AJNR

Quantification of Thalamic Atrophy in MS: From the Multicenter Italian Neuroimaging Network Initiative Data Set to Clinical Application

Loredana Storelli, Elisabetta Pagani, Patrizia Pantano, Antonio Gallo, Nicola De Stefano, Maria A. Rocca, Massimo Filippi and for the INNI Network

Quantification of Thalamic Atrophy in MS: From the Multicenter Italian Neuroimaging Network Initiative Data Set to Clinical Application

Loredana Storelli, Elisabetta Pagani, Patrizia Pantano, Antonio Gallo, Nicola De Stefano, Maria A. Rocca, and Massimo Filippi, for the INNI Network

ABSTRACT

BACKGROUND AND PURPOSE: Thalamic atrophy occurs from the earliest phases of MS; however, this measure is not included in clinical practice. Our purpose was to obtain a reliable segmentation of the thalamus in MS by comparing existing automatic methods cross-sectionally and longitudinally.

MATERIALS AND METHODS: MR images of 141 patients with relapsing-remitting MS (mean age, 38 years; range, 19–58 years; 95 women) and 69 healthy controls (mean age, 36 years; range, 22–69 years; 47 women) were retrieved from the Italian Neuroimaging Network Initiative repository: T1WI, T2WI, and DWI at baseline and after 1 year (136 patients, 31 healthy controls). Three segmentation software programs (FSL-FIRST, FSL-MIST, FreeSurfer) were compared. At baseline, agreement among pipelines, correlations with age, disease duration, clinical score, and T2-hyperintense lesion volume were evaluated. Effect sizes in differentiating patients and controls were assessed cross-sectionally and longitudinally. Variability of longitudinal changes in controls and sample sizes were assessed. False discovery rate–adjusted $P < .05$ was considered significant.

RESULTS: At baseline, FSL-FIRST and FSL-MIST showed the highest agreement in the results of thalamic volume ($R = 0.87$, $P < .001$), with the highest effect size for FSL-MIST (Cohen $d = 1.11$); correlations with demographic and clinical variables were comparable for all software. Longitudinally, FSL-MIST showed the lowest variability in estimating thalamic volume changes for healthy controls (SD = 1.07%), the highest effect size (Cohen $d = 0.44$), and the smallest sample size at 80% power level (15 subjects per group).

CONCLUSIONS: Multimodal segmentation by FSL-MIST increased the robustness of the results with better capability to detect small variations in thalamic volumes.

ABBREVIATIONS: EDSS = Expanded Disability Status Scale; FA = fractional anisotropy; HC = healthy controls; ICC = intraclass correlation coefficient; INNI = Italian Neuroimaging Network Initiative; LV = lesion volumes; RR = relapsing-remitting

The thalamus is a highly organized structure of gray matter nuclei that contains only a modest component of myelinated and unmyelinated white matter. It has a critical role in linking cortical and subcortical circuits, which subserve many neurologic functions.[1]

In patients with MS, a high vulnerability to damage to this strategic structure has been consistently demonstrated.[2] Many studies have reported a volume reduction of the thalamus not only in all clinical MS phenotypes but also in the early phases of the disease. Thalamic atrophy has been found in patients with clinically isolated syndrome,[3] early relapsing-remitting (RR)[4] and primary-progressive MS,[5] and also in pediatric patients with MS.[6] Thalamic atrophy is related not only to the presence of thalamic

lesions but also to the global burden of brain T2-hyperintense and T1-hypointese lesions, supporting its potential to reflect changes secondary to axonal transection of white matter fibers. Notably, the quantification of thalamic damage in MS is also informative for disease evolution, overcoming, for instance, measures of global gray matter involvement and those that reflect intrinsic lesional microstructural damage in explaining changes in Expanded Disability Status Scale (EDSS) after an 8-year follow-up.[7,8] For all these reasons, thalamic atrophy is a candidate biomarker in MS. Thus, it has been already included as an exploratory end point in clinical trials.[9-13]

In the clinical setting, this measure is, however, still not obtained, mainly because of the time-consuming procedure required for its manual segmentation. On the other hand, the existing automatic methods do not provide good enough reproducibility that allows monitoring atrophy changes at a single patient level. This drawback was shown in a recent article[14] using data from the Alzheimer's Disease Neuroimaging Initiative data set and in another study[15] in which the variation from repeated measurements ranged from 1% to 3%, depending on the method used. Although this variability can be considered sufficient for group comparisons, it is not suitable enough for individualized assessments.

A major issue for thalamic segmentation is the precise contouring of lateral boundaries toward the internal capsule, because contrast smoothly degrades due to the presence of white matter fibers terminating in the thalamus. In a previous article,[16] we tried to control for this issue by using the contrast offered by fractional anisotropy (FA) maps from DWI for improving the segmentation. Recently, a multimodal approach for subcortical nuclei segmentation was included in the FSL library.[17,18] Although the published article does not specifically discuss thalamic segmentation, the results obtained for other subcortical structures, such as the striatum and the globus pallidus, encourage the inclusion of FA for a better delineation of these structures.[18]

The Italian Neuroimaging Network Initiative (INNI) supports the creation of a repository in which MR imaging, clinical, and neuropsychological data from patients with MS and healthy controls (HC) are collected from 4 Italian research centers with internationally recognized expertise in MR imaging applied to MS.[19]

Using the large multicenter MR imaging data set from INNI, we aimed to obtain a reliable, automatic segmentation of the thalamus in MS by comparing the results obtained with existing automatic approaches both cross-sectionally and longitudinally. Final suggestions are provided for the application of these pipelines in large studies in MS.

MATERIALS AND METHODS
Ethics Committee Approval
Approval was received from the local ethics standards committee on human experimentation at each Research Center, and written informed consent was obtained from all subjects.

Subjects
We retrospectively analyzed data from 141 patients (center A: 35, center B: 34, center C: 36, center D: 36) with RRMS[20] and 69 HC (center A: 20, center B: 14, center C: 20, center D: 15) collected by INNI from 4 centers identified here as A (where subjects were scanned with 2 different T1-weighted sequences) and B, C, and D (MR imaging data were acquired between January 2008 and July 2017) for pipeline comparison (validation set). A cross-sectional training set was needed for one of the compared pipelines. Thus, we collected from the INNI repository an additional data set including 50 patients with MS (center A: 20, center B: 10, center C: 10, center D: 10) and 50 HC (center A: 20, center B: 10, center C: 11, center D: 9). Inclusion/exclusion criteria for all patients and HC were the following: no contraindications to MR imaging, no history of alcohol or substance abuse, no neurologic diseases (other than MS), and no psychiatric diseases. All participants underwent a clinical and MR imaging evaluation at baseline with rating using the EDSS score and disease duration.

Of the validation set, 136 patients with MS and 31 HC underwent a follow-up re-evaluation 1 year after the baseline visit. At follow-up, all patients had been relapse-free and steroid-free for at least 1 month.

MR Imaging Acquisition
Baseline and follow-up 3D T1-weighted, T2-weighted, and DWI scans were acquired at each center using a local standardized protocol and 3T scanners. Pulse sequence parameters are reported in the Online Supplemental Data and are more extensively described in a previous publication.[21]

Image Analyses
All INNI MR imaging data underwent a standardized preprocessing, including a procedure for quality control described in detail elsewhere.[21] Focal T2-hyperintense white matter lesions had already been manually identified and segmented by each participating center and made available for the quality control to the analyzer center (center A).

DWIs were corrected for movement and eddy current–induced distortions using the FDT tool (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FDT) within the FSL Library.[22,23] The Diffusion Toolbox was estimated by linear regression busing DWI data at $b=0, 900$, or 1000 s/mm^2. Subsequently, FA maps were derived.[24]

The pipelines selected for this study were all freely available and fully-automatic methods for cross-sectional volumetric segmentation of subcortical brain structures on MR images. None of these are specifically proposed and optimized for longitudinal quantification of subcortical tissue volume changes. Thus, 3 software programs were compared:

- FSL-FIRST, Version 5.0.10. (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/FIRST/UserGuide)[25]
- FSL-MIST (β release) (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MIST)[26]
- FreeSurfer, Version 6.0 (http://surfer.nmr.mgh.harvard.edu).[27,28]

The selected approaches were applied on the validation data set at baseline and follow-up for estimation of longitudinal volume change.

3D T1-weighted images after applying the lesion-filling technique[29] were given as input to all the toolboxes. FSL-MIST can also include the FA map coregistered into the T1-weighted space

Main demographic and clinical findings in HC and patients with RRMS at baseline for the training set required by FSL-MIST toolbox

	HC (n = 50)	RRMS (n = 50)	P Value
Female/male	30:20	32:18	.68[a]
Mean age (range) (yr)	38 (19–56)	36 (19–62)	.3[b]
Median disease duration (range) (yr)	–	8 (1–25)	–
Median EDSS (range)	–	1.5 (0–4.5)	–
Median T2 LV (range) (mL)	0.1 (0–0.5)	3.2 (0.3–33.0)	<.001[b]

Note:—The en dash (–) indicates null values.
[a] Pearson χ^2 test;
[b] Mann-Whitney test.

as an optional input to the segmentation method to improve the results of the subcortical masks. Thus, for each subject, FA maps were registered to the T1-weighted space (FLIRT, FSL) and given as an additional input for the FSL-MIST pipeline. Moreover, FSL-MIST is a 2-step approach: one for the training of the model and the second for the test.

For longitudinal thalamic atrophy quantification, the changes between baseline and follow-up cross-sectional volumes obtained by the compared pipelines were assessed as the percentage difference between the 2 MR imaging acquisitions corrected for the baseline volume.

Here is a brief description of the methods:

FSL-FIRST. FIRST is a freely available toolbox for subcortical structure segmentation included within the FSL Library.[25] The method models shapes and appearance of 15 brain structures starting from 336 manually segmented T1-weighted MR images. In particular, FIRST uses the active shape and appearance models[30] that associate the intensities around a deformable shape with the spatial configuration of the shape.

FSL-MIST. MIST is a toolbox included in FSL Library for multimodal subcortical structure segmentation, taking advantage of the different image contrasts.[26] Unlike FIRST, it can use additional information from different MR imaging modalities and is less dependent on manual segmentation learning in an unsupervised fashion from a set of unlabeled training data.

FreeSurfer. FreeSurfer is a freely available software package for the analyses of structural MR imaging data and cortical thickness estimation.[27,28] For subcortical structure segmentation, a probabilistic atlas, which is derived from a number of manual segmentations, is still used, but in this case, the procedure is based on modeling the segmentation as a nonstationary anisotropic Markov random field.

Statistical Analysis
Statistical analysis was performed using the R software package (Version 3.1.1; https://www.r-project.org/). Demographic data and T2 lesion volumes (LV) were compared between groups using the χ^2 Pearson test for categoric variables and the Mann-Whitney U or t test for continuous variables. We evaluated the intraclass correlation coefficient (ICC) to assess the strength of the agreement among the different software packages on the basis of thalamic volume measures. Between-group differences (center-corrected) in thalamic volumes and their longitudinal changes for the different software packages were expressed as effect size,

calculated according to the Cohen d definition.[31] Pearson correlations of thalamic volume with age (separately for MS and HC) and partial correlations (adjusted for age) with clinical and MR imaging variables (EDSS, disease duration, T2 LV) for MS at baseline for each method were estimated. Finally, the variability of the results of the longitudinal changes in thalamic volumes in HC at follow-up for the different pipelines was assessed. The sample size requirements at 80% power level for detecting a significant difference in the rate of thalamic atrophy between HC and MS at the .05 α level was estimated for each software package.

False discovery rate (Benjamini-Hochberg procedure) correction was applied. A P value < .05 was considered statistically significant.

A visual inspection of the thalamic segmentations for all the pipelines was also performed.

RESULTS
Demographic and Clinical Features
The Online Supplemental Data summarize the main demographic and clinical characteristics of the validation set at baseline. Patients with MS and HC were age- and sex-matched, with a higher prevalence of women in both groups. As expected, T2-hyperintense LV were significantly higher in the RRMS group compared with HC.

At follow-up (mean follow-up interval = 1.00 [SD, 0.25] year for MS, 1.06 [SD, 0.20] years for HC; P = .7), the median EDSS score was 1.5 (range, 0–4.5; P = .2 versus baseline) and 7 patients with MS had worsened clinically (EDSS score increase of ≥1.5 when the baseline EDSS was 0, ≥1.0 when the EDSS at baseline was <6.0, and ≥0.5 when EDSS at baseline was ≥6.0).[32]

The Table shows the main demographic and clinical characteristics of the training set used for the FSL-MIST toolbox. Again, patients with MS and HC were age- and sex-matched, with a higher prevalence of women in both groups and higher T2-hyperintense LV in RRMS compared with HC. The validation and training data sets showed the same characteristics.

Baseline Results
At baseline, all software showed a good significant agreement in the results of thalamic volume, with the highest agreement between FSL-FIRST and FSL-MIST (ICC = 0.87, P < .001) and the lowest between FSL-MIST and FreeSurfer (ICC = 0.80, P < .001), for both MS and HC. Online Supplemental Fig 1 shows an example of thalamic segmentations for each software program on a healthy volunteer. However, from a visual inspection of the accuracy of the segmentations, we found that thalamic delineations obtained by FSL-MIST are always underestimated with respect to FSL-FIRST and FreeSurfer, also noticeable from the thalamic volumes at baseline.

All pipelines significantly differentiated patients with MS from HC (P < .001 for all software, Online Supplemental Fig 2), with the highest effect size found for FSL-MIST (Cohen d = 1.11)

and FSL-FIRST (Cohen $d = 1.07$) compared with FreeSurfer (Cohen $d = 0.79$). When we looked at the data, FSL-MIST showed the lowest variability with an increased robustness of the results at baseline (SD for the thalamic volume distribution in HC = 1.36 mL, in MS = 1.48 mL) with respect to the other 2 pipelines (SD for the thalamic volume distribution in HC = 1.74 mL and 1.95 mL, respectively for FSL-FIRST and FreeSurfer).

At baseline, the Pearson correlations between thalamic volumes and subjects' ages were similar and significant for all the compared pipelines, for both MS ($r = -0.33$ for FSL-MIST, $r = -0.32$ for FSL-FIRST, $r = -0.29$ for FreeSurfer, all $P < .001$) and HC ($r = -0.38$ for FSL-MIST, $r = -0.40$ for FSL-FIRST, $r = -0.45$ for FreeSurfer, all $P < .001$).

At baseline in patients with MS, partial correlations (adjusted for age) between thalamic volumes and EDSS were very low or not significant for all the compared pipelines ($r = -0.16$, $P = .06$ for FSL-MIST, $r = -0.3$, $P < .001$ for FSL-FIRST, $r = -0.17$, $P = .04$ for FreeSurfer). Again, at baseline in patients with MS, partial correlations (adjusted for age) between thalamic volumes and disease duration were significant only for FSL-FIRST ($r = -0.2$, $P = .02$), while these were not significant for FSL-MIST ($r = -0.10$, $P = .3$) and FreeSurfer ($r = -0.12$, $P = .1$).

Pearson correlations between thalamic volumes at baseline and T2 LV were all significant and comparable among the different software ($r = -0.46$, $P < .05$ for FSL-MIST, $r = -0.51$, $P < .05$ for FSL-FIRST, and $r = -0.46$, $P < .05$ for FreeSurfer).

Longitudinal Results
At follow-up, in HC, FSL-MIST showed the lowest variability of percentage thalamic volume change (SD = 1.07%) in comparison with the other pipelines (SD = 1.53% for FSL-FIRST and 4.96% for FreeSurfer). Online Supplemental Fig 3 shows the results of thalamic volume changes in HC and patients with MS for the different pipelines.

At follow-up, for percentage thalamic volume changes, FSL-MIST showed a small effect size but a better capability to significantly differentiate between HC and patients with MS (Cohen $d = 0.44$) compared with FSL-FIRST (Cohen $d = 0.09$) and FreeSurfer (Cohen $d = 0.21$).

FSL-MIST and FSL-FIRST showed the smallest sample size requirement for assessment of longitudinal thalamic atrophy at an 80% power level: 15 subjects per arm for FSL-MIST and 29 subjects per arm for FSL-FIRST, while the highest sample size was found for FreeSurfer (105 subjects per group at an 80% level).

DISCUSSION
In this work, we aimed to compare 3 available fully automatic methods for thalamic segmentation and volume quantification on a multicenter data set to obtain a reliable segmentation of the thalamus for a possible future clinical introduction of this measure in MS. Using the INNI data set, we found that the inclusion of FA maps facilitated the automatic identification of thalamic boundaries, increasing the robustness of the results. In particular, the multimodal approach (FSL-MIST) showed a better capability to detect small longitudinal variations of thalamic volumes in patients with MS.

At baseline, we found a good agreement (ICCs ≥ 0.8) among the software for automatic thalamic segmentations analyzed. However, a slightly higher agreement was found between FSL-MIST and FSL-FIRST. This could be because both methods are implemented within the same FSL Library and have some basic methodologic similarities. In detail, both methods use a probabilistic Bayesian approach to fully exploit the relationship between intensity and shape/boundaries and use a generative model for the intensity profiles perpendicular to a deformable mesh. However, for FSL-FIRST, intensities and possible shape variations are derived from a set of 336 manually labeled training segmentations for 15 different subcortical structures based on a T1-weighted scan only, while FSL-MIST can learn in an unsupervised fashion from unlabeled training data, being less dependent on manual segmentations, and can simultaneously combine complementary information from different MR imaging modalities, which also increases the contrast-to-noise ratio. On the other hand, FreeSurfer similarly starts from a probabilistic atlas (as in FSL-MIST and FSL-FIRST), which is derived from a data set of 12 manually labeled segmentations, but it is used as a prior on a Markov random field model. The prior information included both the global spatial information, independent from other information, and the local spatial relationship between anatomic classes. These spatial constraints allow the Markov random field model to segment the image into a large number of classes required to segment the subcortical structures. These technical and implemental differences among the compared software could have been reflected on the observed discrepancies in terms of thalamic volume agreement.

All software could significantly differentiate thalamic volumes of patients with MS from those in HC at baseline, with a slightly higher effect size found for FSL-MIST in comparison with the other tools. Thus, all analyzed pipelines are suitable for analyses at a group level to detect thalamic volume differences between patients with MS and HC. From the standpoint of moving the application of these software programs to the single patient level and personalized medicine, our assessment should also take into account the variability of the results obtained and the measurement errors. Because there are no ground truth segmentations, we could not infer anything about measurement errors. However, by looking at the distributions of thalamic volumes at baseline in HC, we found the lowest variability of the results for FSL-MIST, suggesting more robustness of the measures.

Longitudinal analyses are extremely important to inspect the reliability of a measure and to assess the possibility of applying a tool for measurements on a single-subject level. In this study, after 1 year of follow-up, we expect very small changes in thalamic volumes in HC.[7-9,33] Given these considerations, our longitudinal results on the percentage thalamic volume change in HC confirmed a higher robustness of the measurement for FSL-MIST (showing the lowest variability) in comparison with FSL-FIRST and FreeSurfer. This is also evident at a group level by looking at the results of the effect sizes among the different pipelines on the longitudinal measures of thalamic volume changes. Moreover, we found that thalamic volume changes obtained from FSL-MIST required the smallest sample size compared with the other 2 software programs, making this tool particularly appealing for MS

studies but also for possible use at the individual level and when evaluating treatment effects.

From a visual inspection of the accuracy of the segmentations, we found that thalamic delineations obtained by FSL-MIST are always underestimated with respect to FSL-FIRST and FreeSurfer, also noticeable from the thalamic volumes at baseline. In fact, FSL-MIST segmentation always tends not to include the thalamic region of pulvinar, probably because this subregion is well-contrasted in the FA contrast and is important information used for the FSL-MIST multimodal segmentation. Thus, the tool delineates the final boundary of the thalamus before this region, as shown in Online Supplemental Fig 1. This characteristic is certainly a limitation of the method. However, at a single-patient level, an essential value for a software is the capability to detect small thalamic volume changes with time, and the longitudinal reproducibility demonstrated by FSL-MIST is the most important result in this sense.

Our findings on the association between demographic and clinical variables with baseline thalamic volumes did not show relevant differences among the compared tools. At baseline, correlations of thalamic volumes with age, EDSS, and T2 LV, though significant, were very low and similar among the pipelines, while the association between thalamic volumes and disease duration was slightly significant only for the FSL-FIRST measures. Thus, considering baseline correlation results, none of the software seemed to perform better compared with the others.

Therefore, given the difficulty of modeling structures such as the thalamus, with darker-appearing tissue at the midline and a gradient of brighter intensities as one moves more laterally, the use of a multimodal approach could facilitate this automatic task. From our findings, the inclusion of FA contrast for thalamic segmentation seemed to increase the robustness of the results and demonstrated a better capability to detect small longitudinal variations of thalamic volumes, as shown by FSL-MIST results. However, due to the lack of data on the accuracy and precision in the selection of the appropriate pipeline for automatic thalamic segmentation, it would be important to take into account even other practical aspects. Considering the application context, the lack of FA maps, for example, could prevent the use of multimodal approaches like FSL-MIST in favor of faster approaches (FSL-FIRST). However, an increased complexity could be tolerable for a better longitudinal reproducibility of the results (FSL-MIST).

Limitations

As previously stated, one of the main limitations of this comparative study is the lack of ground truth thalamic segmentation for a fair validation of the software on the accuracy of thalamic volumes. However, manual segmentation of the thalamus is an extremely time-consuming task, especially on high-resolution 3D T1-weighted sequences, and its precise delineation is very difficult to achieve, even for an expert MR imaging reader. Moreover, test-retest data would be helpful to assess precision and measurement errors for the compared software, and further investigations are needed. As already stated, FSL-MIST segmentation always tends not to include the thalamic region of pulvinar, and this exclusion is certainly a limitation in the accuracy of the segmentation at a single visit. Finally, a more clinical setting in terms of

variability and lower quality of input MR imaging data could have allowed us to evaluate the performance and applicability of the methods, even in a clinical framework. However, the large multicenter cohort with no standardized MR imaging sequences and protocols collected for the purpose of this study is a good starting point to evaluate the available pipelines for automatic thalamic segmentation.

CONCLUSIONS

Thalamic atrophy is under investigation as a biomarker in MS. However, its applicability in the clinical setting is still not possible, mainly because of the lack of an automatic and robust segmentation method. Automatic methods are now available that could be optimized and tested in a multicenter context. Because the delimitation of the internal boundary toward the internal capsule is a limiting factor, a multimodal approach that includes FA maps could improve the overall reproducibility of thalamic segmentation. By comparing cross-sectional and longitudinal thalamic segmentations from 3 available automatic methods in a multicenter data set from INNI, we found that the inclusion of FA contrast increased the robustness of the longitudinal results and had a better capability to detect small variations of thalamic volumes, as shown by FSL-MIST results.

Disclosure forms provided by the authors are available with the full text and PDF of this article at www.ajnr.org.

REFERENCES

1. Haber SN. **The primate basal ganglia: parallel and integrative networks.** *J Chem Neuroanat* 2003;26:317–30 CrossRef Medline
2. Henry RG, Shieh M, Okuda DT, et al. **Regional grey matter atrophy in clinically isolated syndromes at presentation.** *J Neurol Neurosurg Psychiatry* 2008;79:1236–44 CrossRef Medline
3. Audoin B, Ranjeva JP, Au Duong MV, et al. **Voxel-based analysis of MTR images: a method to locate gray matter abnormalities in patients at the earliest stage of multiple sclerosis.** *J Magn Reson Imaging* 2004;20:765–71 CrossRef Medline
4. Audoin B, Davies GR, Finisku L, et al. **Localization of grey matter atrophy in early RRMS: a longitudinal study.** *J Neurol* 2006;253:1495–501 CrossRef Medline
5. Sepulcre J, Sastre-Garriga J, Cercignani M, et al. **Regional gray matter atrophy in early primary progressive multiple sclerosis: a voxel-based morphometry study.** *Arch Neurol* 2006;63:1175–80 CrossRef Medline
6. Mesaros S, Rocca MA, Absinta M, et al. **Evidence of thalamic gray matter loss in pediatric multiple sclerosis.** *Neurology* 2008;70:1107–12 CrossRef Medline
7. Mesaros S, Rocca MA, Pagani E, et al. **Thalamic damage predicts the evolution of primary-progressive multiple sclerosis at 5 years.** *AJNR Am J Neuroradiol* 2011;32:1016–20 CrossRef Medline
8. Rocca MA, Mesaros S, Pagani E, et al. **Thalamic damage and long-term progression of disability in multiple sclerosis.** *Radiology* 2010;257:463–69 CrossRef Medline
9. Azevedo CJ, Cen SY, Khadka S, et al. **Thalamic atrophy in multiple sclerosis: a magnetic resonance imaging marker of neurodegeneration throughout disease.** *Ann Neurol* 2018;83:223–34 CrossRef Medline
10. Filippi M, Rocca MA, Pagani E, et al; ALLEGRO Study Group. **Placebo-controlled trial of oral laquinimod in multiple sclerosis: MRI evidence of an effect on brain tissue damage.** *J Neurol Neurosurg Psychiatry* 2014;85:851–58 CrossRef Medline
11. Gaetano L, Haring DA, Radue EW, et al. **Fingolimod effect on gray matter, thalamus, and white matter in patients with multiple sclerosis.** *Neurology* 2018;90:e1324–32 CrossRef Medline

12. Hanninen K, Viitala M, Paavilainen T, et al. **Thalamic atrophy predicts 5-year disability progression in multiple sclerosis.** *Front Neurol* 2020;11:606 CrossRef Medline

13. Schoonheim MM, Ciccarelli O. **The value of including thalamic atrophy as a clinical trial endpoint in multiple sclerosis.** *Neurology* 2018;90:677–78 CrossRef Medline

14. Meijerman A, Amiri H, Steenwijk MD, et al; Alzheimer's Disease Neuroimaging Initiative. **Reproducibility of deep gray matter atrophy rate measurement in a large multicenter dataset.** *AJNR Am J Neuroradiol* 2018;39:46–53 CrossRef Medline

15. Velasco-Annis C, Akhondi-Asl A, Stamm A, et al. **Reproducibility of brain MRI segmentation algorithms: empirical comparison of local MAP PSTAPLE, FreeSurfer, and FSL-FIRST.** *J Neuroimaging* 2018;28:162–72 CrossRef Medline

16. Bisecco A, Rocca MA, Pagani E, et al; MAGNIMS Network. **Connectivity-based parcellation of the thalamus in multiple sclerosis and its implications for cognitive impairment: a multicenter study.** *Hum Brain Mapp* 2015;36:2809–25 CrossRef Medline

17. FSL MIST. https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/MIST. Accessed May 10, 2023

18. Visser E, Keuken MC, Forstmann BU, et al. **Automated segmentation of the substantia nigra, subthalamic nucleus and red nucleus in 7T data at young and old age.** *Neuroimage* 2016;139:324–36 CrossRef Medline

19. Filippi M, Tedeschi G, Pantano P, et al; INNI Network. **The Italian Neuroimaging Network Initiative (INNI): enabling the use of advanced MRI techniques in patients with MS.** *Neurol Sci* 2017;38:1029–38 CrossRef Medline

20. Lublin FD, Reingold SC. **Defining the clinical course of multiple sclerosis: results of an international survey: National Multiple Sclerosis Society (USA) Advisory Committee on Clinical Trials of New Agents in Multiple Sclerosis.** *Neurology* 1996;46:907–11 CrossRef Medline

21. Storelli L, Rocca MA, Pantano P, et al; INNI Network. **MRI quality control for the Italian Neuroimaging Network Initiative: moving towards big data in multiple sclerosis.** *J Neurol* 2019;266:2848–58 CrossRef Medline

22. Andersson JL, Graham MS, Drobnjak I, et al. **Towards a comprehensive framework for movement and distortion correction of diffusion MR images: within volume movement.** *Neuroimage* 2017;152:450–66 CrossRef Medline

23. Andersson JL, Sotiropoulos SN. **An integrated approach to correction for off-resonance effects and subject movement in diffusion MR imaging.** *Neuroimage* 2016;125:1063–78 CrossRef Medline

24. Behrens TE, Woolrich MW, Jenkinson M, et al. **Characterization and propagation of uncertainty in diffusion-weighted MR imaging.** *Magn Reson Med* 2003;50:1077–88 CrossRef Medline

25. Patenaude B, Smith SM, Kennedy DN, et al. **A Bayesian model of shape and appearance for subcortical brain segmentation.** *Neuroimage* 2011;56:907–22 CrossRef Medline

26. Visser E, Keuken MC, Douaud G, et al. **Automatic segmentation of the striatum and globus pallidus using MIST: multimodal image segmentation tool.** *Neuroimage* 2016;125:479–97 CrossRef Medline

27. Fischl B, Salat DH, Busa E, et al. **Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain.** *Neuron* 2002;33:341–55 CrossRef Medline

28. Fischl B, Salat DH, van der Kouwe AJ, et al. **Sequence-independent segmentation of magnetic resonance images.** *Neuroimage* 2004;23 (Suppl 1):S69–84 CrossRef Medline

29. Battaglini M, Jenkinson M, De Stefano N. **Evaluating and reducing the impact of white matter lesions on brain volume measurements.** *Hum Brain Mapp* 2012;33:2062–71 CrossRef Medline

30. Cootes TF, Taylor CJ, Cooper DH, et al. **Active shape models: their training and application.** *Comput Vis Image Underst* 1995;61:38–59 CrossRef

31. Jacob C. *Statistical Power Analysis for the Behavioral Sciences.* Routledge; 1988:579

32. Rio J, Nos C, Tintore M, et al. **Defining the response to interferon-beta in relapsing-remitting multiple sclerosis patients.** *Ann Neurol* 2006;59:344–52 CrossRef Medline

33. Choi EY, Tian L, Su JH, et al. **Thalamic nuclei atrophy at high and heterogenous rates during cognitively unimpaired human aging.** *Neuroimage* 2022;262:119584 CrossRef Medline