AJNR

Considerations for Mean Upper Cervical Cord Area Implementation in a Longitudinal MRI Setting: Methods, Interrater Reliability, and MRI Quality Control

C. Chien, V. Juenger, M. Scheel, A.U. Brandt and F. Paul

Considerations for Mean Upper Cervical Cord Area Implementation in a Longitudinal MRI Setting: Methods, Interrater Reliability, and MRI Quality Control

C. Chien, V. Juenger, M. Scheel, A.U. Brandt, and F. Paul

ABSTRACT

BACKGROUND AND PURPOSE: Spinal cord atrophy is commonly measured from cerebral MRIs, including the upper cervical cord. However, rescan intraparticipant measures have not been investigated in healthy cohorts. This study investigated technical and rescan variability in the mean upper cervical cord area calculated from T1-weighted cerebral MRIs.

MATERIALS AND METHODS: In this retrospective study, 8 healthy participants were scanned and rescanned with non-distortion- and distortion-corrected MPRAGE sequences (11–50 sessions in 6–8 months), and 50 participants were scanned once with distortion-corrected MPRAGE sequences in the Day2day daily variability study. From another real-world observational cohort, we collected non-distortion-corrected MPRAGE scans from 27 healthy participants (annually for 2–4 years) and cross-sectionally from 77 participants. Statistical analyses included coefficient of variation, smallest real difference, intraclass correlation coefficient, Bland-Altman limits of agreement, and paired t tests.

RESULTS: Distortion- versus non-distortion-corrected MPRAGE-derived mean upper cervical cord areas were similar; however, a paired t test showed incomparability ($t = 11.0$, $P = <.001$). Higher variability was found in the mean upper cervical cord areas calculated from an automatic segmentation method. Interrater analysis yielded incomparable measures in the same participant scans ($t = 4.5$, $P = <.001$). Non-distortion-corrected mean upper cervical cord area measures were shown to be robust in real-world data ($t = −1.04$, $P = .31$). The main sources of variability were found to be artifacts from movement, head/neck positioning, and/or metal implants.

CONCLUSIONS: Technical variability in cord measures decreased using non-distortion-corrected MRIs, a semiautomatic segmentation approach, and 1 rater. Rescan variability was within $\pm4.4\%$ for group mean upper cervical cord area when MR imaging quality criteria were met.

ABBREVIATIONS: CoV = coefficient of variation; HP = healthy participants; ICC = intraclass correlation coefficient; MUCCA = mean upper cervical cord area; SC = spinal cord; SCT = Spinal Cord Toolbox; SRD = smallest real difference; UCC = upper cervical cord; DND = Day2day non-distortion corrected; DDC = Day2day distortion corrected cross-sectional; CNDC = clinical non-distortion corrected cross-sectional; D'AD = asymptotic test of equality; LoA = Bland-Altman limits of agreement; SEM = standard error of measurement

Spinal cord (SC) atrophy occurs often in neuroinflammatory diseases and can be measured using structural MRI.[1-3] Several studies have identified cervical SC atrophy as predictive of disease progression and/or disability in patients with multiple sclerosis and related diseases; thus, it is of great interest to measure SC atrophy longitudinally.[1,4,5] One of the most common SC atrophy

measurement methods is the mean upper cervical cord area (MUCCA).[6-8] MUCCA has been found to be robust at the C1–C2 and C2–C3 intervertebral levels.[9,10] However, there is little consensus on how reliable it is longitudinally or whether SC atrophy measures are ready for use as a clinical trial end point and/or for patient monitoring in neuroinflammatory diseases,[11,12] SC injury,[13,14] or degenerative cervical myelopathy,[15] among other diseases affecting the SC. Most studies of MR imaging–based SC

atrophy comparisons of diseased with healthy SC measures have relied on interindividual variability calculated from cross-sectional data.[16] Intraindividual variability in healthy participants (HP) is often ignored or assumed to be equivalent to the mean change (or SD) in SC atrophy measures.[9] The utility of using cerebral 3D MRIs, including the upper cervical cord (UCC), as a source of MUCCA calculations is based on the availability of this sequence in most patients. These scans can be used for both detection of abnormalities in the brain and the retrospective analysis of SC atrophy.[8] Although some methods for reducing variability in MUCCA have been described, these methods require extra calculations and effort that are not implemented in many SC atrophy studies, as of yet.[17] Thus, investigations of raw MUCCA in HP are still required for making considerations for MRI sequence and analysis methods in prospective SC atrophy studies. Technical and rescan variability in HP must be evaluated to understand the pathologic changes in raw MUCCA in a longitudinal setting.

In this study, we hypothesized that the raw MUCCA will differ in HP, dependent on several factors: 1) non-distortion- versus distortion-corrected source MRIs, 2) analysis software used for MUCCA calculation, 3) interrater biases, and 4) MR imaging artifacts/positioning. The aim of our study was to identify factors that impact healthy rescan MUCCA in a longitudinal intraparticipant manner using rescan cerebral distortion- and non-distortion-corrected MPRAGE images, while also accounting for interparticipant variations using cross-sectional cerebral distortion and non-distortion-corrected MPRAGE images.

MATERIALS AND METHODS
The local institutional review board (NeuroCure Clinical Research Center, Charité-Universitätsmedizin Berlin) approved this retrospective study, and written informed consent was obtained from all participants.

Cohort
We performed a retrospective analysis of data of HP collected from the Day2day daily variability study,[18] and an ongoing observational single-center cohort (EA1/163/12). Inclusion criteria were as follows: a minimum of 18 years of age and no presence of any neurologic/psychiatric disorders, contraindications for an MR imaging examination, or incidental MR imaging findings. Participants who had undergone a full MR imaging scan protocol between January 2015 and January 2019 were eligible. Age and sex were collected for each healthy participant.

MR Imaging Acquisition
All images were acquired with 2 distinct 3T systems (Magnetom Trio; Siemens, Erlangen, Germany). From the Day2day study, we collected distortion- and non-distortion-corrected 3D MPRAGE sequences, including the upper cervical cord ($1 \times 1 \times 1$ mm resolution, TR = 2500 ms, TE = 4.77 ms, acquisition time [altogether] = 9:20 minutes, plane = sagittal). Eight HP (female/male = 6:2, mean age = 28.4 ± 2.97 years) were scanned with this MPRAGE protocol weekly for 6–8 months (11–50 scans per healthy participant), and 50 HP (female/male = 50:0, mean age = 24.9 ± 3.10 years) were scanned with the distortion-corrected 3D

MPRAGE sequence once in the Day2day study, all with the same scanner at the Max Planck Institute for Imaging Research, Berlin. From the observational study, we collected a non-distortion-corrected 3D MPRAGE sequence including the UCC ($1 \times 1 \times 1$ mm resolution, TR = 1900 ms, TE = 3.03 ms, acquisition time = 4:26 minutes, plane = sagittal) all with the same scanner at the Berlin Center of Advanced Imaging. Twenty-seven HP (female/male = 18:9, mean age = 36.8 ± 14.3 years) were scanned annually for 2–4 years (2–4 scans per healthy participant), and 77 HP (female/male = 49:28, mean age = 34.5 ± 12.0 years) were scanned using this non-distortion-corrected MPRAGE sequence once in the observational study.

Image Postprocessing
All images were reoriented with fslreorient2std (https://www.rdocumentation.org/packages/fslr/versions/2.24.1/topics/fslreorient2std) to match the Montreal Neurological Institute 152 standard template, which only applies 0°, 90°, 180°, or 270° rotations (FSL; Analysis Group, FMRIB, 2019; Version 5.0.9; open source: https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/), and N4 bias field corrected (Advanced Normalization Tools; 2019; Version 0.0.8; open source: http://stnava.github.io/ANTs/).[19] Longitudinal/rescan images were not coregistered because the SC region is not the center of registration, and in most coregistration pipelines, brain extraction is required, which cuts off the UCC.

Image Analysis
MUCCA was measured in all postprocessed MPRAGE images using the active surface method[20] using Jim software (2019; Version 7.0; http://www.xinapse.com/Manual/cord_intro.html) at the C2/C3 intervertebral space level (MUCCA-Jim), where the mean of cross-sectional areas in 5 consecutive slices of each MPRAGE image were obtained by an experienced rater (C.C.). Three center-of-cord seeds were marked by the rater, 1 at the most superior section of the C2/C3 intervertebral space, 1 after moving 2 slices down (center of C2/C3 intervertebral space), and the last seed after moving 2 more slices down in the cervical cord. Jim Cord Finder tool was set to a nominal cord diameter of 10 mm, with the number of shape coefficients set to 20 coefficients, a default order of longitudinal variation of 6, and the "cord is hypointense to CSF" setting not checked (for a T1 image). The active surface model takes into account the angulation of the cord by calculating cross-sectional areas in a plane perpendicular to the local cord centerline. Unfolded cervical cord images from each MPRAGE image were created using the same 5 consecutive slices, in which the setting for "create an unfolded image of the spinal cord" was checked in the Spinal Cord Finder Tool in Jim 7.0 so that during the calculation of the cross-sectional areas of the cord outline an unfolded image is automatically saved. The unfolded images are straightened cord centerline images, where cord image section planes are perpendicular to the straightened cord centerline.[21] The unfolded cervical cord images were used as input for automatic MUCCA segmentation (MUCCA-SCT) using the Spinal Cord Toolbox (SCT; PropSeg; 2019; version 4.1.0 https://sourceforge.net/p/spinalcordtoolbox/wiki/correction_PropSeg/).[22] The comparison of the 2 segmentation methods is valuable because the

semiautomatic Jim requires roughly 2–3 minutes of manual marking and quality checks per scan after reorienting to standard Montreal Neurological Institute space, while the automatic PropSeg requires only about 3 seconds of computation time after creating unfolded cervical cord images. If no unfolded cervical cord images are created, PropSeg is able to segment full cervical cords in roughly 15 seconds.

For interrater analysis, V.J. (trainee) measured MUCCA-Jim in the same MPRAGE scans from 1 healthy participant, using the same settings and methodology as described above.

Of the 308 rescan MRIs from 8 HP in the Day2day study, 37 (13%) distortion-corrected and 28 (9%) non-distortion-corrected MPRAGE scans were excluded from analyses due to motion artifacts and/or rescan differences in participant positioning in the scanner. From the ongoing observational cohort of 27 HP, we collected 66 longitudinal MRIs, of which 15 (7 HP, 23%) were excluded due to motion and/or metallic implant artifacts and/or differences in positioning in the scanner. In total, 28 HP with 600 longitudinal/rescan MPRAGE scans were included in statistical analysis.

Statistical Analysis

For age and sex comparison between HP from the Day2day study and our center, Kruskal-Wallis rank sum and χ^2 tests were performed, respectively. Group baseline MUCCAs calculated from different MR imaging scans were compared using an ANOVA test. Effect sizes between groups were evaluated with bootstrapped ($n = 5000$) confidence intervals.[23] Comparisons between MUCCAs derived from distortion- and non-distortion-corrected MPRAGE scans and from Jim versus SCT were performed using the coefficient of variation (CoV) asymptotic test of equality,[24] smallest real difference (SRD),[25] Bland-Altman limits of agreement, and paired t tests. The SRD (Equation 1) is a statistical method that estimates a true difference in measures and accounts for the 95% confidence interval of true observed differences among measures (1.96) using the difference in variances of measures ($\sqrt{2}$) and the standard error of measurement (SEM). SEM is calculated using intraclass correlation (ICC) metrics (Equation 2), as shown below.

1)
$$ SRD = 1.96 \times \sqrt{2} \times SEM $$

2)
$$ SEM = \sqrt{(Total\ Variance)(1 - ICC)} $$

Interrater MUCCAs were compared using the CoV, ICC, SRD, and a paired t test. ICCs were calculated with consistency from a 2-way mixed model. Because the number of longitudinal/rescan MUCCAs differed among HP, we calculated the CoV for each individual participant. All statistics and graphs were produced using R (Version 3.4.0; http://www.r-project.org),[26] and statistical significance was set to $P < .05$.

RESULTS

Cohort Demographics

Of the 28 participants with rescan MUCCAs, 21 (75%) were women, and of the 127 participants with cross-sectional MUCCAs, 99 (78%) were women. All relevant demographic data are shown in On-line Table 1.

There were significant differences between the cross-sectional HP' sex ($\chi^2 = 21.3$, $P = <.001$) and age ($\chi^2 = 33.1$, $P = <.001$) distributions. A significant difference in the baseline mean MUCCA-Jim when comparing rescan, longitudinal, and cross-sectional data from the Day2day and clinical participants ($F = 14.7$, $P = <.001$) was observed (On-line Table 1). However, little difference in the effect sizes of mean MUCCA-Jim derived from distortion- versus non-distortion-corrected MPRAGE scans was found (Fig 1).

Larger effect sizes from differences in MUCCA-Jim seemed to be the result of small sample sizes when subtracting the mean MUCCAs of a small cohort of 8 HP from the mean MUCCAs of 20 HP. However, when larger cohorts were compared, MUCCAs from distortion- and non-distortion-corrected MPRAGE scans (Day2day distortion-corrected cross-sectional and clinical non-distortion-corrected cross-sectional scans), a much smaller effect size was seen.

Rescan MUCCA from Distortion versus Non-Distortion-Corrected MPRAGE Sequences

Rescan MUCCA-Jims derived from the Day2day distortion- and non-distortion-corrected MPRAGE scans were compared. The CoVs in repeat MUCCAs for each participant are shown in the Table. An asymptotic test for the equality of the CoV between distortion- and non-distortion-derived MUCCA-Jim was non-significant (asymptotic test of equality = <0.001, $P = .98$). The SRD range in distortion-corrected derived MUCCA-Jim was ± 3.5 mm^2 ($\pm 4.8\%$ of the mean group MUCCA), while it was ± 3.2 mm^2 ($\pm 4.4\%$ of the mean group MUCCA) for non-distortion-corrected derived MUCCA-Jim. Bland-Altman limits of agreement (-2.9 to 6.0 mm^2) showed a predisposition for MUCCA-Jim to have a lower value when measured from non-distortion-corrected MPRAGE scans. This result was corroborated when the rescan MUCCA-Jim was evaluated using a paired t test ($t = 10.98$, $P = <.001$). Fig 2 and On-line Fig 1 illustrate the differences between each rescan MUCCA.

Although CoVs from most non-distortion-corrected MPRAGE-derived MUCCA-Jims were lower than those derived from distortion-corrected derived MUCCA-Jims, it can be seen, along with Fig 1, that participants with a similar number of scans analyzed had similar CoVs from both types of scans (eg, participants 1, 2, 4, 5, 7, and 8). CoVs from a similar number of scans analyzed were mostly lower in non-distortion-corrected derived MUCCA-Jims, and because the mean follow-up scan time per healthy participant was 5.2 days, this finding suggests that these measures are more robust than those from distortion-corrected cerebral scans.

Rescan MUCCA from Semiautomatic-versus-Automatic Segmentation

Variability in non-distortion-corrected MPRAGE-derived MUCCA was tested using the semiautomatic Jim (MUCCA-Jim) and the fully automatic PropSeg (MUCCA-SCT) segmentation methods. The CoVs in MUCCA for each participant from the Day2day cohort are shown in On-line Table 2.

An asymptotic test for the equality of CoV between MUCCA-Jim and MUCCA-SCT was significant (asymptotic test of equality = 130.1, $P = <.001$). The SRD range for MUCCA-Jim was ± 3.2 mm^2 ($\pm 4.4\%$ of the mean group MUCCA-Jim), while it was ± 9.5 mm^2 ($\pm 14.5\%$ of the mean group MUCCA-SCT) for MUCCA-SCT. Bland-Altman limits of agreement (-2.3 to 16.0 mm^2) showed a predisposition for MUCCA-Jim to have a higher value than MUCCA-SCT, which was also evident when evaluating the rescan MUCCA-Jim and MUCCA-SCT with a paired t test ($t = 24.42$, $P = <.001$). Since PropSeg

was optimized on MR imaging acquired with a spine coil instead of a head-neck coil, as used in this study, differences in the CSF to SC contrast-to-noise ratios may cause an underestimation of this segmentation method.[27] On-line Figure 2 shows the Bland-Altman and paired t test results when comparing MUCCA-Jim and MUCCA-SCT, as well as sample segmentations by Jim and PropSeg.

Single- versus Multirater Rescan MUCCA-Jim

To evaluate the robustness of using nondistortion MPRAGE scans and semiautomatic software to measure the MUCCA, we performed an interrater analysis for the same participant with 50 weekly scans from the Day2day study. Rater 1 (C.C.) has several years of experience with MUCCA segmentation, while rater 2 (V.J.) has never segmented MUCCA before this study. The CoVs in rescan MUCCA-Jim for healthy participant the same HP from rater 1 and rater 2 were 1.38% and 1.64%, respectively. An asymptotic test for the equality of CoV was nonsignificant (asymptotic test of equality = 1.5, $P = .23$). The ICC was 0.74, and the SRD range in MUCCA-Jim from the 2 raters was ± 2.8 mm^2 ($\pm 3.8\%$ of the mean group MUCCA-Jim). The rescan MUCCA-Jim when compared using a paired t test still showed a significant difference ($t = 4.54$, $P = <.001$). On-line Figure 3 illustrates the paired t test comparison of rescan MUCCA-Jim, from the same participant, measured by 2 raters.

Real-World Data Confirmation

To validate the findings from our analysis of the Day2day study cohort, we compared, cross-sectionally, 50 Day2day HP MUCCA-Jim derived from distortion-corrected MPRAGE scans with 77 observational HP MUCCA-Jim derived from non-
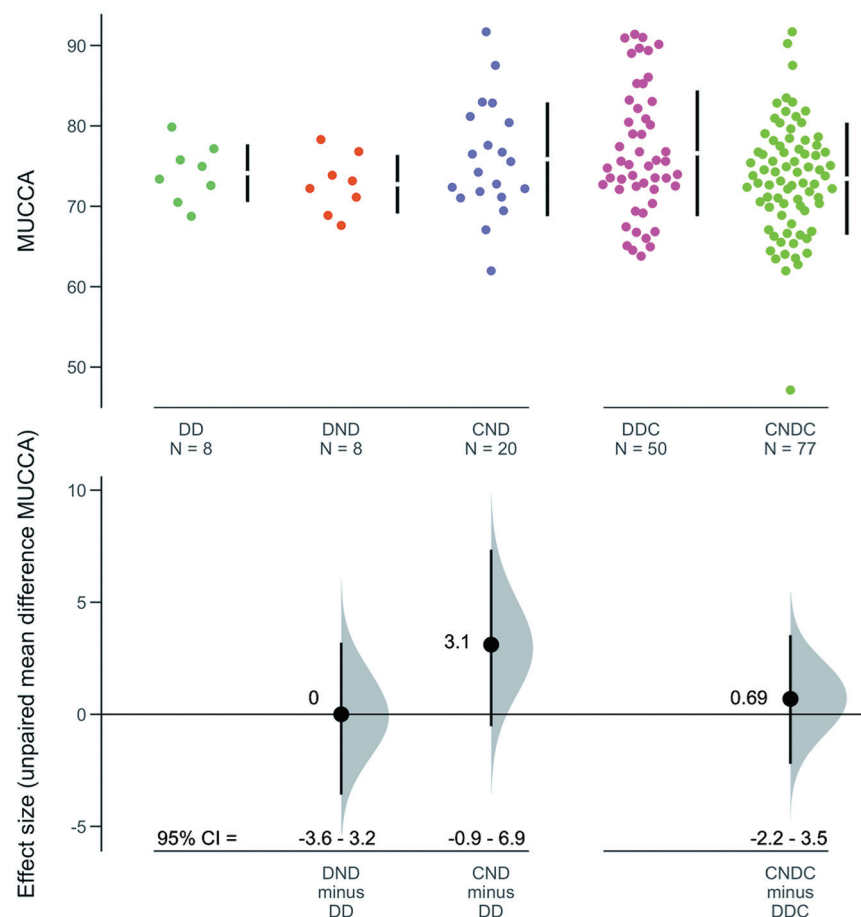

FIG 1. Comparison of effect sizes of differences between baseline MUCCA-Jim derived from the participants with longitudinal Day2day distortion-corrected (DD) DD and Day2day non-distortion-corrected (DND), clinical non-distortion-corrected (CND), Day2day distortion-corrected cross-sectional, and clinical non-distortion-corrected cross-sectional (CNDC) MPRAGE scans. *Vertical lines* beside *dot plots* indicate SDs around the means.

CoVs for each Day2day participant MUCCA-Jim derived from distortion- and non-distortion-corrected MPRAGE scans

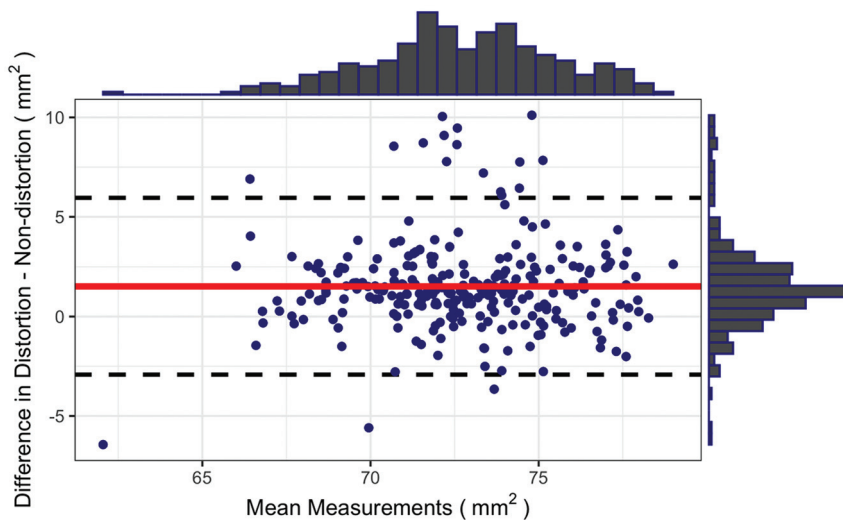| Participant	No. of Rescan MPRAGE Scans Collected	No. of Distortion-Corrected Derived MUCCAs Analyzed	CoV for Distortion-Corrected Derived MUCCA	No. of Non-Distortion-Corrected Derived MUCCAs Analyzed	CoV for Non-Distortion-Corrected Derived MUCCA
1	50	50	1.36	50	1.38
2	13	10	1.89	10	1.72
3	50	37	3.09	43	2.03
4	11	11	1.39	10	1.35
5	45	40	1.40	41	1.33
6	47	36	1.91	42	1.55
7	43	43	1.35	41	1.55
8	49	42	1.55	43	1.40

FIG 2. Bland-Altman graph showing larger distortion-corrected MPRAGE-derived MUCCA-Jim values than non-distortion-corrected MPRAGE derived MUCCA-Jim, with relatively normally distributed mean values.
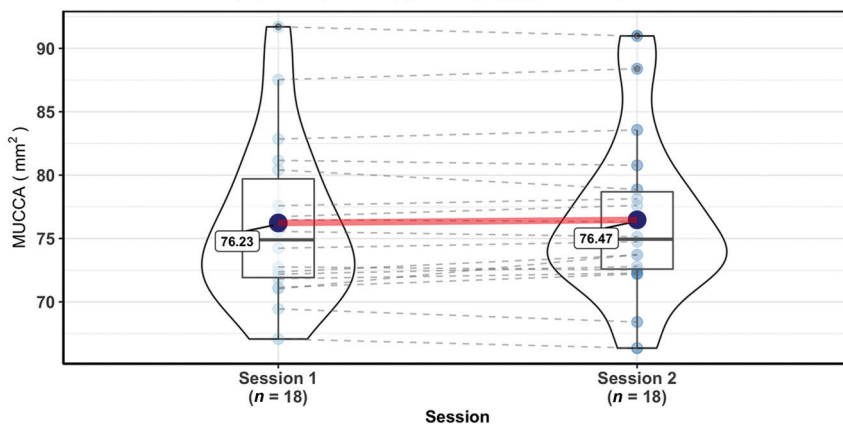

FIG 3. Paired t test showing that the mean rescan MUCCA-Jim is quite stable when measured by an experienced rater from real-world non-distortion-corrected MPRAGE scans obtained, on average, 1 year apart.

distortion-corrected MPRAGE scans. Again, we found that the CoV was smaller in MUCCAs derived from non-distortion-corrected versus distortion-corrected MPRAGE scans (9.5% versus 10.2%, respectively). However, the CoV values were not significantly different when tested with an asymptotic test for equality (asymptotic test of equality = 0.10, $P = .75$). A 2-sample t test showed a slight difference between MUCCA-Jim (derived from distortion-corrected: mean = 76.6 ± 7.8 mm^2; from non-distortion-corrected: mean = 73.4 ± 7.0 mm^2; $t = 2.32$, $P = .02$) from the 2 cohorts. Because the 50 HP from the Day2day study were all women with a small range in age, one would expect less variation in interparticipant MUCCAs in

this cohort. However, our finding that there was a higher (though slightly higher) CoV in the MUCCA-Jim from this cohort suggests that a "normal range" of MUCCA may be larger when using distortion-corrected scans for SC atrophy evaluation.

We also tested rescan MPRAGE MUCCA-Jim from the observational cohort to evaluate the feasibility of using the raw MUCCA in a longitudinal, real-world setting. The CoVs in rescan MUCCA-Jim for each participant are shown in On-line Table 3. The SRD range for MUCCA-Jim was ±2.4 mm^2 (±3.1% of the mean group MUCCA-Jim). The ICC for longitudinal MUCCA-Jim was 0.97. A paired t test of MUCCA-Jim measured between 2 consecutive imaging time points showed that the MUCCA were stable ($t = -1.04$, $P = .31$), as illustrated in Fig 3.

Sources of Exclusion from Analysis (MPRAGE Quality Control)
On inspection of large variations in consecutive MUCCA-Jim from the Day2day and observational cohort, several artifacts were identified to give inadequate rescan MUCCAs for comparison. Participants with largely variable MUCCAs most often had ≥1 of the following MR imaging qualities: 1) motion artifacts seen in the brain or neck region, 2) head and neck positions being different between sessions, and/or 3) metal implants in the mouth causing artifacts (eg, retainer, braces, and so forth). Head and neck positioning differences included head tilt angle and different levels of UCC included in the scan FOV. All of these qualities can be observed in postprocessed MPRAGE images, as shown in Fig 4.

DISCUSSION
We investigated rescan MUCCA variability in HP on the basis of the following: 1) scan settings for source MRIs, 2) segmentation software, 3) interrater bias, and 4) imaging artifacts. Rescan MUCCA was found to be less variable when measured from 3T, non-distortion-corrected cerebral 3D MPRAGE scans. We also showed that using a semiautomatic segmentation approach (Jim 7.0 Cord Finder Tool) gave a robust MUCCA in participants compared with fully automatic segmentation (SCT PropSeg) in
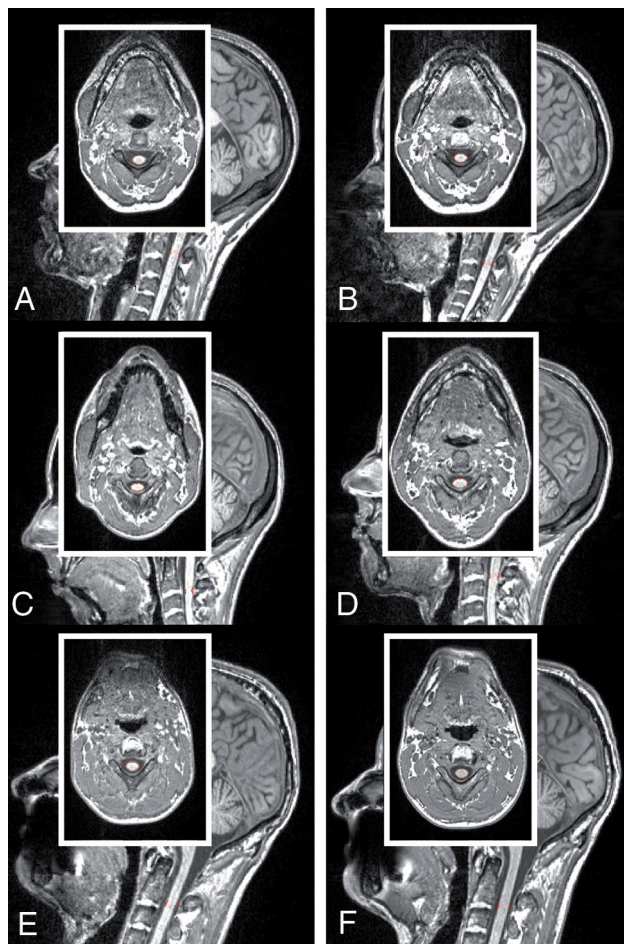
FIG 4. Consecutive sagittal MPRAGE scans. *A* and *B*, Large motion artifacts (session 1 MUCCA-Jim = 67.8 mm^2, session 2 MUCCA-Jim = 53.1 mm^2). *C* and *D*, Different head and neck (vertebrae level cutoff) positioning (session 1 MUCCA-Jim = 47.1 mm^2, session 2 MUCCA-Jim = 64.3 mm^2). *E* and *F*, Metal implants in the mouth region (session 1 MUCCA-Jim = 73.8 mm^2, session 2 MUCCA-Jim = 79.7 mm^2). Inlays show axial MUCCA regions, where the SC does not immediately seem different, but surrounding head/neck regions are often sliced differently. *Red outline* in images indicates the location of MUCCA segmentation.

the exact same MR imaging slices. Interrater bias evaluation revealed that MUCCA from a single rater is more consistent, and if measured by different raters, a systematic bias could occur. Finally, several sources of highly variable rescan MUCCAs were found through quality control analysis of postprocessed MPRAGE scans. These included imaging artifacts in the brain/mouth/neck regions and, most noticeably, when participants were positioned differently (ie, different head tilt, different levels of vertebrae in the FOV) and/or had metal implants in the mouth.

Few studies publish whether MR imaging sequences are distortion- or non-distortion-corrected using a phantom. MUCCA has been shown to be more robust when normalized with a factor calculated from a cylindric phantom that represents the UCC.[17] However, we found that raw MUCCA derived from distortion-corrected MPRAGE scans yielded more variable (higher CoVs and SDs) within-participant rescan measures than those derived

from non-distortion-corrected MPRAGE scans. This finding suggests that without an SC-specific phantom for distortion correction, non-distortion-corrected cerebral scans yield less variable and more robust raw MUCCAs for longitudinal SC atrophy evaluation. We would recommend, in retrospective studies, using non-distortion-corrected cerebral MPRAGE for higher longitudinal reproducibility, as seen from the results of our real-world clinical data. Furthermore, to reduce variability in repeat MUCCAs as much as possible, when using cerebral MPRAGE scans, we propose a strict quality control step before inclusion in a study, which involves the removal of participants with any metallic implants and scans that have any brain-movement artifacts. It would be also beneficial to standardize the positioning of patients in the scanner to achieve cervical cord placement as straight as possible and include a minimum number of vertebral levels.

We compared 2 common SC segmentation tools used to calculate the MUCCA: the semiautomatic Jim Cord Finder and the fully automatic SCT PropSeg. Jim uses an active surface method,[20] and SCT uses an iterative propagation of a deformable adaptive contrast method.[22] One study found good correlation between Jim and PropSeg using a normalization factor,[28] while another study found a systematic difference in the 2 outputs.[27] We confirmed this systematic bias in raw MUCCA, in which PropSeg yielded significantly smaller and more variable segmentations than Jim, with higher intraparticipant variability and low consistency in paired measures. Because we segmented MUCCAs using the exact same MR imaging slices for this analysis, our study shows that raw MUCCAs measured by these 2 methods are not comparable. SCT has another functionality, namely DeepSeg (https://github.com/neuropoly/spinalcordtoolbox), which is based on a convolutional neural network algorithm that segments the spinal cord, with the capability of also segmenting intramedullary multiple sclerosis lesions.[29] This method was shown to better segment MR imaging scans, including the brain, than PropSeg; however, raw MUCCAs were still significantly different between DeepSeg and Jim.[7] Thus, when one compares results from different studies using different segmentation methods, absolute MUCCA cannot be used; rather, we propose the evaluation of effect sizes from the change in the MUCCA from each study.

Many studies have used Jim for MUCCA segmentation, but they had either 1 experienced rater for multiple centers[30] or no raters were mentioned.[31] We found only a moderate intraclass correlation between MUCCAs segmented by 2 raters. The less experienced rater also had higher variability in the MUCCA, as evidenced by large differences in paired measures. These results highlight the need for disclosure of raters in MUCCA studies, standardized training of individuals performing SC segmentation, and/or using 1 rater for longitudinal studies.

We validated our findings in real-world clinical data. In distortion versus non-distortion MPRAGE-derived MUCCA, stability of the rescan MUCCA was shown to be a smallest real difference of roughly ±4% change. This is in line with a recent study that found high reproducibility of the MUCCA within 1 scanner with 1 measurement method, but no comparability among scanners or methods.[27] One large metadata study found,

on average, that patients with multiple sclerosis lose 1.78%/year of their cervical SC area.[16] We do not need to assume that SC atrophy during 1 year in HP would be zero; however, a large study with >1200 asymptomatic participants spanning 20–80 years in age found that in both men and women the C2/C3 intervertebral level cross-sectional area had a minimum decrease of about 1% and a maximum decrease of roughly 4.5% spanning several decades.[32] Thus, we can be fairly certain that during the course of a 1-year follow-up, normal SC atrophy would, most likely, not surpass 1.5%. Our study shows that annual MUCCA may not give enough signal over noise; thus, we suggest that MUCCA be monitored for >24 months to assess SC atrophy.

The limitations of our study include the small sample size of HP scanned longitudinally. Because our study was performed with retrospective data, we could not evaluate dedicated SC sequences, which may give more robust SC atrophy measurements using fast, automated methods such as PropSeg or DeepSeg. A future direction would be to collect a large amount of healthy participant SC MR imaging for further longitudinal testing. It would be ideal to scan many HP with a wide age range, to truly validate whether there is an effect of age or sex on the MUCCA. One study did find an effect of age, sex, height, and weight on the lower cervical volume of HP, however, with very small effect sizes for their models,[33] We were also not able to investigate the time of day of MR imaging acquisition or hydration levels of HP, which may lead to further fluctuations in the MUCCA. These factors have been found to affect brain morphometric measures[34] and SC area.[35] However, due to the robustness of our real-world longitudinal MUCCA analysis, we can postulate that these factors play a smaller role in MUCCA variability.

CONCLUSIONS

Our study illustrates some technical and rescan factors important in longitudinal SC atrophy studies using MUCCA: type and correction of source MR imaging, segmentation method, rater training, and MR imaging quality controls should be addressed, because all these factors contribute to increased rescan variability in the MUCCA. With these considerations, the MUCCA has the potential to be used in a longitudinal setting to evaluate and track SC atrophy.

REFERENCES

1. Brownlee WJ, Altmann DR, Alves Da Mota P, et al. **Association of asymptomatic spinal cord lesions and atrophy with disability 5 years after a clinically isolated syndrome.** *Mult Scler* 2017;23:665–74 CrossRef Medline

2. Tsagkas C, Magon S, Gaetano L, et al. **Preferential spinal cord volume loss in primary progressive multiple sclerosis.** *Mult Scler* 2019; 25:947–57 CrossRef Medline

3. Chien C, Scheel M, Schmitz-Hübsch T, et al. **Spinal cord lesions and atrophy in NMOSD with AQP4-IgG and MOG-IgG associated autoimmunity.** *Mult Scler* 2019;25:1926–36 CrossRef Medline

4. Zeydan B, Gu X, Atkinson EJ, et al. **Cervical spinal cord atrophy: an early marker of progressive MS onset.** *Neurol Neuroimmunol Neuroinflamm* 2018;5:e435 CrossRef Medline

5. Ciccarelli O, Cohen JA, Reingold SC, et al; International Conference on Spinal Cord Involvement and Imaging in Multiple Sclerosis and Neuromyelitis Optica Spectrum Disorders. **Spinal cord involvement in multiple sclerosis and neuromyelitis optica spectrum disorders.** *Lancet Neurol* 2019;18:185–97 CrossRef Medline

6. Sinnecker T, Granziera C, Wuerfel J, et al. **Future brain and spinal cord volumetric imaging in the clinic for monitoring treatment response in MS.** *Curr Treat Options Neurol* 2018;20:17 CrossRef Medline

7. Alcaide-Leon P, Cybulsky K, Sankar S, et al. **Quantitative spinal cord MRI in radiologically isolated syndrome.** *Neurol Neuroimmunol Neuroinflamm* 2018;5:e436 CrossRef Medline

8. Chien C, Brandt AU, Schmidt F, et al. **MRI-based methods for spinal cord atrophy evaluation: a comparison of cervical cord cross-sectional area, cervical cord volume, and full spinal cord volume in patients with aquaporin-4 antibody seropositive neuromyelitis optica spectrum disorders.** *AJNR Am J Neuroradiol* 2018;39:1362–68 CrossRef Medline

9. Moccia M, Ruggieri S, Ianniello A, et al. **Advances in spinal cord imaging in multiple sclerosis.** *Ther Adv Neurol Disord* 2019;12: 1756286419840593 CrossRef Medline

10. Valsasina P, Aboulwafa M, Preziosa P, et al. **Cervical cord T1-weighted hypointense lesions at MR imaging in multiple sclerosis: relationship to cord atrophy and disability.** *Radiology* 2018;288: 234–44 CrossRef Medline

11. Filippi M, Rocca MA, Ciccarelli O, et al; MAGNIMS Study Group. **MRI criteria for the diagnosis of multiple sclerosis: MAGNIMS consensus guidelines.** *Lancet Neurol* 2016;15:292–303 CrossRef Medline

12. Prados F, Barkhof F. **Spinal cord atrophy rates: ready for prime time in multiple sclerosis clinical trials?** *Neurology* 2018;91:157–58 CrossRef Medline

13. Lammertse D, Dungan D, Dreisbach J, et al; National Institute on Disability and Rehabilitation. **Neuroimaging in traumatic spinal cord injury: an evidence-based review for clinical practice and research.** *J Spinal Cord Med* 2007;30:205–14 CrossRef Medline

14. Seif M, Wheeler-Kingshott CA, Cohen-Adad J, et al. **Guidelines for the conduct of clinical trials in spinal cord injury: neuroimaging biomarkers.** *Spinal Cord* 2019;57:717–28 CrossRef Medline

15. Nouri A, Martin AR, Mikulis D, et al. **Magnetic resonance imaging assessment of degenerative cervical myelopathy: a review of structural changes and measurement techniques.** *Neurosurg Focus* 2016; 40:E5 CrossRef Medline

16. Casserly C, Seyman EE, Alcaide-Leon P, et al. **Spinal cord atrophy in multiple sclerosis: a systematic review and meta-analysis.** *J Neuroimaging* 2018;28:556–86 CrossRef Medline

17. Papinutto N, Bakshi R, Bischof A, et al; North American Imaging in Multiple Sclerosis Cooperative (NAIMS). **Gradient nonlinearity effects on upper cervical spinal cord area measurement from 3D T1-weighted brain MRI acquisitions.** *Magn Reson Med* 2018;79: 1595–601 CrossRef Medline

18. Filevich E, Lisofsky N, Becker M, et al. **Day2day: investigating daily variability of magnetic resonance imaging measures over half a year.** *BMC Neurosci* 2017;18:65 CrossRef Medline

19. Tustison NJ, Avants BB, Cook PA, et al. **N4ITK: improved N3 bias correction.** *IEEE Trans Med Imaging* 2010;29:1310–20 CrossRef Medline

20. Horsfield MA, Sala S, Neema M, et al. **Rapid semi-automatic segmentation of the spinal cord from magnetic resonance images: application in multiple sclerosis.** *Neuroimage* 2010;50:446–55 CrossRef Medline

21. Valsasina P, Horsfield MA, Rocca MA, et al. **Spatial normalization and regional assessment of cord atrophy: voxel-based analysis of cervical cord 3D T1-weighted images.** *AJNR Am J Neuroradiol* 2012;33:2195–200 CrossRef Medline

22. De Leener B, Kadoury S, Cohen-Adad J. **Robust, accurate and fast automatic segmentation of the spinal cord.** *NeuroImage* 2014;98: 528–36 CrossRef Medline

23. Ho J, Tumkaya T, Aryal S, et al. **Moving beyond P values: data analysis with estimation graphics.** *Nat Methods* 2019;16:565–66 CrossRef Medline

24. Feltz CJ, Miller GE. **An asymptotic test for the equality of coefficients of variation from k populations.** *Stat Med* 1996;15:646–58 CrossRef Medline

25. Beckerman H, Roebroeck ME, Lankhorst GJ, et al. **Smallest real difference, a link between reproducibility and responsiveness.** *Qual Life Res* 2001;10:571–78 CrossRef Medline

26. Core Team R. *R: A language and environment for statistical computing.* Vienna; R Foundation for Statistical Computing; 2019

27. Weeda MM, Middelkoop SM, Steenwijk MD, et al. **Validation of mean upper cervical cord area (MUCCA) measurement techniques in multiple sclerosis (MS): high reproducibility and robustness to lesions, but large software and scanner effects.** *Neuroimage Clin* 2019;24:101962 CrossRef Medline

28. Yiannakas MC, Mustafa AM, De Leener B, et al. **Fully automated segmentation of the cervical cord from T1-weighted MRI using PropSeg: application to multiple sclerosis.** *Neuroimage Clin* 2016;10:71–77 CrossRef Medline

29. Gros C, De Leener B, Badji A, et al. **Automatic segmentation of the spinal cord and intramedullary multiple sclerosis lesions with convolutional neural networks.** *Neuroimage* 2019;184:901–15 CrossRef Medline

30. Liu Y, Lukas C, Steenwijk MD, et al. **Multicenter validation of mean upper cervical cord area measurements from head 3D T1-weighted MR imaging in patients with multiple sclerosis.** *AJNR Am J Neuroradiol* 2016;37:749–54 CrossRef Medline

31. Valsasina P, Rocca MA, Horsfield MA, et al. **A longitudinal MRI study of cervical cord atrophy in multiple sclerosis.** *J Neurol* 2015;262:1622–28 CrossRef Medline

32. Kato F, Yukawa Y, Suda K, et al. **Normal morphology, age-related changes and abnormal findings of the cervical spine, Part II: magnetic resonance imaging of over 1,200 asymptomatic subjects.** *Eur Spine J* 2012;21:1499–507 CrossRef Medline

33. Yanase M, Matsuyama Y, Hirose K, et al. **Measurement of the cervical spinal cord volume on MRI.** *J Spinal Disord Tech* 2006;19:125–29 CrossRef Medline

34. Trefler A, Sadeghi N, Thomas AG, et al. **Impact of time-of-day on brain morphometric measures derived from T1-weighted magnetic resonance imaging.** *Neuroimage* 2016;133:41–52 CrossRef Medline

35. Wang C, Tam RC, Mackie E, et al. **Dehydration affects spinal cord cross-sectional area measurement on MRI in healthy subjects.** *Spinal Cord* 2014;52:616–20 CrossRef Medline