## Surfing the Hippocampus Wave

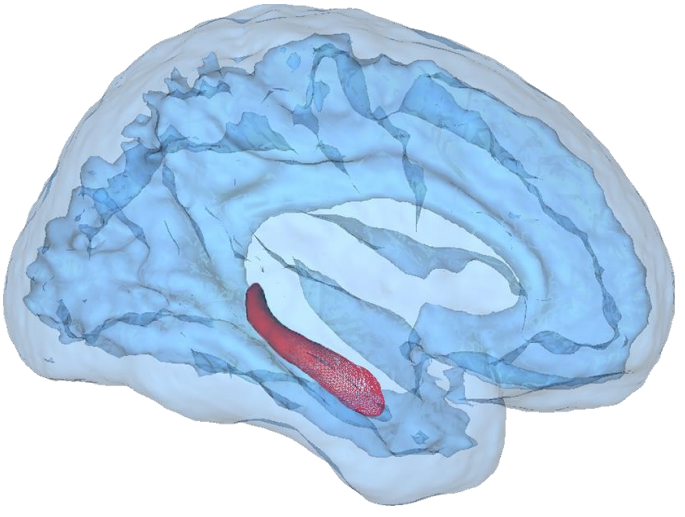
Bartel, F.

2019

***document version***
Publisher's PDF, also known as Version of record

# Chapter 5

## Creating accurate reference segmentations of deep GM structures in MS patients by fast semi-automated outlining



Alexandra de Sitter
Fabian Bartel
Miklos Palotai
Jessica Burggraaff
Yaou Liu
Jorge Simoes
Serena Ruggieri
Katharina Schregel
Alfredo Morales Pinzon
Stefan Ropele
Maria A. Rocca
Claudio Gasperini
Antonio Gallo
Menno Schoonheim
Michael Amann
Marios Yiannakas
Mike P. Wattjes
Jaume Sastre-Garriga
Ludwig Kappos
Massimo Filippi
Christian Enzinger
Olga Ciccarelli
Jette Frederiksen
Frederik Barkhof
Charles R.G. Guttmann
Jan C. de Munck
Hugo Vrenken

Submitted to Multiple Sclerosis Journal June 2018

## Abstract

**Background:** To create reference segmentations for training and evaluating automated brain segmentation methods, manual outlining is the “gold standard”; however, it is challenging and time consuming. In this study we investigated whether FASTSURF (FAst Segmentation Through SURface Fairing), a surface-fairing technique based on a bi-Laplacian mesh operator to reconstruct structures from sparse delineations, is accurate enough to create reference segmentations of deep GM structures in MS.

**Method:** Three independent raters outlined caudate nucleus, putamen, and thalamus bilaterally on 35 multi-center, 3D-T1-weighted scans. Optimal FASTSURF settings were determined from a training set (n=17). Agreement of FASTSURF with manual segmentations was assessed in a test set (n=18) by evaluating mean volumes, intra-class correlation coefficients (ICC), and generalized conformity index ($CI_{gen}$) between manual and FASTSURF segmentations.

**Results:** In the test set, mean volumes of left and right caudate were 3.78±0.63mL for FASTSURF versus 3.85±0.67mL for manual; putamen: 4.63±1.01mL versus 4.65±1.02mL; and thalamus 6.72±1.52mL versus 6.75±1.51mL. Absolute agreement ICCs were excellent: 0.979 (caudate), 0.999 (putamen), and 0.999 (thalamus). The good inter-rater overlap between manual segmentations, with $CI_{gen}$ values of 0.74±0.05 (caudate), 0.74±0.06 (putamen) and 0.75±0.06 (thalamus), was well reproduced by FASTSURF (0.74±0.05, 0.73±0.06 and 0.75±0.06, respectively).

**Conclusions:** FASTSURF reconstructions exhibit excellent volumetric agreement with manual segmentations. Moreover, inter-rater overlap was almost identical for FASTSURF and for manual segmentations. We conclude that accurate reference segmentations could be created at a strongly reduced workload using this semi-automated method, which generates opportunities for developing improved automated deep GM atrophy measurement methods for MS and other neurological diseases.

## 5.1. Introduction

Multiple sclerosis (MS) is an inflammatory and neurodegenerative disease of the central nervous system. MS patients exhibit damage to the grey matter (GM), including focal lesions [1,2] and atrophy (volume loss) [3]. GM atrophy can be quantified from longitudinal structural magnetic resonance images (MRI) of the brain and has become an important aspect of MS pathology and outcome measures.

In particular, atrophy of deep GM structures such as caudate nucleus, putamen and thalamus has become of interest as it has been shown to correlate with important clinical outcome measures such as cognition [4–8]. A better understanding of the pathophysiologic basis for cognitive dysfunction in MS could lead to more sensitive markers for individual patients and improved therapies to slow or prevent cognitive decline in MS.



It is relevant to consider atrophy measures of these GM structures as potential biomarkers for cognitive deterioration, but current semi-automatic and automatic measurement techniques still suffer from substantial reproducibility limitations [9], which may be partly due to their impaired performance in the presence of other MS pathologies, such as lesions [10,11]. Therefore, existing semi-automatic and automatic methods need to be improved. To optimize these methods a reliable “gold standard” reference set with segmentations of GM structures in MS patients is needed. Such a reference set can be used for training automated methods and for a systematic comparison and ranking of methods.

Such “gold standard” reference sets are usually created from manual segmentations by trained experts in the field of neuroradiology. However, these manual outlines can suffer from considerable inter- and intra-observer variability. Moreover, establishing a “gold standard” reference with manual delineations is labor-intensive [12–14] and therefore time consuming.

Two promising approaches have been recently proposed to reduce the time to create a “gold standard” reference set. Firstly, it has been proposed to reduce time-investment and workload of experts by letting non-experts, i.e., a radiologist in training or even the general public, outline structures (in the brain), without substantial loss of quality of the reference set [15,16]. A similar approach has been applied to the task of identifying anatomical subdivisions of the cerebellum and is currently being tested for the segmentation of lesions [17]. Secondly, Bartel et al. (2017) developed a semi-automatic segmentation method, FAst Segmentation Through SURface Fairing (FASTSURF) to speed up the delineation of the hippocampus [18,19]. This study demonstrated the proof-of-concept of the FASTSURF method, showing promising results, with hippocampus segmentations

using FASTSURF being closer to manual segmentations than fully automated hippocampus segmentation methods (FSL-FIRST [20] and FreeSurfer [14,21]). Furthermore, the study illustrated that FASTSURF segmentation is less time consuming than fully manual segmentations. The method is based on mesh processing procedures and requiring only a few manually outlined contours and no atlases. If FASTSURF could be used to generate large reference segmentation datasets with substantially lower workload, this could provide an important impetus to the improvement of deep GM atrophy measurement in MS and, by extension, other diseases. Because the three deep GM structures of interest in our study have geometrically fewer complex shapes than the hippocampus, we expect FASTSURF to perform well for those structures too. Ultimately, this study could lead into a (research) crowdsourcing experiment in which non-experts create partial contours that are completed by FASTSURF to produce a "gold standard" reference set within a short time frame and without the need for expert segmentation.

In this study, therefore, we aimed to investigate the performance of FASTSURF in MS patients' images for three deep GM structures: caudate nucleus, putamen and thalamus. Because the ultimate goal is to use this approach to create a widely applicable set of reference segmentations, we used a multi-center dataset. We quantitatively assessed agreement with fully manual outlines by comparing mean volumes, by quantifying intra-class correlation coefficients for absolute agreement, and by investigating how well the overlap between the outlines of three raters was reproduced.

## 5.2. Materials and Method

*Dataset and MRI acquisition*

Brain MRI scans of 12 healthy controls (eight females) and 23 MS patients (12 females) from nine centers of the MAGNIMS Study Group (www.magnims.eu) and acquired as part of two previous MAGNIMS studies [22,23], were retrospectively included. Demographics of these patients are shown in Table 1. The institutional review boards of each center approved the original study in which patients participated and written informed consent had been obtained from all subjects. MR imaging was performed on 3.0 Tesla whole-body MR systems and near-isotropic, ~1mm$^3$ voxel size, 3D T1 images were included. More detailed image acquisition details are listed in Table 2.

**Table 1: Demographics of healthy controls and MS patients**

| Set | Type | N[b] | Age years[a] | Disease types | DD years | EDSS[c] |
|---|---|---|---|---|---|---|
| Total | HC | 12 [67%] | 38.4±7.8 | | | |
| | Patient | 23 [57%] | 42.9±9.9 | 11 RR, 5 SP, 7 PP | 11.8±8.5 | 3.3±1.8 |
| Training | HC | 5 [100%] | 34.7±8.0 | | | |
| | Patient | 12 [57%] | 44.4±11.9 | 7 RR, 2 SP, 3 PP | 12.9±10.1 | 3.0±1.7 |
| Test | HC | 7 [43%] | 41.1±7.1 | | | |
| | Patient | 11 [55%] | 41.3±7.4 | 4 RR, 3 SP, 4 PP | 10.6±6.9 | 3.7±2.0 |

[a]Mean ± standard deviation; [b]Number of subjects [%female]; [c]Median (range); *HC* Healthy control, *DD* Disease duration, *EDSS* Expanded disability status scale, *CIS* Clinically isolated syndrome *RR* Relapsing-remitting, *SP* Secondary-progressive, *PP* Primary-progressive



**Table 2: An overview of the acquisition parameters for each center**

| Center | Scanner brand, scanner type | TR (ms) | TE (ms) | TI (ms) | FA (◦) | Acquisition (Voxel size ($mm^3$)) |
|---|---|---|---|---|---|---|
| A | GE, Signa HDxt | 7.8 | 3 | 450 | 12 | 256x256x188 (0.976x0.976x1) |
| B | Siemens, Trio | 2300 | 2.98 | 900 | 9 | 232x256x176 (1x1x1) |
| C | Siemens, Trio | 1570 | 2.70 | 900 | 9 | 160x256x256 (1x1x1) |
| D | Philips, Achieva | 6.9 | 2.78 | 831 | 9 | 160x240x240 (1x1x1) |
| E | Siemens, Trio | 1900 | 2,1 | 900 | 9 | 224x256x176 (1x1x1) |
| F | Siemens, Trio | 2200 | 2,94 | 900 | 10 | 256x192x192 (1x1x1) |
| G | Philips, Achieva | 8,3 | 3,72 | 1000 | 8 | 256x256x192 (1x1x1) |
| H | GE, Signa HDxt | 5,5 | 1,76 | 450 | 10 | 256x256x188 (1x1x1) |
| I | Philips, Achieva | 8,3 | 3,72 | 1000 | 8 | 256x256x192 (1x1x1) |

*TR* Repetition time, *TE E*cho time, *TI* Inversion time, *FA* Flip angle

The dataset was divided into a training set (N = 17) and a test set (N = 18). In both groups, the different centers and the numbers of patients and controls were equally distributed. The training set was used to find optimal settings for the parameters of FASTSURF and the test set was used to study the performance of optimized FASTSURF compared to the fully manual segmentation.

### *Manual segmentations*

Three equally trained raters (J.B, J.S., S.R.), who were blinded to clinical data, manually segmented the caudate nucleus, putamen and thalamus for both hemispheres on all 35 images. The segmentations were performed using a standardized protocol. Structures were outlined by using the Wacom Intuos digital pen within the online framework of Structured Planning and Implementation of New

Explorations (SPINE, https://spinevirtuallab.org/)). SPINE is an online research tool developed by the Brigham and Women's Hospital and Harvard Medical School, which can be used for manual delineations of regions-of-interest on MR images.

*FASTSURF method*

The semi-automated segmentation method FASTSURF is based on mesh processing procedures using a surface fairing technique [18,19]. Briefly, to reduce the delineation time for manual observers, only a few contours have to be outlined, at regular slice intervals. First, these sparse contours are interpolated so that each contour has the same number of points. A closed mesh is then constructed by placing intermediate contours between the known contours. Vertex positions for the intermediate contours are obtained by solving the following bi-Laplacian system of equations for the unknown *x*, *y* and *z*-coordinates of the vertices of the intermediate contours:

$$\sum_m {L_{n,m}}^2 \boldsymbol{x}_m = \sum_m {L_{n,m}}^2 \boldsymbol{y}_m = \sum_m {L_{n,m}}^2 \boldsymbol{z}_m = 0 \qquad (1)$$

in which the Laplacian filter $L_{n,m}$ represents the connectivity graph with *n* and *m* being the mesh vertices. Solving these equation leads to a smooth surface mesh passing through the delineated points with minimum curvature.

*Sparse contour simulation and training of FASTSURF*

Similar as described by [18,19] we extracted sparse contours from fully manually segmented structures. The segmented structures were converted to meshes using the marching cubes algorithm and sparse contours were extracted at regular intervals serving as input for FASTSURF.

FASTSURF segmentations were compared to the fully manually segmented structures to evaluate FASTSURF's performance. These segmentations were transformed to meshes using the marching cubes algorithm which were compared to FASTSURF meshes using the method described by [18,19]. A few parameters can be set in FASTSURF and optimal settings may vary depending on the shape of the structure: 1) The orientation of outlining planes (axial, coronal or sagittal), 2) the number of the outlined contours (Ncontours), 3) the number of intermediate contours added by FASTSURF between two outlined contours (Ninter), and 4) the number of points used for each contour (Npoints).

We used the training set (see, section 2.1. Dataset and MRI acquisition) to determine optimal parameter settings for FASTSURF, by maximizing the mean

volume overlap, with the Jaccard $\left(\frac{V_{i\cap j}}{V_{i\cup j}}\right)$, between the FASTSURF and fully manual segmented structures. The training has been performed in three steps:

i. Find the optimal plane in which sparse contours should be delineated, with given parameters: Ncontours = {4, 6, …,10}, Ninter = 4 and Npoints = 100.
ii. Find the optimal Ninter for FASTSURF for each structure with the use of the optimal plane and given parameters: Ncontours = {4, 6, …,10} and Npoints = 100.
iii. Find the optimal Npoints ranging from {30, 50, …, 300} and optimal Ncontours (4, 6, 8 and 10) with the use of the optimal plane and Ninter.

*Statistical analysis*

Using the optimal settings obtained from the training set results, FASTSURF's performance was validated on the test set. Volumes were calculated from FASTSURF and from fully manual segmentations for all three raters. The volumes from FASTSURF and manual outlines were correlated with the intra-class correlation coefficient (ICC) for absolute agreement [24]. We used Altman's criteria to interpret the ICCs: <0.40 was considered as poor reliability, 0.40 to 0.74 was considered fair to good, and ≥0.75 was considered excellent [25]. Variability of outlines between raters was expressed with the generalized conformity index ($CI_{gen}$) [26]:

$$\mathrm{CI}_{\mathrm{gen}} = \frac{\sum_{\text{pairs i>j}} \mathrm{Vol}(O_{\mathrm{i}} \cap O_{\mathrm{j}})}{\sum_{\text{pairs i>j}} \mathrm{Vol}(O_{\mathrm{i}} \cup O_{\mathrm{j}})} \quad (2)$$

Here $O_i$ and $O_j$ represent meshes according to delineations by individual raters, Vol() the volumes of their intersections and unions and the summations run over all different pairs of delineations. Perfect overlap yields a $CI_{gen}$ of 1, and no overlap yields a $CI_{gen}$ of 0. The $CI_{gen}$ between raters was calculated for the manual segmented sets and the FASTSURF segmentations.

## 5.3. Results

The dataset of 35 brain images was divided into a training set (N = 17) and a test set (N = 18). The training set was used to find optimal settings for the parameters of FASTSURF. The results of the separate training steps are provided in Tables A, B, C and D of the supplementary data. Table 3 presents the resulting optimal settings for each structure (both hemispheres combined) derived from the training set. Fig. 1 displays examples of the fully manual segmented, FASTSURF segmented structures and the comparison.

**Table 3: Optimal settings derived in training.**

| Structure | Plane | Ninter | Ncontours | Npoints |
|---|---|---|---|---|
| Caudate nuclues | Axial | 3 | 10 | 100 |
| Putamen | Coronal | 3 | 10 | 150 |
| Thalamus | Sagittal | 3 | 10 | 300 |

*Ninter* Amount of intermediate lines used between the sparse contours for reconstruction, *Npoints* Amount of points on a sparse contour for reconstruction, *Ncontours* Amount of extracted sparse contours.

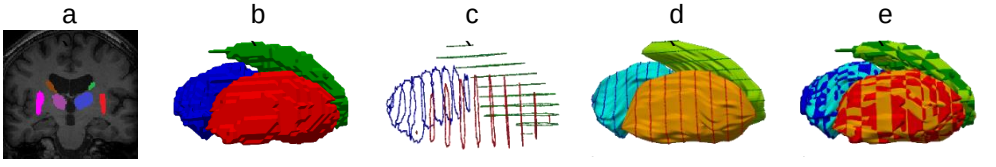



**Fig. 1: The FASTSURF method interpolates a 3D surface by minimizing its curvature under the constraint that it passes through the points contained in sparse outlines provided as input. Linearization and discretization of this mathematical problem boils down to solving a large sparse system of equations representing the bi-Laplacian of the unknown part of the surface [18,19]. (a) Coronal 3D T1-weighted MRI slice with outlines, (b) 3D view of full manual outlines, (c) 3D view of sparse manual outlines, (d) full outlines reconstructed by FASTSURF with lines for the sparse contours, (e) comparisons between reconstructed and manual outlines of caudate nucleus (green/dark green), putamen (red/orange), thalamus (blue/dark blue).**

The test set was used to study the performance of optimized FASTSURF compared to the fully manual segmentation. In Fig. 2 and 3 and Tables 4 and 5, the results of the test set are shown. In Table 4, the mean volumes of the manual segmentation and the reconstructed segmentations are shown. Grouping all three raters' segmentations together showed that the method was accurate. The average total (bilateral) volume of the caudate nucleus was 3.78 ± 0.67 mL and 3.85 ± 0.63 mL for FASTSURF segmentations and manual segmentations, respectively. For putamen, the average reconstructed and manual volumes were 4.63 ± 1.01 mL and 4.65 ± 1.02 mL, and for thalamus they were 6.72 ± 1.52 mL and 6.75 ± 1.51 mL, respectively. These volumes are the pooled volumes of the right and left hemisphere, the volumes for each hemisphere can be observed in Table 4. The volumetric agreement of FASTSURF and manual segmentations was excellent (Table 5 and Figure 2), with total bilateral volume giving ICCs for absolute agreement of 0.979 for caudate nucleus, 0.999 for putamen and 0.999 for thalamus. In Figure 2, volumes of all three structures from FASTSURF segmentations are plotted against volumes from manually segmented structures, showing the left and right structures and the different raters as separate markers.

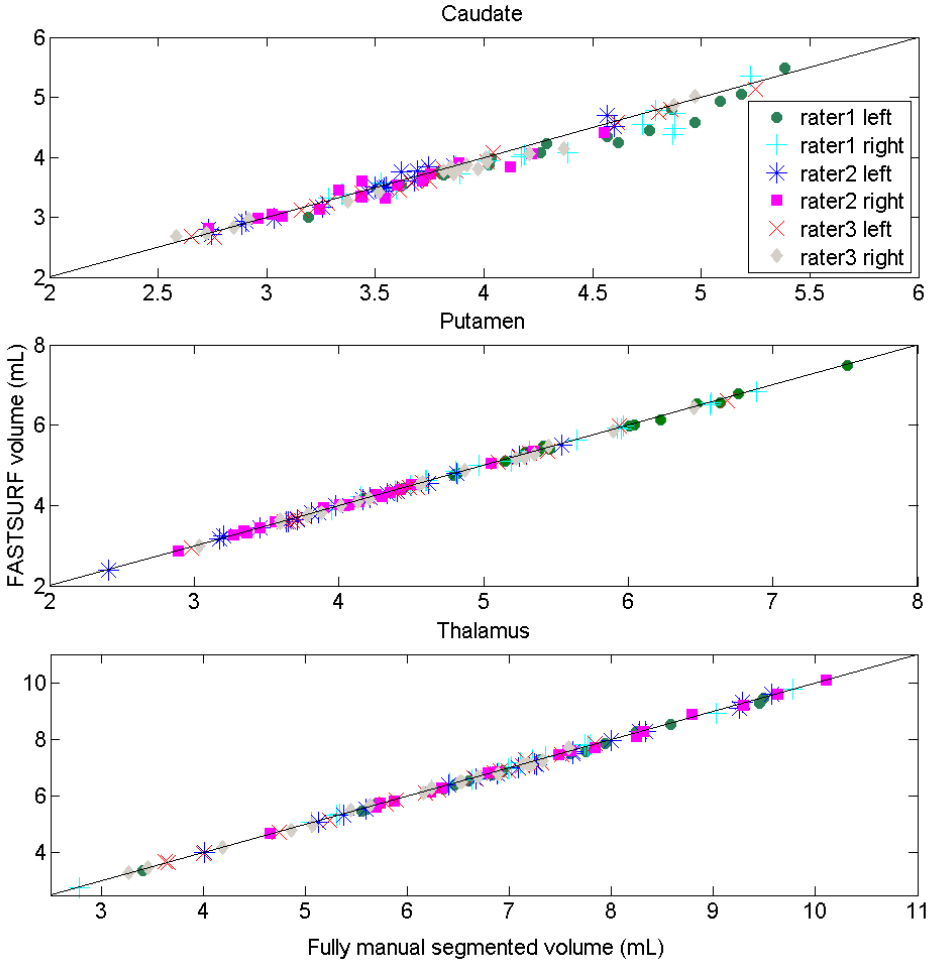



**Fig. 2: Scatter plots of volumes of FASTSURF-reconstructed segmentations (vertical axes) versus full manual segmentations (horizontal axes) for caudate nucleus (top panel), putamen (middle panel) and thalamus (bottom panel), separated by rater and hemisphere, with the identity line for reference.**

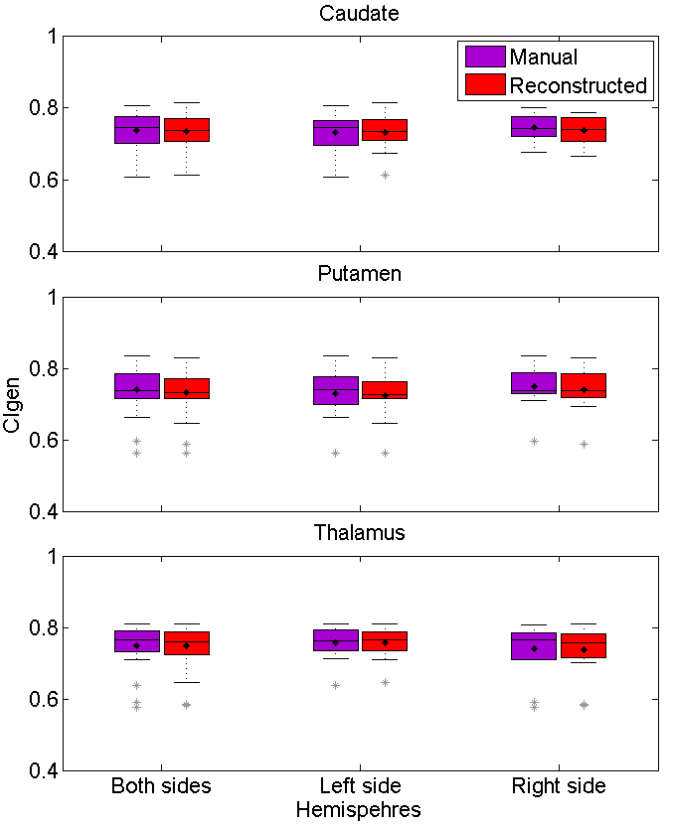



**Fig. 4: Boxplots of generalized conformity index (CIgen), representing the overlap between outlines generated by three raters, for full manual segmentations (purple) and FASTSURF-reconstructed segmentations (red) for caudate nucleus (top panel), putamen (middle panel) and thalamus (bottom panel), for left and right hemispheres separately as well as both hemispheres combined.**

**Table 4: Mean volume across the three raters and the generalized conformity index (CIgen) reflecting overlap between the raters separated by structure, hemisphere and segmentation method.**

| Structure | Set | Hemispheres | Volume (mL)[a] | $CI_{gen}$[a] |
|---|---|---|---|---|
| Caudate | Both | Manual | 3.85±0.67 | 0.74±0.05 |
| nucleus | | Reconstructed | 3.78±0.63 | 0.74±0.05 |
| | Left | Manual | 3.87±0.71 | 0.73±0.05 |
| | | Reconstructed | 3.81±0.68 | 0.73±0.05 |
| | Right | Manual | 3.82±0.64 | 0.75±0.04 |
| | | Reconstructed | 3.74±0.59 | 0.74±0.04 |
| Putamen | Both | Manual | 4.65±1.02 | 0.74±0.06 |
| | | Reconstructed | 4.63±1.01 | 0.73±0.06 |
| | Left | Manual | 4.69±1.08 | 0.73±0.06 |
| | | Reconstructed | 4.67±1.08 | 0.73±0.06 |
| | Right | Manual | 4.61±0.96 | 0.75±0.05 |
| | | Reconstructed | 4.59±0.95 | 0.74±0.05 |
| Thalamus | Both | Manual | 6.75±1.51 | 0.75±0.06 |
| | | Reconstructed | 6.72±1.52 | 0.75±0.06 |
| | Left | Manual | 6.78±1.50 | 0.76±0.04 |
| | | Reconstructed | 6.73±1.49 | 0.76±0.04 |
| | Right | Manual | 6.72±1.56 | 0.74±0.07 |
| | | Reconstructed | 6.70±1.55 | 0.74±0.07 |

[a]Mean ± standard deviation; *$CI_{gen}$* Generalized conformity index

**Table 5: ICC between volumes of reconstructed and fully manual segmented.**

| Structure | Hemispheres | ICC |
|---|---|---|
| Caudate | Both | 0.979 |
| Nucleus | Left | 0.984 |
| | Right | 0.973 |
| Putamen | Both | 0.999 |
| | Left | 0.999 |
| | Right | 0.999 |
| Thalamus | Both | 0.999 |
| | Left | 0.999 |
| | Right | 0.999 |

*ICC* Intraclass correlation coefficient (two-way mixed model with absolute agreement)

For spatial agreement the $CI_{gen}$ of the three raters was calculated for both the manually outlined structures and the reconstructed structures. For manual segmentations, both left and right hemisphere pooled together, the average $CI_{gen}$ for the caudate nucleus, putamen and thalamus were 0.74±0.05, 0.74±0.06 and 0.75±0.06 respectively. FASTSURF showed highly similar results, with $CI_{gen}$ of 0.74±0.05, 0.73±0.06 and 0.75±0.06, respectively. The $CI_{gen}$ for each hemisphere are provided in Table 4 and as boxplots in Figure 3. The highly similar $CI_{gen}$ of FASTSURF compared to manual segmentations across all three structures and for both hemispheres indicates that the overlap between raters was almost identical for FASTSURF and manual outlining.

## 5.4. Discussion

In this study, we showed that FASTSURF could be used to create accurate reference segmentations of the caudate nucleus, putamen and thalamus. The inter-rater overlap of three raters was almost identical for FASTSURF compared to fully manual delineations, and volumetric agreement with manual segmentations was excellent. Therefore, FASTSURF could be an adequate tool to create a reference segmentation set with considerably less effort than manual segmentations, thus opening up possibilities for improving automated methods for deep GM atrophy studies in MS and other neurological diseases.

Several studies have noted that a “gold standard” for testing atrophy measuring methods is currently lacking and that is urgently needed [27,28]. A reliable reference data set would help to improve development of accurate measurement of deep GM atrophy in MS patients. Such data sets are already available for the hippocampus, aimed at research on Alzheimer’s Disease such as the Harmonized Hippocampal Protocol (HarP) [29–31]. Several studies have since used the HarP gold standard data set as ground truth for improvement of volumetric measurements of the hippocampus [32–34].

Although the advantage of using FASTSURF would be the drastic reduction in time needed to create a reference dataset, its quality is crucial. In the hippocampus segmentation study of Bocchetta and colleagues, the inter-rater agreement was high, with an average similarity coefficient of 0.73 between five raters and excellent ICC of 0.88 [25,31]. In this study we observed similarly high levels of overlap: the average $CI_{gen}$ of FASTSURF structures was 0.74± 0.05, 0.73±0.06 and 0.75±0.06 for caudate nucleus, putamen and thalamus, respectively, indicating that our reference dataset was of comparable quality as that of Bocchetta et al.. Moreover, the high agreement between manual and FASTSURF inter-rater $CI_{gen}$ values

indicates that use of FASTSURF will not compromise the quality of the reference data.

Our results showed that FASTSURF segmentation were highly correlated with manual segmentations for all three structures (Fig.2 and Table 5), but the ICC of the caudate nucleus is slightly lower compared to putamen and thalamus. We suggest that this is due to the different shapes of the structures. The tail of the caudate is substantially more elongated and curved compared to the other structures, which would make it more challenging to delineate manually as well as with FASTSURF.

Near-isotropic, ~1mm$^3$ voxel size, 3D images were used, requiring on average 53, 43 and 35 contours in the coronal plane for fully manual segmentations of the caudate nucleus, putamen and thalamus respectively (Table E of the supplementary data provides the average numbers of contours in the axial and sagittal planes). In theory, when outlining only 10 contours in one plane for each structure, segmentation time could be reduced 3-5 times. However, in clinical MRI scans for MS diagnostics, images are not always acquired with near-isotropic 1mm$^3$ voxels [35], and the time-reductions for each structure might differ.



The training set was used to find optimal settings for FASTSURF. Different settings, e.g. orientation, showed sometimes extremely small differences in the Jaccard between the FASTSURF and the manually outlined structure. For example, the mean volume overlap of the thalamus was 0.918±0.025 mL when delineating contours in the axial plane and 0.922±0.023 mL when delineating them in the sagittal plane (see Tables A-D of the Supplementary materials for all results obtained on the training data). Therefore, FASTSURF could perform similarly well with different settings than the optimal settings. This has important practical advantages because it may be more convenient for the raters to outline all structures in the same orientation rather than having to use different viewing orientations for each structure. Similarly, we expect that drawing 6 or 8, instead of 10 contours, would lead to a more time efficient protocol without much loss of accuracy as can be seen from the training set results in Table D in the Supplementary materials.

A possible limitation of this study is that we extracted the contours from the fully manual outlines instead of making new sparse delineations with a given number of contours. This possible source of bias was previously quantified for FASTSURF segmentation of the hippocampus using scan-rescan data, showing this bias to be small [18]. Unfortunately, given the relatively low number of patients and given the lack of scan-rescan data, a bias estimation was not feasible for our study. Nevertheless, we expect that for geometrically simpler structures, with larger volumes such as the putamen and the thalamus, the bias is even smaller than for the hippocampus tested in [18]. In future prospective studies, partial contours should

be generated independently of full contours to investigate FASTSURF accuracy without any possible bias. Lastly, while FASTSURF can be used to train improved cross-sectional segmentation and volume quantification techniques, whether reference data for capturing volume change over time (atrophy) can also be reliably created in this manner remains to be investigated for FASTSURF as well as for other methods.

## 5.5. Conclusion

In this study, we showed that FASTSURF segmentations exhibit excellent agreement with manual segmentations. Moreover, the optimized settings (viewing orientation, number of slices, etc.) derived from the training dataset can be incorporated into a protocol to sparsely segment these structures, which can be implemented in the FASTSURF workflow. Therefore, we conclude that using this semi-automated method, a set of reference segmentations could be created at a considerably reduced workload compared to manual outlining, thus generating opportunities for improving deep GM atrophy measurement in MS and other neurological diseases.

# Supplementary data

**Table A: Jaccard of reconstructed FASTSURF and fully manual segmented structures with Ninter = 4 and Npoints = 100.**

| Structure | Ncontours | Coronal plane [a] | Axial plane [a] | Sagittal plane [a] |
|---|---|---|---|---|
| Caudate nucleus | 4 | 0.586±0.044 | 0.709±0.047 | 0.619±0.121 |
| | 6 | 0.764±0.033 | 0.820±0.030 | 0.749±0.107 |
| | 8 | 0.842±0.028 | 0.842±0.029 | 0.787±0.079 |
| | 10 | 0.862±0.031 | 0.868±0.032 | 0.819±0.062 |
| Putamen | 4 | 0.712±0.038 | 0.809±0.037 | 0.804±0.053 |
| | 6 | 0.867±0.022 | 0.867±0.032 | 0.856±0.041 |
| | 8 | 0.884±0.024 | 0.891±0.036 | 0.888±0.037 |
| | 10 | 0.901±0.027 | 0.895±0.032 | 0.891±0.038 |
| Thalamus | 4 | 0.803±0.033 | 0.819±0.036 | 0.823±0.046 |
| | 6 | 0.879±0.028 | 0.878±0.027 | 0.886±0.030 |
| | 8 | 0.903±0.022 | 0.911±0.024 | 0.912±0.030 |
| | 10 | 0.918±0.017 | 0.918±0.025 | 0.922±0.023 |

[a] Mean±standard deviation; *Ninter* Amount of intermediate lines used between the sparse contours for reconstruction, *Npoints* Amount of points on a sparse contour for reconstruction, *Ncontours* Amount of extracted sparse contours.

**Table B: Jaccard of reconstructed FASTSURF and fully manual structures with optimal plane orientation (see Table A) and Npoints= 100.**

| Structure | Ncontours | Ninter = 3 [a] | Ninter = 4 [a] | Ninter = 5 [a] |
|---|---|---|---|---|
| Caudate nucleus | 4 | 0.732±0.048 | 0.731±0.047 | 0.731±0.048 |
| | 6 | 0.827±0.030 | 0.826±0.030 | 0.827±0.030 |
| | 8 | 0.840±0.030 | 0.840±0.029 | 0.840±0.029 |
| | 10 | 0.873±0.032 | 0.873±0.032 | 0.870±0.031 |
| Putamen | 4 | 0.814±0.037 | 0.813±0.037 | 0.807±0.038 |
| | 6 | 0.860±0.032 | 0.859±0.032 | 0.855±0.032 |
| | 8 | 0.881±0.036 | 0.880±0.036 | 0.878±0.036 |
| | 10 | 0.895±0.027 | 0.894±0.027 | 0.893±0.026 |
| Thalamus | 4 | 0.822±0.046 | 0.823±0.046 | 0.822±0.045 |
| | 6 | 0.886±0.030 | 0.886±0.030 | 0.885±0.028 |
| | 8 | 0.912±0.030 | 0.911±0.030 | 0.910±0.030 |
| | 10 | 0.924±0.022 | 0.922±0.023 | 0.921±0.023 |

[a] Mean±standard deviation; *Ninter* Amount of intermediate lines used between the sparse contours for reconstruction, *Npoints* Amount of points on a sparse contour for reconstruction, *Ncontours* Amount of extracted sparse contours.

**Table C: Jaccard of reconstructed FASTSURF and fully manual segmented structures with optimal plane orientation and Ninter (see Table A and B).**

| Struct. | Ncont. | Npoints [a] | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | 30 | 50 | 80 | 100 | 150 | 200 | 250 | 300 |
| Caud. nucl. | 4 | 0.586± 0.044 | 0.709± 0.047 | 0.619± 0.121 | 0.710± 0.048 | 0.712± 0.048 | 0.713± 0.050 | 0.714± 0.050 | 0.714± 0.051 |
| | 6 | 0.764± 0.033 | 0.820± 0.030 | 0.749± 0.107 | 0.821± 0.030 | 0.821± 0.029 | 0.821± 0.030 | 0.820± 0.029 | 0.821± 0.030 |
| | 8 | 0.842± 0.028 | 0.842± 0.029 | 0.787± 0.079 | 0.842± 0.029 | 0.841± 0.030 | 0.841± 0.029 | 0.840± 0.029 | 0.840± 0.028 |
| | 10 | 0.862± 0.031 | 0.868± 0.032 | 0.819± 0.062 | 0.869± 0.032 | 0.869± 0.031 | 0.869± 0.033 | 0.869± 0.033 | 0.869± 0.033 |
| Puta. | 4 | 0.792± 0.044 | 0.807± 0.041 | 0.807± 0.041 | 0.811± 0.037 | 0.813± 0.037 | 0.814± 0.037 | 0.814± 0.036 | 0.815± 0.036 |
| | 6 | 0.842± 0.032 | 0.863± 0.032 | 0.869± 0.032 | 0.869± 0.032 | 0.870± 0.033 | 0.871± 0.032 | 0.872± 0.031 | 0.872± 0.032 |
| | 8 | 0.857± 0.032 | 0.883± 0.034 | 0.890± 0.034 | 0.892± 0.036 | 0.891± 0.036 | 0.892± 0.037 | 0.892± 0.037 | 0.892± 0.037 |
| | 10 | 0.875± 0.024 | 0.893± 0.024 | 0.899± 0.024 | 0.902± 0.027 | 0.903± 0.025 | 0.903± 0.025 | 0.903± 0.025 | 0.903± 0.025 |
| Thal. | 4 | 0.829± 0.045 | 0.828± 0.044 | 0.822± 0.045 | 0.823± 0.046 | 0.821± 0.046 | 0.822± 0.046 | 0.823± 0.046 | 0.824± 0.046 |
| | 6 | 0.874± 0.027 | 0.885± 0.027 | 0.887± 0.029 | 0.886± 0.030 | 0.886± 0.030 | 0.887± 0.029 | 0.887± 0.030 | 0.888± 0.029 |
| | 8 | 0.888± 0.031 | 0.902± 0.029 | 0.913± 0.030 | 0.912± 0.030 | 0.914± 0.030 | 0.913± 0.031 | 0.914± 0.031 | 0.914± 0.031 |
| | 10 | 0.900± 0.021 | 0.915± 0.021 | 0.925± 0.022 | 0.924± 0.022 | 0.926± 0.025 | 0.925± 0.025 | 0.926± 0.025 | 0.927± 0.025 |

[a] Mean±standard deviation; *Struct.* Structure, *Caud. nucl.* Caudate nucleus, *Puta.* Putamen, *Thal.* Thalamus, *Ninter* Amount of intermediate lines used between the sparse contours for reconstruction, *Npoints* Amount of points on a sparse contour for reconstruction, *Ncont.* Amount of extracted sparse contours.

**Table D: Jaccard of reconstructed FASTSURF and fully manual segmented structures with optimal plane orientation, Ninter and Npoints (see Table A, B and C).**

| | Ncontours [a] | | | |
|---|---|---|---|---|
| Structure | 4 | 6 | 8 | 10 |
| Caudate nucleus | 0.714±0.05 | 0.821±0.03 | 0.842±0.029 | 0.869±0.032 |
| Putamen | 0.815±0.036 | 0.872±0.032 | 0.892±0.037 | 0.903±0.025 |
| Thalamus | 0.824±0.046 | 0.888±0.029 | 0.914±0.031 | 0.927±0.025 |

[a] Mean±standard deviation; *Ninter* Amount of intermediate lines used between the sparse contours for reconstruction, *Npoints* Amount of points on a sparse contour for reconstruction, *Ncontours* Amount of extracted sparse contours.

**Table E: Mean number of slices for each structure per plane orientation for the left and right hemisphere.**

| Structure | Side | Axial plane | Coronal plane | Sagittal plane |
|---|---|---|---|---|
| Caudate nucleus | Left | 19 | 52 | 29 |
| | Right | 20 | 54 | 31 |
| Putamen | Left | 25 | 42 | 21 |
| | Right | 20 | 43 | 26 |
| Thalamus | Left | 17 | 36 | 28 |
| | Right | 26 | 33 | 26 |

# References


[1] Geurts JJG, Pouwels PJW, Uitdehaag BMJ, Polman CH, Barkhof F, Castelijns JA. Intracortical lesions in multiple sclerosis: improved detection with 3D double inversion-recovery MR imaging. Radiology 2005;236:254–60.

[2] Bagnato F, Butman JA, Gupta S, Calabrese M, Pezawas L, Ohayon JM, et al. In vivo detection of cortical plaques by MR imaging in patients with multiple sclerosis. AJNR Am J Neuroradiol 2006;27:2161–7.

[3] Fisher C, Lindhorst H, Matthews T, Munroe DJ, Paulin D, Scott D. Nursing staff attitudes and behaviours regarding family presence in the hospital setting. J Adv Nurs 2008;64:615–24.

[4] Bermel R a, Innus MD, Tjoa CW, Bakshi R. Selective caudate atrophy in multiple sclerosis: a 3D MRI parcellation study. Neuroreport 2003;14:335–9.

[5] Bishop CA, Newbould RD, Lee JSZ, Honeyfield L, Quest R, Colasanti A, et al. Analysis of ageing-associated grey matter volume in patients with multiple sclerosis shows excess atrophy in subcortical regions. NeuroImage Clin 2017;13:9–15.

[6] Houtchens MK, Benedict RHB, Killiany R, Sharma J, Jaisani Z, Singh B, et al. Thalamic atrophy and cognition in multiple sclerosis. Neurology 2007;69:1213–23.

[7] Pubmed E, M F. Regional brain atrophy evolves differently in patients with multiple sclerosis according to clinical phenotype . Clinical and MRI outcome after autologous hematopoietic stem cell transplantation in MS . MRI / MRS of corpus callosum in patients with clinic. Magn Reson Imaging 2005;c:1–21.

[8] Schoonheim MM, Hulst HE, Brandt RB, Strik M, Wink AM, Uitdehaag BMJ, et al. Thalamus structure and function determine severity of cognitive impairment in multiple sclerosis. Neurology 2015;84:776–83.

[9] Meijerman A, Amiri H, Steenwijk MD, Jonker MA, Van Schijndel RA, Cover KS, et al. Reproducibility of deep gray matter atrophy rate measurement in a large multicenter dataset. Am J Neuroradiol 2018;39:46–53.

[10] Gelineau-Morel R, Tomassini V, Jenkinson M, Johansen-Berg H, Matthews PM, Palace J. The effect of hypointense white matter lesions on automated gray matter segmentation in multiple sclerosis. Hum Brain Mapp 2012;33:2802–14.

[11] Popescu V, Ran NCG, Barkhof F, Chard DT, Wheeler-Kingshott CA, Vrenken H. Accurate GM atrophy quantification in MS using lesion-filling with co-registered 2D lesion masks. NeuroImage Clin 2014;4:366–73.

[12] Tams PS, Mcdonald WI, Miller DH. Original QUANTIFICATION OF MRI LESION LOAD IN MULTIPLE SCLEROSIS : A COMPARISON. Magn Reson Imaging 1996;14:495–505.

[13] Paty C. D.; Grochowski, E.; Palmer, M. R.; Oger, J.; Kastrukoff, L. F. DW. I. Magnetic resonance imaging (MRI) in multiple sclerosis (MS): a serial study in relapsing and remitting patients with quantitative measurements of lesion size. Neurology 1986;36:177.

[14] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55.

[15] Bogovic JA, Jedynak B, Rigg R, Du A, Landman BA, Prince JL, et al. Approaching expert results using a hierarchical cerebellum parcellation protocol for multiple inexpert human raters. Neuroimage 2013;64:616–29.

[16] Landman BA, Asman AJ, Scoggins AG, Bogovic JA, Stein JA, Prince JL. Foibles, follies, and fusion: Web-based collaboration for medical image labeling. Neuroimage 2012;59:530–9.

[17] SPINE donne le feu vert à la science participative - Université de Bordeaux n.d. https://www.u-bordeaux.fr/Actualites/De-la-recherche/SPINE-donne-le-feu-vert-a-la-science-participative (accessed May 16, 2018).

[18] Bartel F, Vrenken H, van Herk M, de Ruiter M, Belderbos J, Hulshof J, et al. Fast Segmentation Through Surface Fairing (FASTSURF): A novel semi-automatic hippocampus segmentation method. Under Review. PLoS One 2018.

[19] Bartel F, Vrenken H, van Herk M, de Ruiter MB, Belderbos J, Hulshof J, et al. Semi-automatic hippocampus delineation algorithm using surface fairing. IFMBE Proc., vol. 65, Springer, Singapore; 2017, p. 956–9.

[20] Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage 2011;56:907–22.

[21] Reuter M, Schmansky NJ, Rosas HD, Fischl B. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 2012;61:1402–18.

[22] Rocca MA, Valsasina P, Hulst HE, Abdel-Aziz K, Enzinger C, Gallo A, et al. Functional correlates of cognitive dysfunction in multiple sclerosis: A multicenter fMRI Study. Hum Brain Mapp 2014;35:5799–814.

[23] Ropele S, Kilsdonk ID, Wattjes MP, Langkammer C, De Graaf WL, Frederiksen JL, et al. Determinants of iron accumulation in deep grey matter of multiple sclerosis patients. Mult Scler J 2014;20:1692–8.

[24] Koch GG. Intraclass correlation coefficient. In: Samuel K. and Norman L.J. Encyclopedia of Statistical Sciences 4. New York, John Wiley & Sons; 213–17 1982.

[25] Cicchetti D V. Multiple Comparison Methods: Establishing Guidelines for their Valid Application in Neuropsychological Research. J Clin Exp Neuropsychol 1994;16:155–61.

[26] Kouwenhoven E, Giezen M, Struikmans H. Measuring the similarity of target volume delineations independent of the number of observers. Phys Med Biol 2009;54:2863–73.

[27] Amiri H, de Sitter A, Bendfeldt K, Battaglini M, Wheeler-Kingshott CA, Calabrese M, et al. Urgent challenges in quantification and interpretation of grey matter atrophy in multiple sclerosis, Under Review 2017.

[28] Rocca MA, Battaglini M, Benedict RHB, De Stefano N, Geurts JJG, Henry RG, et al. Brain MRI atrophy quantification in MS. Neurology 2017;88:403–13.

[29] Boccardi M, Bocchetta M, Apostolova LG, Barnes J, Bartzokis G, Corbetta G, et al. Delphi definition of the EADC-ADNI Harmonized Protocol for hippocampal segmentation on magnetic resonance. Alzheimer's Dement 2015;11:126–38.

[30] Boccardi M, Bocchetta M, Morency C, Collins DL, Nishikawa M, Ganzola R, et al. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol , for the EADC-ADNI Working Group on The Harmonized Protocol for Manual Hippocampal Segmentation and for the Alzheimer's Disease Neuroimaging Initiative. Alzheimer's Dement 2015;11:175–83.

[31] Bocchetta M, Boccardi M, Ganzola R, Apostolova LG, Preboske G, Wolf D, et al. Harmonized benchmark labels of the hippocampus on magnetic resonance: The EADC-ADNI project. Alzheimer's Dement 2015;11:151–160.e5.

[32] Ahdidan J, Raji CA, DeYoe EA, Mathis J, Noe KO, Rimestad J, et al. Quantitative neuroimaging

software for clinical assessment of hippocampal volumes on MR imaging. J Alzheimer's Dis 2016;49:723–32.

[33] Inglese P, Amoroso N, Boccardi M, Bocchetta M, Bruno S, Chincarini A, et al. Multiple RF classifier for the hippocampus segmentation: Method and validation on EADC-ADNI Harmonized Hippocampal Protocol. Phys Medica 2015;31:1085–91.

[34] Weiner MW, Veitch DP, Aisen PS, Beckett LA, Cairns NJ, Green RC, et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's Dement 2017;13:e1–85.

[35] Vrenken H, Jenkinson M, Horsfield MA, Battaglini M, Van Schijndel RA, Rostrup E, et al. Recommendations to improve imaging and analysis of brain lesion load and atrophy in longitudinal studies of multiple sclerosis. J Neurol 2013;260:2458–71.