Published in final edited form as:

Magn Reson Imaging. 2017 June ; 39: 110–122. doi:10.1016/j.mri.2017.02.002.

An Improved FSL-FIRST Pipeline for Subcortical Gray Matter Segmentation to Study Abnormal Brain Anatomy Using Quantitative Susceptibility Mapping (QSM)

Xiang Feng¹, Andreas Deistung^{1,2,3}, Michael G. Dwyer⁴, Jesper Hagemeier⁴, Paul Polak⁴, Jessica Lebenberg⁵, Frédérique Frouin⁶, Robert Zivadinov^{4,7}, Jürgen R. Reichenbach^{1,8,9}, and Ferdinand Schweser^{4,7}

¹Medical Physics Group, Institute of Diagnostic and Interventional Radiology, Jena University Hospital - Friedrich Schiller University Jena, Jena, Germany

²Section of Experimental Neurology, Department of Neurology, Essen University Hospital, Essen, Germany

³Erwin L. Hahn Institute for Magnetic Resonance Imaging, University Duisburg-Essen, Essen, Germany

⁴Buffalo Neuroimaging Analysis Center, Dept. of Neurology, Jacobs School of Medicine and Biomedical Sciences, State University of New York at Buffalo, Buffalo, NY, United States

⁵UNATI, CEA DRF/I2BM, INSERM, Université Paris-Sud, Université Paris-Saclay, NeuroSpin center, Gif/Yvette, France

⁶Inserm/CEA/Université Paris Sud/CNRS, CEA/I2BM/SHFJ, Laboratoire IMIV, Orsay, France

⁷MRI Molecular and Translational Imaging Center, Buffalo CTRC, State University of New York at Buffalo, Buffalo, NY, United States

⁸Center of Medical Optics and Photonics, Friedrich Schiller University Jena, Germany

⁹Michael Stifel Center for Data-driven and Simulation Science Jena, Friedrich Schiller University Jena, Germany

Abstract

Accurate and robust segmentation of subcortical gray matter (SGM) nuclei is required in many neuroimaging applications. FMRIB's Integrated Registration and Segmentation Tool (FIRST) is one of the most popular software tools for automated subcortical segmentation based on T₁-weighted (T1w) images. In this work, we demonstrate that FIRST tends to produce inaccurate SGM segmentation results in the case of abnormal brain anatomy, such as present in atrophied brains, due to a poor spatial match of the subcortical structures with the training data in the MNI

Corresponding author Xiang Feng, M.Sc., Medical Physics Group, Institute of Diagnostic and Interventional Radiology, Jena University Hospital - Friedrich Schiller University Jena, Philosophenweg 3, 07743 Jena, Germany, xiang.feng@med.uni-jena.de.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

space as well as due to insufficient contrast of SGM structures on T1w images. Consequently, such deviations from the average brain anatomy may introduce analysis bias in clinical studies, which may not always be obvious and potentially remain unidentified. To improve the segmentation of subcortical nuclei, we propose to use FIRST in combination with a special Hybrid image Contrast (HC) and Non-Linear (nl) registration module (HC-nlFIRST), where the hybrid image contrast is derived from T1w images and magnetic susceptibility maps to create subcortical contrast that is similar to that in the Montreal Neurological Institute (MNI) template. In our approach, a nonlinear registration replaces FIRST's default linear registration, yielding a more accurate alignment of the input data to the MNI template. We evaluated our method on 82 subjects with particularly abnormal brain anatomy, selected from a database of more than 2000 clinical cases. Qualitative and quantitative analyses revealed that HC-nlFIRST provides improved segmentation compared to the default FIRST method.

Keywords

Quantitative susceptibility mapping; MRI; brain; segmentation; FIRST; subcortical gray matter

1. Introduction

Robust and accurate brain image segmentation is an essential prerequisite for a sensitive and unbiased identification of differences in MR image intensity or volume of brain structures between normal and pathologically affected brains. Clinical imaging studies encompassing large cohorts of patients and controls commonly apply automatic image segmentation tools. To date, several sophisticated software packages have been developed for segmenting subcortical gray matter (SGM) [1–4], most of them relying on high-resolution T₁-weighted (T1w) images. Among them, FMRIB's Integrated Registration and Segmentation Tool (FIRST) [3] of the FMRIB Software Library (FSL) is one of the most popular packages. FIRST has been applied in multiple brain imaging studies to investigate volume and shape changes of subcortical brain structures that may be associated with normal aging [5, 6] or neurodegenerative diseases, like Alzheimer's disease [7, 8], schizophrenia [9], multiple sclerosis [10], or epilepsy [11]. It has also been shown quite recently that FIRST performs similarly well regarding hippocampal segmentation to predict the progress in Alzheimer's Disease dementia when compared to other volumetric methods (e.g., Statistical Parametric Mapping) [12]. The scan-rescan reliability of SGM segmentation with FIRST on the same scanning platform [13] as well as between different scanning platforms [14] has been confirmed, which warrants the applicability of FIRST in large-scale longitudinal and multisite studies.

Brain atrophy is a hallmark of many neurological diseases, including stroke, traumatic brain injury or Alzheimer's disease [15–17], and has long been one of the primary targets of neuroimaging research. Specifically, it is correlated with disability in multiple sclerosis (MS) [18]. Consequently, if suboptimal and less accurate segmentation results occur with brain anatomies that show deviations from the anatomy of the training datasets, an analysis bias may consequently follow in clinical studies, which commonly compare patients with abnormal brains with matched controls with normal brains. This bias may not always be

obvious and may even remain unidentified, especially in large-cohort studies. In this study, we therefore aim to improve the accuracy of FIRST-based segmentation of subjects with abnormal brain anatomy due to severe atrophy (see Fig. 1).

FIRST is a model-based subcortical brain segmentation tool, which models each subcortical structure as a surface mesh within a Bayesian framework, using shape and signal intensity information of manually segmented subcortical structures in 336 T1w datasets (including, but not limited to, normal brains and cases of schizophrenia and Alzheimer's diseases) as a prior [3]. The individual T1w images are linearly (affine) registered to the Montreal Neurological Institute (MNI) space using a two-stage registration with weighting of subcortical structures during the second stage. Subsequently, applying the inverse transformation brings the Bayesian model into the native space of the individual T1w images. Small inaccuracies resulting from the affine registration step are usually overcome by FIRST by applying sophisticated fine-tuning of the surface mesh. Because it may be difficult or even impossible to straighten out fully severe registration inaccuracies by the surface mesh optimization, there is consequently an intrinsic demand to transform as accurately as possible the individual subject data into the MNI space. While it is usually possible to transform brain images of healthy adults relatively accurately to the MNI space, affine registration of brain images of patients or individuals with substantial brain atrophy may produce residual misalignment in various brain structures (see Fig. 2). One reason for this is that the MNI template originates from a healthy adult population (age 18.5–43.5 years) [19, 20] and that affine registration is unable to deform and resolve local shape dissimilarities. Recently, Amann *et al.* [21] reported that linear registration to MNI space within the FIRST pipeline is prone to fail with T1w images of patients showing global brain atrophy, and suggested to replace the MNI template by a new template that includes advanced atrophy. However, while such a modification might result in a better matching of abnormal cases, it does not account for the need to affect FIRST's internal data modelling of the residual shape variance. The use of sub-group specific templates may also introduce further types of bias in the analysis of different subject groups. In addition, region-specific and/or pathology-related brain atrophy may occur [22–24], which cannot be covered by only a single template. Therefore, a more general approach is required to account for individual shape dissimilarities in cases of diseased brains.

Another important issue regarding FIRST-based SGM segmentation is the observation that, depending on T1w sequence parameters of the clinical imaging protocols, SGM nuclei may exhibit deviating or even poor contrast on T1w images compared to that in FIRST's training datasets, which may further affect the segmentation accuracy [25]. While the MNI template shows detailed delineation of SGM structures with excellent contrast, a similar contrast is desirable to have available with FIRST to improve inferior segmentation results in cases of insufficient contrast on T1w images. Against this background, we aim to enhance the contrast of T1w images to render their appearance more similar to the MNI template by combining T1w images with a second MRI dataset that displays SGM nuclei with distinct contrast. Quantitative susceptibility mapping (QSM) is indeed able to provide such a contrast [26–31].

In this contribution, we thus propose and evaluate an improved FIRST segmentation pipeline that replaces the affine transformation to the MNI space by a nonlinear deformation registration to overcome individual brain tissue variations and utilizes composite images with increased SGM contrast created from T1w images and quantitative susceptibility maps.

2. Materials and Methods

2.1. Data acquisition

We retrospectively identified eighty-two subjects with particularly small brain volumes from our research image database at the Buffalo Neuroimaging Analysis Center with more than 2000 subjects who had undergone both T1w imaging and QSM. All subjects had been involved in studies approved by the Institutional Review Board of the University at Buffalo. Informed written consent was available from each individual. We determined the global brain volume for all subjects in the database from the T1w images using SIENAX (Version 2.6) of FSL [32] and subsequently selected the 41 multiple sclerosis (MS) patients and the 41 normal subjects with the smallest brain volumes. Children and adolescents (age < 18 years) were excluded to avoid their inclusion in this brain-volume based selection. All data had been acquired with the same 3T whole-body GE MRI scanner (Signa Excite HD 12.0; GE Healthcare, Milwaukee, Wisconsin) and identical sequence parameters, software, and receiver coils. T1w data had been acquired with a 3D inversion-recovery prepared spoiled gradient echo sequence (IR-SPGR) with the following sequence parameters: echo time (TE) = 2.8 ms, repetition time (TR) = 5.9 ms, inversion time (TI) = 900 ms, flip angle (FA) = 10°, bandwidth (BW) = 240 Hz/px, field of view (FOV) = 256 mm×256 mm×180 mm, matrix size = 256×256×180, no acceleration and no averaging, resulting in an isotropic voxel size of 1 mm³ and an acquisition time of 7:14 min:sec. Data for QSM were acquired using an unaccelerated 3D single-echo spoiled gradient echo (GRE) sequence with first-order flow compensation in read and slice directions, a matrix of 512×192×64 and a nominal resolution of 0.5mm×1mm×2mm (FOV = 256×192×128 mm³), flip angle = 12°, TE/TR=22ms/40ms, bandwidth=13.89 kHz, and no averaging, resulting in an acquisition time of 8:46 min:sec. Raw *k*-space data had been saved for each channel for off-line reconstruction of images.

2.2. Data Preprocessing

Non-parametric non-uniform intensity normalization (N4ITK) [33] was applied to retrospectively correct intensity inhomogeneity on the T1w images (bias-field correction). Subsequently, the signal intensity of the T1w images was normalized (scaled) to a predefined mean value of 110 in white matter (WM). The scaling factor was extracted from the mean signal value of WM on the bias-field corrected T1w images after warping the predefined WM mask from the MNI template into the T1w native space using nonlinear ANTs registration (see below). All warped WM masks were inspected carefully to ensure that they did not include hypointense WM lesions in case of the MS subjects. The target normalization value was chosen to be consistent with that of the intensity normalization tool ‘*mri_normalize*’ of the Freesurfer library (<http://surfer.nmr.mgh.harvard.edu/>). This normalization step was used to make the weighting coefficients (w_1 , w_2) in the hybrid contrast (see below) insensitive to the arbitrary signal scaling of the T1w images. The bias-

field corrected and intensity normalized T1w images will be referred to hereinafter as *T1n* images.

The complex GRE raw data were used to reconstruct the susceptibility maps. To achieve isotropic in-plane resolution, *k*-space was zero-padded in phase-encoding direction resulting in a 512×512×64 matrix. After Fourier transform, the single channel magnitude and phase images were combined using the sum-of-squares approach [34] and scalar phase matching [35], respectively. In-plane distortions due to imaging gradient non-linearity were compensated for [36]. Phase images were unwrapped with a best-path algorithm [37], background-field corrected with V-SHARP [38, 39] (radii 0.5 mm to 5 mm; TSVD threshold 0.05), and converted to magnetic susceptibility maps using homogeneity enabled incremental dipole inversion (HEIDI) [40]. Magnetic susceptibility was referenced to the average susceptibility of the brain (0 ppm). All QSM-related data processing was performed with in-house developed algorithms in MATLAB (2013b, The MathWorks, Natick, MA).

2.3. HC-nIFIRST

In this section, we first describe the preparation of manual reference ROIs and the generation of the hybrid contrast before providing a detailed description of the proposed modification of the default FIRST segmentation framework. Fig. 3 contains a flowchart of the proposed framework.

2.3.1. Manual reference ROIs—We randomly selected eight subjects (four normal subjects and four MS patients) from the 82 subjects to serve as training datasets. Using Freeview of the Freesurfer library, manual segmentation (by X.F., 3 years training experience) of SGM structures (caudate nucleus, putamen, accumbens, hippocampus) was performed using *T1n* images, while the thalamus and globus pallidus were delineated by viewing both *T1n* and QSM images simultaneously. We used FMRIB's Automated Segmentation Tool (FAST) [41] to generate ROIs for WM, cortical gray matter (CGM), and cerebrospinal fluid (CSF) based on *T1n* images. The CGM ROI was created by excluding (masking out) the manually delineated SGM structures in the FAST segmented gray matter volume. These automatically determined ROIs were inspected visually and corrected manually where necessary. Subsequently, all brain structures (WM, CGM, CSF, and the six SGM regions) were outlined manually on the FSL's MNI template with 1 mm spatial resolution.

2.3.2. Creation of the HC images—With QSM providing superior contrast for SGM structures compared to T1w images, we hypothesize that combining susceptibility contrast with T1w contrast results in an image contrast that improves SGM segmentation accuracy without having to modify the training data of FIRST. Such a hybrid image contrast (HC) can be created by linearly combining T1w images and quantitative susceptibility maps according to

$$HC = w_1 \cdot T1w + w_2 \cdot QSM, \quad [1]$$

where w_1 and w_2 are weighting coefficients. Having no better reference available, we used the contrast in the MNI template as the target for the coefficient optimization. Although this reference may not be identical to the average contrast in the FIRST training data, it represents a typical T1 contrast with excellent delineation of SGM structures. Thus, we propose to choose w_1 and w_2 in a way that the contrast of the HC image becomes similar to that of the MNI template.

Optimized values of w_1 and w_2 can be derived by determining the mean signal intensities of specific brain regions on the $T1n$ images, susceptibility maps and the MNI template, and solving the following minimization (weighted least square) problem

$$\operatorname{argmin}_{w_1, w_2} \left\| \Phi \begin{bmatrix} I_1^{T1n} & I_1^{QSM} \\ \vdots & \vdots \\ I_N^{T1n} & I_N^{QSM} \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} - \begin{bmatrix} I_1^{MNI} \\ \vdots \\ I_N^{MNI} \end{bmatrix} \right\|_2 \quad [2]$$

where N is the number of brain regions, Φ is an $N \times N$ diagonal weighting matrix, I_i^{T1n} , I_i^{QSM} , and I_i^{MNI} are the mean values of the $T1n$ data, susceptibility maps, and MNI template in the i -th region, respectively. As the variances of the input data I_i^{T1n} , I_i^{QSM} are known we incorporated the weighting matrix, Φ , to increase the system's stability (robustness) over ordinary least squares solutions. The diagonal elements Φ_{ii} in the weighting matrix are given by summing up the inverse variances of the $T1n$ data and susceptibility maps in the i th region.

The mean intensity values were then determined in each manual reference ROI on the $T1n$ images, quantitative susceptibility maps, and the MNI template. The weighting coefficients for the HC were calculated for each of the four subjects of the two training groups by using Eq. (2) and subsequently applied to generate individual HC images. Since image intensities of the $T1n$ images and the susceptibility maps can be different between normal subjects and MS patients, the calculated coefficients can be also distinctly different between these two groups. To address this potential problem, the weighting coefficients were averaged across the normal control and MS group, respectively, and subsequently used to create HC images according to Eq. (1) using the $T1n$ images and the susceptibility maps from the corresponding group of the remaining 74 subjects.

2.3.3. HC-nFIRST pipeline—To account for potential subject movement between the two MRI acquisitions, the GRE magnitude images were registered to the $T1n$ images by using a linear, rigid body model and normalized mutual information as the cost function [42]. Based on the resulting transformation matrix, the susceptibility maps were then transformed into the T1w image space. Rigid registration was considered to be sufficient in aligning QSM to the T1w image space as both data sets share the exact same anatomy and head motion between scans can be mastered with linear transformation. Following generation of the HC images in the T1w image space (see Sec. 2.3.1), we registered the HC images to the MNI template using Advanced Normalization Tools (ANTs, [*Magn Reson Imaging*. Author manuscript; available in PMC 2018 June 01.](http://</p>
</div>
<div data-bbox=)

stnava.github.io/ANTs/) with an initial affine registration step followed by a non-linear registration employing a non-rigid diffeomorphic registration scheme [43]. We decided to use ANTs as it had been shown to provide the most consistent and accurate registration results among 14 different registration methods [44]. Subsequently, we applied FIRST (using the FSL script *run_first_all*) to the hybrid contrast images in MNI space with a disabled FLIRT (FMRIB's Linear Image Registration Tool) module, which was achieved by supplying an identity matrix as the transformation matrix to MNI space (in the mentioned FSL script above). Finally, we inversely warped each resulting ROI from the MNI space to the individual T1w image space with nearest neighbor interpolation.

2.4 Data analysis

2.4.1 Similarity between HC images and the MNI template—To assess the similarity between the HC and the *T1n* images with respect to the MNI template, we calculated the structural similarity (SSIM) index. With the assumption that the human visual system is highly adapted to extract structural information from its visual field, SSIM is based on three characteristics including luminance, contrast and structure similarity between images, and provides an objective method for assessing perceptual image quality [45]. We used the Wilcoxon signed rank test to compare statistically the SSIM indices between *T1n* images and the MNI template as well as between the HC images and the MNI template.

2.4.2 Effect of non-linear registration—Since registration to the MNI space is crucial for accurate FIRST segmentation, we systematically compared the original (linear) FLIRT registration module with the nonlinear ANTs registration (as specified in Sec. 2.3.3). The default FLIRT approach is a two-stage affine transformation. In the first step, the whole data set is linearly registered to the MNI template with 12 degrees of freedom. In the second step, the registration is refined by optimizing the transformation for subcortical regions. We applied both the FLIRT and ANTs registration approaches separately to T1w, *T1n* and the HC images to investigate the influence of image contrast on registration.

Similar to the evaluation measures in [44] we used the Dice coefficient and volume overlap to assess how well individual SGM regions and total brain volumes were registered to the MNI template. We transformed the manual reference ROIs from the T1w image space to the MNI space using either only the transformation matrix (for FLIRT) or both the transformation matrix and the warping field (for ANTs) with nearest neighbor interpolation. The Dice coefficient and volume overlap were calculated as measures for accuracy and sensitivity, respectively, as follows:

$$\text{Dice} = 2 \frac{|V_{\text{man}} \cap V_{\text{ref}}|}{|V_{\text{man}}| + |V_{\text{ref}}|} \quad [3]$$

and

$$\text{overlap} = \frac{|V_{\text{man}} \cap V_{\text{ref}}|}{|V_{\text{ref}}|}, \quad [4]$$

where $|V|$ denotes the size (number of voxels) of the ROI V , V_{man} and V_{ref} are the ROIs of the warped manual reference ROI in MNI space and the reference ROI of the MNI template, respectively. With each brain structure analyzed individually, we finally averaged the Dice coefficients and volume overlap over the multiple SGM structures.

2.4.3 Comparison of segmentation methods—To compare the performance of HC-nlFIRST with non-linear registration based segmentation, we analyzed the segmentation results obtained by applying inverse warping and transformation of the manually outlined reference regions (see Fig. 4) in MNI space to the individual T1w image space, where the inverse warping field and the transformation matrix were generated within the HC-nlFIRST pipeline. In the following, we will refer to this solely nonlinear registration based segmentation approach as *atlas-based method*. To assess the potential benefit of using HC images compared to $T1n$ images as the input for FIRST, we applied FIRST with non-linear registration to $T1n$ images, which we refer to as *T1n-nlFIRST* in the following. We also applied default FIRST on both $T1n$ and HC images in the training datasets to investigate whether the usage of HC images improves FIRST segmentation.

To assess the differences among the different segmentation approaches, we qualitatively and quantitatively characterized their performance. First, segmentation results were visually inspected for all 82 subjects by overlaying the outlines of the segmented ROIs onto the $T1n$ images and the correspondingly aligned susceptibility maps. We assessed the segmentation quality according to the matching degree of (i) the positions of all segmented ROIs (caudate nucleus, hippocampus, thalamus, GP, putamen, accumbens) with that of the SGM structures visible on the $T1n$ and QSM images, and (ii) of the outlines of all ROIs with the surfaces of individual SGM brain structures. We rated all six SGM structures together using a 6-point rating scale, similar to the approach in visualization of the internal globus pallidus [46] as follows: 0 – no segmentation results obtained (FIRST aborted); 1 – wrong (unacceptable) segmentation; 2 – poor segmentation; 3 – moderate segmentation; 4 – good segmentation; 5 – perfect segmentation.

For the quantitative assessment of the segmentation quality, Dice coefficients were calculated using Eq. (3) and false negative rates (FNR) were calculated as follows,

$$FNR = \frac{|V_{auto} \cap \overline{V_{man}}|}{|V_{man}|}, \quad [5]$$

where $|V|$ denotes the size (number of voxels) of the ROI V , V_{auto} and V_{man} are the ROIs of the automated and manual reference segmentation, respectively, and $\overline{V_{man}}$ is the region outside V_{man} . We used the manually delineated ROIs of the SGM structures of the training data set as the reference (V_{man}) and computed for each structure the Dice coefficient and FNR in the original T1w image space. A high Dice coefficient indicates high overlap of the automatically segmented region (V_{auto}) and the manually defined reference region (V_{man}), whereas a low FNR indicates that the automatically segmented region less likely includes voxels that do not belong to the reference region. Note that the quantitative Dice and FNR

analyses were carried out only for the eight training subjects, for whom manual reference ROIs existed.

2.4.4 Non-supervised ranking of segmentation methods—Evaluating and ranking the performance of different segmentation methods is important, whereupon one usually needs knowledge of the ground truth (true segmentation). In general, manual segmentation by trained experts is considered as a reference standard that nevertheless may be imperfect and may not exactly represent the true results. However, since we undertook manual segmentation only in a subset of the data and other independent ground truth or reference information was not available for the remaining subject groups, we decided to evaluate and rank the performance of the segmentation methods in a non-supervised manner. To this end, we applied the extended version of the Regression Without Truth (eRWT) approach which can be used to rank different segmentation methods in the absence of a gold standard [47]. The eRWT approach is based on the probability distribution of the true volume of the segmented SGM regions in the population and the ranking of the methods is based on computing a figure-of-merit for each segmentation method. The eRWT approach assumes a parametric relationship between the true value (volume in our case) and the estimated value and relies on the following three hypotheses: (i) the statistical distribution of the true value for the study group has a finite support, (ii) each segmentation method yields an estimated value, which is linearly related to the true value, and (iii) the error terms for each segmentation method are independent. The ranking and the associated subsequent rank analysis is based on applying a bootstrap approach (1000 replicates) to the available database. To describe the probability distribution of the SGM volumes the true beta distribution was adjusted according to the different experimental distributions and the *figure-of-merit* was calculated for each segmentation method as the expectation of the difference between the true values and the estimated values. The smaller the *figure-of-merit* the more accurate (closer to ground truth) the segmentation method. Each beta distribution was refined in order to minimize the sum of the *figures-of-merit* over the different segmentations. The SGM specific configuration parameters for the eRWT analysis are summarized in Table S1 and Fig. S1 of the Supplementary Material.

2.4.5 Statistical analysis—Since multiple methods were evaluated with the same database and the same brain regions, a linear mixed effects model was applied to evaluate differences between the registration methods as well as segmentation methods. In the evaluation of registration, Dice and volume overlap were treated as dependent variables, registration method (FLIRT and ANTs) and image contrast (T1w, T1n or HC) were used as fixed factors, and the subject as a random effect. In the evaluation of segmentation, Dice and FNR values were used as dependent variables, segmentation method (default FIRST, T1n-nlFIRST, HC-nlFIRST, atlas-based) and anatomical regions (putamen, caudate, accumbens, hippocampus, globus pallidus, thalamus) were modeled as fixed factors, and the subject was incorporated as a random effect. The significance of the segmentation method term was evaluated via a likelihood ratio test [48]. Pairwise post-hoc comparison between segmentation methods was performed via Tukey's HSD (honest significant difference) test [49]. To assess the interaction between the anatomical region and the segmentation method, we avoided a complex mixed effect interaction model and opted to perform simply

additional post-hoc tests restricting the analysis to each region in turn. Multiple comparisons were corrected for within regional analyses, but not across them in order to avoid an over-conservative analysis. Multiple comparisons were performed with a dedicated software (R, v3.2.0., <https://www.R-project.org>), while the Wilcoxon signed rank test was calculated with MATLAB (2013b, The MathWorks, Natick, MA). If not explicitly specified otherwise, a corrected p -value below 0.05 was considered as statistically significant.

2.4.6 Automatic quality assessment of segmented results—In large cohort studies, quality assessment of segmentation results is required to track potential failures in the post-processing pipeline to avoid errors in the analysis or misinterpretation. One well established qualitative method is visual inspection of the segmented ROIs, whereas for quantitative assessment by using, e.g., Dice coefficients or other metrics, manually delineated reference ROIs are necessary. However, visual inspection of each subject or manual delineation of ROIs is usually not feasible in large cohort studies. Instead, it might be more efficient to apply an automatic method for the quality assessment. Using Dice coefficients as the metric, we analyzed the quality of segmentation in the MNI space with the smallest possible human involvement. To this end, we used the previously defined subcortical ROIs (Fig. 4) in MNI space as reference. As both T1n-nlFIRST and HC-nlFIRST generate the segmentation results in MNI space, we only needed to transfer the default FIRST segmentation results from the T1w image space to the MNI space. To perform this warping, we applied the nonlinear transformation with nearest neighbor interpolation, where the transformation matrix and the deformation field were generated from the nonlinear ANTs registration of the hybrid contrast image to the MNI space. Then we calculated and compared the Dice coefficients in MNI space for the three segmentation methods of default FIRST, T1n-nlFIRST and HC-nlFIRST. Finally, referring to visual inspection, we checked whether the calculated Dice coefficients were able to reveal the difference between successful and failed segmentations or not and set a threshold to automatically determine whether the segmentation was considered successful or failed. Such measurements only serve to identify automatically failed segmentation cases, but not to disentangle minor differences between successful segmentations.

3. Results

3.1 HC weighting coefficients and similarity to MNI template

The calculated weighting coefficients for creating the HC images in the two training groups ($n=4$ each) are summarized in Table 1. Mean values and standard deviations were $w_1 = 1.48 \pm 0.04$ and $w_2 = (-89.11 \pm 22.81) \text{ ppm}^{-1}$ for MS patients, and $w_1 = 1.57 \pm 0.03$ and $w_2 = (-72.42 \pm 25.57) \text{ ppm}^{-1}$ for normal control subjects. The mean coefficients for the two groups were used to calculate HC images for all remaining 74 subjects not belonging to one of the two training datasets.

Fig. 5 illustrates the contrast enhancement of SGM when combining T1w images and susceptibility maps compared to T1w images. While the globus pallidus and the thalamus were hardly discernible on T1w and T1n images (Fig. 5a,b), these structures became clearly visible with respect to the surrounding white matter on the HC images (Fig. 5d). The mean

SSIM between the HC images and the MNI template ($SSIM_{HCvsMNI} = 0.411 \pm 0.021$, mean and standard deviation) averaged over all 82 subjects was significantly higher ($p < 0.001$) than that between the *T1n* and MNI template images ($SSIM_{T1nvsMNI} = 0.406 \pm 0.019$), indicating that HC images reflect the contrast of the MNI template more closely than *T1n* images.

3.2 Effect of the non-linear registration

Visual comparison of the different registration methods revealed that the default FLIRT module applied to T1w, *T1n* and HC images resulted in misalignment of the brain (i.e., the brain was rotated with respect to the MNI template) in one of the 82 subjects (see Supplementary Fig. S3). Using nonlinear ANTs, all 82 subjects were registered successfully with T1w, *T1n* and HC images. Fig. 6 illustrates the improved alignment to the MNI template with nonlinear registration compared to FLIRT. We found that compared to FLIRT, nonlinear registration with ANTs significantly ($p < 0.001$) improved the accuracy of SGM registration in terms of Dice coefficients and volume overlap, independent of the selected input contrast image (T1w, T1n or HC). While the mean Dice coefficients and FNRs did not significantly vary across the image contrasts, the use of HC images with nonlinear ANTs registration, however, revealed the smallest standard deviations for the Dice coefficients (0.036, compared to 0.048 for both T1w and T1n) and volume overlaps (0.046, compared to 0.060 for both T1w and T1n). We also found that both registration methods consistently returned slightly improved performance (i.e., higher Dice and overlap) for all image contrasts in control subjects compared to MS patients (see Supplementary Table S2).

3.3 Comparison of segmentation methods

The default FIRST method completed the segmentation procedure in 81 of the 82 subjects successfully, and aborted segmentation in one subject (1.2%) with no result. The abortion of FIRST was caused by a failed FLIRT registration. Using T1n-nlFIRST, HC-nlFIRST, and the atlas-based method segmentation succeeded for all 82 subjects with each method. Visual rating of the segmentation results revealed that HC-nlFIRST was superior compared to the other approaches (Fig. 7).

Fig. 8 illustrates the segmentation results for two subjects of the training dataset. Fig. S5 in the Supplementary Material shows the results for two other subjects of the non-training dataset. Best visual conformity to the underlying anatomy was obtained with HC-nlFIRST for the MS patient with severe brain atrophy and considerably enlarged ventricles (Fig. 8a), whereas FIRST, T1n-nlFIRST and atlas-based segmentations returned less accurately segmented regions (white arrows). Though not reaching statistical significance (Fig. 9, Table 2), HC-nlFIRST nevertheless resulted in the best segmentation results, i.e., closest to manual reference. The same observation was found for a control subject (Fig. 8b), where FIRST was unable to depict the putamen, globus pallidus, caudate and accumbens due to insufficient linear registration of the images to be segmented and the MNI template. Again, HC-nlFIRST returned conform depiction of these subcortical structures.

The mean values and standard deviations of the Dice coefficients and FNRs computed using manually delineated ROIs as reference of the training dataset ($n=8$) are shown as bar plots in Fig. 9 and as spreadsheet in Table S3 (in the Supplementary Material) for the four different

segmentation methods. HC-nlFIRST revealed the highest Dice coefficients for all SGM structures compared to the other segmentation methods, indicating that the ROIs provided by HC-nlFIRST match the manual reference ROIs most closely. For most of the structures (except left accumbens), T1n-nlFIRST returned the second highest Dice coefficients, followed by the atlas-based and default FIRST methods. The atlas-based method always exhibited the highest FNRs, indicating that more voxels of a segmented region did not belong to the manual reference region. HC-nlFIRST revealed substantially higher Dice and lower FNR values than the atlas-based method, which is associated with the surface mesh optimization of the FIRST algorithm. The higher FNRs in case of the accumbens compared to the other subcortical structures might arise from the small volume of interest, which is more likely affected by imperfect registration and partial volume effects when transforming the segmentation from the MNI to the individual T1w image space. Reduced Dice coefficients and increased FNR of the accumbens compared to the other subcortical structures have also been observed with the default FIRST approach [14]. Using linear mixed model analysis, we found that the choice of the subcortical segmentation method (FIRST, T1n-nlFIRST, HC-nlFIRST, and atlas-based) significantly affected Dice coefficients ($\chi^2(3) = 77.45$, $p < 0.001$) and FNRs ($\chi^2(3) = 110.39$, $p < 0.001$). The post-hoc comparisons (Tukey's Multiple Comparison) between HC-nlFIRST and the other segmentation methods, summarized in Table 2, revealed significant improvement of HC-nlFIRST in terms of Dice coefficients in both hemispheres of caudate, accumbens, hippocampus and thalamus, and left putamen with respect to default FIRST, while no significances were observed with respect to T1n-nlFIRST. Significant higher Dice coefficients were observed in the left caudate, left accumbens, and right hippocampus and significantly lower FNRs were detected for the putamen, caudate, accumbens, hippocampus and thalamus in both hemispheres with respect to the atlas-based method. In addition, we found that all segmentation approaches revealed comparable mean values of Dice and FNR between control and MS subjects (see Supplementary Table S4).

Figure S6 and Table S3 in the Supplementary Material summarize the segmentation performance of the default FIRST approach with *T1n* and HC images as input contrasts. Only minor and non-significant differences in Dice coefficients and FNRs were measured, suggesting that using HC images instead of T1n images does not significantly improve the outcome of FIRST.

Fig. 10 illustrates the distribution of the obtained *figure-of-merit* F for the different segmentation methods by applying the *extended Regression Without Truth* approach (eRWT). The rank analysis based on the *figure-of-merit* performance of the different methods resulted in the following (descending) order: HC-nlFIRST/T1n-nlFIRST, FIRST, and atlas-based segmentation with the distribution of the *figure-of-merit* being very similar between HC-nlFIRST and T1n-nlFIRST.

3.4 Automatic quality assessment of segmented results

Fig. 11 shows the averaged Dice coefficients for all subjects resulting from the automatic quality assessment. The plots reflect in general the same trend across all the different segmentation methods for all SGM structures when compared to the manually defined

overlap measures obtained from the training dataset (Fig. 9). We observed a substantial gap of the Dice coefficients between successful and unsuccessful (failed) segmentations. The black arrows in Fig. 11 indicate two examples where the segmentation with default FIRST aborted (Dice coefficient equaled to zero; subject 9) or resulted in substantial segmentation errors (Dice = 0.312; subject 6, see also Fig. 8b). These findings are consistent with the results of the visual inspection. Consequently, the automatically determined averaged Dice coefficient may serve as a rough criterion to automatically characterize the quality of the SGM segmentation and more importantly to identify data sets with failed segmentation. Visual inspection confirmed that the third lowest Dice coefficients (equal to 0.601, subject 17) still revealed successful segmentation results with fair quality. Thus, we empirically selected a Dice coefficient of 0.6 as a threshold to determine whether SGM segmentation was successful or not.

4. Discussion

In our study, we demonstrated that the widely applied FSL FIRST pipeline tends to produce inaccurate subcortical segmentations (or may even fail) in individuals with abnormal brain anatomy. To address this issue, we propose a modification of the FIRST pipeline (HC-nlFIRST) by incorporating nonlinear registration and using a dedicated hybrid image contrast created by combining standard T1w images and quantitative susceptibility maps. By investigating Dice coefficients and FNRs with respect to manual reference ROIs in a test dataset and performing the eRWT analysis across all subjects, we were able to demonstrate that HC-nlFIRST outperforms other approaches including default FIRST, T1n-nlFIRST, and atlas-based segmentation. The Dice coefficients obtained using HC-nlFIRST were generally comparable to or even better than those reported by other authors using advanced segmentation methods [13, 14, 50, 51]. Lim *et al.* [50], for example, created a brain atlas based on five subjects by incorporating diffusion tensor images, T1w images and susceptibility maps to segment automatically subcortical structures by non-linear registration of the self-created atlas to individual susceptibility maps. The authors measured an overall average Dice coefficient of 0.88 ± 0.03 across their five subjects. Applying HC-nlFIRST we obtained an overall average Dice coefficient of 0.90 ± 0.07 (highest Dice coefficient in the thalamus [0.94 ± 0.01]; lowest in the accumbens [0.79 ± 0.07]), indicating excellent agreement with the manually outlined reference. It should, however, be noted that different manual reference ROI delineation protocols may have been applied in the different studies, hampering direct comparison among these studies.

As linear registration is not able to transform accurately subcortical structures of abnormal brains into the MNI space (Fig. 2), we replaced the FLIRT module with a nonlinear registration module, which substantially improved the alignment of registered images to the MNI template (see Fig. 6) and was rather insensitive to the different image contrasts (T1w, T1n and HC). For all these contrasts, higher mean values and smaller standard deviations of the Dice coefficients and volume overlaps resulted with nonlinear registration compared to FLIRT, indicating the consistently satisfying performance of the former. More specifically, the smallest standard deviations of Dice coefficients and volume overlaps were observed with HC images, suggesting the potential benefit of using HC with nonlinear registration. Since nonlinear registration using ANTs relies on the symmetric diffeomorphic deformation

model, anatomical topology even in regions with large deformations is preserved, making HC-nlFIRST suitable for large cohort studies. Comparing the control and MS group resulted in a consistently slightly higher registration performance of the former group (see Supplementary Table S2), which is likely due to the presence of less abnormal brain anatomy in control subjects. Our study is not the first to incorporate nonlinear registration to improve FIRST. In a previous study, Goodro *et al.* used FSL's FNIRT (FMRIB's Nonlinear Image Registration Tool) to register images to MNI space and then applied FIRST to investigate age effects on subcortical structures in healthy adults [52]. As in the evaluation work of 14 different registration methods [44], a direct comparison between ANTs and FNIRT indicated that ANTs provides more accurate registration results which adds confidence in producing accurate segmentations by using our proposed HC-nlFIRST method. While the FIRST training data also contain patients suffering from conditions such as Alzheimer's disease, schizophrenia and other diseases, the severe atrophy (structural abnormalities) in our study might not be well represented in that training cohort. One concern may be that since the FIRST algorithm has been trained using linear registration, nonlinear registration might introduce a bias in its shape model fitting. This potential bias, however, can be outweighed by the data-driven components of FIRST's underlying Bayesian framework with improved alignment of SGM structures using nonlinear registration in abnormal brain anatomy. One drawback, however, of incorporating non-linear registration is the substantial prolongation of the computation time from approximately 5 minutes per subject with FIRST to approximately 3 hours with T1n-nlFIRST, HC-nlFIRST and the atlas-based segmentation approach when using the same computer hardware (Octa-core 1.6 GHz processor and 24 GB memory). The computation speed may be increased in the future using faster nonlinear registration algorithms.

Another contribution to improve subcortical segmentation was the use of our dedicated hybrid contrast that combines contrast of T1w images and susceptibility maps to overcome the limited contrast of SGM structures on the T1w images and to increase the signal intensity similarity between individual data and FIRST training data. Recently, Cobzas *et al.* [53] used QSM together with T1w images in a multi-atlas approach, where more than one atlas was used and the segmented labels were fused into one final segmentation, yielding higher subcortical segmentation accuracy than by applying FIRST to T1w images. Unlike the multi-atlas approach and the multimodal image segmentation tool (MIST) [54], our proposed approach of replacing the T1w image by a dedicated hybrid contrast image does not require creating atlases or modifying the FIRST algorithm including its sophisticated training data. The proposed hybrid contrast approach is not restricted to the use of susceptibility maps and can, in principle, incorporate also other image contrasts with distinct SGM contrast (e.g., R_2^* maps, T_2 or T_2^* weighted images, fractional anisotropy [FA] maps from diffusion tensor imaging [DTI]).

The use of hybrid contrast images instead of T1w and $T1n$ images increased segmentation accuracy (Figs. 8 and 9). In general, even though quantitative comparison between $T1n$ and HC contrast did not reach statistical significance (Table 2), it should be noted that if structures are indiscernible on T1w images, FIRST's segmentation results are arbitrary, i.e., the resulting segmentation depends on numerical implementation aspects of the algorithm and, thus, most likely falls back to the initialization solution. While this initialization

segmentation may lead accidentally to a high Dice coefficient relative to the manual reference ROIs for some subjects, there is no anatomical basis for the segmentation and, hence, has to be considered arbitrary. In contrast, HC-based FIRST segmentations actually rely on anatomical contrast, as long as one of the image contrasts used to generate the HC images clearly delineates the anatomical region.

Regarding the generation of HC images, another motivation is that enhanced iron accumulation, which often occurs in neurodegenerative diseases, decreases the T1w contrast and may thus potentially introduce a bias in the FIRST segmentation. QSM, on the other hand, being an important component to the HC, shows increasing intensity with iron load and delivers a supreme contrast. Intuitively, with more iron load and the negative weighting coefficients (see Table 1), the increased intensity in QSM becomes decreased in the reversed QSM, which seems to have an adverse effect on the HC images. However, by using Eq. (2) to adjust the weighting contributions from both T1w and QSM, the HC images can homogenize the data and should thus mitigate this potential bias. From Table 1, we see that the fitted values of w_1 are relatively similar across the subjects whereas the values of w_2 show substantial variation. The latter reflects the sensitivity of tissue susceptibility against individual variations, such as those due to age or different pathological conditions. Nevertheless, the individual susceptibility maps are relatively robust against changes of sequence parameters, which is contrary to sequence parameter changes in T1w imaging. In our study, we determined the optimal weighting coefficients based on randomly selected training subjects of both normal controls and MS patients that were supposed to be representative for the whole dataset. The advantage of this method is that the weighting coefficients need to be calculated only once for each group of subjects. The weighting coefficients, however, are not universal as they depend on T1w image contrast, which is influenced mainly by the scan parameters (e.g. echo time, repetition time, flip angle, inversion time). Signal intensity in T1w images and susceptibility maps can also vary substantially across different subjects due to age [55–57] and pathological alterations [55, 56, 58]. Hence, the use of constant weighting coefficients might not be optimal in studies involving heterogeneous control and patient groups, potentially alleviating the benefit of hybrid contrast in some of these participants. This drawback may, however, be overcome by determining the weighting coefficients for each subject individually. To this end, we suggest an automatic approach to define the ROIs to estimate the mean signal intensities for Eq. (2) by relying on FSL-FAST to define regions of WM, CGM, and CSF and performing an initial run of T1n-nlFIRST to estimate the subcortical regions with subsequent morphological operations (i.e., erosion). The evaluation of such an approach will be the subject of future research. Another interesting aspect of the HC concept is its application to different scanners to decrease scanner-related variations of T1w images (e.g., different pulse sequence parameters and scanner field strength) in large cohort studies and will be also investigated in the future.

We used the eRWT approach to evaluate and rank the performance of different segmentation methods in the absence of a ground truth or reference in a non-supervised manner. Our results indicate that HC-nlFIRST and T1n-nlFIRST are the two best-suited segmentation methods, i.e., yielding results being more close to the unknown ground truth. While eRWT is able to rank methods based on biomedical scalar parameters (in our study the volume of

each subcortical structure), it cannot take into account complex information such as, e.g., volume location or morphological information (structure shape) [47]. Thus, eRWT results in the ranking of the performance of different segmentation methods based on simplified information (i.e. volume) rather than a comprehensive analysis.

We also addressed the issue of automatic quality assessment of segmentation results, which is inevitably necessary in large cohort studies. For this purpose, we propose to evaluate the segmentation results in a common space (e.g., the MNI space), where the reference ROIs need to be prepared only once, instead of relying on manual delineations on each subject. Since this approach strongly relies on the accuracy of non-linear registration into a common space, which may also be subject to errors, it is important as a first step to evaluate the registration accuracy between the individual subject data and the template using e.g. volume overlap measures within the target regions. Such a procedure can be implemented easily in an automatic manner, and a critical threshold may be defined depending on the visual impression of the results. Overlap measures such as Dice coefficients can be calculated between the individual and the reference segmentation in the common space. By using the overall region of SGM as reference, we were able to select an empirical threshold of 0.6 that indicated severe segmentation errors (confirmed by visual inspection) below this numerical value. However, there is a limitation in such quality assessments that it can only be used to automatically pick up the failure, but not being able to distinguish between or compare the minor differences among different segmentation methods. Other segmentation approaches may also be characterized with minimal manual interventions as far as the corresponding reference segmentations are available. Our proposed automatic evaluation in common space might be an alternative elegant way to assess segmentation quality in large cohort studies.

5. Conclusion

We presented an extension to the established FIRST pipeline, referred to as HC-nlFIRST, for automated subcortical segmentation that incorporates non-linear registration and dedicated image contrast combination of T1w images and quantitative susceptibility maps. This method overcomes insufficiencies of the linear registration in the default FIRST pipeline, particularly in the case of abnormal brain anatomy. HC-nlFIRST does not require any modifications of the underlying FIRST algorithm or its training data, as it is based solely on pre- and post-processing of the input images and the segmentation results, respectively. HC-nlFIRST thus provides a robust and accurate segmentation approach of subcortical brain structures particularly in abnormal brains, opening the door to minimize potential bias in large cohort studies.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The study was supported by the German Research Foundation (DFG, RE 1123/9-2, DE 2516/1-1), by a stipend (Landesgraduiertenstipendium) of the Graduate Academy of the Friedrich Schiller University Jena awarded to XF, a seed grant awarded to A.D. by the Interdisciplinary Center for Clinical Research (IZKF) in Jena, Germany, and seed grants awarded to F.S. by the International Society for Magnetic Resonance in Medicine (ISMRM) and the

Friedrich Schiller University Jena. It was also supported by the bilateral PPP-USA 2015/2016 program (project 57131995 – Quantitative MRI for Neurological Studies) of the German Academic Exchange Service (DAAD) awarded to J.R.R. The authors are grateful to Dr. Daniel Güllmar (Jena, Germany) and Dr. Niels Bergsland (Buffalo, NY, USA) for insightful discussions about the FIRST algorithm. Research reported in this publication was funded by the National Center for Advancing Translational Sciences of the National Institutes of Health under award Number UL1TR001412. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

Abbreviations

QSM	quantitative susceptibility mapping
SGM	subcortical gray matter
FIRST	FMRIB's Integrated Registration and Segmentation Tool
T1w	T ₁ -weighted
T1n	bias-field corrected and intensity normalized T ₁ -weighted images
MNI	Montreal Neurological Institute
FSL	FMRIB Software Library
HC	hybrid contrast
nl	non-linear
MS	multiple sclerosis
V-SHARP	sophisticated harmonic artefact reduction for phase data with varied radii
HEIDI	homogeneity enabled incremental dipole inversion
FAST	FMRIB's Automated Segmentation Tool
FLIRT	FMRIB's Linear Image Registration Tool
ANTs	Advanced Normalization Tools
SSIM	structural similarity
eRWT	extended Regression Without Truth
CC	cross correlation

References

1. 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(3):341–55. [PubMed: 11832223]
2. Fischl B, Salat DH, van der Kouwe AJ, Makris N, Segonne F, Quinn BT, et al. Sequence-independent segmentation of magnetic resonance images. *NeuroImage*. 2004; 23(Suppl 1):69–84.
3. Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. *NeuroImage*. 2011; 56(3):907–22. [PubMed: 21352927]

4. Wang J, Vachet C, Rumple A, Gouttard S, Ouziel C, Perrot E, et al. Multi-atlas segmentation of subcortical brain structures via the AutoSeg software pipeline. *Front Neuroinform.* 2014; 8:7. [PubMed: 24567717]
5. Hughes EJ, Bond J, Svrckova P, Makropoulos A, Ball G, Sharp DJ, et al. Regional changes in thalamic shape and volume with increasing age. *NeuroImage.* 2012; 63(3):1134–42. [PubMed: 22846656]
6. Hagemeyer J, Dwyer MG, Bergsland N, Schweser F, Magnano CR, Heininen-Brown M, et al. Effect of age on MRI phase behavior in the subcortical deep gray matter of healthy individuals. *AJNR Am J Neuroradiol.* 2013; 34(11):2144–51. [PubMed: 23721902]
7. Alves GS, O'Dwyer L, Jurcoane A, Oertel-Knochel V, Knochel C, Prvulovic D, et al. Different patterns of white matter degeneration using multiple diffusion indices and volumetric data in mild cognitive impairment and Alzheimer patients. *PloS one.* 2012; 7(12):e52859. [PubMed: 23300797]
8. Cheng WC, Cheng PE, Liou M. Group factor analysis for Alzheimer's disease. *Comput Math Methods in Med.* 2013; 2013:428385. [PubMed: 23533539]
9. Felsky D, Voineskos AN, Lerch JP, Nazeri A, Shaikh SA, Rajji TK, et al. Myelin-associated glycoprotein gene and brain morphometry in schizophrenia. *Front Psychiatry.* 2012; 3:40. [PubMed: 22563322]
10. Zivadinov R, Heininen-Brown M, Schirda CV, Poloni GU, Bergsland N, Magnano CR, et al. Abnormal subcortical deep-gray matter susceptibility-weighted imaging filtered phase measurements in patients with multiple sclerosis: a case-control study. *NeuroImage.* 2012; 59(1): 331–9. [PubMed: 21820063]
11. Lin JJ, Riley JD, Hsu DA, Stafstrom CE, Dabbs K, Becker T, et al. Striatal hypertrophy and its cognitive effects in new-onset benign epilepsy with centrotemporal spikes. *Epilepsia.* 2012; 53(4): 677–85. [PubMed: 22360313]
12. Suppa P, Hampel H, Kepp T, Lange C, Spies L, Fiebach JB, et al. Performance of Hippocampus Volumetry with FSL-FIRST for Prediction of Alzheimer's Disease Dementia in at Risk Subjects with Amnesic Mild Cognitive Impairment. *J Alzheimers Dis.* 2016:1–7. (Preprint).
13. Morey RA, Selgrade ES, Wagner HR 2nd, Huettel SA, Wang L, McCarthy G. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Hum Brain Mapp.* 2010; 31(11):1751–62. [PubMed: 20162602]
14. Nugent AC, Luckenbaugh DA, Wood SE, Bogers W, Zarate CA, Drevets WC. Automated subcortical segmentation using FIRST: Test-retest reliability, interscanner reliability, and comparison to manual segmentation. *Hum Brain Mapp.* 2013; 34(9):2313–29. [PubMed: 22815187]
15. Schmidt R, Ropele S, Enzinger C, Petrovic K, Smith S, Schmidt H, et al. White matter lesion progression, brain atrophy, and cognitive decline: the Austrian stroke prevention study. *Ann Neurol.* 2005; 58(4):610–6. [PubMed: 16178017]
16. Rudick R, Fisher E, Lee J-C, Simon J, Jacobs L. Use of the brain parenchymal fraction to measure whole brain atrophy in relapsing-remitting MS. *Neurology.* 1999; 53(8):1698. [PubMed: 10563615]
17. Jack C, Shiung M, Gunter J, O'Brien P, Weigand S, Knopman D, et al. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology.* 2004; 62(4): 591–600. [PubMed: 14981176]
18. Jones BC, Nair G, Shea CD, Crainiceanu CM, Cortese ICM, Reich DS. Quantification of multiple-sclerosis-related brain atrophy in two heterogeneous MRI datasets using mixed-effects modeling. *Neuroimage Clin.* 2013; 3:171–9. [PubMed: 24179861]
19. Mazziotta JC, Toga AW, Evans A, Fox P, Lancaster J. A probabilistic atlas of the human brain: theory and rationale for its development. *NeuroImage.* 1995; 2(2):89–101. [PubMed: 9343592]
20. Fonov V, Evans AC, Botteron K, Almli CR, McKinstry RC, Collins DL, et al. Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage.* 2011; 54(1):313–27. [PubMed: 20656036]
21. Amann M, Andelova M, Pfister A, Mueller-Lenke N, Traud S, Reinhardt J, et al. Subcortical brain segmentation of two dimensional T1-weighted data sets with FMRIB's Integrated Registration and Segmentation Tool (FIRST). *NeuroImage Clin.* 2015; 7:43–52. [PubMed: 25610766]

22. Rosen HJ, Gorno-Tempini ML, Goldman W, Perry R, Schuff N, Weiner M, et al. Patterns of brain atrophy in frontotemporal dementia and semantic dementia. *Neurology*. 2002; 58(2):198–208. [PubMed: 11805245]
23. Miller DH, Barkhof F, Frank JA, Parker GJ, Thompson AJ. Measurement of atrophy in multiple sclerosis: pathological basis, methodological aspects and clinical relevance. *Brain*. 2002; 125(8): 1676–95. [PubMed: 12135961]
24. Bermel RA, Bakshi R. The measurement and clinical relevance of brain atrophy in multiple sclerosis. *Lancet Neurol*. 2006; 5(2):158–70. [PubMed: 16426992]
25. Derakhshan M, Caramanos Z, Giacomini PS, Narayanan S, Maranzano J, Francis SJ, et al. Evaluation of automated techniques for the quantification of grey matter atrophy in patients with multiple sclerosis. *NeuroImage*. 2010; 52(4):1261–7. [PubMed: 20483380]
26. Li W, Wu B, Liu C. Quantitative susceptibility mapping of human brain reflects spatial variation in tissue composition. *NeuroImage*. 2011; 55(4):1645–56. [PubMed: 21224002]
27. Deistung A, Schafer A, Schweser F, Biedermann U, Turner R, Reichenbach JR. Toward in vivo histology: a comparison of quantitative susceptibility mapping (QSM) with magnitude-, phase-, and R2*-imaging at ultra-high magnetic field strength. *NeuroImage*. 2013; 65:299–314. [PubMed: 23036448]
28. Schweser F, Deistung A, Reichenbach JR. Foundations of MRI phase imaging and processing for Quantitative Susceptibility Mapping (QSM). *Z Med Phys*. 2016; 26(1):6–34. [PubMed: 26702760]
29. Deistung A, Schweser F, Reichenbach JR. Overview of quantitative susceptibility mapping. *NMR Biomed*. 2016
30. Haacke EM, Liu S, Buch S, Zheng W, Wu D, Ye Y. Quantitative susceptibility mapping: current status and future directions. *Magn Reson Imaging*. 2015; 33(1):1–25. [PubMed: 25267705]
31. Sun H, Wilman AH. Quantitative susceptibility mapping using single-shot echo-planar imaging. *Magn Reson Med*. 2015; 73(5):1932–8. [PubMed: 24938830]
32. Smith SM, Zhang Y, Jenkinson M, Chen J, Matthews PM, Federico A, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *NeuroImage*. 2002; 17(1):479–89. [PubMed: 12482100]
33. Tustison NJ, Avants BB, Cook PA, Zheng Y, Egan A, Yushkevich PA, et al. N4ITK: improved N3 bias correction. *IEEE Trans Med Imaging*. 2010; 29(6):1310–20. [PubMed: 20378467]
34. Roemer PB, Edelstein WA, Hayes CE, Souza SP, Mueller OM. The NMR phased array. *Magn Reson Med*. 1990; 16(2):192–225. [PubMed: 2266841]
35. Hammond KE, Lupo JM, Xu D, Metcalf M, Kelley DA, Pelletier D, et al. Development of a robust method for generating 7.0 T multichannel phase images of the brain with application to normal volunteers and patients with neurological diseases. *NeuroImage*. 2008; 39(4):1682–92. [PubMed: 18096412]
36. Polak, P., Zivadinov, R., Schweser, F. Gradient unwarping for phase imaging reconstruction. *Proc Intl Soc Mag Reson Med*; 2015; Toronto.
37. Abdul-Rahman HS, Gdeisat MA, Burton DR, Lalor MJ, Lilley F, Moore CJ. Fast and robust three-dimensional best path phase unwrapping algorithm. *Appl Opt*. 2007; 46(26):6623–35. [PubMed: 17846656]
38. Schweser F, Deistung A, Lehr BW, Reichenbach JR. Quantitative imaging of intrinsic magnetic tissue properties using MRI signal phase: an approach to in vivo brain iron metabolism? *NeuroImage*. 2011; 54(4):2789–807. [PubMed: 21040794]
39. Wu B, Li W, Guidon A, Liu C. Whole brain susceptibility mapping using compressed sensing. *Magn Reson Med*. 2012; 67(1):137–47. [PubMed: 21671269]
40. Schweser F, Sommer K, Deistung A, Reichenbach JR. Quantitative susceptibility mapping for investigating subtle susceptibility variations in the human brain. *NeuroImage*. 2012; 62(3):2083–100. [PubMed: 22659482]
41. Zhang YY, Brady M, Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging*. 2001; 20(1):45–57. [PubMed: 11293691]
42. Ashburner J, Friston K. Multimodal image coregistration and partitioning--a unified framework. *NeuroImage*. 1997; 6(3):209–17. [PubMed: 9344825]

43. Avants BB, Tustison NJ, Song G, Cook PA, Klein A, Gee JC. A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*. 2011; 54(3):2033–44. [PubMed: 20851191]
44. Klein A, Andersson J, Ardekani BA, Ashburner J, Avants B, Chiang MC, et al. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*. 2009; 46(3):786–802. [PubMed: 19195496]
45. Wang Z, Bovik AC, Sheikh HR, Simoncelli EP. Image quality assessment: from error visibility to structural similarity. *IEEE Trans Image Process*. 2004; 13(4):600–12. [PubMed: 15376593]
46. Nölte IS, Gerigk L, Al-Zghloul M, Groden C, Kerl HU. Visualization of the internal globus pallidus: sequence and orientation for deep brain stimulation using a standard installation protocol at 3.0 Tesla. *Acta neurochirurgica*. 2012; 154(3):481–94. [PubMed: 22167532]
47. Lebenberg J, Buvat I, Lalande A, Clarysse P, Casta C, Cochet A, et al. Nonsupervised ranking of different segmentation approaches: application to the estimation of the left ventricular ejection fraction from cardiac cine MRI sequences. *IEEE Trans Med Imaging*. 2012; 31(8):1651–60. [PubMed: 22665506]
48. Satorra A, Saris WE. Power of the likelihood ratio test in covariance structure analysis. *Psychometrika*. 1985; 50(1):83–90.
49. Hochberg, Y., Tamhane, AC. Multiple comparison procedures. John Wiley & Sons, Inc; 1987.
50. Lim IA, Faria AV, Li X, Hsu JT, Airan RD, Mori S, et al. Human brain atlas for automated region of interest selection in quantitative susceptibility mapping: application to determine iron content in deep gray matter structures. *NeuroImage*. 2013; 82:449–69. [PubMed: 23769915]
51. Tang X, Crocetti D, Kuttan K, Ceritoglu C, Albert MS, Mori S, et al. Segmentation of brain magnetic resonance images based on multi-atlas likelihood fusion: testing using data with a broad range of anatomical and photometric profiles. *Front Neurosci*. 2015; 9:61. [PubMed: 25784852]
52. Goodro M, Sameti M, Patenaude B, Fein G. Age effect on subcortical structures in healthy adults. *Psychiatry Res*. 2012; 203(1):38–45. [PubMed: 22863654]
53. Cobzas D, Sun H, Walsh AJ, Lebel RM, Blevins G, Wilman AH. Subcortical gray matter segmentation and voxel-based analysis using transverse relaxation and quantitative susceptibility mapping with application to multiple sclerosis. *J Magn Reson Imaging*. 2015
54. Visser E, Keuken MC, Douaud G, Gaura V, Bachoud-Levi A-C, Remy P, et al. Automatic segmentation of the striatum and globus pallidus using MIST: Multimodal Image Segmentation Tool. *NeuroImage*. 2016; 125:479–97. [PubMed: 26477650]
55. Li W, Wu B, Batrachenko A, Bancroft-Wu V, Morey RA, Shashi V, et al. Differential developmental trajectories of magnetic susceptibility in human brain gray and white matter over the lifespan. *Hum Brain Mapp*. 2014; 35(6):2698–713. [PubMed: 24038837]
56. Persson N, Wu J, Zhang Q, Liu T, Shen J, Bao R, et al. Age and sex related differences in subcortical brain iron concentrations among healthy adults. *NeuroImage*. 2015; 122:385–98. [PubMed: 26216277]
57. Lu H, Nagae-Poetscher LM, Golay X, Lin D, Pomper M, van Zijl P. Routine clinical brain MRI sequences for use at 3.0 Tesla. *J Magn Reson Imaging*. 2005; 22(1):13–22. [PubMed: 15971174]
58. Liu C, Li W, Tong KA, Yeom KW, Kuzminski S. Susceptibility-weighted imaging and quantitative susceptibility mapping in the brain. *J Magn Reson Imaging*. 2015; 42(1):23–41. [PubMed: 25270052]

Highlights

- Default FSL-FIRST tends to generate inaccurate subcortical segmentation in brains with extremely abnormal anatomy.
- Combining T₁-weighted images and quantitative susceptibility maps provides a new hybrid contrast (HC) with improved delineation of subcortical brain nuclei and structures.
- A new HC-nlFIRST framework that uses hybrid contrast images and nonlinear registration improves the accuracy of subcortical segmentation compared to FSL's default FIRST approach.

Fig. 1.
Representative examples of T1w images of eight subjects who were included in the study showing particularly abnormal brain anatomy (atrophy).

Fig. 2.

Linear registration to the MNI space of a T1w image of one subject with severe atrophy. (a) shows an axial view of the original T1w image, (b) the linearly registered T1w image, and (c) the corresponding MNI template. While the position misalignment between the original T1w image (a) and the MNI template (c) can be resolved by applying a translation during linear registration, the mismatch of the ventricles and the caudate nuclei, however, remain between the registered image (b) and the MNI template (c).

Fig. 3.

Flow chart of the proposed **HC-nlFIRST** segmentation framework. T1bc is the abbreviation of **b**ias field **c**orrected T1w image, *T1norm* stands for the intensity normalized, bias field corrected T1w image, GRE abbreviates gradient echo, and QSM denotes the quantitative susceptibility map.

Fig. 4.

Three-dimensional rendering of the subcortical reference ROIs in MNI space from four different views of (a) left, (b) right, (c) top, and (d) bottom. Abbreviations in coordinate system: S – superior, R – right, A – anterior, P – posterior. Different colors are used to indicate the subcortical GM structures: pink – putamen, cyan – caudate, orange – accumbens, yellow – hippocampus, blue – globus pallidus, and green – thalamus.

Fig. 5.

Exemplary slices of subject #2 at the level of the basal ganglia. (a) T1w image, (b) $T1n$ image, (c) quantitative susceptibility map, and (d) the HC image, which was obtained by combining the $T1n$ image and the susceptibility map ($w_1=1.50$ and $w_2=-111.19 \text{ ppm}^{-1}$). Improved SGM contrast with respect to the surrounding white matter is discernible on the HC image (d) compared to the other T1w images (a,b). The arrows point to the globus pallidus (red) and thalamus (yellow) whose contrast relative to the internal capsule was substantially enhanced in the HC image (d). Bias field correction successfully suppressed low frequency, non-uniform intensity noticeable in the center of the T1w image, but did not affect the SGM contrast. See Fig. S2 (in Supplementary Material) for a different subject of the non-training dataset.

Fig. 6. Spatial registration to the MNI template. (a) MNI template and manual reference ROIs overlaid on the MNI template in (d). The registered HC images using FLIRT and nonlinear ANTs are presented in (b) and (c), and the subcortical reference labels warped in MNI space using FLIRT and nonlinear ANTs are shown in (e) and (f), respectively. The mean and standard deviations of the Dice coefficients and volume overlaps of the warped SGM labels with respect to those of the MNI template are displayed in the bottom row in (g) and (h), respectively.

Fig. 7. Visual assessment of subcortical segmentation with the different segmentation methods averaged over all 82 subjects. The error bars indicate the standard deviation. A significant difference in the rating scores between HC-nlFIRST and one of the other segmentation methods is denoted by asterisks (***) stands for $p < 0.001$, * stands for $p < 0.05$).

Fig. 8.

Comparison between manual and automated subcortical segmentation (pink – putamen, cyan – caudate, orange – accumbens, yellow – hippocampus, blue – globus pallidus, green – thalamus) of (a) an MS patient with severe brain atrophy (subject #1) and (b) a healthy subject (subject #6). The axial and coronal views of the hybrid contrast images, shown in the left column, are superposed from left to right with the boundaries of the segmentation results due to manual delineation, FIRST, T1n-nlFIRST, HC-nlFIRST and atlas-based segmentations, respectively. The arrows indicate inaccurately segmented regions compared to the manual reference. Segmentation results overlaid on the susceptibility map are displayed in Fig. S4 (see Supplementary Material).

Fig. 9.

Quantitative analysis of the training dataset (n=8) for the different segmentation approaches. The mean and standard deviations of Dice coefficients and FNRs are shown in the top and bottom rows, respectively. For the outliers, the standard deviations of the Dice coefficients of the default FIRST method are 0.266 for the right putamen, 0.257 for the right caudate, 0.323 and 0.305 for the left and right hippocampus. The standard deviations of FNRs of atlas-based method are 0.230 and 0.389 for the left and right accumbens, respectively. (Putamen – putamen; Caud – caudate; Accu – accumbens; Hipp – hippocampus; GP – globus pallidus; Thal – thalamus)

Fig. 10.

Boxplots of the calculated *figure-of-merit* (F) after bootstrapping for each segmentation method and SGM structure (horizontal axis: '1' – FIRST '2' – T1n-nlFIRST '3' – HC-nlFIRST '4' – atlas-based method). The lower *figure-of-merit* indicates that the segmentation method is more accurate. The title of each plot shows the respective SGM structure, where the prefixes L and R stand for the left and right hemisphere, respectively. (Putamen – putamen; Caudate – caudate; Accumbens – accumbens; Hippocampus – hippocampus; Globus pallidus – globus pallidus; Thalamus – thalamus)

Fig. 11.

Averaged Dice coefficients across all SGM structures plotted for all individual subjects.

(blue squares – FIRST, orange crosses – T1n-nlFIRST, black dots – HC-nlFIRST).

Substantial wrong or aborted segmentations are indicated by the black arrows. The dashed red line indicates an empirically selected threshold to identify failed and successful segmentations.

Table 1

Weighting coefficients for hybrid contrast images of MS patients (n=4) and normal controls (n=4) of the two training groups.

category	Subject #	w_1 for TIn	w_2 [ppm ⁻¹] for QSM
MS patients	1	1.49	-105.74
	2	1.50	-111.19
	3	1.49	-74.66
	4	1.42	-64.86
Controls	5	1.57	-86.84
	6	1.54	-53.43
	7	1.60	-48.41
	8	1.55	-101.00

Table 2

The p -values of Tukey's multiple comparison for the Dice coefficients and FNRs between HC-nlFIRST and the other applied segmentation methods in the subcortical regions of the training dataset. Significant differences ($p < 0.05$, marked in bold) represent higher Dice and lower FNR coefficients in HC-nlFIRST.

	HC-nlFIRST vs. FIRST		HC-nlFIRST vs. T1n-nlFIRST		HC-nlFIRST vs. Atlas-based	
	Dice	FNR	Dice	FNR	Dice	FNR
left Putamen	0.004	0.996	0.913	0.625	0.907	< 0.001
right Putamen	0.071	0.608	0.961	0.533	0.887	< 0.001
left Caudate	< 0.001	0.294	0.946	0.983	0.016	< 0.001
right Caudate	0.034	0.429	0.970	0.994	0.160	< 0.001
left Accumbens	0.023	0.996	0.426	0.509	< 0.001	< 0.001
right Accumbens	0.001	0.955	0.804	0.877	0.076	< 0.001
left Hippocampus	0.017	0.549	0.997	0.982	0.083	< 0.001
right Hippocampus	0.020	0.928	0.997	0.995	0.029	< 0.001
left Globus Pallidus	0.108	0.938	0.966	0.965	0.837	0.966
right Globus Pallidus	0.080	0.401	0.939	0.757	0.905	0.510
left Thalamus	0.024	0.581	0.969	0.911	0.216	0.002
right Thalamus	0.009	0.315	0.980	0.932	0.285	0.004