

Measurement of Whole-Brain and Gray Matter Atrophy in Multiple Sclerosis: Assessment with MR Imaging

Loredana Storelli, MSc • Maria A. Rocca, MD • Elisabetta Pagani, MSc • Wim Van Hecke, PhD • Mark A. Horsfield, PhD • Nicola De Stefano, MD, PhD • Alex Rovira, MD • Jaume Sastre-Garriga, MD • Jacqueline Palace, MD • Diana Sima, PhD • Dirk Smeets, PhD • Massimo Filippi, MD • for the MAGNIMS Study Group

From the Neuroimaging Research Unit (L.S., M.A.R., E.P., M.F.) and Department of Neurology, Institute of Experimental Neurology, Division of Neuroscience (M.A.R., M.F.), San Raffaele Scientific Institute, Vita-Salute San Raffaele University, Via Olgettina 60, 20132 Milan, Italy; Department of Research and Development, Icometrix, Leuven, Belgium (W.V.H., D. Sima, D. Smeets); Xinapse Systems, Colchester, England (M.A.H.); Department of Medicine, Surgery and Neuroscience, University of Siena, Siena, Italy (N.D.S.); Section of Neuroradiology, Department of Radiology, Hospital Universitari Vall d'Hebron, Barcelona, Spain (A.R.); Unit of Clinical Neuroimmunology, CEM-Cat, Hospital Universitari Vall d'Hebron, Barcelona, Spain (J.S.G.); and Nuffield Department of Clinical Neurosciences, University of Oxford, Oxford, England (J.P.). Received October 23, 2017; revision requested December 6; final revision received January 18, 2018; accepted January 25. Address correspondence to M.F. (e-mail: filippi.massimo@hsr.it).

Conflicts of interest are listed at the end of this article.

Radiology 2018; 288: 554–564 • <https://doi.org/10.1148/radiol.2018172468> • Content code: **NR**

Purpose: To compare available methods for whole-brain and gray matter (GM) atrophy estimation in multiple sclerosis (MS) in terms of repeatability (same magnetic resonance [MR] imaging unit) and reproducibility (different system/field strength) for their potential clinical applications.

Materials and Methods: The softwares ANTs-v1.9, CIVET-v2.1, FSL-SIENAX/SIENA-5.0.1, Icometrix-MSmetrix-1.7, and SPM-v12 were compared. This retrospective study, performed between March 2015 and March 2017, collected data from (a) eight simulated MR images and longitudinal data (2 weeks) from 10 healthy control subjects to assess the cross-sectional and longitudinal accuracy of atrophy measures, (b) test-retest MR images in 29 patients with MS acquired within the same day at different imaging unit field strengths/manufacturers to evaluate precision, and (c) longitudinal data (1 year) in 24 patients with MS for the agreement between methods. Tissue segmentation, image registration, and white matter (WM) lesion filling were also evaluated. Multiple paired *t* tests were used for comparisons.

Results: High values of accuracy (0.87–0.97) for whole-brain and GM volumes were found, with the lowest values for MSmetrix. ANTs showed the lowest mean error (0.02%) for whole-brain atrophy in healthy control subjects, with a coefficient of variation of 0.5%. SPM showed the smallest mean error (0.07%) and coefficient of variation (0.08%) for GM atrophy. Globally, good repeatability ($P > .05$) but poor reproducibility ($P < .05$) were found for all methods. WM lesion filling technique mainly affected ANTs, MSmetrix, and SPM results ($P < .05$).

Conclusion: From this comparison, it would be possible to select a software for atrophy measurement, depending on the requirements of the application (research center, clinical trial) and its goal (accuracy and repeatability or reproducibility). An improved reproducibility is required for clinical application.

© RSNA, 2018

Online supplemental material is available for this article.

Multiple sclerosis (MS) is a heterogeneous inflammatory disease of the central nervous system, characterized by a chronic and unpredictable clinical course, with development of disability over time, which can differ among clinical phenotypes. Pathologic studies have consistently shown that neurodegeneration is one of the pathologic hallmarks of MS (1,2) and because of this, MS is no longer considered to be solely an inflammatory, demyelinating disease (3,4). Magnetic resonance (MR) imaging measurements of atrophy and its progression over time are among the best-studied and accepted methods for quantifying neurodegeneration not only in MS but also in other neurologic and psychiatric conditions (5,6). In MS, substantial evidence has emerged over several years that gray matter (GM) is heavily affected by neurodegenerative phenomena causing neuronal and axonal damage (7). Using MR imaging techniques, many research studies have consistently demonstrated that brain atrophy is

more pronounced in patients with progressive MS than in those with relapsing MS phenotypes (6). Furthermore, GM atrophy is more relevant than whole-brain or white matter (WM) atrophy in explaining clinical disability and cognitive impairment in MS (8–10). Measuring GM atrophy also has prognostic value for the subsequent clinical evolution of patients with MS (ie, progression of disability, cognitive impairment, and evolution toward the more severe clinical phenotypes) (11,12). Halting neurodegeneration and promoting neuroprotection are probably the most important goals of current therapeutic strategies (13,14). There is a need to improve image analysis techniques to enable the reliable use of whole-brain and GM atrophy assessment techniques in the clinical setting as a basis for individualized treatment decisions (15,16). It has been recommended that brain atrophy assessment be included in the “no evidence of disease activity” (NEDA) criteria, which are defined as an absence of MR

Abbreviations

CoV = coefficient of variation, DSC = Dice similarity coefficient, EDSS = Expanded Disability Status Scale, FLAIR = fluid-attenuated inversion recovery, GM = gray matter, ICC = intraclass correlation coefficient, MS = multiple sclerosis, NEDA = no evidence of disease activity, 3D = three-dimensional, WM = white matter

Summary

From the comparison between the atrophy methods, the identification of a single best software is currently not possible; the selection of atrophy measurement software should be based on whether its performance is satisfactory with regard to the specific requirements of the application.

Implications for Patient Care

- Different software for calculation of whole-brain and gray matter volume quantification showed comparable repeatability when consistent patient positioning and pulse sequence are used with a single MR imaging unit.
- Different methods for determining whole-brain and gray matter atrophy quantification were not sufficiently reproducible for application in the routine clinical setting.
- Given the poor reproducibility of all software, MR imaging unit changes should be avoided if atrophy is to be quantified meaningfully by using these methods, and more standardized MR imaging acquisition protocols are needed.

imaging activity (appearance of new T2 and gadolinium-based contrast agent-enhancing T1 lesions), relapses, and disability progression (NEDA-3), leading to NEDA-4 (17). Many improvements in MR imaging technology have been made over the past few years, and several software tools (either free to use or commercial) are currently available with different levels of interactivity required. A number of these software tools have already been applied in academic research (18–23).

For use in clinical practice, there are at present no validated techniques for monitoring atrophy in patients with MS using MR imaging (24,25). Measurement of disease-related changes in MS is hindered by the difficulties imposed by the diseased brain. In particular, WM lesions can affect the accuracy of GM and WM segmentation, which may result in greater variability in cross-sectional tissue volume measurements (26,27). On longitudinal atrophy measures, the impact of WM lesions needs to be investigated further. Because of these issues, it is particularly important to validate techniques for whole-brain and GM atrophy measurement using images from patients with MS.

Since brain and particularly GM tissue loss are clinically relevant in MS, the evaluation of automatic techniques for their measurement is essential. Here, we aimed to compare different available methods for whole-brain and GM atrophy estimation in MS and to analyze the feasibility of a clinical application.

Materials and Methods

Image processing was performed by L.S. and E.P. (with 4 and 20 years of experience in MR imaging, respectively).

Icometrix company provided the test-retest data set for our study. The authors who were not employees or consultants for Icometrix had control of inclusion of any data and information that might present a conflict of interest for the authors working for the cited company.

Image Data Set

Ethics committee approval.—Approval of the institutional review board was obtained, along with written informed consent from all participants. Our retrospective study, performed between March 2015 and March 2017, collected available data sets consisting of whole-brain MR fluid-attenuated inversion recovery (FLAIR) image (for MS lesion segmentation) and three-dimensional (3D) T1-weighted nonenhanced sequences (for volumetric analysis).

Subset 1—simulated data.—T1-weighted and FLAIR MR imaging sequences of two MS brains with mild (0.42 mL) and severe (10.1 mL) WM lesion loads were simulated. The details on the implementation of the MR imaging simulator are described in Appendix E1 (online). The sequence parameters used for the simulation are shown in Table 1.

Subset 2—test-retest data sets.—The MR imaging protocol consisted of a 3D T1-weighted and a 3D FLAIR sequence. MR images for all patients were acquired twice on each imaging unit on the same day, removing the patient from the unit between the two acquisitions. For each examination, the patient was positioned in the MR imaging unit according to the same positioning protocol. Because test and retest MR studies were performed within few hours and because there was no dehydrating exercise between studies, physiologic fluctuations of brain volumes due to biologic factors should not have appreciably affected atrophy measurements. Subset 2 included all subjects described in subsets 2a and 2b.

Subset 2a—test-retest data set using different field strengths.—Nineteen patients with MS (12 with relapsing-remitting MS, six with secondary progressive MS, and one with primary progressive MS (28) (mean age, 40 years; range, 21–63 years; female/male ratio, 14/4; mean age of women, 38 years [range, 21–53 years]; mean age of men, 42 years [range, 28–63 years]; mean Expanded Disability Status Scale [EDSS] scale score, 3.0 [range, 1.5–6.5]; mean disease duration, 10 years [range, 3–25 years]) were imaged by using two Philips Healthcare MR imaging systems: a 1.5-T Intera and a 3.0-T Achieva imaging unit (Philips Medical Systems, Best, the Netherlands). The details of the MR imaging protocol for each imaging unit are reported in Table 1.

Subset 2b—test-retest data set with different 3.0-T imaging units.—Ten patients with MS underwent test-retest MR imaging studies performed on the same day by using a 3.0-T Philips system (Achieva; Philips Medical Systems, Best, the Netherlands), a 3.0-T Siemens system (Skyra; Siemens Medical Systems, Erlangen, Germany), and a 3.0-T GE system (DISCOVERY MR750w; GE Healthcare, Chicago, Ill). The MR imaging protocol is detailed in Table 1.

Subset 3—longitudinal data set of patients with MS.—Twenty-four patients (11 with clinically isolated syndromes suggestive of MS and 13 patients with relapsing-remitting

Table 1: MR Imaging Acquisition Protocols for the Different Data Sets

Imaging Type and Parameter	Simulated Data	TRT Data for Different Field Strengths	TRT Data for Different Manufacturers	MS Longitudinal Data	HC Longitudinal Data
3D T1-weighted					
MR imaging unit	MR imaging simulator	(a) 1.5-T Philips; (b) 3.0-T Philips	(a) 3.0-T Philips; (b) 3.0-T Siemens; (c) 3.0-T GE	(a) 3.0-T Siemens; (b) 3.0-T GE	3.0-T Philips
MR imaging sequence	FFE	TFE	(a) FSPGR; (b) MPRAGE; (c) FSPGR	MPRAGE	FFE
Repetition time (msec)	22	(a) 8.8; (b) 9.8	(a) 4.93; (b) 2300; (c) 7.32	2300	25
Echo time (msec)	10	(a) 4.2; (b) 4.6	(a) 3.4; (b) 2.29; (c) 3.14	2.98	4.6
Inversion time (msec)	None	None	(a) None; (b) 1100; (c) none	900	None
Acquisition voxel size (mm ³)	1 × 1 × 1	(a) 0.87 × 1.25 × 1.2; (b) 0.88 × 1.19 × 1	(a) 0.53 × 0.53 × 0.5; (b) 0.94 × 0.94 × 0.94; (c) 1 × 1 × 1	1 × 1 × 1	1 × 1 × 1
Field of view (mm ²)	181 × 217	236 × 236	(a) 230 × 230; (b) 240 × 240; (c) 220 × 220	240 × 256	230 × 170
Flip angle (degrees)	10	8	(a) 8; (b) 8; (c) 12	9	30
FLAIR					
MR imaging sequence	3D-IR	3D-IR	3D-IR	2D-IR	...
Repetition time (msec)	22 ms	(a) 11000; (b) 10000	(a) 4800; (b) 5000; (c) 9500	9000	...
Echo time (msec)	10 ms	(a) 140; (b) 140	(a) 371; (b) 387; (c) 135.78	93	...
Inversion time (msec)	2100	(a) 2800; (b) 2750	(a) 2100; (b) 1800; (c) 2428	2500	...
Acquisition voxel size (mm ³)	1 × 1 × 1	(a) 1.36 × 1.77 × 1.5; (b) 1.31 × 1.31 × 1.34	(a) 1.04 × 1.04 × 0.56; (b) 1 × 1 × 1; (c) 1.05 × 0.5 × 3 × 1 × 1
Field of view (mm ²)	181 × 217	(a) 240 × 192; (b) 230 × 167	(a) 240 × 240; (b) 230 × 230; (c) 240 × 240	400 × 512	...
Flip angle (degrees)	90	90	90	120	...

Note.—Seven patients from the test-retest data set were selected to provide a data set to check the outputs produced by each method. Data used in this step were not subsequently used for testing. FFE = fast field echo, FSPGR = fast spoiled gradient-recalled echo, HC = healthy control (subjects), IR = inversion recovery, MPRAGE = magnetization prepared rapid acquisition gradient echo, MS = multiple sclerosis, TRT = test-retest, 3D = three-dimensional, TFE = turbo field echo, 2D = two-dimensional.

MS; mean T2 lesion load, 4.9 mL ± 3.9 [standard deviation]; mean age, 37 years; range, 19–61 years; female/male ratio, 12/12; mean age of women, 34 years [range, 24–49 years]; mean age of men, 38 years [range, 19–61 years]; mean EDSS score, 2.4 [range, 0–5]; mean disease duration of relapsing-remitting MS, 2.3 years [range, 0.3–5 years]) were examined by using a 3.0-T Siemens (TrioTim, Siemens Medical Systems, Erlangen, Germany) or a 3.0-T GE imaging unit (Signa HDxt; GE Healthcare, Chicago, Illinois) at two different centers (Barcelona and Oxford) within the MAGNIMS consortium (a European network for the study of MS using MR imaging techniques) at baseline and after a mean follow-up of 12 months (range, 9–15 months). The MR imaging protocol is shown in Table 1. Image distortion correction was managed using internal MR software and no phantom was used for harmonization among centers.

Subset 4—longitudinal data set of healthy control subjects.—Ten healthy subjects (mean age, 30 years; range, 27–34 years; female/male ratio, 6/4; mean age of women, 29 years [range, 27–33 years]; mean age of men, 31 years [range, 28–34 years]) underwent 3D T1-weighted MR imaging performed with a 3.0-T Philips imaging unit (Achieva; Philips Medical Systems, Best, the Netherlands) in Milan at baseline and after 2 weeks. The MR imaging protocol is given in Table 1.

Image Analysis

The software packages included in our study were specifically developed for volumetric brain tissue segmentation and atrophy assessment from MR images. Surface-based methods (eg, FreeSurfer) and methods that provide only the final atrophy result were not included. Five packages were found (four free to use, and one commercial) that were suitable for whole-brain and GM atrophy assessment in MS:

1. Advanced Normalization Tools (ANTs) version 1.9, University of Philadelphia, Pa; <http://stnava.github.io/ANTs/>;

2. Corticometry Analysis Tool-CIVET version 2.1, Montreal Neurologic Institute, Canada; <http://www.bic.mni.mcgill.ca/ServicesSoftware/CIVET/>;

3. FMRIB Software Library (FSL) version 5.0.1, University of Oxford, England; <https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/SIENA>;

4. MSmetrix version 1.3, Icometrix Leuven, Belgium; <https://icometrix.com>;

5. Statistical Parametric Mapping (SPM) toolbox version 12, University College London, England; <http://www.fil.ion.ucl.ac.uk/spm/software/spm12>.

The methods implemented by each software are described in Appendix E1 (online).

Validation

Each estimated measure includes some uncertainty, which is considered to have two components: a systematic error (or bias) and a random error (29). To standardize terminology, it was recommended by the RSNA-Quantitative Imaging Biomarker Alliance Metrology Working Group to define these components as "accuracy" and "precision," respectively, and in our study both were evaluated (29).

Assessment of accuracy requires comparison to a reference value (see Appendix E2 [online]), that in this case is the true tissue volume being known for the brain digital phantoms of the simulated data set. These data (subset 1) were used to estimate measurement bias in whole-brain and GM volumes for each software. The bias in longitudinal whole-brain and GM atrophy was assessed by using the healthy control longitudinal data (subset 4), by assuming that no atrophy occurred between the test and retest.

Precision is assessed by making repeated measures. We evaluated both the repeatability and the reproducibility, which are two different aspects included into the general term of precision: repeatability is a measurement of precision that occurs with identical or near-identical conditions; reproducibility, in contrast, is a measurement of precision when measuring system, or other factors differ (29). Specifically, the repeatability of each method was assessed by comparing tissue volumes between test and retest studies performed with the same MR imaging unit (subset 2). The reproducibility was evaluated by comparing tissue volumes between MR imaging units with different field strengths (subset 2a) and between MR imaging units from different manufacturers at 3.0 T (subset 2b).

Table 2: Differences between the True and the Calculated Values for Whole-Brain and Gray Matter Volume Quantification for the Different Methods Estimated in the Simulated Data Set (Subset 1)

Software	Accuracy		DSC	
	Brain	Gray Matter	Brain	Gray Matter
ANTs	0.96 ± 0.003	0.96 ± 0.03	0.96 ± 0.001	0.89 ± 0.01
CIVET	0.92 ± 0.003	0.96 ± 0.003	NA	NA
FSL-SIENAX	0.96 ± 0.005	0.95 ± 0.06	0.95 ± 0.002	0.84 ± 0.03
MSmetrix	$0.89 \pm 0.01^*$	$0.87 \pm 0.02^*$	$0.87 \pm 0.005^*$	$0.59 \pm 0.01^*$
SPM	0.96 ± 0.01	0.97 ± 0.005	0.97 ± 0.003	0.90 ± 0.01

Note.—Data are means \pm standard deviations. In the last two columns, Dice similarity coefficient (DSC) values for whole-brain and GM tissue segmentation for the different methods, compared with the ground truth of the simulated data set (Equation [E10] [online]). DSC values were not available for CIVET, because tissue segmentations were not in patient coordinate space. Paired *t* tests were performed to compare mean differences between the results of the different methods. NA = not available.

* *P* < .05 for the comparison with the other methods.

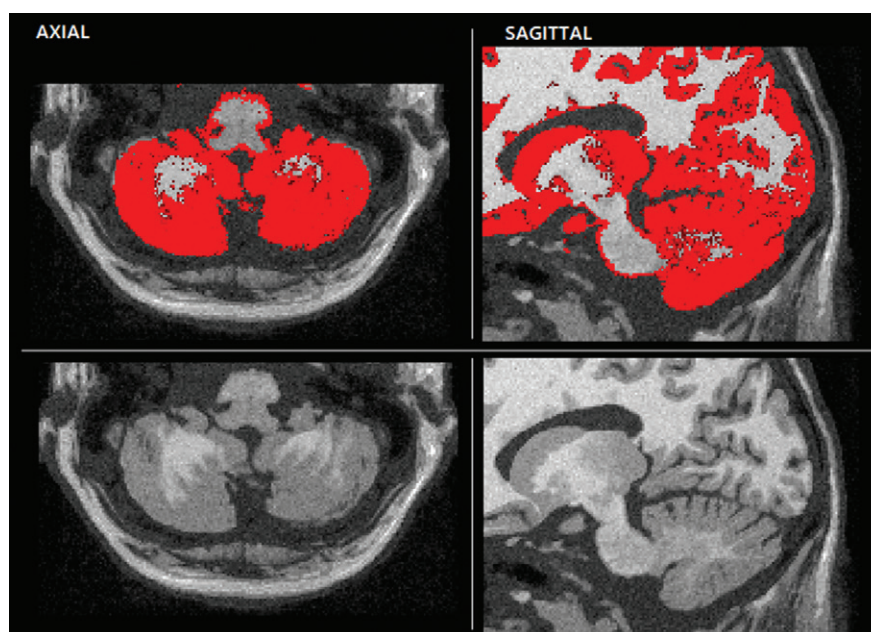


Figure 1: Images in subset 1. First row: An example of inaccurate gray matter segmentation (in red) for the MSmetrix software within the cerebellum of simulated MR images (in the axial and sagittal planes). A possible reason could be the fact that MSmetrix uses the fluid-attenuated inversion recovery image to impose the border between the gray matter and cerebrospinal fluid. But simulated images might be not realistic enough, and this could have caused this bias in the segmentation. Second row: Corresponding simulated T1-weighted images.

Brain and GM tissue segmentation, image registration, and the effect of WM lesion filling on longitudinal atrophy estimation were evaluated for each image processing. See Appendix E2 (online) for a detailed description of the metrics implemented.

Subset 3 (1 year of follow-up) was used to assess the sensitivity in detecting whole-brain and GM atrophy changes.

Statistical Analysis

Statistical analysis was performed by using the R software package, version 3.1.1, New Jersey. Paired *t* tests were used to evaluate (a) mean differences between repeated measures

of whole-brain and GM volumes, (b) to evaluate mean differences across different MR field strengths and across different manufacturers of 3.0-T MR imaging units (pairwise between the three MR imaging units), (c) to compare the accuracy and Dice similarity coefficient (DSC) results between the different methods, and (d) to evaluate the effect of WM lesion filling on longitudinal atrophy measures. Bonferroni correction was applied to adjust significance levels for multiple comparisons. Subjects were matched for age and sex within each subgroup. Coefficients of variation (CoVs) were used to evaluate the variability of longitudinal atrophy results and the variability between test and retest whole-brain and GM volumes for the same imaging unit and to evaluate variability across different MR manufacturers and field strengths. The intraclass correlation coefficient (ICC) was computed to evaluate the agreement between the different software for whole-brain and GM atrophy results. $P < .05$ was considered to indicate a statistically significant result.

Results

Subset 1—Cross-Sectional Accuracy

In Table 2 the accuracy of whole-brain and GM volume estimates from the simulated data are reported for the different methods, with the DSCs for the assessment of brain and GM segmentation steps. All software systems showed high values for both accuracy and DSCs ($P > .05$), except for MSmetrix, whose accuracy values were significantly lower than the other methods ($P < .05$) for both measures (Fig 1).

Subset 2—Precision

Figure 2 shows the differences in whole-brain and GM volumes between test and retest studies for each method. No software showed volume differences between test and retest for both whole-brain and GM measures ($P > .05$). MSmetrix showed the smallest mean difference (8.8 mL) and the lowest variability (CoV = 0.5%) between test and retest for whole-brain volume ($P > .05$). CIVET showed a significant ($P < .05$) highest mean difference (26.2 mL) and variability (CoV = 2.6%) between test and retest for whole-brain volume, but a significant smallest mean difference (0.5 mL) and lowest variability (CoV = 0.2%) between test and retest for GM volume ($P < .05$). SIENAX showed highest mean difference (11.3 mL) and variability (CoV = 2.1%) between test and retest for GM volumes ($P > .05$). For each method, comparable variability in estimating whole-

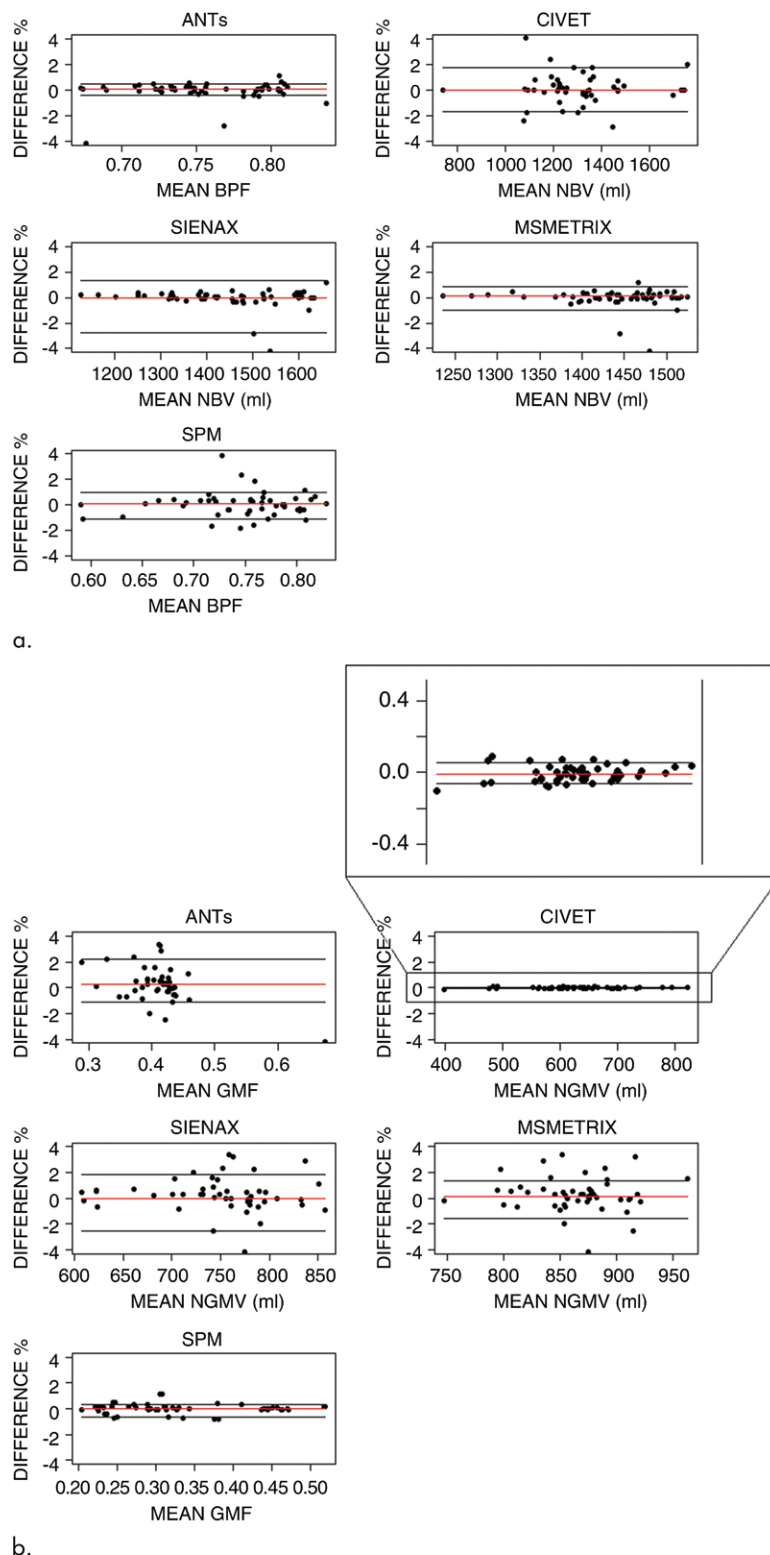


Figure 2: Graphs in subset 2. (a) Bland-Altman plots of the differences between test and retest whole-brain volumes (in subset 2) for the different methods. (b) Similar plots for gray matter volumes. Lines = median and the 5th and 95th percentiles. The CIVET normalized gray matter volumes are also shown on a larger scale in b. The poorer results for CIVET whole-brain volume in comparison with the high precision for gray matter volumes are likely due to a lower precision in white matter segmentation. BPF = brain parenchymal fraction, GMF = gray matter fraction, NBV = normalized brain volume (in milliliters), NGMV = normalized gray matter volume (in milliliters).

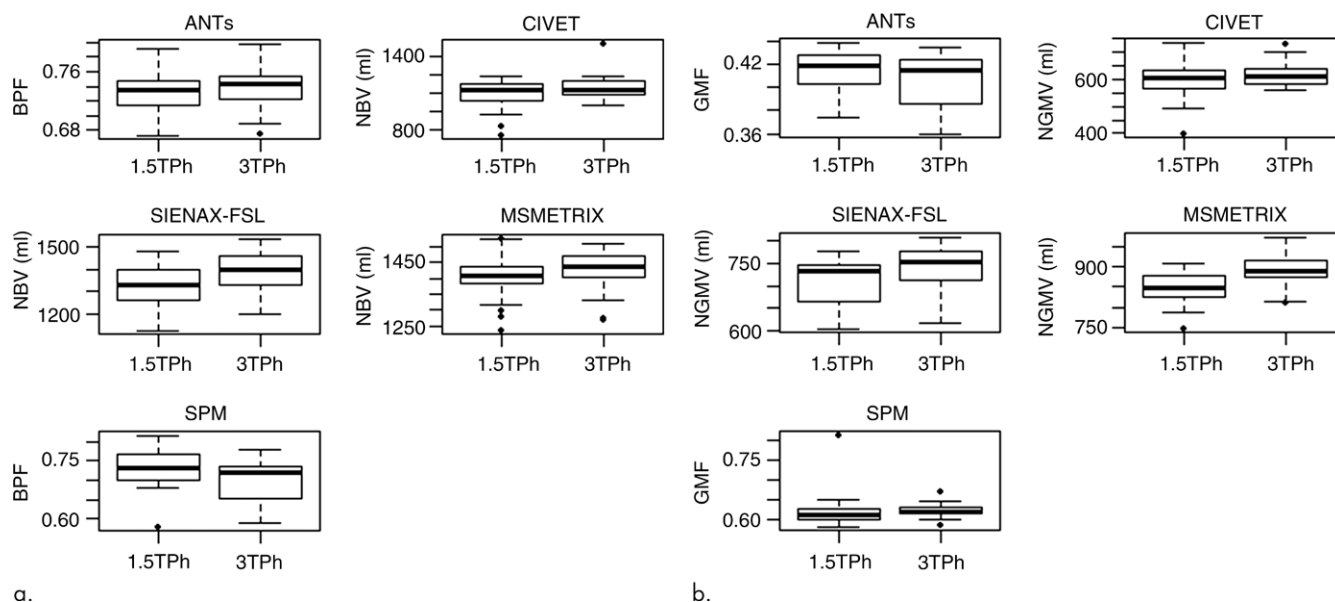


Figure 3: Graphs in subset 2a. The distributions of **(a)** whole-brain and **(b)** gray matter fractions and volumes for the 1.5-T Intera and 3.0-T Achieva Philips (Ph) MR imaging units (subset 2a) estimated by using the different softwares. BPF = brain parenchymal fraction, GMF = gray matter fraction, NBV = normalized brain volume (in milliliters), NGMV = normalized gray matter volume (in milliliters).

brain volumes and GM volumes at test-retest was found, except for CIVET, whose variability was about four times higher when estimating whole-brain volumes. This was due to a higher variability in WM versus CSF volume estimation compared with GM volume.

Figure 3 shows the reproducibility across different MR field strengths for each software (subset 2a). CIVET and SPM showed no difference ($P > .5$) between 1.5-T and 3.0-T Philips whole-brain volumes, with a mean difference between test and retest whole-brain volumes of 83 mL and the lowest CoV (1.4%) for CIVET. For GM volumes, no differences were found again for CIVET ($P = .7$) with a mean GM volume difference of 23.3 mL, and for SPM ($P = .9$) with a mean GM fraction difference of 0.04 between the two field strengths. The lowest CoV for GM volume estimation between field strengths was for CIVET (CoV = 2%, $P = .06$), while the highest CoVs were for ANTs whole-brain volumes (8%) and GM volumes (10%). Figure 4 shows the segmented GM for the same patient for the different methods, at 1.5-T and 3.0-T MR imaging for one image section.

The reproducibility across different manufacturers of 3.0-T MR imaging units (subset 2b) is shown in Figure 5. For whole-brain volume estimation, only ANTs and CIVET showed no significant differences ($P > .05$) between 3.0-T unit results, with a low mean volume difference (14.1 mL) and coefficient of variation (3.3%) for CIVET. For GM volumes, all methods showed significant mean differences ($P < .05$) and high variability (CoV > 3%) across different 3.0-T MR imaging units.

Subset 3—Agreement between Methods

Figure 6 shows a comparison of whole-brain and GM volume changes over 1 year for the different methods applied to patients with MS. Two patients who were imaged with a different unit at follow-up were removed from this data set. Significant

agreement was found for whole-brain volume change between SIENA and SPM (ICC = 0.52, $P = .05$) and between ANTs and MSmetrix (ICC = 0.51, $P = .03$). No agreement was found for GM volume change across the different methods (ICC < 0.5, $P > .05$).

Subset 4—Longitudinal Accuracy

Figure 7 shows the errors (assuming no real change in volume) on whole-brain and GM longitudinal volume changes estimated for the healthy controls data set for the different software. Given the “true value” (about 0% of atrophy between the two studies), it was possible to estimate the accuracy, in terms of error for whole-brain and GM volume change assessment for the different methods. ANTs showed no significant smallest mean error in estimating percentage whole-brain volume changes (0.02%), but significantly higher mean error (−0.4%) for GM atrophy (Fig 8) compared with the other methods ($P < .05$). Moreover, ANTs also had the highest mean variability for both whole-brain (CoV = 0.5%, $P > .05$) and GM tissue volume change (CoV = 1%, $P < .05$). The lowest mean variability of the errors for whole-brain volume change estimation was found for SIENA and SPM (CoV = 0.1%, $P > .05$). Again, SPM showed no significant lowest variability in longitudinal GM volume change estimates (CoV = 0.1%, $P > .05$). All methods had a comparable variation between the estimation of whole-brain and GM volume changes, except for ANTs.

Processing Steps Evaluation

In Table 3, the normalized mutual information values used to evaluate the registration of T1-weighted images of the subject to brain atlas and baseline to follow-up are listed for the software evaluated. Normalized mutual information values were

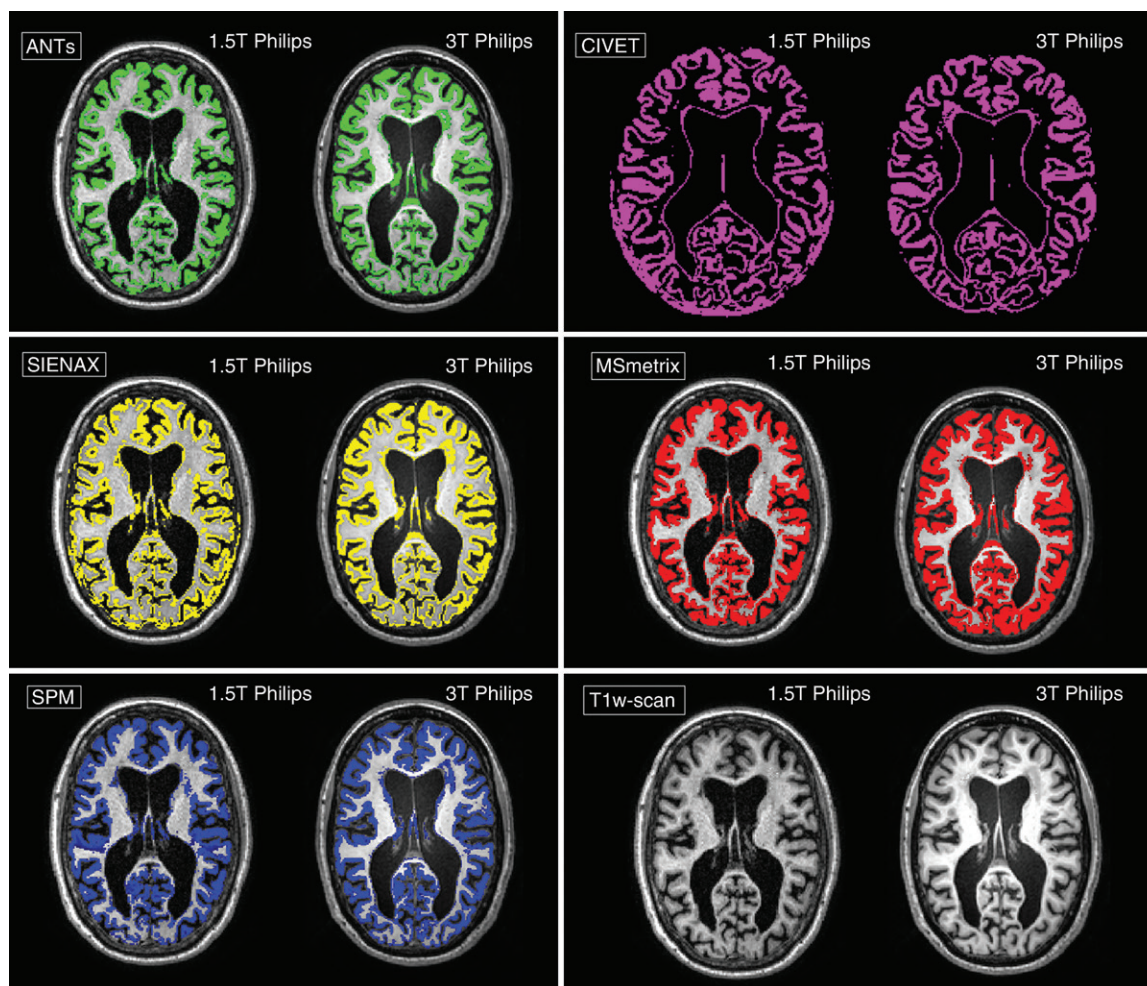


Figure 4: Images in subset 2a. Gray matter segmentation masks (with different colors) estimated by the different methods for images acquired on the 1.5-T Intera and on the 3.0-T Achieva MR imaging units (Philips Medical Systems, Best, the Netherlands), for the same patient. At the bottom right are the non-annotated T1-weighted (T1w) images. CIVET segmentation has a different brain shape compared with the other methods because it is shown in atlas space rather than in patient image space.

similar and showed no significant differences ($P > .05$) between the different software evaluated.

Filling of WM lesions on T1-weighted studies with intensities similar to those of WM is used to improve volume estimations (26,27). The results from FSL-SIENA were not significantly affected by WM lesion filling ($P > .05$); while GM atrophy results for ANTs, MSmetrix, and SPM ($P < .05$) were significantly influenced.

Accuracy and Precision

High values of accuracy (0.87–0.97) for whole-brain and GM volumes were found for all methods. ANTs and SPM showed the highest longitudinal accuracy (mean error = 0.02% and 0.07%) in estimating whole-brain and GM atrophy in healthy control subjects. On average, good repeatability ($P > .05$), but poor reproducibility ($P < .05$), were found for all software systems. Table 4 summarizes our study as a set of guidelines to help in the selection of software suitable for whole-brain and GM atrophy analyses.

Discussion

Since brain and particularly GM tissue loss are clinically relevant in MS (5,15,17,30), the evaluation of automatic techniques for their measurements is an active area of research. Over the past few years, several studies have compared brain atrophy measurement techniques in MS (20,24,31,32). All assessments so far have compared only the final whole-brain atrophy measures for the different methods (16,24,33). In our study, five freely available or commercial software packages were included for the comparison of both whole-brain and GM atrophy quantification specifically in MS brains (16,24,33). The performances of the main steps (tissue segmentation, image registration and lesion filling) were also evaluated to identify the strengths and weaknesses of each method.

A study about the agreement between MSmetrix and the other established methods has recently been published (34). The authors described similar results using a simulated MR imaging data set. However, for the in vivo data set, they

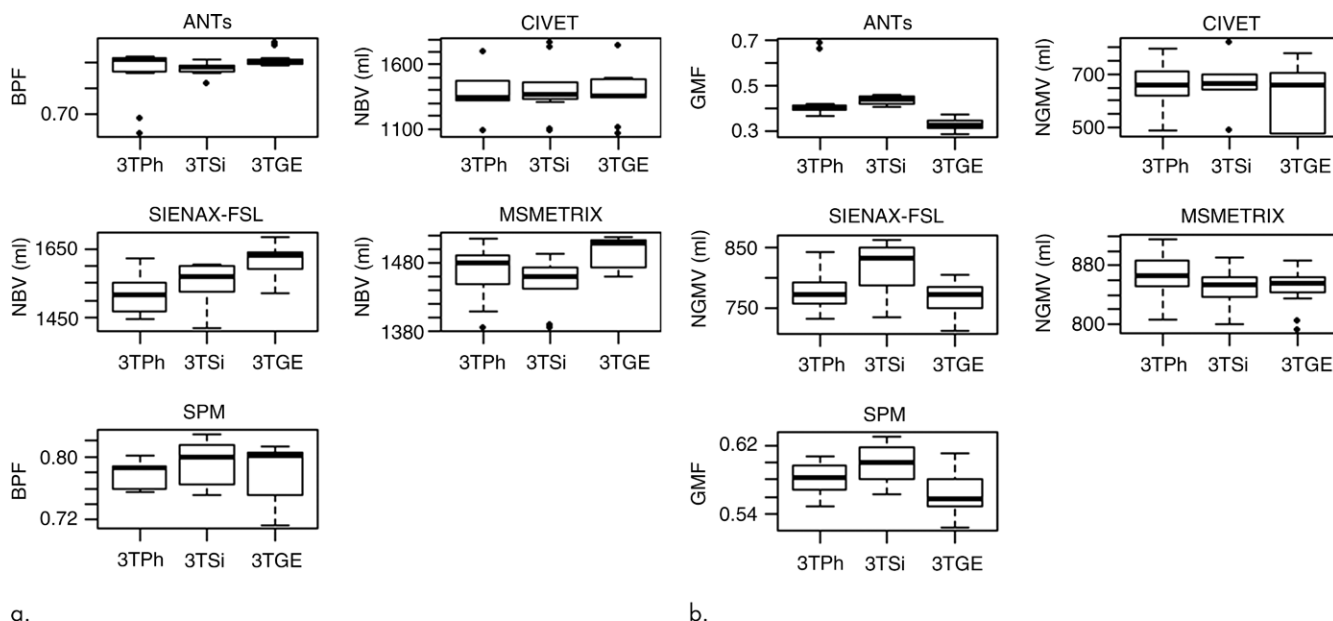


Figure 5: Graphs in subset 2b. The distributions of (a) whole-brain and (b) gray matter fractions and volumes for 3.0-T Philips Achieva, 3.0-T Siemens Skyra, and 3.0-T GE DISCOVERY MR750w MR imaging units (subset 2b) estimated by using the different softwares. BPF = brain parenchymal fraction, GMF = gray matter fraction, NBV = normalized brain volume (in milliliters), NGMV = normalized gray matter volume (in milliliters), Ph = Philips, Si = Siemens.

included studies acquired before and after an MR imaging unit hardware upgrade, which may have affected the results of longitudinal measurements and the reliability of the comparison.

It is important to demonstrate, for any new method, that it can be used to make measurements in a reliable manner, and that it is robust to the sorts of variations in the inputs likely to be encountered in real-life deployment. We evaluated whole-brain and GM volume quantification when the input data were carefully controlled (consistent patient positioning and pulse sequence in a single imaging unit), but also when MR imaging acquisition varied across field strengths and imaging unit manufacturers. However, even the method with the best repeatability for cross-sectional analysis showed a mean difference between test and retest whole-brain volume estimation (8.8 mL) that is about the 0.6% (8.8 of 1500) of the average adult whole-brain volume (approximately 1500 mL). This value is within the expected range of pathologic brain atrophy change over 1 year in patients with MS (about 0.5%–1%) (35). This pathologic cutoff was considered here as the threshold to determine unacceptable levels of variability to use the different atrophy methods in clinical setting for monitoring single patient. The precision for all methods was much poorer when changing MR imaging unit, with a mean difference between

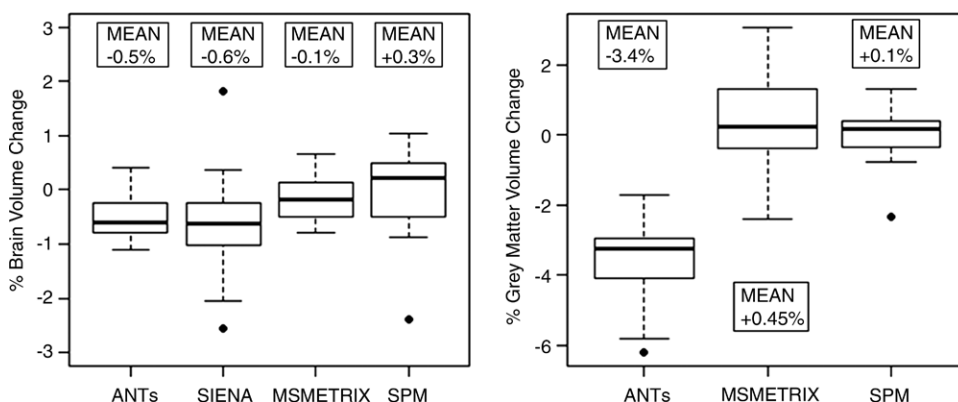
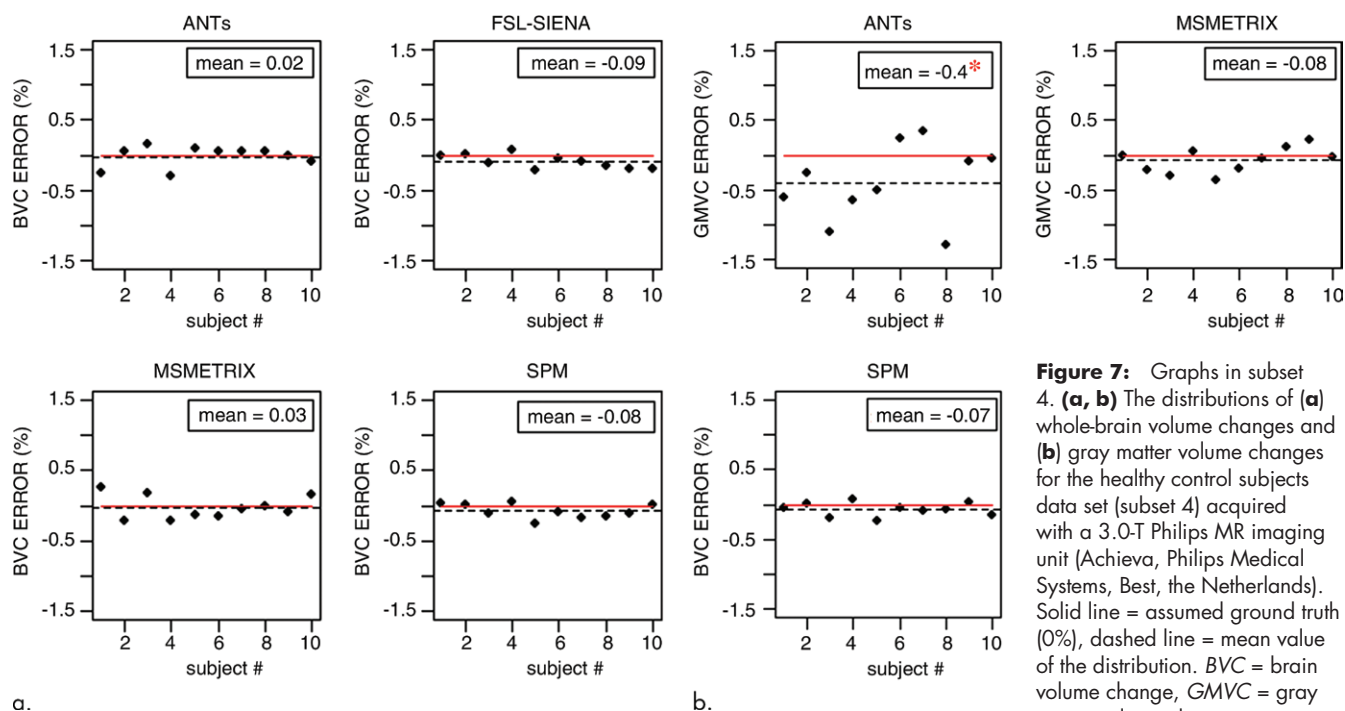


Figure 6: Graphs in subset 3. On the left are the distributions of the whole-brain volume changes for the different softwares estimated for the patients with multiple sclerosis with 1 year of follow-up (subset 3), while on the right are the distributions of the gray matter volume changes. The mean values for each distribution are also reported.

1.5-T and 3.0-T results that is about the 5% (83/1500 mL) of an average adult whole-brain volume (approximately 5 years of brain atrophy change). Similar results were found when comparing tissue volumes across 3.0-T units from different manufacturers. As a consequence, the individual differences in tissue volume estimation due to the different imaging unit acquisitions could cover the real individual pathologic variations.

The true value of a measure is often not available for in vivo imaging studies. However, using simulated data, we were able to quantify the systematic errors on cross-sectional atrophy measurements. For longitudinal methods, systematic bias could arise when applying a longitudinal atrophy method if the algorithm treats one of the pair of the images differently at the two time-points (ie, it is not symmetric), or there is some bias in the quality or contrast in the input images. We investigated



the possibility of systematic bias in whole-brain and GM volume change estimates using longitudinal data from healthy controls, where no substantial changes in brain volumes are expected. The variability of the volume changes was different among the methods and this affects the sample size needed to evaluate a treatment effect in clinical studies or trials.

For the longitudinal data from patients with MS, we looked at the agreement between the results of the different image processing, since the true degree of atrophy was not known. A pairwise agreement was found only for whole-brain atrophy between SIENA and SPM, and between ANTs and MSmetrix.

Our study had some limitations. First, the implemented MR simulator did not take into account possible different radiofrequency coils, artifacts, and pulse sequences, thus simulating a condition that is quite different from the real one. Moreover, a limited number of digital brain phantoms is available for the simulations. The comparison between different imaging units was performed only between 3.0-T MR systems, without considering different 1.5-T MR units, which are more commonly used in a clinical setting. Finally, the number of subjects included in each analysis was imposed by its availability.

In conclusion, from the comparison between the atrophy methods, the identification of a single best software is currently not possible; the selection of atrophy measurement

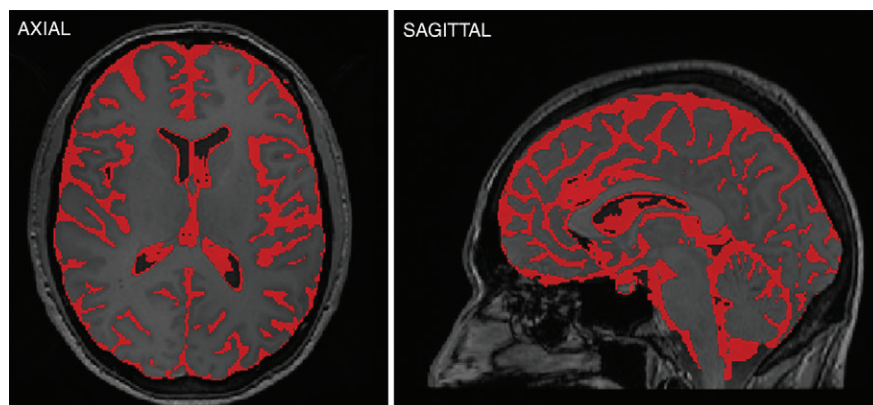


Figure 8: Images in subset 4. Example of gray matter segmentation failure (in red) for the ANTs tool in a healthy control subject in both the axial and sagittal planes. From a visual check of the results, we found that the higher errors for the ANTs tool were mainly due to inaccurate gray matter segmentation caused by poor tissue contrast in the spin-echo T1-weighted images (without an inversion pulse) in this data set.

Table 3: Normalized Mutual Information Values for Registration between Subject T1-weighted Images to the Brain Atlas and Baseline to Follow-up T1-weighted Studies for ANT, FSL-SIENAX, and MSmetrix

Software	Normalized Mutual Information Value	
	Subject-to-Atlas	Baseline-to-Follow-up
ANTs	1.2 ± 0.005	1.4 ± 0.02
FSL-SIENAX	1.1 ± 0.01	1.2 ± 0.03
MSmetrix	1.2 ± 0.02	1.3 ± 0.05

Note.—Data are means ± standard deviations. Normalized mutual information was not quantified for CIVET and SPM because they do not provide registered images as output.

Table 4: Summary of Guidelines for Selecting a Suitable Software for Whole-Brain and Gray Matter Atrophy Analysis according to the Results of Our Study

Software	Freely Available	Manual White Matter Lesion Filling Required	Expert User Only	Cross-Sectional Analysis		Longitudinal Analysis	
				Brain	Gray Matter	Brain	Gray Matter
ANTs	YES	YES	YES	YES (for single-center and multicenter application)	YES (only for single-center application)	YES (possibility to build a tool)	NO (possibility to build a tool)
CIVET	YES	YES	NO	YES (for single-center and multicenter application)	YES (for single-center and multicenter application)	NO	NO
FSL-SIENA (X)	YES	YES (not mandatory for longitudinal analysis)	YES	YES (only for single-center application)	YES (only for single-center application)	YES	NO
MSmetrix	NO	NO	NO	YES (only for single-center application)	YES (only for single-center application)	YES	YES
SPM	YES	YES	YES	YES (for single-center and multicenter application)	YES (for single-center and multicenter application)	YES (possibility to build a tool)	YES (possibility to build a tool)

software should be based on whether its performance is satisfactory with regard to the specific requirements of the application. In the context of a single-center research study, for example, the input images would have low variability, since the same MR imaging unit and sequences would be used for image acquisition, and the user would have considerable expertise. Thus, high repeatability and accuracy on atrophy measures are sufficient to satisfy requirements for this setting. On the other hand, routine clinical application in MS is much more demanding because the input images would have a much higher variability: The MR imaging acquisition protocol and even the MR imaging unit may change at patient follow up. Moreover, a high level of automation is required in the clinical context. Given the poor reproducibility of all software, changes of the MR imaging unit should be avoided if atrophy is to be quantified meaningfully using these methods, and more standardized MR imaging acquisition protocols are needed. Moreover, a lower variability of the methods is required for the aim of individualized medicine.

Author contributions: Guarantors of integrity of entire study, L.S., J.S.G., M.F.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; manuscript final version approval, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, L.S., M.A.R., E.P., W.V.H., M.A.H.; clinical studies, W.V.H., A.R., J.S.G., J.P.; experimental studies, L.S., E.P., J.S.G., D. Smeets; statistical analysis, L.S.; and manuscript editing, L.S., M.A.R., W.V.H., M.A.H., N.D.S., A.R., J.S.G., J.P., D. Sima, D. Smeets, M.F.

Disclosures of Conflicts of Interest: L.S. disclosed no relevant relationships. M.A.R. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for Biogen and Merck Serono; is on the speakers bureau of Biogen, Teva, Genzyme, Novartis, and Merck Serono. Other relationships: disclosed no relevant relationships. E.P. Activities related to the present article: disclosed no relevant

relationships. Activities not related to the present article: has been paid for lectures by the Accademia Nazionale di Medicina, Excellence in Medical Education, Biogen Italia, and Università di Firenze. Other relationships: disclosed no relevant relationships. W.V.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is an employee of and owns stock or stock options in Icometrix. Other relationships: disclosed no relevant relationships. M.A.H. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a director and employee of and shareholder in Xinapse Systems. Other relationships: disclosed no relevant relationships. N.D.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for Schering, Biogen-Idec, Teva, Novartis, Sanofi-Genzyme, Roche, and Merck-Serono; has grants or grants pending from FISM and Novartis, is on the speakers bureaus of Biogen-Idec, Teva, Novartis, Sanofi-Genzyme, Roche, and Merck-Serono; has received travel funds from Teva, Novartis, Sanofi-Genzyme, Roche, and Merck-Serono. Other relationships: disclosed no relevant relationships. A.R. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for Novartis and Sanofi-Genzyme; is on the speakers bureaus of Roche, Stendhal, Sanofi-Genzyme, Biogen, Novartis, and Bayer. Other relationships: disclosed no relevant relationships. J.S. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: has grants or grants pending from Sanofi-Genzyme; is on the speakers bureaus of Teva, Almirall, Novartis, Biogen, Roche, Merck, and Sanofi-Genzyme (this includes payments both to J.S. and J.S.'s institution); has been paid for travel by Novartis. Other relationships: disclosed no relevant relationships. J.P. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: paid for advisory work for pharma; institution has grants or grants pending from MSS, the Guthy-Jackson Foundation, and pharma grants for MR imaging; has been paid by Novartis, Biogen, Merck-Serono, and Medimmune for lectures for academically organized meetings (not let by pharma except for Medimmune); Teva paid ECTRIMS conference costs this year. Other relationships: disclosed no relevant relationships. D. Sima disclosed no relevant relationships. D. Smeets Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is an employee of Icometrix. Other relationships: disclosed no relevant relationships. M.F. Activities related to the present article: disclosed no relevant relationships. Activities not related to the present article: is a consultant for Biogen Idec, Merck-Serono, Novartis, and Teva Pharmaceutical Industries; has grants or grants pending from Biogen Idec, Merck-Serono, Novartis, Teva Pharmaceutical Industries, and Roche; is on the speakers bureau of Biogen Idec, Merck-Serono, Novartis, and Teva Pharmaceutical Industries; is the editor in chief of the *Journal of Neurology*. Other relationships: disclosed no relevant relationships.

References

1. Evangelou N, Esiri MM, Smith S, Palace J, Matthews PM. Quantitative pathological evidence for axonal loss in normal appearing white matter in multiple sclerosis. *Ann Neurol* 2000;47(3):391–395.
2. Popescu BF, Lucchinetti CF. Pathology of demyelinating diseases. *Annu Rev Pathol* 2012;7(1):185–217.
3. Charil A, Filippi M. Inflammatory demyelination and neurodegeneration in early multiple sclerosis. *J Neurol Sci* 2007;259(1-2):7–15.
4. Filippi M, Rocca MA. MRI evidence for multiple sclerosis as a diffuse disease of the central nervous system. *J Neurol* 2005;252(5 Suppl 5):v16–v24.
5. Bermel RA, Bakshi R. The measurement and clinical relevance of brain atrophy in multiple sclerosis. *Lancet Neurol* 2006;5(2):158–170.
6. Rocca MA, Battaglini M, Benedict RH, et al. Brain MRI atrophy quantification in MS: From methods to clinical application. *Neurology* 2017;88(4):403–413.
7. Filippi M, Rocca MA. MR imaging of gray matter involvement in multiple sclerosis: implications for understanding disease pathophysiology and monitoring treatment efficacy. *AJNR Am J Neuroradiol* 2010;31(7):1171–1177.
8. Damjanovic D, Valsasina P, Rocca MA, et al. Hippocampal and deep gray matter nuclei atrophy is relevant for explaining cognitive impairment in MS: a multicenter study. *AJNR Am J Neuroradiol* 2017;38(1):18–24.
9. De Stefano N, Matthews PM, Filippi M, et al. Evidence of early cortical atrophy in MS: relevance to white matter changes and disability. *Neurology* 2003;60(7):1157–1162.
10. Fisher E, Lee JC, Nakamura K, Rudick RA. Gray matter atrophy in multiple sclerosis: a longitudinal study. *Ann Neurol* 2008;64(3):255–265.
11. Filippi M, Preziosa P, Copetti M, et al. Gray matter damage predicts the accumulation of disability 13 years later in MS. *Neurology* 2013;81(20):1759–1767.
12. Pravatà E, Rocca MA, Valsasina P, et al. Gray matter atrophy, cognitive impairment, and depression in patients with multiple sclerosis. *Mult Scler* 2017;23(14):1864–1874.
13. Sormani MP, Arnold DL, De Stefano N. Treatment effect on brain atrophy correlates with treatment effect on disability in multiple sclerosis. *Ann Neurol* 2014;75(1):43–49.
14. Vidal-Jordana A, Sastre-Garriga J, Rovira A, Montalban X. Treating relapsing-remitting multiple sclerosis: therapy effects on brain atrophy. *J Neurol* 2015;262(12):2617–2626.
15. Sastre-Garriga J, Pareto D, Rovira À. Brain atrophy in multiple sclerosis: clinical relevance and technical aspects. *Neuroimaging Clin N Am* 2017;27(2):289–300.
16. Wang C, Beadnall HN, Hatton SN, et al. Automated brain volumetrics in multiple sclerosis: a step closer to clinical application. *J Neurol Neurosurg Psychiatry* 2016;87(7):754–757.
17. Kappos L, De Stefano N, Freedman MS, et al. Inclusion of brain volume loss in a revised measure of ‘no evidence of disease activity’ (NEDA-4) in relapsing-remitting multiple sclerosis. *Mult Scler* 2016;22(10):1297–1305.
18. Ad-Dab'bagh Y, Einarson D, Lyttelton O, et al. The CIVET image-processing environment: a fully automated comprehensive pipeline for anatomical neuroimaging research. In: Corbetta M, ed. *Proceedings of the 12th Annual Meeting of the Organization for Human Brain Mapping*. Florence, Italy: NeuroImage, 2006.
19. Avants BB, Tustison NJ, Wu J, Cook PA, Gee JC. An open source multi-variate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 2011;9(4):381–400.
20. Smeets D, Ribbens A, Sima DM, et al. Reliable measurements of brain atrophy in individual patients with multiple sclerosis. *Brain Behav* 2016;6(9):e00518.
21. Smith SM, De Stefano N, Jenkinson M, Matthews PM. Normalized accurate measurement of longitudinal brain change. *J Comput Assist Tomogr* 2001;25(3):466–475.
22. Smith SM, Zhang Y, Jenkinson M, et al. Accurate, robust, and automated longitudinal and cross-sectional brain change analysis. *Neuroimage* 2002;17(1):479–489.
23. Zijdenbos AP, Forghani R, Evans AC. Automatic “pipeline” analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Trans Med Imaging* 2002;21(10):1280–1291.
24. Derakhshan M, Caramanos Z, Giacomini PS, et al. Evaluation of automated techniques for the quantification of grey matter atrophy in patients with multiple sclerosis. *Neuroimage* 2010;52(4):1261–1267.
25. Enzinger C, Fazekas F. Measuring gray matter and white matter damage in MS: why this is not enough. *Front Neurol* 2015;6:56.
26. 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(9):2062–2071.
27. Nakamura K, Fisher E. Segmentation of brain magnetic resonance images for measurement of gray matter atrophy in multiple sclerosis patients. *Neuroimage* 2009;44(3):769–776.
28. Polman CH, Reingold SC, Banwell B, et al. Diagnostic criteria for multiple sclerosis: 2010 revisions to the McDonald criteria. *Ann Neurol* 2011;69(2):292–302.
29. Sullivan DC, Obuchowski NA, Kessler LG, et al. Metrology standards for quantitative imaging biomarkers. *Radiology* 2015;277(3):813–825.
30. De Stefano N, Battaglini M, Smith SM. Measuring brain atrophy in multiple sclerosis. *J Neuroimaging* 2007;17(Suppl 1):10S–15S.
31. Nakamura K, Jones S, Van Hecke W, et al. Comparison of brain atrophy measurement techniques in a longitudinal study of multiple sclerosis patients with frequent MRIs. *Neurology* 2017;88(16 Suppl):P4.376.
32. Mendrik AM, Vincken KL, Kuijff HJ, et al. MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comput Intell Neurosci* 2015;2015:813696.
33. Kazemi K, Noorizadeh N. Quantitative comparison of SPM, FSL, and BrainSuite for brain MR image segmentation. *J Biomed Phys Eng* 2014;4(1):13–26.
34. Steenwijk MD, Amiri H, Schoonheim MM, et al. Agreement of MSmetrix with established methods for measuring cross-sectional and longitudinal brain atrophy. *Neuroimage Clin* 2017;15:843–853.
35. De Stefano N, Stromillo ML, Giorgio A, et al. Establishing pathological cut-offs of brain atrophy rates in multiple sclerosis. *J Neurol Neurosurg Psychiatry* 2016;87(1):93–99.