Manual and automated tissue segmentation confirm the impact of thalamus atrophy on cognition in multiple sclerosis: A multicenter study

Jessica Burggraaff, Yao Liu, Juan C. Prieto, Jorge Simoes, Alexandre de Sitter, Serena Ruggieri, Iman Brouwer, Birgit I. Lissenberg-Witte, Mara A. Rocca, Paola Valsasina, Stefan Ropele, Claudio Gasperini, Antonio Gallo, Deborah Pareto, Jaume Sastre-Garriga, Christian Enzinger, Massimo Filippi, Nicola De Stefano, Olga Ciccarelli, Hanneke E. Hulst, Mike P. Wattjes, Frederik Barkhof, Bernard M. J. Uitdehaag, Hugo Vrenken, Charles R.G. Guttmann, on behalf of the MAGNIMS Study Group

Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, and Neurology Unit, Vita-Salute San Raffaele University, Via Olgettina, 58, 20132 Milano MI, Italy

Institute of Pathological Anatomy and Neurosciences, University of Siena, Via Olgettina, 58, 53100 Siena SI, Italy

Department of Neurology, Medical University of Graz, Auenbruggerplatz 22, 8036 Graz, Austria

Department of Neurology, University Hospital iValldHebron, Autonomous University of Barcelona, Passeig de la Vall d’Hebron 119-129, 08035 Barcelona, Spain

Division of Neurology and 3T MRI Research Center, Department of Advanced Medical and Surgical Sciences, University of Campania ‘Luigi Vanvitelli’, Viale Abramo Lincoln, 5, 81100 Caserta, CE, Napoli, Italy

Section of Neuroradiology and MRI Unit, Department of Radiology, University Hospital Vall d’Hebron, Autonomous University of Barcelona, Passeig de la Vall d’Hebron 119-129, 08035 Barcelona, Spain

Division of Neuroimaging, Department of Radiology, Brigham and Women’s Hospital, Harvard Medical School, 1249 Boylston Street, Boston, MA 02215, USA

Department of Neurosciences, Sagissa University of Rome, Piazzale Aldo Moro, 5, 00185 Roma RM, Italy

Department of Neurosciences, San Camillo Forlanini Hospital, Circonvallazione Gianicolica, 87, 00152 Roma RM, Italy

Department of Epidemiology and Biostatistics, Amsterdam UMC, Location VUMc, De Boelelaan 1089a, 1081 HV Amsterdam, the Netherlands

Neuroimaging Research Unit, Institute of Experimental Neurology, Division of Neuroscience, and Neurology Unit, San Raffaele Scientific Institute, Via Olgettina, 58, 20132 Milano MI, Italy

Department of Radiology and Nuclear Medicine, Academic Medical Center, Amsterdam University Medical Center, De Boelelaan 1117, 1118, 1081 HV Amsterdam, The Netherlands

Department of Human Neurosciences, “Sapienza” University of Rome, Piazzale Aldo Moro, 5, 00185 Roma RM, Italy

Department of Radiology and Nuclear Medicine, Amsterdam UMC, Location VUMc, De Boelelaan 1117, 1118, 1081 HV Amsterdam, The Netherlands

Department of Neurology, MS Center Amsterdam, Amsterdam Neuroscience, Amsterdam UMC, Location VUMc, De Boelelaan 1117, 1118, 1081 HV Amsterdam, The Netherlands

Department of Anatomy and Neurosciences, Amsterdam Neuroscience, MS Center Amsterdam, Amsterdam UMC, Vrije Universiteit Amsterdam, De Boelelaan 1108, P. O. Box 7057, 1007 MB, Amsterdam, The Netherlands

Department of Diagnostic and Interventional Neuroradiology, Hannover Medical School, Hannover, Carl-Neuberg-Straße, 30625 Hannover, Germany

Abbreviations: BBBN, Brief Repeatable Battery of Neuropsychological Tests; CAT12, Computational Anatomy Toolbox for Statistical Parametric Mapping 12; CI, cognitively impaired and preserved (CP); CII, cognitive impairment index; CNN, contrast-to-noise ratio; CP, cognitively preserved; EDSS, Expanded Disability Status Scale; eTV, estimated total intracranial volume; FSL-FIRST, FMRIB Integrated Registration and Segmentation Tool; GIF, Geodesic Information Flows; GM, grey matter; GMV, grey matter volume; ICC, intraclass correlation coefficient; MS, Multiple Sclerosis; NBV, Normalized brain volume; NMGV, Normalized grey matter volume; NWMV, Normalized white matter volume; IPS, information processing speed; HC, healthy control; PASAT, Paced Auditory Serial Addition Test; RRMS, Relapsing-Remitting Multiple Sclerosis; SD, standard deviations; SDMT, Symbol Digit Modalities Test; SPM12, Statistical Parametric Mapping 12; SRT, Selective Reminding Test; 10/36 SRT, 10/36 Spatial Recall Test; WCST, Wisconsin Card Sorting Test; WLG, Word List Generation; WM, white matter; WMV, white matter volume; VoilBrain, MRI Brain Volumetry System.

* Corresponding author.

E-mail addresses: jburggraaff@amsterdamumc.nl (J. Burggraaff), yaolu180@163.com (Y. Liu), juanprieto@gmail.com (J.C. Prieto), j.p.simoes@live.com (J. Simoes), a.desitter@amsterdamumc.nl (A. de Sitter), serena.ruggieri@gmail.com (S. Ruggieri), i.brouwer2@amsterdamumc.nl (I. Brouwer), b.lissenberg@amsterdamumc.nl (B.I. Lissenberg-Witte), rocca.mara@hsr.it (M.A. Rocca), valusina.paula@hsr.it (P. Valsasina), stefan.ropele@medunigraz.at (S. Ropele), c.gasperini@libero.it (C. Gasperini), antonio.gallo@unicampania.it (A. Gallo), deborah.pareto.idi@genecat.cat (D. Pareto), jsastre-garriga@cem.cat.org (J. Sastre-Garriga), chris.enzinger@medunigraz.at (C. Enzinger), filippi.massimo@hsr.it (M. Filippi), destefano@uni.it (N. De Stefano), o.ciccarelli@ucu.ac.uk (O. Ciccarelli), h1.hurol@vumc.nl (H.E. Huult), Wartjes.Mike@nh-hannover.de (M.P. Wattjes), f.barkhof@amsterdamumc.nl (F. Barkhof), bmj.uitdehaag@amsterdamumc.nl (B.M.J. Uitdehaag), h.vrenken@amsterdamumc.nl (H. Vrenken), guttmann@bwh.harvard.edu (C.R.G. Guttmann).

Available online 25 December 2020
2213-1582/© 2020 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).
Background and rationale: Thalamus atrophy has been linked to cognitive decline in multiple sclerosis (MS) using various segmentation methods. We investigated the consistency of the association between thalamus volume and cognition in MS for two common automated segmentation approaches, as well as fully manual outlining.

Methods: Standardized neuropsychological assessment and 3-Tesla 3D-T1-weighted brain MRI were collected from 57 MS patients and 17 healthy controls. Thalamus segmentations were generated manually and using five automated methods. Agreement between the algorithms and manual outlines was assessed with Bland-Altman plots; linear regression assessed the presence of proportional bias. The effect of segmentation method on the separation of cognitively impaired (CI) and preserved (CP) patients was investigated through Generalized Estimating Equations; associations with cognitive measures were investigated using linear mixed models, for each method and vendor.

Results: In smaller thalami, automated methods systematically overestimated volumes compared to manual segmentations \((p = (-0.42) - (-0.76), p-values < 0.001)\). All methods significantly distinguished CI from CP MS patients, except manual outlines of the left thalamus \((p = 0.23)\). Poorer global neuropsychological test performance was significantly associated with smaller thalamus volumes bilaterally using all methods. Vendor significantly affected the findings.

Conclusion: Automated and manual thalamus segmentation consistently demonstrated an association between thalamus atrophy and cognitive impairment in MS. However, a proportional bias in smaller thalami and choice of MRI acquisition system might impact the effect size of these findings.

1. Introduction

Cognitive deficits are present in up to 70% of patients with multiple sclerosis (MS) and have a significant effect on their activities of daily living and quality of life (Amato et al., 2010; Chiaravalloti & DeLuca, 2008; Rao et al., 1991). Disturbances in the domains of attention, information processing speed (IPS), memory and executive skills are major features of the MS cognitive profile and can often be detected already early in the disease course (Amato et al., 2010; Rao et al., 1991; Rogers & Panegyres, 2007).

In MS patients, there is increasing evidence of the relationship between cognitive dysfunction and damage to deep grey matter (GM) structures, which is typically measured in vivo from structural magnetic resonance imaging (MRI) (Amiri et al., 2018; Geurts, Calabrese, Fisher, & Rudick, 2012). Especially thalamus atrophy seems strongly associated with cognitive decline (Filippi et al., 2014; Houchtens et al., 2007; Minagar et al., 2013; Schoonheim et al., 2015, 2012). Therefore, thalamus volume is a potential surrogate outcome measure for cognition in multicenter observational and treatment studies. However, when using different segmentation approaches a considerable amount of variability is found in the measurement of thalamus volume, leading to inconclusive results regarding the correlation with cognitive tests (Amiri et al., 2018; Derakhshan et al., 2010; Houchtens et al., 2007; Popescu et al., 2016).

Currently, several software packages are available for measurement of thalamus volume, most of which employ an atlas-based segmentation approach based on information from healthy control (HC) images (Amiri et al., 2018; Geurts et al., 2012). These have been widely applied in MS, but their accuracy and consistency are impacted by various sources of error related to technical factors (e.g., variations in image intensity and tissue contrast due to different MRI hardware and acquisition parameters), variability due to disease related changes (white matter lesion, parenchymal atrophy, etc.) and other physiological / pathological factors (e.g., age, sex, hydration, vascular risk factors etc.) (Amiri et al., 2018; de Sitter et al., 2020; Gelineau-Morel et al., 2012; Rocca et al., 2017a, 2017b; Sastre-Garriga et al., 2020). Given the previously reported limitations of image analysis methods, it is important to understand how consistent and reliable the association between thalamus atrophy and cognition is when using different segmentation approaches in MS patients.

Therefore, the primary aim of this study was to assess the replicability and consistency of the association between thalamus volume and cognitive scores for five automated segmentation methods and fully manual outlining, in a large multi-center cohort of relapsing-remitting MS (RRMS) patients. We chose to compare software packages that are well established, freely available, and widely used throughout the neuroimaging MS research community in order to ensure that our findings would be relevant for future MS neuroimaging studies.

2. Materials and methods

This study was approved by the Local Ethical Committees on human studies in each participating center and all subjects gave written informed consent prior to study participation.

2.1. Subjects

Subjects were recruited from January 2009 to May 2012 as part of a project on imaging correlates of cognitive impairment in MS at 7 European centers (Biscecco et al., 2015; Damjanovic et al., 2017; Preziosa et al., 2016; Rocca et al., 2014; Tillema et al., 2016). Patients had to have a diagnosis of RRMS (Lublin et al., 2014; Polman et al., 2011), no relapse or corticosteroids treatment within the month before scanning and no history of psychiatric conditions, including major depression. Further inclusion criteria for this study required all subjects to be right-handed and aged between 20 and 65 years.

Since manually delineating the thalamus is labor-intensive and time-consuming, a subset of the full multicenter dataset was selected for...
automated and manual tissue segmentation of the thalamus. A random sample of patients and HCs was selected by H.V., matched on age and sex, using a computer-generated list of random numbers. The final dataset included 57 RRMS patients [37 females; age 38.9 ± 8.5 (mean ± standard deviations (SD) years); 13.0 (7.0–20.0) (median (range)) years of education] and 17 HCs [12 females; age 40.5 ± 6.6 (mean ± SD) years; 17.0 (8.0–20.0) (median (range)) years of education]. See Table 1 for demographic and clinical variables. Patients had a median (range) disease duration of 6.0 (2.0–33.0) years, and a median (range) Expanded Disability Status Scale (EDSS) score of 2.0 (0.0–6.0). Age, sex and education did not differ between HCs and MS patients (p = 0.47; p = 0.66 and p = 0.12, respectively).

2.2. Clinical and cognitive evaluation

Within 48 hours of the MRI acquisition, MS patients underwent a neurological evaluation including EDSS score and a neuropsychological assessment (see Table 1), performed at each participating site by experienced neurologists and neuropsychologists, unaware of the MS results, using validated translations of the neuropsychological tests. For all patients, cognitive performance was assessed by using the Brief Repeatable Battery of Neuropsychological Tests (BBB-N) (Rao et al., 1999), which includes the Selective Reminding Test (SRT) to assess verbal memory; the 10/36 Spatial Recall Test (10/36 SRT) to assess visuospatial memory; the Symbol Digit Modalities Test (SDMT) and verbal memory; the Selective Reminding Test (SRT) to assess visuospatial memory; the Symbol Digit Modalities Test (SDMT) and the Wisconsin Card Sorting Test (WCST) (Heaton et al., 1990), which includes the Selective Reminding Test (SRT) to assess verbal memory; the 10/36 Spatial Recall Test (10/36 SRT) to assess visuospatial memory; the Symbol Digit Modalities Test (SDMT) and the Wisconsin Card Sorting Test (WCST) (Heaton et al., 1993) were considered cognitively impaired (CI), as previously described (Damjanovic et al., 2017; Preziosa et al., 2016). In all MS patients, a cognitive impairment index (CII) was determined as an overall measure of cognitive dysfunction for each patient. Briefly, the CII is a continuous variable obtained by a grading system applied to each patient’s score on every cognitive test, dependent on the number of SDs below the mean normative value (Amato et al., 2006; Camp et al., 1999). Hence, the higher the grade, the greater the patient’s impairment.

2.3. MRI acquisition

MR images were acquired on 3 T scanners (Amsterdam and Naples: Signa, GE Healthcare, Milwaukee, Wisconsin; Barcelona and London: Magnetom Trio, Siemens, Erlangen, Germany; Milan and Siena: Philips Inter, Best, the Netherlands). The brain imaging sequences included: (a) a dual-echo turbo-spin-echo (TSE) T2-weighted scan: TR = 4000–5580 ms; TE1 = 10–20 ms; TE2 = 90–102 ms; echo-train length = 5–11; 44 contiguous, 3-mm-thick axial sections parallel to the anterior/posterior commissure plane; matrix = 256 × 192; FOV = 240 × 180 mm2 (rectangular FOV = 75%); (b) three-dimensional (3D) T1-weighted scan: TR = 5.5–8.3 ms (for GE Healthcare/Philips Inter scanners) or 1900–2300 ms (for Siemens scanners); TE = 1.7–3.0 ms; flip angle = 8°–12°; 176–192 sagittal sections with thickness = 1 mm and in-plane resolution = 1 × 1 mm. All scans were visually inspected for quality.

2.4. MRI analysis of lesions and global atrophy

The analysis of lesions and global atrophy on structural MRI data was done centrally at the Neuroimaging Research Unit (Milan, Italy) by experienced observers under supervision of a neurologist (M.A.R.) with 20 years of experience, blinded to the subjects’ identity. T2 hyperintense lesion volumes (LV) were measured on dual-echo TSE images in a semi-automated fashion using a local thresholding segmentation technique (Jim 6.0 software; Xinapse Systems, Colchester, UK). Normalized brain (NBV), normalized white matter (WM) and grey matter (GM) volumes were measured on 3D T1-weighted scans using the SIENAX software (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/SIENAX) (Smith et al., 2002), after WM lesion-filling with LEAP (Chard, Jackson, Miller, & Wheeler-Kingshott, 2010), using co-registration of the T2 lesion masks to the 3D T1-weighted scans (Popescu et al., 2014).

2.5. Thalamus volume measurements

Manual and automated volumetric analyses of the thalamus were performed on 3D T1-weighted data sets.

2.5.1. Manual delineations

Manual volumetric analysis was performed within the online framework of the SPINE virtual laboratory (https://spinevirtuallab.org/), developed by the Center for Neurological Imaging at Brigham and Women’s Hospital, which can be used for manual tracing of regions-of-interest on MRI. This web-based program allows visualization of MR images in axial, coronal, and sagittal orientations to facilitate 3D anatomical interpretation. The delineations were performed according to a standardized protocol (see supplementary material for a detailed description of the anatomical definitions and detailed outlining instructions) and the voxel-wise labeling process was completely manual; that is, it involved no thresholding, seed-growing, shape fitting or other automated interference. One expert reader manually delineated the whole thalamus on axial slices, in a slice-by-slice manner. To assess the long-term test–retest reliability, a random subset of thalami for nine MR images (4 HCs and 5 MS patients) were delineated in a separate session more than three months later. The reader was a neurologist (J.B.), with specialized training and experience in the anatomical labeling of deep GM structures on MRI, supervised by a neuroradiologist (F.B. with more than 30 years of experience). The reader was blinded to the subject’s clinical characteristics.

2.5.2. Automated segmentation methods

FreeSurfer’s (http://surfer.nmr.mgh.harvard.edu/) (Dale, Fischl, & Sereno, 1999; Fischl, Sereno, & Dale, 1999) volume-based stream is designed to preprocess MRI volumes and label subcortical structures. The stream consists of multiple stages: in brief, the first stage is an affine registration with Talairach space specifically designed to be insensitive to pathology and to maximize the accuracy of the final segmentation. This is followed by an initial tissue classification and correction of the variation in intensity resulting from the B1 bias field. Finally, there is a high-dimensional nonlinear volumetric alignment to the Talairach atlas, where the final segmentation takes place. The manual editing steps that are recommended for FreeSurfer to adjust for cortical reconstructions were excluded here, since we are focusing on the subcortical output; FreeSurfer was applied as a fully automated software, without the addition of any manual editing steps.

FIRST (Patenaude, Smith, Kennedy, & Jenkinson, 2011) is a model-based segmentation tool also part of FSL (http://www.fmrib.ox.ac.uk/fsl/first/index.html) (Smith et al., 2002). Subcortical brain...
segmentation is performed using Bayesian shape and appearance models constructed from a set of 336 manually-labeled T1-weighted MR images. FIRST models the outer surface of each deep GM structure as a mesh, using models derived from the reference images and the local intensity profiles around the mesh. Finally, it assigns each voxel in the image an appropriate structure label, taking into account local variations in both surface and shape, as well as the presence of neighboring structures.

The CAT12 toolbox (the successor of VBM8) is an extension to SPM12 (http://www.fil.ion.ucl.ac.uk/spm/software/spm12/) to provide computational anatomy (Mutsaerts et al., 2020). The algorithm allows local variations in the tissue intensity distributions, making it more robust to the presence of pathology such as WM lesions.

GIF software (part of NiftySeg: http://cmictig.cs.ucl.ac.uk/niftyweb/) uses manually created atlases for segmentation of the input images (http://www.neuromorphometrics.com/) (Klein & Tourville, 2012). GIF captures the local variation in morphology and standard space locations, and has been recommended in previous studies on (deep) GM atrophy in MS (Eshaghi et al., 2018). With the use of an iterative geodesic minimization algorithm and the manual labels, more accurate segmentations are expected (Cardoso et al., 2015).

The VolBrain fully automated pipeline provides volumetric brain information at different scales (Manjon & Coupé, 2016). The proposed pipeline is based on a library of manually labeled templates to perform the segmentation process, constructed from subjects from different publicly available datasets (normal adults, Alzheimer disease and pediatric datasets), including subcortical structure segmentation as proposed by Coupé et al. 2011 (Coupé et al., 2011).

2.6. Normalization

To correct for the influence of head size, thalamus volumes were multiplied by the head-normalization factor derived from SIENAX for all segmentation methods, including the manual segmentations. Alternatively, FreeSurfer segmentations were divided by the estimated total intracranial volume (eTIV) from FreeSurfer. The unnormalized data were used for the evaluation of agreement between methods; the normalized data served for the association analyses with cognitive outcomes.

2.6.1. Contrast-to-noise ratio

To assess whether there were different tissue contrasts in the T1-weighted images obtained at different sites in this multi-center study, as well as to assess if this was related to the observed relation with cognitive measures, we quantified the contrast-to-noise ratio (CNR) for each thalamus (left and right separately, in each subject). This was done as follows: The mean signal in the thalamus was calculated by eroding the manual thalamus outline once using a 3x3x3 kernel (to avoid any chance of partial volume effects from thalamus) and applying this as a mask on the N3-corrected T1-weighted image, and calculating the mean signal intensity in that region. The mean signal intensity in the WM bordering the thalamus was obtained similarly, but in this case the mask was created by first dilating the manual thalamus mask once, using a 3x3x3 kernel. The border region was then masked with the SIENAX WM mask and then creating a border region around that expanded thalamus mask by dilating three times using a 3x3x3 kernel. The border region was then masked with the SIENAX WM mask and with the inverse of the lesion mask, to exclude GM, CSF and lesions. This WM border mask was then applied on the N3-corrected T1-weighted image and the mean signal intensity was calculated. Subsequently, the standard deviation of the image noise was approximated by taking the standard deviation of the signal in the ventricular CSF. The FreeSurfer ventricles segmentation, after excluding choroid plexus, was eroded once using a 3x3x3 kernel to avoid partial volume effects, and then applied as a mask on the N3-corrected T1-weighted image, and the standard deviation was calculated. Finally, the CNR for that thalamus was calculated by dividing the absolute difference between the mean thalamus signal intensity and the mean border WM signal intensity, by the standard deviation of the ventricles.

2.7. Statistical analysis

All data analysis was done using SPSS for Windows version 22.0 (Armonk, NY: IBM Corp.). The normality of each variable’s distribution was assessed using histograms and normality plots. Group differences of the demographical and clinical variables, as well as the volumetric MRI quantities and scanner type were evaluated using independent sample T-tests for normally distributed variables, non-parametric analysis (Mann-Whitney) for non-normally distributed variables, and Chi² for categorical variables. Brain T2 and T1 LV were log-transformed due to their skewed distribution. Mean and standard deviation of CNR values were reported both per site and per vendor / scanner type.

Volumetric agreement of the manually and automatically generated thalamus segmentations was evaluated through the intraclass correlation coefficient (ICC) based on a two-way mixed effects model, where people effects are random and measure effects are fixed (McGraw & Wong, 1996). The absolute agreement (ICC “type A”) and consistency (ICC “type C”) were reported. Further, to describe the agreement between different segmentation methods, Bland-Altman plots were created in which the difference of two paired measurements (A-B) was plotted against the average of the two measurements [(A + B)/2] (Altman & Bland, 1983; Giavarina, 2015). We ran a One-Sample T-Test to examine whether the mean of the difference equals 0, and a linear regression (Pearson rho (ρ)) to evaluate whether a proportional bias was present. In the Bland-Altman plot this bias will be reflected in the scatter points with a trend to high or low values of the difference across the range of values of the average.

Intra-rater reliability of the manual delineations was evaluated through the ICC as described above, reporting the absolute agreement. We used Koo’s criteria to interpret the ICCs: values < 0.5 are indicative of ‘poor’ reliability, values between 0.5 and 0.75 indicate ‘moderate’ reliability, values between 0.75 and 0.9 indicate ‘good’ reliability, and values greater than 0.90 indicate ‘excellent’ reliability (Koo & Li, 2016).

The ability of the thalamus volumes to distinguish between CI and CP MS patients was compared between different segmentation methods by using Generalized Estimating Equations with logit link function and an unstructured covariance matrix, corrected for age. Correlations of cognition with thalamus volumes were investigated using linear mixed models CII and cognitive domain Z-scores as the dependent variables, adjusting for age and with random effects for subject and center, comparing the results between the different segmentation methods. Sex and education were not significantly different between CI and CP patients and were not retained in the models. To assess the influence of vendor, we additionally performed the same general linear regression analysis with CII as the dependent variable for each method, per vendor.

A p-value of <0.05 was considered statistically significant. As the main goal of our study was to investigate the replicability of the association between thalamus volume and cognitive scores using different automated segmentation methods, we did not correct for multiple comparisons to address possible type I errors.

3. Results

3.1. Subject characteristics

Table 2 summarizes the main demographic, clinical and MRI characteristics of the HCs and MS patients, as well as CP and CI MS patient subgroups. Twenty-two (39%) MS patients were classified as CI. Compared with CP, CI patients were older (p = 0.01) and had a higher EDSS score (p = 0.025); whereas no difference was found for sex (p = 0.33), education (p = 0.52) and disease duration (p = 0.83). As a consequence, age was included as nuisance covariate in the regression models. Compared to HCs, MS patients had lower NBV (p = 0.001), NWMV (p = 0.01) and NGMV (p < 0.05). Except NWMV (p = 0.33), all
MRI volumes were more altered in CI than in CP patients (all p-values < 0.05), including T2 LV (p < 0.01). The cognitive domains most frequently affected were attention / IPS (32% of the MS patients), executive function (23%), verbal memory (19%), visuospatial memory (16%) and verbal fluency (16%). The distribution of vendors across the HC and MS patient subgroups was similar (MS vs HC: p = 0.08; CI vs CP: p = 0.51). Table 3 lists the number of subjects per center and MR scanner type. CNR values by site and hemisphere are also included, displaying some heterogeneity between sites in this multi-center study.

### 3.2. Analysis of volumetric agreement

#### 3.2.1. Intraclass correlation analysis

Fig. 1 shows examples of the segmentations for each method. In terms of consistency, the agreement between the automatically and manually generated left and right thalamus volumes was good for FreeSurfer and FSL-FIRST, with ICC values ≥ 0.77, and moderate for CAT12, GIF and VolBrain (ICC: 0.61–0.75) (Table 4). In terms of absolute agreement, ICC values were good for FreeSurfer (≥0.79), and moderate for FSL-FIRST (≥0.68). Poor absolute agreement was found for left and right thalamus volume measurements from CAT12 (ICC 0.20 and 0.21), GIF (ICC 0.44 and 0.47) and VolBrain (ICC 0.39 and 0.42).

#### 3.2.2. Bland-Altman scatter plots and analysis

Fig. 2 and Table 5 describe the results of the Bland-Altman scatter plots and analysis of the unnormalized thalamus volume measurements: automated minus the manual methods. On average, FreeSurfer left thalamus volumes were similar to the manual output, while right thalamus were larger [mean difference (SD): left thalamus: −0.09 (0.85), p = 0.39; right thalamus: 0.36 (0.79), p < 0.001]. FSL-FIRST obtained larger thalamus volumes for both hemispheres [left thalamus: 0.69 (0.92), p < 0.001; right thalamus: 0.60 (0.88), p < 0.001]. In comparison, the software packages CAT12, GIF and VolBrain obtained smaller thalamus volumes bilaterally (all p-values < 0.001). Except for CAT12, a proportional difference with a negative trend was observed in all scatter plots showing the agreement between the automated and manual thalamus volume measurements. In smaller thalami the automated methods appeared to systematically overestimate the thalamus volumes compared to manual outlines, whereas in larger thalami the reverse was found. Qualitatively, the areas with the most disagreement occurred in the superior and inferior parts of the thalami, including the geniculate bodies (see Fig. 1).

### 3.3. Reproducibility of manual thalamus outlining

The long-term intra-rater reliability of the manual output, assessed on the images of 9 subjects, was moderate with a median ICC (absolute agreement) of 0.62 (p < 0.01) for the left thalamus and 0.63 (p < 0.001) for the right thalamus.

### 3.4. Relation of thalamus volume measures with cognition

#### 3.4.1. Thalamus volumes

Table 6 lists the normalized left and right thalamus volumes obtained through manual tracings and automated techniques in CI and CP MS patients. Compared to CP patients, CI patients had smaller thalami based on all methods, excepted the left thalamus volumes obtained through manual outlining (p = 0.18) and marginally significant for left thalamus volumes from GIF (p = 0.05). All segmentation methods consistently demonstrated smaller thalami in MS patients than in HCs (all p-values < 0.001; not shown in the table). In both HCs and MS subjects, the right thalamus were smaller than the left thalami for all methods. This difference in left and right thalamus volumes was not statistically significant between methods (p = 0.79 for both HCs and MS patients; not shown in the table).

#### 3.4.2. Consistency of discrimination between cognitively impaired and preserved patients

Table 6 summarizes the results of the binary logistic regression analysis for the discrimination between CI and CP MS patients, using the...
6

For example, CII is expected to increase by 1.45 (lower left and right thalamus volumes using all segmentation methods, logical test performance (higher CII) was significantly associated with patients (odds ratios: 0.44). Patients were significantly more likely to have smaller thalami than CP. The effect was found for all segmentation methods, indicating that CI patients consistently demonstrated a relationship between the degree of thalamic atrophy and cognitive dysfunction, which suggests that the observed association is truly a manifestation of the disease. However, the robustness of these associations was systematically affected by scanner. Somewhat surprisingly, our results showed that images with lower CNR resulted in more significant correlations with cognitive measures, warranting further and more systematic studies of these.

### 3.4.3. Analysis of correlations with cognition

After normalization through SIENAX, poorer global neuropsychological test performance (higher CII) was significantly associated with lower left and right thalamus volumes using all segmentation methods, (table 7). For example, CII is expected to increase by 1.45 (p = 0.021), 1.26 (p = 0.002), 1.22 (p = 0.002), 1.06 (p = <0.001), 1.05 (p = 0.013) and 0.65 points (p = 0.032), when the left thalamus volume decreases by one centimeter when obtained through GIF, FreeSurfer, VolBrain, CAT12, FSL-FIRST and manual outlining, respectively. Normalization through FreeSurfer (eTIV) also resulted in significant correlations between CII and thalamus volumes for FreeSurfer. Table 8 shows the associations between CII and thalamus volume measurements for each method, for each scanner vendor (GE, Philips or Siemens) separately. Volumes that were obtained with Siemens scanners resulted in significant correlations for all methods (p-values: 0.001–0.031). Philips scans only showed significant correlations when analyzed with CAT12 (bilaterally: p = 0.007 and 0.038), FreeSurfer (right thalamus: p = 0.045) and FSL-FIRST (left thalamus: p = 0.043). No associations were found for any of the methods when applied to GE images. These correlations seem to be in contradiction with the CNR results by vendor, listed at the bottom of Table 8, which show that in fact the CNR values were lowest for Siemens and highest for GE.

### 3.4.4. Analysis of correlations with performance scores on separate cognitive domains

Looking at the correlation with cognitive domain z-scores (table 7), thalamus volume loss was associated with visuospatial memory and attention / IPS based on all methods, excepted a lack of statistically significant association between manually segmented left thalamus volume and visuospatial memory. Based on CAT12, right thalamus volume was associated with verbal fluency (p = 0.044) and executive function (p = 0.045). No associations were found with the other cognitive domain z-scores. Similar results were found for the normalized (eTIV) FreeSurfer thalamus volume measurements, except that a significant correlation between left thalamus volume loss and verbal memory was also found using this method (p = 0.03).

### 4. Discussion

In this multi-center cohort, RRMS patients with relatively mild physical disability and overt CI showed severe thalamus atrophy based on all automated segmentation techniques, as was also evidenced by a unique set of manually defined reference outlines in which the whole thalamus was segmented. Automated and manual tissue segmentation consistently demonstrated a relationship between the degree of thalamus atrophy and cognitive dysfunction, which suggests that the observed association is truly a manifestation of the disease. However, the robustness of these associations was systematically affected by scanner. Somewhat surprisingly, our results showed that images with lower CNR resulted in more significant correlations with cognitive measures, warranting further and more systematic studies of these.

### Table 4

| Intraclass Correlation | Freesurfer – Manual | FSL-FIRST – Manual | CAT12 – Manual | GIF – Manual | VolBrain – Manual |
|------------------------|---------------------|------------------|---------------|-------------|------------------|
|                       | Absolute            | Consistency      | Absolute      | Consistency | Absolute          |
| Left thalamus          | 0.81                | 0.80             | 0.68          | 0.77        | 0.20             |
| Right thalamus         | 0.79                | 0.82             | 0.69          | 0.77        | 0.21             |

Abbreviations: Absolute = absolute agreement; a,b. Two-way mixed effects model where people effects are random and measures effects are fixed, single measures. Intraclass correlation coefficients are displayed; b p = <0.001 for all variables.

---

Fig. 1. 3D T1-weighted images and thalamus segmentations of manual tracing, FreeSurfer, FSL-FIRST, CAT12, GIF and VolBrain. Segmentations of the thalamus bilaterally in the axial plane of two MS patients, revealing the inferior portion of the thalamus of one cognitively impaired patient (top row) and the middle part of the thalamus of one cognitively preserved patient (bottom row) for: manual tracings (A), FreeSurfer (B), FSL-FIRST (C), CAT12 (D), GIF (E) and VolBrain (F) segmentations.
issues. The differential bias present in smaller and larger thalami should be taken into account when evaluating treatment response of therapeutic interventions.

To our knowledge, this is the first multicenter study that compared automated thalamus segmentation methods and manual outlining, and evaluated their influence on the association of thalamus volume with cognition in MS patients in the presence of MS-related pathologies. Earlier research on this topic considered single-scanner data only.

![Bland Altman scatter plots of the unnormalized thalamus volume measurements of the MS patients.](image)

The difference of two paired measurements \(\frac{\text{automated} - \text{manual}}{\text{average}}\) was plotted against the average of the two measurements \(\frac{\text{automated} + \text{manual}}{2}\). Except for CAT12, a proportional bias was observed between the automated and manual thalamus volume measurements, indicated by a trend [linear regression (Pearson rho ($\rho$))] to high and low values of the difference across the range of values of the average.
the present study that expert manual outlining, by and large, resulted in
et al., 2010; Sastre-Garriga et al., 2020). Taken together, the finding of
thology such as WM lesions and atrophy (Amiri et al., 2018; Derakhshan
Derakhshan et al., 2010; Popescu et al., 2016). When aiming to fully
demonstrated more severe thalamus damage in CI patients (Benedict
scores across all segmentation software methods, further studies are
required to more systematically study the interplay between image
occurrence of lowest CNR and significant correlations with cognitive
threshold at 0.05. (a) Data are mean (SD) for normally distributed variables; b Thalamic volumes were multiplied by the head
normalization factor derived from SIENAX; c Thalamic volumes were divided by the estimated total intracranial volume (eTIV) from FreeSurfer; p-values in bold represent significant values.
measurements. Impaired IPS is a common and highly invalidating deficit
in MS, which can occur at the earliest stages of the disease (Amato et al.,
2010; Chiaravalloti & DeLuca, 2008; Rao et al., 1991). With its extensive
afferent and efferent interconnections with the midbrain and the cerebro
thalamic degeneration is likely to contribute to IPS dysfunction (Minagar et al.,
Although the present work confirms that the thalamus is of great
clinical relevance to cognitive processes in MS, considerable variations
were observed between software packages and scanners, which co
incides with the variability reported by previous investigators (Amiri et al.,
2018; Glaister et al., 2017; Popescu et al., 2016). In line with an
earlier report by Glaister et al, visual inspection of our data showed that
the areas with most disagreement occurred in the superior and inferior
parts of the thalami, including the geniculate bodies (Glaister et al.,
2017). This is probably due to their low contrast compared to sur
rounding tissue in T1-weighted MRI, which makes it more complicated
to trace the edges of the thalamus in these subregions, also manually.

### Table 5
Pairwise Bland-Altman comparisons between segmentation methods.

| Measure       | µ diff | SD  | SEµ | p- Value | Proportional bias | φ (rho) | t  | p- Value |
|---------------|--------|-----|-----|----------|-------------------|---------|----|----------|
| Freesurfer -  |        |     |     |          |                   |         |    |          |
| Manual        |        |     |     |          |                   |         |    |          |
| Thalamus Left | −0.09  | 0.85| 0.10| 0.391    | −0.44             | −4.14   | <0.001 |          |
| Thalamus Right| 0.36   | 0.79| 0.09| <0.001   | −0.42             | −3.98   | <0.001 |          |
| FSL - first   |        |     |     |          |                   |         |    |          |
| Manual        |        |     |     |          |                   |         |    |          |
| Thalamus Left | 0.69   | 0.92| 0.11| <0.001   | −0.44             | −4.12   | <0.001 |          |
| Thalamus Right| 0.60   | 0.88| 0.10| <0.001   | −0.48             | −4.58   | <0.001 |          |
| CAT12 -       |        |     |     |          |                   |         |    |          |
| Manual        |        |     |     |          |                   |         |    |          |
| Thalamus Left | −2.87  | 1.25| 0.15| <0.001   | −0.17             | −1.46   | 0.15  |          |
| Thalamus Right| −2.75  | 1.18| 0.14| <0.001   | −0.16             | −1.33   | 0.19  |          |
| GIF - Manual  |        |     |     |          |                   |         |    |          |
| Thalamus Left | −1.04  | 1.08| 0.13| <0.001   | −0.74             | −9.30   | <0.001 |          |
| Thalamus Right| −1.02  | 0.98| 0.11| <0.001   | −0.76             | −9.82   | <0.001 |          |
| VollBrain -   |        |     |     |          |                   |         |    |          |
| Manual        |        |     |     |          |                   |         |    |          |
| Thalamus Left | −1.65  | 1.05| 0.12| <0.001   | −0.42             | −3.92   | <0.001 |          |
| Thalamus Right| −1.63  | 0.91| 0.11| <0.001   | −0.47             | −4.48   | <0.001 |          |

### Table 6
Normalized thalamic volume measurements and summary of results of the bi
nary logistic regression analysis for cognitively impaired versus cognitively
preserved MS patients

| Measure       | µ diff | SD  | SEµ | p- Value | Normalization |
|---------------|--------|-----|-----|----------|---------------|
| Thalamus Left |        |     |     |          | SIEANAX²      |
| Thalamus Right|        |     |     |          | Manual outlines |
| Thalamus Left | 8.99   | 8.24| 0.18| 0.65     | 0.85          |
| Thalamus Right| 9.06   | 7.91| 0.033| 0.72 | 0.52         |
| Thalamus Left | 9.02   | 8.11| 0.012| 0.64 | 0.42         |
| Thalamus Right| 8.59   | 7.54| 0.002| 0.53 | 0.37         |

### Abbreviations
- µ diff = mean difference; SD = standard deviation; SEµ = standard error of µ; φ (rho) = Pearson correlation; t = t-test statistic; a Correlation of the volume difference and mean between two measurements; p-value in bold represent significant values.
- CI = cognitively impaired; Conf int = confidence interval; CP = cognitively preserved; OR = odds ratio.
- a, b Data are mean (SD) for normally distributed variables; b Thalamic volumes were multiplied by the head-normalization factor derived from SIENAX; c Thalamic volumes were divided by the estimated total intracranial volume (eTIV) from FreeSurfer; p-values in bold represent significant values.

(Glaister et al., 2017; Houchens et al., 2007; Popescu et al., 2016); or
compared automated techniques without including manual outlining
(Derakhshan et al., 2010; Popescu et al., 2016). When aiming to fully
understand the relationship between thalamus atrophy and cognitive
decline, automated methods may present a biased picture or reflect
spurious correlations, since there have been reports that the algorithms
may yield measurement errors that increase with increasing MS
pathology such as WM lesions and atrophy (Amiri et al., 2018; Derakhshan et al., 2010; Sastre-Garriga et al., 2020). Taken together, the finding of
the present study that expert manual outlining, by and large, resulted in
the same associations with cognition as automated methods, is an
important confirmation of many earlier reports that have consistently
demonstrated more severe thalamus damage in CI patients (Benedict et al., 2013; Houchens et al., 2007; Minagar et al., 2013; Popescu et al., 2016; Rocca et al., 2018; Schoonheim et al., 2015, 2012). Of note, attention to variations in image characteristics, in particular the CNR
between target structure (thalamus) and surrounding tissue, between
different scanners and protocols is essential, especially when attempting
to minimize the number of patients and observations needed to
adequately power clinical trials relying on MRI-derived measurements.
Based on our results, which for Siemens showed an unexpected co
occurrence of lowest CNR and significant correlations with cognitive
scores across all segmentation software methods, further studies are
required to more systematically study the interplay between image
contrast, image noise and thalamus segmentation quality.

Similarly to previous studies (Batista et al., 2012; Benedict et al.,
2013; Houchens et al., 2007; Schoonheim et al., 2015, 2012), impaired
performance on the domains of attention / IPS and visuospatial memory
were associated with thalamus degeneration bilaterally, which was also
confirmed through manual outlining. In contrast, we did not find a
correlation with executive function, except using CAT12 right thalamus

### Abbreviations
- µ diff = mean difference; SD = standard deviation; SEµ = standard error of µ; φ (rho) = Pearson correlation; t = t-test statistic; a Correlation of the volume difference and mean between two measurements; p-value in bold represent significant values.
Table 7
Summary of results of the general linear regression analysis with cognitive scores as the dependent variables\(^a\)

| Thalamic volumes | Cognitive Impairment Index (CII) | Verbal memory | Visually memory | Attention | Fluency | Executive function\(^b\) |
|------------------|---------------------------------|---------------|----------------|----------|---------|-------------------------|
|                  | B (95% CI) p-Value              | B (95% CI) p-Value | B (95% CI) p-Value | B (95% CI) p-Value | B (95% CI) p-Value | B (95% CI) p-Value |
| Normalization    |                                 |               |                 |          |         |                         |
| Sienax\(^c\)     |                                 |               |                 |          |         |                         |
| Left thalamus    | –0.65 (-1.23 – (-0.06)) 0.032 | –0.06 (0.08 – 0.21) 0.38 | 0.15 (0.002 – 0.30) 0.053 | 0.16 (0.04 – 0.28) 0.010 | 0.07 (-0.07 – 0.22) 0.31 | 0.19 (-1.44 – 1.83) 0.19 |
| Right thalamus   | –0.72 (-1.32 – (-0.13)) 0.017 | 0.09 (-0.06 – 0.23) 0.22 | 0.17 (0.02 – 0.32) 0.032 | 0.18 (0.06 – 0.30) 0.004 | 0.11 (-0.03 – 0.26) 0.11 | –0.08 (-1.74 – 1.58) 0.92 |
| Freesurfer       |                                 |               |                 |          |         |                         |
| Left thalamus    | –1.26 (-2.05 – (-0.47)) 0.002 | 0.17 (-0.03 – 0.37) 0.091 | 0.33 (0.12 – 0.53) 0.002 | 0.32 (0.16 – 0.47) <0.001 | 0.18 (-0.02 – 0.38) 0.076 | –0.45 (-2.74 – 1.85) 0.67 |
| Right thalamus   | –1.36 (-2.18 – (-0.53)) 0.002 | 0.11 (-0.10 – 0.32) 0.29 | 0.40 (0.20 – 0.60) <0.001 | 0.33 (0.16 – 0.49) <0.001 | 0.19 (-0.02 – 0.40) 0.070 | 0.63 (1.76 – 3.03) 0.60 |
| FSL-First        |                                 |               |                 |          |         |                         |
| Left thalamus    | –1.05 (-1.87 – (-0.23)) 0.013 | 0.09 (-0.12 – 0.29) 0.40 | 0.32 (0.12 – 0.53) 0.003 | 0.31 (0.15 – 0.47) <0.001 | 0.15 (-0.06 – 0.35) 0.16 | 1.36 (-0.91 – 3.63) 0.23 |
| Right thalamus   | –0.93 (-1.81 – (-0.05)) 0.039 | 0.05 (-0.17 – 0.26) 0.67 | 0.34 (0.12 – 0.55) 0.003 | 0.30 (0.13 – 0.47) 0.001 | 0.15 (-0.06 – 0.37) 0.15 | 1.22 (-1.18 – 3.62) 0.31 |
| CAT12            |                                 |               |                 |          |         |                         |
| Left thalamus    | –1.06 (-1.63 – (-0.49)) <0.001 | 0.07 (-0.08 – 0.22) 0.34 | 0.25 (0.10 – 0.39) 0.002 | 0.26 (0.14 – 0.37) <0.001 | 0.13 (-0.01 – 0.28) 0.072 | 1.58 (-0.05 – 2.32) 0.57 |
| Right thalamus   | –1.12 (-1.71 – (-0.53)) <0.001 | 0.08 (-0.07 – 0.24) 0.29 | 0.27 (0.12 – 0.43) 0.001 | 0.27 (0.16 – 0.39) <0.001 | 0.16 (0.004 – 0.31) 0.044 | 1.74 (-1.74 – 1.58) 0.045 |
| GIF              |                                 |               |                 |          |         |                         |
| Left thalamus    | –1.45 (-2.67 – (-0.23)) 0.021 | 0.17 (-0.13 – 0.47) 0.26 | 0.40 (0.10 – 0.71) 0.010 | 0.48 (0.24 – 0.71) <0.001 | 0.27 (-0.03 – 0.57) 0.074 | 0.26 (-3.16 – 3.67) 0.88 |
| Right thalamus   | –1.36 (-2.61 – (-0.11)) 0.033 | 0.14 (-0.17 – 0.44) 0.38 | 0.39 (0.07 – 0.70) 0.016 | 0.45 (0.21 – 0.70) <0.001 | 0.30 (0.0004 – 0.60) 0.050 | 0.40 (-3.86 – 0.82) 0.82 |
| VolBrain         |                                 |               |                 |          |         |                         |
| Left thalamus    | –1.22 (-1.98 – (-0.46)) 0.002 | 0.12 (-0.07 – 0.32) 0.20 | 0.35 (0.16 – 0.54) <0.001 | 0.34 (0.19 – 0.49) <0.001 | 0.15 (-0.04 – 0.34) 0.13 | 1.27 (-0.90 – 3.45) 0.25 |
| Right thalamus   | –1.30 (-2.10 – (-0.51)) 0.002 | 0.12 (-0.08 – 0.32) 0.24 | 0.37 (0.17 – 0.57) <0.001 | 0.34 (0.19 – 0.50) <0.001 | 0.17 (-0.03 – 0.37) 0.10 | 1.63 (-0.62 – 3.88) 0.15 |
| Fraction of eTV |                                 |               |                 |          |         |                         |
| Freesurfer\(^d\) |                                 |               |                 |          |         |                         |
| Left thalamus    | –2.51 (-4.13 – (-0.89)) 0.003 | 0.45 (-0.03 – 0.37) 0.03 | 0.59 (0.16 – 1.03) 0.009 | 0.65 (0.34 – 0.96) <0.001 | 0.10 (-0.32 – 0.52) 0.63 | 1.03 (-3.67 – 5.73) 0.66 |
| Right thalamus   | –2.58 (-4.24 – (-0.93)) 0.003 | 0.30 (-0.12 – 0.71) 0.16 | 0.73 (0.32 – 1.14) 0.001 | 0.65 (0.32 – 0.97) <0.001 | 0.11 (-0.32 – 0.53) 0.62 | 3.03 (-1.69 – 7.75) 0.20 |

Abbreviations: B = unstandardized regression coefficient; CI = confidence interval; \(^a\) All regression analysis were corrected for center and age; \(^b\) WCST number of perseverative errors; \(^c\) Thalamic volumes were multiplied by the head-normalization factor derived from Sienax; \(^d\) Thalamic volumes were divided by the estimated total intracranial volume (eTV) from Freesurfer; p-values in bold represent significant values.

The Bland Altman plots revealed that thalamic volumes were on average overestimated by FSL-FIRST and Freesurfer (excepted left thalamus measurements), while they were systematically underestimated by CAT12, GIF and VolBrain, which is in line with an earlier publication on this topic (de Sitter et al., 2020). It appeared that the absolute agreement for CAT12 (ICC: 0.20–0.21), GIF and VolBrain (ICCs between 0.39 and 0.47) in our study were much worse than previously reported by de Sitter et al. (2020). However, different study populations and combined manual segmentations created by majority voting were used in previous work. Further investigations are needed to unravel in more detail the mechanisms leading to the observed differences between different segmentation pipelines.

Furthermore, the analysis of agreement between the software packages and manual outlines revealed important insights into how MS pathological changes may affect the association between thalamus atrophy and cognitive outcome. First, Bland-Altman revealed a proportional bias with a negative trend of differences between virtually all automated segmentation techniques included in this study (excepted CAT12) and manually derived thalamus measurements, proportional to the magnitude of thalamus size. It seems therefore that the algorithms...
inconsistently localized effects. Nevertheless, automated thalamus segmentation yielded larger effect sizes for the separation of CI vs CP MS patients than manually derived volumes. These discrepancies are most likely explained by the higher level of variability present in the manual data (as indicated by the higher SD, especially for the left thalamus) and a worse level of agreement (ICC) between repeated measures. Future algorithmic developments should be directed towards minimizing proportional bias, since this is likely to significantly influence the statistical power of experiments measuring thalamus volumes.

A discernible amount of variability was found in the manual tracing of the thalamus as evidenced by the intra-rater ICC’s (Derakhshan et al., 2010; Fischl et al., 2002; Houtchens et al., 2007). Owing to the complexity of the cerebral anatomy combined with imaging artefacts (partial volume, intensity inhomogeneity, noise, etc.) present in MRI data, manual outlining is difficult, labor-intensive and time consuming. This particularly applies to the thalamus, which is an agglomeration of smaller nuclei, which leads to an ill-defined boundary of the overall thalamus on conventional MRI, especially in the presence of neurodegeneration. In order to minimize error and reduce variability, we decided to solicit a single expert reader trained in manual tracing on MRI to obtain the highest quality thalamus outlines possible. We did not limit the number of patients or slices and decided to generate thalamus segmentations on each slice, which increases the relevance of this study. Importantly, by using this dataset we were able to objectively compare some of the most widely applied automated segmentation techniques in a multi-center setting, considering the sampling from a large cohort of patients, representative of the full range of a typical RMS population. Moreover, we have created a valuable set of full manual thalamus outlines of all subjects to provide reference correlations with the cognitive scores.

### 4.1. Limitations

Our study has several limitations, including the absence of a neuropsychological evaluation of the HCs, as well as the assessment of thalamus damage only, which did not allow us to investigate other patterns of microstructural tissue and (deep) GM damage that likely contribute to CI (Damjanovic et al., 2017; Preziosa et al., 2016; Schoonheim et al., 2015, 2012). The choice of the thalamus as a region of interest was motivated by the abundance of literature showing a relationship between damage to the thalamus and cognitive dysfunction in MS patients. As a result, we cannot rule out the possibility that other patterns of more diffuse pathological processes contributed to CI in our MS patients, and a multi-structure imaging and measurement approach is likely needed (Damjanovic et al., 2017; Sastre-Garriga et al., 2020). Concerning image acquisition, (near)isotropic 3D T1-weighted images with similar acquisition parameters were used to obtain thalamus atrophy. In this work we addressed the potential effect of between-center heterogeneity in MRI acquisition in the regression analyses, however, remaining differences between scanners can systematically affect the robustness of the association between deep GM atrophy measurements and cognition across methods (Amiri et al., 2018). A more detailed evaluation of the interaction between MRI acquisition parameters and different thalamus segmentation methods (i.e., the robustness of the various segmentation methods with regards to MRI acquisition parameters) transcended the scope of this study, but should be addressed in future work.

### 4.2. Conclusion

This multi-center study helps to shed light on some previously reported differences between various automated segmentation techniques and how these might influence the relationship between thalamus volume measurements and cognition in MS. It supports the notion that thalamus atrophy is associated with a worse cognitive profile in MS patients. However, one should be cautious when interpreting these findings given the proportional biases that might be present in automated volumetry, especially in smaller and larger thalamus, as well as the impact of differences in scanners and acquisition protocols. The approaches work in a multi-center setting, but statistical power is increased by appropriate matching of algorithms with optimal scanners and MRI acquisition parameters. Further research is needed to account for these potential sources of error and ensure the accuracy of these methods in the real-world clinical evaluation of MS patients.

---

Table 8

Summary of results of the general linear regression analysis with CII as the dependent variable, for each vendor.

| Volumes     | GE (N = 18) | Philips (N = 18) | Siemens (N = 21) |
|-------------|-------------|-----------------|------------------|
| Manual lines |             |                 |                  |
| Left thalamus | 0.32 | 0.63 | 0.47 | 0.36 | 0.32 | 1.32 | 0.001 |
| Right thalamus | -0.08 | 0.90 | -0.58 | 0.32 | 1.15 | 0.003 |
| FreeSurfer Left thalamus | -0.48 | 0.48 | -1.79 | 0.06 | 1.48 | 0.005 |
| Right thalamus | -0.57 | 0.44 | -1.65 | 0.045 | 1.68 | 0.005 |
| FSL-FIRST Left thalamus | -0.05 | 0.94 | -2.11 | 0.043 | -1.40 | 0.009 |
| Right thalamus | -0.12 | 0.86 | -1.56 | 0.17 | 1.34 | 0.020 |
| CAT12 Left thalamus | -0.27 | 0.56 | -2.00 | 0.007 | -1.38 | 0.001 |
| Right thalamus | -0.34 | 0.49 | -1.51 | 0.038 | -1.53 | 0.001 |
| GIFT Left thalamus | -0.82 | 0.39 | -0.253 | 0.22 | -1.62 | 0.031 |
| Right thalamus | -0.90 | 0.35 | -1.48 | 0.37 | 1.71 | 0.029 |
| VolumSurf Left thalamus | -0.22 | 0.73 | -1.83 | 0.06 | -1.53 | 0.002 |
| Right thalamus | -0.24 | 0.75 | -1.59 | 0.07 | -1.45 | 0.001 |
| Contrast-to-noise ratio |             |                 |                  |
| Left thalamus | 2.11 ± 0.66 | 1.74 ± 0.48 | 1.04 ± 0.35 |
| Right thalamus | 2.11 ± 0.64 | 1.78 ± 0.47 | 1.18 ± 0.37 |

Abbreviations: B = unstandardized regression coefficient; CI = confidence interval; CII = Cognitive Impairment Index. *All regression analysis were corrected for center and age; † Thalamus volumes were multiplied by the head-normalization factor derived from SIENAX; p-values in bold represent significant values.
Announcement of financial support and other

The members of the MAGNIMS Study Group Steering Committee are: F. Barkhof and H. Vrenken (VU University Medical Center, Amsterdam, Netherlands), O. Ciccarelli and T. Yousry (Queen Square MS Centre, UCL Institute of Neurology, London, UK), N. De Stefano (University of Siena, Siena, Italy), C. Enzinger (Department of Neurology, Medical University of Graz, Graz, Austria), M. Filippi and M.A. Rocca (San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Milan, Italy), C. Gasperini (San Camillo-Forlanini Hospital, Rome, Italy), L. Kappos (University of Basel, Basel, Switzerland), J. Palace (University of Oxford Hospitals Trust, Oxford, UK), A. Rovira and J. Sastre-Garriga (Hospital Universitari Vall d’Hebron, Universitat Autonoma de Barcelona, Barcelona, Spain). The study was funded by the Nauta funds through a travel grant. The MS Center Amsterdam is supported by the Dutch MS Research Foundation through a program grant (current grant 18–3586). D.B. is supported by project PI18/00823 from the “Fondo de Investigación Sanitaria Carlos III”. F.B. and O.C. are supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. The acquisition of data in London was funded by supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre. A sincere thank you to Tom Verhoeven for his editing of the figures.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.nicl.2020.102549.

References

Altman, D.G., Bland, J.M., 1983. Measurement in Medicine: The Analysis of Method Comparison Studies. J. Royal Stat. Soc. Ser. D Stat. 32 (3), 307–317. https://doi.org/10.2307/2987097.

Amato, M.P., Battaglini, M., Ruggieri, S., Gasperini, C., Filippi, M., Rocca, M.A., Rovira, A., Battaglia, G., Giannini, M., Pastore, L., 2020. Reduced accuracy of MRI deep grey matter segmentation in multiple sclerosis: an evaluation of four automated methods against manual reference segmentations. J. Neurol. 267 (12), 3541–3554. https://doi.org/10.1007/s00415-020-09023-1.

Derakhshan, M., Caramanos, Z., Gioacchino, P.S., Narayanan, S., Maranjan, F., Francis, S.J., Arnold, D.L., Collin, D.L., 2010. Evaluation of automated techniques for quantification of grey matter atrophy in patients with multiple sclerosis. NeuroImage 52 (4), 1261–1267. https://doi.org/10.1016/j.neuroimage.2010.05.029.

Eghabhi, A., Prados, F., Brownlee, W.J., Altmann, D.R., Tur, C., Cardoso, M.J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., Ourselin, S., 2017. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954. https://doi.org/10.1016/j.neuroimage.2010.09.019.

Dale, A.M., Fischl, B., Sereno, M.I., 1999. Cortical Surface-Based Analysis. NeuroImage 9 (2), 179–194. https://doi.org/10.1016/S1053-8119(98)00035-9.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A.M., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32 (1), 223–228. https://doi.org/10.1002/jmri.22214.

Chiaravalloti, N.D., Deluca, J., 2008. Cognitive impairment in multiple sclerosis. The Lancet Neurology 7 (12), 1139–1151. https://doi.org/10.1016/S1474-4420(08)70299-9.

Coupé, P., Manjon, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L., 2011. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. NeuroImage 54 (2), 940–954. https://doi.org/10.1016/j.neuroimage.2010.09.019.

Chard, D.T., Jackson, J.S., Miller, D.H., Wheeler-Kingshott, C.A.M., 2010. Reducing the impact of white matter lesions on automated measures of brain gray and white matter volumes. J. Magn. Reson. Imaging 32 (1), 223–228. https://doi.org/10.1002/jmri.22214.
