       | Segment | 0.70   | 0.67 | 0.10 | 0.35-0.87 | 0.78 | 0.76 | 0.08 | 0.46-0.88 |
|                   | ANTs   | 0.72   | 0.72 | 0.06 | 0.52-0.83 | 0.75 | 0.75 | 0.07 | 0.47-0.88 |
|                   | Inter-Rater | 0.76   | 0.75 | 0.09 | 0.54-0.88 | 0.80 | 0.79 | 0.05 | 0.70-0.88 |
| SD mm             | Segment | 0.45   | 0.48 | 0.14 | 0.21-1.19 | 0.45 | 0.48 | 0.16 | 0.25-1.05 |
|                   | ANTs   | 0.38   | 0.38 | 0.07 | 0.24-0.72 | 0.49 | 0.49 | 0.12 | 0.22-0.87 |
|                   | Inter-Rater | 0.41   | 0.45 | 0.20 | 0.21-1.00 | 0.45 | 0.49 | 0.14 | 0.31-0.83 |
| Volume mm³        | Segment | 117    | 119  | 12  | 96-152 | 372  | 377  | 34 | 311-473 |
|                   | ANTs   | 147    | 148  | 15  | 118-193 | 397  | 406  | 40 | 340-546 |
|                   | Inter-Rater | 119   | 120  | 20 | 68-168 | 363  | 359  | 50 | 231-473 |

Table 4

| Method                | MacBook Pro | Desktop PC |
|-----------------------|-------------|------------|
| 1 SPM New Segment     | 10          | 5          |
| 2 SHOOT               | 15          | 8          |
| 3 DARTEL              | 24          | 13         |
| 4 ANTs SyN mid Variance | 31         | 8          |
| 5 ANTs SyN low Variance | 33         | 9          |
| 6 ANTs SyN high Variance | 36         | 8          |
| 7 ANTs SyN low Variance with subcortical refine | 42 | 13 |
| 8 ANTs BSplineSyN high Variance | 43 | 12 |
| 9 ANTs BSplineSyN mid Variance | 45 | 12 |
| 10 FNIRT              | 126         | 35         |
| 11 ANTs BSplineSyN low Variance | 1090 | 197 |

Table 3b

| PSEUDO-Clinical Cohort (PC) | STN | Median | Mean | Std. | Range | GPI | Median | Mean | Std. | Range |
|-----------------------------|-----|--------|------|------|-------|-----|--------|------|------|-------|
| Dice Coeff.                | Segment | 0.62   | 0.60 | 0.12 | 0.16-0.79 | 0.68 | 0.64 | 0.14 | 0.23-0.85 |
|                            | ANTs   | 0.67   | 0.65 | 0.09 | 0.41-0.82 | 0.71 | 0.70 | 0.09 | 0.44-0.86 |
|                            | Inter-Rater | 0.66  | 0.66 | 0.10 | 0.41-0.81 | 0.65 | 0.65 | 0.06 | 0.51-0.79 |
| SD mm                      | Segment | 0.52   | 0.55 | 0.20 | 0.21-1.27 | 0.73 | 0.79 | 0.38 | 0.30-1.80 |
|                            | ANTs   | 0.61   | 0.64 | 0.19 | 0.31-1.16 | 0.75 | 0.82 | 0.30 | 0.41-1.75 |
|                            | Inter-Rater | 0.64  | 0.72 | 0.25 | 0.40-1.30 | 0.82 | 0.86 | 0.16 | 0.53-1.18 |
| Volume mm³                 | Segment | 110    | 112  | 12 | 83-142 | 356  | 357  | 32 | 279-428 |
|                            | ANTs   | 117    | 118  | 10 | 93-135 | 372  | 380  | 36 | 305-460 |
|                            | Inter-Rater | 113  | 112  | 19 | 74-154 | 324  | 324  | 62 | 157-434 |
electrodes, as shown in B-D, may form a false sense of certainty when presented in isolation without the raw data. This is especially true for fully-automated platforms and highlights the importance of incorporating reports on registration accuracy or an interface for manual confirmation (Husch et al., 2018a, 2018b; D’Haese et al., 2010). However, depending on the chosen method, results differed meaningfully. The exemplary results of Fig. 6 underline the relevance of our findings, the importance of choosing the right method, and emphasize the importance of manual performance checks even for our best performing algorithms.

No registration pair in the sample analyzed here failed in terms of a gross misalignment between subject and template. However, there was variability in registration quality, in particular with the three-step linear approach. The reason for generally high robustness could be that HCP subject data had been linearly prealigned to the MNI space and subjects off the IXI dataset were scanned with rough alignment to the AC/PC line. The two best and two worst performing automated segmentation results for each data set and method are shown in Supplementary Figs. S2a and S2b.

**Fig. 6.** Does the choice of the normalization method matter? A) shows an axial view of the postoperative T2 image of a DBS patient with leads implanted into left and right Gpi (left). The image is windowed for maximum Gpi contrast which exaggerates the DBS lead artifact. The right-hand side shows an enlarged view of the subcortex at the level of the Gpe, Gpi and the putamen. Left and right Gpi are manually outlined with white dotted lines and the leads are marked with white stars. The preoperative patient imaging consisted of T1 and T2-weighted MRI images that were normalized multimodally to MNI space using three different methods. The electrodes were localized in native patient space using the Lead-DBS toolbox. The deformation field estimated during the normalization was subsequently applied to electrode coordinates in native space to visualize the electrodes in MNI space and assess their spatial relationship to the Gpi. B) Normalization results for ANTs SyN with the preset “Low Variance with Subcortical Refine,” C) for ANTs BSplineSyN with preset “Low Variance,” and D) for SPM SHOOT. For panels B, C and D the left column shows the normalized T2-weighted patient image with boundaries of the anatomical structures of the MNI template (red wires) overlaid to demonstrate fit of patient image to MNI template. Middle and right columns show a reconstruction and 3D visualization of the warped patient electrodes in MNI space. Normalization results deteriorate from rows B to D, with the third method showing the electrodes far off their actual anatomical target of the Gpi. Comparing the 3D reconstruction in MNI space to the position of the electrode artifacts in the native patient image, the first method reflects the ground truth most accurately.
4.2. Results of atlas-based segmentations compared with other automated segmentation methods

Others have proposed alternative strategies of automatic segmentation of DBS structures (Garzón et al., 2017; Visser et al., 2016b, 2016a; Xiao et al., 2014; Chakravarty et al., 2013; Haegelen et al., 2012). In a recent paper, Garzón et al. use a new automated segmentation method for the midbrain target regions STN, SN and RN. Their algorithm constructs a map of spatial priors based on a training set of manual labels nonlinearly registered to the midbrain. For the STN, they reported Dice scores of 0.66 for 3T QSM-, 0.57 for R2* - and 0.59 for FLAIR-weighted MRI 0.59.

Visser et al. recently introduced a new method called MIST (Multi-modal Image Segmentation Tool) which uses multiple image modalities simultaneously while using one single reference segmentation for the initialization. On 7T MRI data they reached median Dice scores for their automated segmentation of around 0.60 for the STN (Visser et al., 2016b). On a different 1.5T clinical dataset they reached Dice scores for the whole globus pallidus of around 0.75 and median mean mesh distances (compare to our mean surface distances) of around 0.8 mm (Visser et al., 2016a).

Chakravarty et al. presented a multi-atlas-based approach that was extended by the use of an automatically generated set of templates derived from different brains (Chakravarty et al., 2013). This “template layer” introduces an additional level of redundant registrations, in a first step warping the atlas brain(s) to all template brains and in a second step all of the template brains to the individual, to-be-segmented brain. This results in multiple versions of automatic segmentations in the individual brain that are then combined in a final segmentation. In theory this will distribute and thus eliminate errors from single warps due to registration and resampling errors. Using this method they reported a Dice coefficient of 0.75 for the whole globus pallidus.

In the young cohort for our best-performing method (ANTS SyN with preset “Low Variance with Subcortical Refine”) we were able to reach a mean Dice score for STN and GPI of 0.72 and 0.75 respectively and a mean SD of 0.38 mm and 0.49 mm. This approach performs similarly to the best of the previously reported methods (Husch et al., 2018b; Garzón et al., 2017; Visser et al., 2016b, 2016a; Chakravarty et al., 2013; Haegelen et al., 2012). In fact, the results for mean SD (0.44 mm for both nuclei) are below the actual image resolution of the datasets (0.7 mm), consistent with a high level of accuracy.

4.3. Extension of methods to other structures

Here we have focused on the two most common DBS targets, STN and GPI, though this method may also be applicable to other small subcortical structures such as thalamic nuclei or the nucleus accumbens. The borders of these structures lack distinct MRI contrasts which may reduce the ability to accurately align them to template space. Still, our results show that the STN can still be segmented well even when using T1-weighted MRI as an input, where STN contrast is very poor. This is accomplished by registering surrounding structures with sufficient image contrast as accurately as possible. Given the phylogenetic age of subcortical structures, displacement across subjects is likely to be less than it is for cortical folds. For the thalamus specifically, the methods proposed here could potentially adjust for alterations in the overall size and shape of the thalamus, as well as potentially align intrathalamic structures like the internal medullary lamina which are discernible on MRI. Still, absent histology, we lack for a gold standard to segment the thalamus into its subnuclei, which limits the ability to explicitly assess the performance of automated segmentation algorithms for these structures.

4.4. Volumes of segmented STN and GPI

We found a marginal but significant difference in volumes of the segmented STNs across all YC subjects with a slightly higher volume for the left STN (mean volume left STN = 122 mm³, right STN = 118 mm³). This left-right asymmetry has previously been reported in larger cohorts (e.g., 120 subjects in Shen and Gao, 2009 and 168 STNs in Nowinski et al., 2005). For the PC we could not replicate this finding, perhaps due to a lower number of segmented subjects and resulting reduced statistical power to discriminate such small absolute volume differences. Our STN volumes are in line with other histology studies with volumes ranging from 119 to 139 mm³ for the STN (Zwirner et al., 2016; Mai et al., 2015; Hardman et al., 2002). In addition to the high inter-rater overlap, these findings give us additional confidence in the quality of our manual segmentations.

4.5. Limitations

The study has several limitations. There is no definitive test to assess whether an algorithm is performing optimally. As such, we cannot rule out that further optimizations of ANTs, SPM or other algorithms (e.g., FNIRT) could outperform the results reported here. We therefore are careful not to claim that the performance ranking reported here could not be altered by, for example, further adapting the preset parameters of the FSL method for better subcortical alignment. Second, we used datasets of healthy subjects for this study, although we included an age- and quality-matched cohort to resemble a typical Parkinson’s disease. Nevertheless, we have not directly assessed the performance of these algorithms in a clinical cohort. Furthermore, here, we focused on routinely available acquisitions (T1-, T2-, PD-) instead of specialized basal ganglia sequences such as QSM or FGATIR that are increasingly utilized in DBS surgical planning. Of note, however, our final results still performed competitively with respect to results of a recent study that used these sequences for automated segmentation (Garzón et al., 2017).

Lastly, the present results show how well automatic segmentation reproduces manual segmentations of basal ganglia structures as visible on MRI. Although these results suggest that modern automatic procedures are (nearly) able to reproduce the work of expert humans, there remains a gap between the MRI-defined structure and the true boundary of the structures as demonstrated on histology. The size of the subthalamic nucleus and pallidum are both larger on histology than on MRI (Dormont et al., 2004; Richter et al., 2004; Schäfer et al., 2012; de Hollander et al., 2014; Massey et al., 2012). For the STN, much of the MRI imaging contrast derives from iron deposition which is not uniform within the nucleus. Rather, it exhibits a gradient that decays from anterior to posterior similar to the gradient in cellular density (de Hollander et al., 2014; Marani et al., 2008). As we did not have histologic data for the subjects studied here, we can not address these discrepancies directly.

It is possible that ongoing technical refinements in MRI including quantitative susceptibility mapping (Alkemade et al., 2017) and high-field MRI imaging (Forstmann et al., 2017a, b) may narrow this gap in the future.

4.6. Conclusions

In conclusion, we have presented a detailed, validated approach to i) select nonlinear warping algorithms including their parameters and presets and ii) justify the use of atlas-based segmentations for automated definition of DBS targets. We see that several future innovations could further improve on these results. For one, incorporation of the aforementioned specialized basal ganglia sequences (QSM and FGATIR) may further improve accuracy and robustness.

We have released our 103-brain manually segmented dataset as an evaluation tool that can be used to assess the performance of future iterations of image registration algorithms. We plan to integrate this dataset into the open source toolbox Lead-DBS, of which the code is accessible through Github (https://github.com/leadlbs/leadlbs/). The optimal pipeline evaluated in the present study is included as the default preset in Lead-DBS. Code to perform evaluations of new methods similar
to the present study will be included as an addition to the toolbox. We hope that other groups will build on our results to further optimize approaches to atlas-based segmentation.

Funding

This research was supported in part by NINDS grant K23NS099380 and an American Academy of Neurology/American Brain Foundation Clinical Research Training Fellowship to TMI. The project was further supported by the German Research Council, DFG grant KFO247 and SPP-2041 (AAK).

Conflicts of interest

All authors report no conflicts of interest.

Acknowledgements

The authors would especially like to thank Dr. Brian Edlow and Dr. Bruce Fischl for providing us with high-resolution post-mortem human MRI scans that showed the target structures in detail and guided our manual segmentation.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.neuroimage.2018.09.061.

References

Alkemade, A., de Hollander, G., Keulen, M.C., Schafer, A., Ott, D.V.M., Schwarz, J., Weise, D., Kotz, S.A., Forstmann, B.U., 2017. Comparison of T2-weighted and QSM contrasts in Parkinson's disease to visualize the STN with MRI. PLoS One 12, 1-13. https://doi.org/10.1371/journal.pone.0176130.e0176130-13

Andersson, J.L.R., Jenkinson, M., Smith, S., 2010. Non-linear registration, aka spatial normalisation. FMRI technical report TR07JA2.

Ashburner, J., 2007. A fast diffeomorphic image registration algorithm. Neuroimage 38, 1458-1468. https://doi.org/10.1016/j.neuroimage.2007.07.007.

Ashburner, J., Friston, K.J., 2011. Diffeomorphic registration using geodesic shooting and Newton’s method. Neuroimage 54, 2654-2666. https://doi.org/10.1016/j.neuroimage.2010.07.033.

Evans, A.C., Janke, A.L., Collins, D.L., Bartlett, S., 2012. Brain templates and atlases 62, 219-265. https://doi.org/10.1016/j.neuroimage.2012.01.024.

Fischl, B., 2012. FreeSurfer 62, 774-781. https://doi.org/10.1016/j.neuroimage.2011.07.051.

Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinstry, R.C., Collins, D.L., Group, T.B.D.C., 2011. Unbiased average age-appropriate atlases for pediatric studies. Neuroimage 54, 313-327. https://doi.org/10.1016/j.neuroimage.2010.07.053.

Friston, B.U., Isaacs, B.R., Temel, Y., 2017a. Ultra high field MRI-guided deep brain stimulation. Trends Biotechnol. 35 (10), 904-907. https://doi.org/10.1016/j.tibtech.2017.06.010.

Friston, B.U., de Hollander, G., van Maanen, L., Alkemade, A., Keulen, M.C., 2017b. Towards a mechanistic understanding of the human subcortex. Nat. Rev. Neurosci. 18, 57-65. https://doi.org/10.1038/nrn.2016.163.

Glasser, M.F., Menon, R., Uddin, L.K., Sridharan, M., Sabuncu, M.R., Shehzad, Z., Fischl, B., 2016. Multimodal parcellation of human cerebral cortex. Nature Publishing Group 536, 171-178. https://doi.org/10.1038/nature18933.

Glasser, M.F., Sotiropoulos, S.N., Wollen, J.A., Cansol, G.S., Fischl, B., Anderson, J.L., Xu, J., Shabih, M., Webster, M., Polimeni, J.R., Van Essen, D.C., 2013. A multi-modal parcellation of human cerebral cortex. Nature 501, 205-209. https://doi.org/10.1038/nature12587.

Glasser, M.F., Fischl, B., 2016. The minimal preprocessing pipelines for the Human Connectome Project 80, 105-124. https://doi.org/10.1016/j.neuroimage.2013.04.127.

Haegelen, C., Coupi, P., Fonov, V., Guizard, N., Jannin, P., Morandi, X., Collins, D.L., 2012. Automated segmentation of basal ganglia and deep brain structures in MRI of Parkinson's disease. Int J CARS 8, 99-110. https://doi.org/10.1007/s11782-012-9078-8.

Hardman, C.D., Henderson, J.M., Finkelstein, D.I., Horne, M.K., Paxinos, G., Halliday, G.M., 2002. Comparison of the basal ganglia in rats, marmosets, macaques, baboons, and humans: volume and neuronal number for the output, internal relay, and striatal modulating nuclei. J. Comp. Neurol. 445, 238-255. https://doi.org/10.1002/cne.10165.

Helmke, G., Draganski, B., Frackowiak, R., Ashburner, J., Weiskopf, N., 2009. Improved Segmentation of Deep Brain Grey Matter Structures Using Magnetization Transfer (MT) Parameter Maps, vol. 47, pp. 194-198. https://doi.org/10.1016/j.neuroimage.2009.03.053.

Horn, A., Blankenburg, F., 2016. Toward a standardized structural-functional group connectome in MRI space. Neuroimage 124, 310-322. https://doi.org/10.1016/j.neuroimage.2015.08.048.

Horn, A., Kuhn, A.A., 2015. Lead-DBS: A Toolbox for Deep Brain Stimulation Electrode Localizations and Visualizations, vol. 107, pp. 127-135. https://doi.org/10.1016/j.jir.2014.12.002.

Horn, A., Kuhn, A.A., Merkl, A., Shih, L., Alterman, R., Fox, M.A., 2017a. Prognostic Discrimination of the Expected Brain Stimulation Outcome in Parkinson Disease. J. Neurosurg. 127, 150-154. https://doi.org/10.3171/2015.12.JNS15615.

Horn, A., Kuhn, A.A., Merkl, A., Shih, L., Alterman, R., Fox, M.A., 2017b. Prognostic Discrimination of the Expected Brain Stimulation Outcome in Parkinson Disease. J. Neurosurg. 127, 150-154. https://doi.org/10.3171/2015.12.JNS15615.

Horn, A., Kuhn, A.A., Merkl, A., Shih, L., Alterman, R., Fox, M.A., 2017b. Prognostic Discrimination of the Expected Brain Stimulation Outcome in Parkinson Disease. J. Neurosurg. 127, 150-154. https://doi.org/10.3171/2015.12.JNS15615.

Horn, A., Kuhn, A.A., Merkl, A., Shih, L., Alterman, R., Fox, M.A., 2017b. Prognostic Discrimination of the Expected Brain Stimulation Outcome in Parkinson Disease. J. Neurosurg. 127, 150-154. https://doi.org/10.3171/2015.12.JNS15615.

Horn, A., Kuhn, A.A., Merkl, A., Shih, L., Alterman, R., Fox, M.A., 2017b. Prognostic Discrimination of the Expected Brain Stimulation Outcome in Parkinson Disease. J. Neurosurg. 127, 150-154. https://doi.org/10.3171/2015.12.JNS15615.
Jenkinson, M., Bannister, P., Brady, M., Smith, S., 2002. Improved optimization for the robust and accurate linear registration and motion correction of brain images. Neuroimage 17, 825–841. https://doi.org/10.1016/S1053-8119(02)00033-9.

Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M., 2012. FSL 6.2, 782–790. https://doi.org/10.1016/j.neuroimage.2011.09.015.

Keuken, M.C., Bazin, P.L., Crown, L., Hootsmans, J., Laufer, A., Müller-Ax, C., Sier, R., van der Putten, E.J., Schafer, A., Turner, R., Forstmann, B.U., 2014. Quantifying inter-individual anatomical variability in the subcortex using 7 T structural MRI. Magn. Reson. Imaging 94, 40–46. https://doi.org/10.1016/j.mri.2014.03.032.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.-C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Seng, 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, vol. 46, pp. 786–802. https://doi.org/10.1016/j.neuroimage.2008.12.037.

Lim, I.M., Faria, A.V., Li, X., Hsu, J.T.C., Airan, R.D., Mori, S., van Zijl, P.C.M., 2013. Pallidal, and Thalamic Segmentation Methods: Comparing Automated Approaches to Improved Optimization for the Pallidal and Thalamic Segmentation Methods. Comparison of manual and automated approaches to segmentation of the pallidum and thalamus in healthy control subjects and patients with Parkinson’s disease. Hum. Brain Mapp. 33, 586–598. https://doi.org/10.1002/hbm.23427.

Weber et al. NeuroImage 184 (2019) 586–598