HHS Public Access

Author manuscript

AJNR Am J Neuroradiol. Author manuscript; available in PMC 2018 February 01.

Published in final edited form as:

AJNR Am J Neuroradiol. 2017 August ; 38(8): 1501–1509. doi:10.3174/ajnr.A5254.

Volumetric Analysis From A Harmonized Multi-Site Brain MRI Study of a Single-Subject with Multiple Sclerosis

Russell T. Shinohara[1], **Jiwon Oh**[2,3], **Govind Nair**[4], **Peter A. Calabresi**[2], **Christos Davatzikos**[5], **Jimit Doshi**[5], **Roland G. Henry**[6], **Gloria Kim**[7], **Kristin A. Linn**[1], **Nico Papinutto**[6], **Daniel Pelletier**[8], **Dzung L. Pham**[9], **Daniel S. Reich**[2,4], **William Rooney**[10], **Snehashis Roy**[9], **William Stern**[6], **Subhash Tummala**[7], **Fawad Yousuf**[7], **Alyssa Zhu**[6], **Nancy L. Sicotte**[11], **Rohit Bakshi**[7,12], and **the North American Imaging in Multiple Sclerosis Cooperative**[13]

[1]Department of Biostatistics and Epidemiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[2]Department of Neurology, the Johns Hopkins University School of Medicine, Baltimore, MD, USA

[3]St. Michael's Hospital, University of Toronto, Toronto, Ontario, Canada

[4]Translational Neuroradiology Section, National Institute of Neurological Disorders and Stroke, National Institutes of Health, Bethesda, MD, USA

[5]Department of Radiology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA

[6]Department of Neurology, University of California - San Francisco, San Francisco, CA, USA

[7]Laboratory for Neuroimaging Research, Partners Multiple Sclerosis Center, [7]Department of Neurology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

[8]Department of Neurology, Yale Medical School, New Haven, CT, USA

[9]Henry M. Jackson Foundation for the Advancement of Military Medicine, Bethesda, MD, USA

[10]Advanced Imaging Research Center, Oregon Health & Science University, Portland, OR, USA

[11]Department of Neurology, Cedars-Sinai Medical Center, Los Angeles, CA, USA

[12]Department of Radiology, Brigham and Women's Hospital, Harvard Medical School, Boston, MA, USA

Abstract

Background and Objectives—Magnetic resonance imaging (MRI) can be used to measure structural changes in the brain of people with multiple sclerosis (MS), and is essential for diagnosis, longitudinal monitoring, and therapy evaluation. The North American Imaging in Multiple Sclerosis Cooperative (NAIMS) steering committee developed a uniform high-resolution

Corresponding Author: Russell T Shinohara, Department of Biostatistics and Epidemiology, University of Pennsylvania Perelman School of Medicine, 514 Blockley Hall, 423 Guardian Drive, Philadelphia, PA, USA 19104, rshi@mail.med.upenn.edu.

[13]A complete list of the NAIMS participants is provided in the acknowledgment section.

This work was presented in preliminary form at the 2016 annual meeting of the European Committee on Treatment and Research in Multiple Sclerosis (ECTRIMS) in London, UK.

3T MRI protocol relevant to the quantification of cerebral lesions and atrophy and implemented it at seven sites across the United States. To assess inter-site variability in scan data, a volunteer with relapsing-remitting MS was imaged with a scan-rescan at each site.

Materials and Methods—All imaging was acquired on Siemens scanners (4 Skyra, 2 TIM Trio, and 1 Verio). Expert segmentations were manually obtained for T1-hypointense and T2 (FLAIR)-hyperintense lesions. Several automated lesion detection and whole-brain, cortical, and deep gray matter volumetric pipelines were applied. Statistical analyses were conducted to assess variability across sites, as well as systematic biases in the volumetric measurements that were site-related.

Results—Systematic biases due to site differences in expert-traced lesion measurements were significant ($p<0.01$ for both T1 and T2 lesion volumes), with site explaining over 90% of the variation (range = 13.0–16.4ml in T1 and 15.9–20.1ml in T2) in lesion volumes. Site also explained more than 80% of the variation in most automated volumetric measurements. Output measures clustered according to scanner models, with similar results from the Skyra vs. the other two units.

Conclusions—Even in multi-center studies with consistent scanner field strength and manufacturer, after protocol harmonization, systematic differences can lead to severe biases in volumetric analyses.

Introduction

Conventional magnetic resonance imaging (MRI) is an established tool for measuring CNS lesion and tissue compartment volumes *in vivo* in people with multiple sclerosis (MS). In the brain and spinal cord, inflammatory demyelinating lesions appear hyperintense on T2-weighted (T2) images. Total cerebral T2 lesion volume (T2LV) is a key metric for the longitudinal monitoring of disease severity, as well as a standard outcome in clinical trials of MS therapeutics[1–3]. Many T2 lesions exhibit pulse-sequence-dependent hypointensity on T1-weighted images (T1), which has been shown to be associated with more severe (destructive) histopathology and worse clinical outcomes[4–8]. MRI is also used to measure cerebral atrophy, a commonly-used supportive outcome measure of the neurodegenerative aspects of the diseases in both relapsing-remitting and progressive forms of MS[9–18]. Together, lesion and atrophy measures provide complementary quantitative information about disease progression that are considered central to patient assessment[19].

Unfortunately, differences in acquisition methods have the potential to bias MRI metrics. Factors such as equipment manufacturer, magnetic field strength, and acquisition protocol can affect image contrast and resultant volumetric data. Indeed, several groups have investigated the reliability of volumetric measurements across scanners[20–27], but little is understood about the variability in volumetric measurements of lesions and atrophy in people with MS. Furthermore, many automated segmentation algorithms depend on statistical atlases or models that are built using healthy volunteers or that depend on registration, which can be compromised by the presence of MS pathology[28].

The North American Imaging in Multiple Sclerosis (NAIMS) cooperative was established to accelerate the pace of imaging research. As a consortium, our first aim was to facilitate multi-center imaging study by creating harmonized MRI protocols across sites. In this

manuscript, we describe initial results from our pilot study, which tested the feasibility of multi-site standardization of MRI acquisition for the quantification of lesion and tissue volumes. We compared inter- to intra-site scan-rescan variability in various MRI output metrics, using consistently acquired 3T acquisitions.

Materials and Methods

Participant

A 45-year-old man with clinically stable relapsing-remitting MS and mild to moderate physical disability was imaged at seven NAIMS sites across the United States (Table 1). He developed the first symptoms of the disease 13 years before study enrollment, and had been relapse-free in the previous year after starting dimethyl fumarate. His last intravenous corticosteroid administration was five years previously. Timed 25-foot walk at study entry was 5.3 seconds. Expanded Disability Status Scale score was 3.5, both at study entry and exit, without any intervening relapses on-study. The participant signed informed consent for this study, which was approved by each site's institutional review board.

Scan acquisition

Through consensus agreement in the cooperative, the NAIMS cooperative developed a standardized high-resolution 3T MRI brain scan protocol. All imaging was acquired using Siemens scanners, which at the time of the study were used by the majority of NAIMS sites. Scan-rescan pairs were acquired on these scanners using the acquisition protocol shown in Table 1. At each site, the scan-rescan experiment was performed on the same day, with the participant removed and repositioned between scans. To replicate a "real world" clinical trial setting, none of the participant's scans were co-registered to each other. The volunteer was also imaged at the National Institutes of Health (NIH) NAIMS site at the beginning and end of the study (five months later) to assess disease stability. Raw MRI scans were distributed to 4 NAIMS sites for post-processing.

Expert lesion tracing

De-identified images underwent manual quantification to assess total cerebral T1-hypointense lesion volume (T1LV) and T2LV from the native 3D FLAIR and T1 images by the consensus of trained observers (FK, FY) under the supervision of an experienced observer (ST). For T2LV, this involved manually identifying all lesions on the FLAIR images. For T1LV, lesions were required to show hypointensity on T1-weighted images and at least partial hyperintensity on FLAIR images. The lesions were then segmented by one observer (GK) using a semi-automated edge-finding tool in Jim (v. 7.0) to determine lesion volumes. Images were presented to the same reading panel for all of the above steps in random order in one batch, and mixed into a stack of 50 other MS images to reduce scan-to-scan memory effects and to preserve blinding.

Automated analysis

Several fully automated pipelines were also used to estimate T2LV, and volume of total brain, normal-appearing white matter (WM), and both cortical and deep gray matter structures. To prevent overfitting, all pipelines were used with their default settings.

According to published recommendations for each method separately, where appropriate images were inhomogeneity corrected, rigidly aligned across sequences from each scan session, processed for removal of extracerebral voxels for all processing pipelines, and intensity normalized. For lesional measurements, several algorithms were applied by the laboratories that developed or co-developed the various methods: Lesion-TOADS[29], a fuzzy c-means-based segmentation technique with topological constraints, OASIS[30], a logistic-regression-based segmentation method leveraging statistical intensity normalization, S3DL[31], a patch-based dictionary learning multi-class method, and WMLS[32], a local support vector machine-based segmentation algorithm developed for vascular lesions that also employs corrective learning. To estimate the volume of gray matter structures, Lesion-TOADS, FSL-FIRST[33] (a Bayesian appearance method), MaCRUISE[34] (a combined multi-atlas segmentation and cortical reconstruction algorithm), and MUSE[35] (an ensemble multi-atlas label fusion method) were used. The FSL-FIRST[33] analysis was applied directly to the raw T1 images according to common practice, and OASIS[30] was applied to the T1, FLAIR, T2, and PD images after preprocessing; all other images were applied to appropriately preprocessed T1 and FLAIR imaging. Not all algorithms measured volumes of the same set of structures. Lesion filling was not performed. Lesion-TOADS, MaCRUISE, and MUSE also yielded estimates for total brain volume.

Statistical Analysis

All statistical analysis was conducted in the R software environment[36]. To compare estimated volumes within and across sites, mean volumes and standard deviations were computed. T-tests were also used to test for differences in within-site average between scanner platforms. Correlations between these averages across segmentation algorithms were also explored. The proportion of variation explained by site was computed and the association with site was assessed using permutation testing. The coefficients of variation were also estimated across sites. To assess associations between session-average measured total brain and lesional volumes and time of day (morning versus afternoon), Wald testing within a linear model framework was employed both marginally and adjusting for scanner platform.

Results

The participant was found to be stable regarding cerebral lesion load during the study. Comparing images acquired at the NIH at study entry and exit, the manually measured T2LV in the participant was similar (17.9ml in September 2015 versus 17.8ml in February 2016). The T1LV was also stable (15.5ml versus 15.1ml). This imaging stability paralleled his clinical stability (see Methods).

The manually estimated T1LV and T2LV for each scan is shown in Figure 1. Site explained 95% of the variation observed in the estimated T2LV, and 92% of the variation in the estimated T1LV, indicating significant scanner-to-scanner differences despite protocol harmonization, which clearly exceed scan-rescan variability within sites. The range of T2LVs was 15.9ml to 20.1 ml, indicating that differences of up to 25% of the lesion volume were observed across sites. The range of T1LVs was similarly wide, ranging from 13.0ml to

16.4ml. Further inspection of these volumes across platforms indicated that Skyra scanners showed larger lesion volumes compared to other Siemens platforms both on T1LV (Skyra mean T1 = 15.2ml compared with non-Skyra mean T1 =13.8ml, $p<0.05$) and T2LV (Skyra mean T2 = 18.9ml compared with non-Skyra mean T2 = 16.6ml, $p<0.01$). A visual example of the segmented lesions across scanners is provided in Figure 2.

Results from the automated techniques for delineating and measuring T2LV are shown in Figure 3. The automated lesion segmentations showed marked disagreement in the average lesional volume measurements compared with the manually assessed volumes, and all methods showed large site-to-site differences (in some cases up to 7.5ml, or almost 50% of the manually measured lesion volume), except for LesionTOADS (range 10.5ml to 11.0ml) which was more stable. For all methods, site explained more than 50% of the observed variation; 53% of the variation was explained by site (permutation $p = 0.36$) for S3DL, 54% for Lesion-TOADS ($p=0.41$), 44% for OASIS ($p=0.57$), and 83% for WMLS ($p=0.002$) which clearly was most prone to site-related variation.

To measure brain structure volumes, several automated methods were used. As an example, results for the thalamus are shown in Figures 4 and 5. While LesionTOADS estimated smaller volumes, MUSE, FIRST, and MaCRUISE yielded similar average measurements. Nonetheless, site was strongly associated with measured thalamic volume, explaining 96% of the LesionTOADS volume variation ($p<0.01$), 89% of MUSE ($p<0.01$), 84% of FIRST ($p=0.04$), and 65% of MaCRUISE ($p=0.17$). Similar results for the putamen, caudate, cortical gray matter, normal-appearing white matter, and total brain volume were found, as provided in Supplementary Figures 1–5. Summaries of the coefficient of variation give an intuitive measure of the scale of the combined scan-rescan and across-site variation as shown in Figure 6. Finally, the proportion of variation explained by site is shown in Figure 7. Note that, in almost all cases, site explained more than 50% of the variation, with the majority of measurement techniques exhibiting more than 80% variation due to site for all structures assessed.

While all images were acquired on 3T Siemens scanners, the model type appeared to influence the results; there was evidence of systematic differences in many measurements between Skyra and non-Skyra scanners. Figure 8 shows the negative log p-value for the comparison of volumes averaged across scan-rescan measurements, with larger values indicating more systematic differences between platforms. The largest platform-associated differences were observed in MaCRUISE measurements of normal-appearing white matter, cortical gray matter, and, consequently, total brain volume. LesionTOADS also showed large differences in total brain volume attributable to cortical gray matter, as did S3DL for T2 lesion volume measurements. MUSE showed major differences in thalamic volume across scanner model, and FIRST showed similar discrepancies in the thalamus and caudate. The correlation between site-averaged measurements varied dramatically, especially for lesional and total brain volume measurements (see Supplementary Figure 6), indicating that site differences resulted in contrasting effects on outputs from the different algorithms. While the other measurements showed less scanner model-related variation, most still showed prominent differences between Skrya and non-Skrya scanners.

The time of day of scan acquisition was not associated with manually segmented T1 (t=0.45) or T2 lesion volumes (t=0.38), or total brain volume, as measured by any of the automated algorithms (see Supplementary Figures 7 and 8).

Discussion

Clinical MS therapeutic trials have traditionally employed 1.5T MRI platforms to provide metrics on cerebral lesions and atrophy as supportive outcome measures. However, there is a growing interest in the use of high-resolution 3T imaging to assess disease activity and disease severity in MS. Such 3T imaging brings the potential for increased sensitivity to lesions[37,38] and atrophy,[39] higher reliability,[39,40] and closer relationships to clinical status,[38,39] when compared to scanning at 1.5T. The purpose of this study was to evaluate the consistency of metrics obtained from a single MS participant using a high-resolution 3T brain MRI protocol distributed to seven sites. The results of our study indicate that even in multi-center acquisitions from the same scanner vendor after careful protocol harmonization, systematic differences in images led to severe biases in volumetric analyses. These biases were present in manually and automatically measured volumes of white matter lesions, as well as in automatically measured volumes of whole brain, gray and white matter structures. These biases were also highly dependent on scanning equipment, which resulted from a significantly higher sensitivity to lesions in newer scanners from the same manufacturer compared with earlier models even at the same field strength.

In comparison to past estimates of reliability of volumetric measurements of brain structures, our findings point to higher between-site variation than previously documented. In particular, Cannon et al.[27] reported that between 3% and 26% of the observed variation in global and subcortical volumes was attributable to site; this was a study of 8 healthy participants imaged on two successive days across 8 sites using 3T Siemens and GE scanners. However, it is important to note that the proportion of explained variation has a different interpretation from that reported here. The total variation in Cannon et al. consisted of four contributors to variance: first, across-site differences; second, across-scan differences; third, across-day differences; and fourth, across-subject differences. In our single-participant study, we isolated only the first two variance components which allowed us to compare variation as it is relevant for precision medicine (subject-specific) applications. Previous work indicated that the observed variation attributable to scanning occasion was small[25,27]; indeed, Cannon et al. found this to constitute less than 1% of the variation. Thus, we did not scan our participant on subsequent days but rather simply repositioned the participant between scans during the same imaging session. A notable difference between our study and that of Cannon et al. is that we did not use data from a standardized phantom concurrently acquired for correction of between-scanner variations in gradient non-linearity and scaling. Cannon et al. found this correction to improved between-site intra-class correlations and greatly reduced differences between scanner manufacturers. Similarly, Gunter et al.[41] reported the usefulness of a phantom for scanner harmonization and quality control in the Alzheimer's Disease Neuroimaging Initiative. In future studies, we will focus on applying phantom calibrations across NAIMS sites to extend our current observations. Despite the growing literature on the importance of diurnal variation and hydration status for volumetric analyses [42–45], we found no significant associations between

time of day and measured volumes. This may indicate that in single participant analyses, time of day and day-to-day variation may be of less concern than the much larger source of variation of scanner platform. Interestingly, Cannon et al. also found that measurements acquired using scanners from the same manufacturer and similar receive coils had higher reliability. In our study, we found that even scanner models (i.e. Skyra versus non-Skyra) from the same manufacturer varied markedly in their estimates of lesion volume, highlighting the importance of between-scanner differences for assessing MS-related structural changes.

To assess differences across processing pipelines, we used a variety of techniques for automated segmentation of lesion and white and gray matter volumes. Different segmentation algorithms showed a range of variability in their estimates, as well as their sensitivity to differences between scanners. For example, LesionTOADS showed much less variable lesion measurements than any other technique, and was not as sensitive to differences in scanner platform. Lesion-TOADS was the only unsupervised lesion segmentation technique employed. Contrast differences between the participant data and the training data of the other supervised methods could be associated with the greater sensitivity to scanner differences, and this might be mitigated by specific (albeit potentially laborious) tuning to individual platforms. However, while sensitivity to biological change is generally higher for methods yielding less noisy estimates, as only a single individual was studied here, our data cannot be taken to indicate that LesionTOADS is superior to other methods of estimating thalamic volume, for example. Additionally, both purely intensity-based segmentation algorithms, OASIS and WMLS, appeared to be more sensitive to site differences, which may indicate that methods that rely more on topology, shape, or spatial context may be more stable across scanners. This indicates that across-scanner differences may be driven by contrast differences rather than geometric distortions. Future investigation to extend these findings could involve quantitative contrast-to-noise and signal-to-noise comparisons across scanners. Allowing segmentation parameters to vary across sites could also help stability.

A limitation of this study is its single-subject and single-time-point design, which makes the generalizability of the findings dependent on further investigation. In particular, the degree to which across-site differences might vary by lesion burden and degree of atrophy, as well as demographic variables, requires additional study. Future larger studies of multiple participants across disease stages including longitudinal measurements are necessary for understanding the implications of the biases described in this pilot study. Indeed, such studies would also allow for the assessment of the tradeoff between stability in measures across sites with sensitivity to biological differences. Differences between scanning equipment and scanner software versions have also been noted in past studies of reliability[23,25,27,46,47], but their implications for the assessment of pathology remain unclear. In particular, repeated acquisitions on scanners with different receive coils could provide additional insight concerning reliability. In addition, our study was from a single time point across scanners, whereas clinical trials rely on the quantification of intrasubject longitudinal change[48]. Each participant is typically scanned on the same platform, which may limit the variability in on-study change between participant. Further studies are necessary to assess

whether scan platform introduces the same level of acquisition-related variability when assessing longitudinal changes.

Given the inter-site differences observed in lesional measurements, across-site inference statistical adjustment for site is clearly necessary when analyzing volumetrics from multi-site studies, even when images are acquired using a harmonized protocol on 3T scanners produced by the same manufacturer. From a single participant, it is unclear what the role of differential sensitivity to lesions might be across people with heterogeneity in lesion location. For example, while lesion detection in the supratentorial white matter might be more straightforward and comparable across people, detection of lesions in the brain stem, cerebellum, and spinal cord may be more sensitive to differences in equipment. New statistical methods for measuring and correcting systematic biases are warranted, especially for studies in which patient populations may differ across sites. Indeed, intensity normalization and scan-effect removal techniques[49–55] (akin to batch-effect-removal methods in genomic studies[56]) are an active area of methodological research and promise to improve comparability of volumetric estimates from automated segmentation methods. After volumes are measured, statistical techniques for modeling estimated volumes from multi-center studies are also rapidly evolving[18,57]. These techniques bring the potential to mitigate site-to-site biases in group-level analyses, with better external validity at the cost of increased sample size.

By imaging the same subject with stable relapsing-remitting MS over a period of five months, we assessed scanner-related biases in volumetric measurements at seven NAIMS centers. Despite careful protocol harmonization and the acquisition of all imaging at 3T on Siemens scanners, we found significant differences in lesion and structural volumes. These differences were especially pronounced when comparing Skyra scanners to other Siemens 3T platforms. The results from this study highlight the potential for inter-scanner and inter-site differences that, unless properly accounted for, might confound MRI volumetric data from multi-center studies of brain disorders.

We conclude that our findings raise a key issue with the interpretability of MRI measurements in the context of personalized medicine, even in carefully controlled studies using harmonized imaging protocols.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Major support for this study was provided by the Race to Erase MS. Additional support came from RO1NS085211, R21NS093349, R01EB017255, and S10OD016356 from the National Institutes of Health and RG-1507-05243 from the National Multiple Sclerosis Society. The study was also partially supported by the Intramural Research Program of the National Institute of Neurological Disorders and Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

The following is a full list of individuals who contributed to this NAIMS study: Brigham and Women’s Hospital, Harvard Medical School (Boston, MA): Rohit Bakshi, Renxin Chu, Gloria Kim, Shahamat Tauhid, Subhash Tummala, Fawad Yousuf; Cedars-Sinai Medical Center (Los Angeles, CA): Nancy L. Sicotte; Henry M. Jackson Foundation for the Advancement of Military Medicine (Bethesda, MD): Dzung Pham, Snehashis Roy; National

Institutes of Health (Bethesda, MA): Frances Andrada, Irene C.M. Cortese, Jenifer Dwyer, Rosalind Hayden, Haneefa Muhammad, Govind Nair, Joan Ohayon, Daniel S. Reich, Pascal Sati, Chevaz Thomas; Johns Hopkins (Baltimore, MD): Peter A. Calabresi, Sandra Cassard, Jiwon Oh; Oregon Health and Science University (Portland, OR): William Rooney, Daniel Schwartz, Ian Tagge; University of California (San Francisco, CA): Roland G. Henry, Nico Papinutto, William Stern, Alyssa Zhu; University of Pennsylvania (Philadelphia, PA): Christos Davatzikos, Jimit Doshi, Guray Erus, Kristin Linn, Russell Shinohara; University of Toronto (Ontario, Canada): Jiwon Oh; Yale University (New Haven, CT): R. Todd Constable, Daniel Pelletier.

References

1. García-Lorenzo D, Francis S, Narayanan S, Arnold DL, Collins DL. Review of automatic segmentation methods of multiple sclerosis white matter lesions on conventional magnetic resonance imaging. Med Image Anal. 2013; 17(1):1–18. DOI: 10.1016/j.media.2012.09.004 [PubMed: 23084503]
2. Lublin FD, Reingold SC, Cohen JA, et al. Defining the clinical course of multiple sclerosis The 2013 revisions. Neurology. May.2014 doi: 10.1212/WNL.0000000000000560
3. Simon JH, Jacobs LD, Campion M, et al. Magnetic resonance studies of intramuscular interferon beta-1a for relapsing multiple sclerosis. The Multiple Sclerosis Collaborative Research Group. Ann Neurol. 1998; 43(1):79–87. DOI: 10.1002/ana.410430114 [PubMed: 9450771]
4. Bagnato F, Jeffries N, Richert ND, et al. Evolution of T1 black holes in patients with multiple sclerosis imaged monthly for 4 years. Brain J Neurol. 2003; 126(Pt 8):1782–1789. DOI: 10.1093/brain/awg182
5. Sahraian MA, Radue E-W, Haller S, Kappos L. Black holes in multiple sclerosis: definition, evolution, and clinical correlations. Acta Neurol Scand. 2010; 122(1):1–8. DOI: 10.1111/j.1600-0404.2009.01221.x [PubMed: 20003089]
6. Giorgio A, Stromillo ML, Bartolozzi ML, et al. Relevance of hypointense brain MRI lesions for long-term worsening of clinical disability in relapsing multiple sclerosis. Mult Scler Houndmills Basingstoke Engl. 2014; 20(2):214–219. DOI: 10.1177/1352458513494490
7. van Walderveen MA, Kamphorst W, Scheltens P, et al. Histopathologic correlate of hypointense lesions on T1-weighted spin-echo MRI in multiple sclerosis. Neurology. 1998; 50(5):1282–1288. [PubMed: 9595975]
8. Truyen L, van Waesberghe JH, van Walderveen MA, et al. Accumulation of hypointense lesions (“black holes”) on T1 spin-echo MRI correlates with disease progression in multiple sclerosis. Neurology. 1996; 47(6):1469–1476. [PubMed: 8960729]
9. 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. [PubMed: 10716264]
10. Evangelou N, Konz D, Esiri MM, Smith S, Palace J, Matthews PM. Regional axonal loss in the corpus callosum correlates with cerebral white matter lesion volume and distribution in multiple sclerosis. Brain J Neurol. 2000; 123(Pt 9):1845–1849.
11. Sastre-Garriga J, Ingle GT, Chard DT, et al. Grey and white matter volume changes in early primary progressive multiple sclerosis: a longitudinal study. Brain J Neurol. 2005; 128(Pt 6):1454–1460. DOI: 10.1093/brain/awh498
12. Sanfilipo MP, Benedict RHB, Weinstock-Guttman B, Bakshi R. Gray and white matter brain atrophy and neuropsychological impairment in multiple sclerosis. Neurology. 2006; 66(5):685–692. DOI: 10.1212/01.wnl.0000201238.93586.d9 [PubMed: 16534104]
13. Ge Y, Grossman RI, Udupa JK, Babb JS, Nyúl LG, Kolson DL. Brain atrophy in relapsing-remitting multiple sclerosis: fractional volumetric analysis of gray matter and white matter. Radiology. 2001; 220(3):606–610. DOI: 10.1148/radiol.2203001776 [PubMed: 11526256]
14. Fisher E, Lee J-C, Nakamura K, Rudick RA. Gray matter atrophy in multiple sclerosis: a longitudinal study. Ann Neurol. 2008; 64(3):255–265. DOI: 10.1002/ana.21436 [PubMed: 18661561]
15. Fisniku LK, Chard DT, Jackson JS, et al. Gray matter atrophy is related to long-term disability in multiple sclerosis. Ann Neurol. 2008; 64(3):247–254. DOI: 10.1002/ana.21423 [PubMed: 18570297]

16. 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. [PubMed: 12682324]
17. Losseff NA, Wang L, Lai HM, et al. Progressive cerebral atrophy in multiple sclerosis. A serial MRI study. Brain J Neurol. 1996; 119(Pt 6):2009–2019.
18. Keshavan A, Paul F, Beyer MK, et al. Power estimation for non-standardized multisite studies. NeuroImage. Apr.2016 doi: 10.1016/j.neuroimage.2016.03.051
19. Bakshi R, Thompson AJ, Rocca MA, et al. MRI in multiple sclerosis: current status and future prospects. Lancet Neurol. 2008; 7(7):615–625. DOI: 10.1016/S1474-4422(08)70137-6 [PubMed: 18565455]
20. Agartz I, Okuguwa G, Nordström M, Greitz D, Magnotta V, Sedvall G. Reliability and reproducibility of brain tissue volumetry from segmented MR scans. Eur Arch Psychiatry Clin Neurosci. 2001; 251(6):255–261. DOI: 10.1007/PL00007542 [PubMed: 11881838]
21. Bartzokis G, Mintz J, Marx P, et al. Reliability of in vivo volume measures of hippocampus and other brain structures using MRI. Magn Reson Imaging. 1993; 11(7):993–1006. DOI: 10.1016/0730-725X(93)90218-3 [PubMed: 8231683]
22. Maclaren J, Han Z, Vos SB, Fischbein N, Bammer R. Reliability of brain volume measurements: A test-retest dataset. Sci Data. 2014; 1:140037.doi: 10.1038/sdata.2014.37 [PubMed: 25977792]
23. Morey RA, Selgrade ES, Wagner HR, Huettel SA, Wang L, McCarthy G. Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. Hum Brain Mapp. 2010; 31(11): 1751–1762. DOI: 10.1002/hbm.20973 [PubMed: 20162602]
24. Schnack HG, van Haren NEM, Hulshoff Pol HE, et al. Reliability of brain volumes from multicenter MRI acquisition: a calibration study. Hum Brain Mapp. 2004; 22(4):312–320. DOI: 10.1002/hbm.20040 [PubMed: 15202109]
25. Jovicich J, Marizzoni M, Sala-Llonch R, et al. Brain morphometry reproducibility in multi-center 3T MRI studies: a comparison of cross-sectional and longitudinal segmentations. NeuroImage. 2013; 83:472–484. DOI: 10.1016/j.neuroimage.2013.05.007 [PubMed: 23668971]
26. Schnack HG, van Haren NEM, Brouwer RM, et al. Mapping reliability in multicenter MRI: voxel-based morphometry and cortical thickness. Hum Brain Mapp. 2010; 31(12):1967–1982. DOI: 10.1002/hbm.20991 [PubMed: 21086550]
27. Cannon TD, Sun F, McEwen SJ, et al. Reliability of neuroanatomical measurements in a multisite longitudinal study of youth at risk for psychosis. Hum Brain Mapp. 2014; 35(5):2424–2434. DOI: 10.1002/hbm.22338 [PubMed: 23982962]
28. Eloyan A, Shou H, Shinohara R, et al. Health effects of lesion localization in multiple sclerosis: spatial registration and confounding adjustment. PloS One. 2014; 9(9):e107263.doi: 10.1371/journal.pone.0107263 [PubMed: 25233361]
29. Shiee N, Bazin P-L, Ozturk A, Reich DS, Calabresi Pa, Pham DL. A topology-preserving approach to the segmentation of brain images with multiple sclerosis lesions. NeuroImage. 2010; 49(2): 1524–1535. DOI: 10.1016/j.neuroimage.2009.09.005 [PubMed: 19766196]
30. Sweeney EM, Shinohara RT, Shiee N, et al. OASIS is Automated Statistical Inference for Segmentation, with applications to multiple sclerosis lesion segmentation in MRI. NeuroImage Clin. 2013; 2:402–413. DOI: 10.1016/j.nicl.2013.03.002 [PubMed: 24179794]
31. Roy S, He Q, Sweeney E, et al. Subject-Specific Sparse Dictionary Learning for Atlas-Based Brain MRI Segmentation. IEEE J Biomed Health Inform. 2015; 19(5):1598–1609. DOI: 10.1109/JBHI.2015.2439242 [PubMed: 26340685]
32. Lao Z, Shen D, Liu D, et al. Computer-assisted segmentation of white matter lesions in 3D MR images using support vector machine. Acad Radiol. 2008; 15(3):300–313. DOI: 10.1016/j.acra.2007.10.012 [PubMed: 18280928]
33. Patenaude B, Smith SM, Kennedy DN, Jenkinson M. A Bayesian model of shape and appearance for subcortical brain segmentation. NeuroImage. 2011; 56(3):907–922. DOI: 10.1016/j.neuroimage.2011.02.046 [PubMed: 21352927]
34. Huo, Y., Carass, A., Resnick, SM., Pham, DL., Prince, JL., Landman, BA. Combining multi-atlas segmentation with brain surface estimation. Styner, MA., Angelini, ED., editors. 2016. p. 97840E

35. Doshi J, Erus G, Ou Y, et al. MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. NeuroImage. 2016; 127:186–195. DOI: 10.1016/j.neuroimage.2015.11.073 [PubMed: 26679328]
36. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2016. https://www.R-project.org/
37. Sicotte NL, Voskuhl RR, Bouvier S, Klutch R, Cohen MS, Mazziotta JC. Comparison of multiple sclerosis lesions at 1.5 and 3.0 Tesla. Invest Radiol. 2003; 38(7):423–427. DOI: 10.1097/01.RLI.0000065426.07178.f1 [PubMed: 12821856]
38. Stankiewicz JM, Glanz BI, Healy BC, et al. Brain MRI lesion load at 1.5T and 3T versus clinical status in multiple sclerosis. J Neuroimaging Off J Am Soc Neuroimaging. 2011; 21(2):e50–56. DOI: 10.1111/j.1552-6569.2009.00449.x
39. Chu R, Tauhid S, Glanz BI, et al. Whole Brain Volume Measured from 1.5T versus 3T MRI in Healthy Subjects and Patients with Multiple Sclerosis. J Neuroimaging Off J Am Soc Neuroimaging. 2016; 26(1):62–67. DOI: 10.1111/jon.12271
40. Bakshi R. Deep gray matter segmentation from 1.5T vs. 3T MRI in normal controls and patients with multiple sclerosis. 2016
41. Gunter JL, Bernstein MA, Borowski BJ, et al. Measurement of MRI scanner performance with the ADNI phantom. Med Phys. 2009; 36(6):2193–2205. DOI: 10.1118/1.3116776 [PubMed: 19610308]
42. Duning T, Kloska S, Steinsträter O, Kugel H, Heindel W, Knecht S. Dehydration confounds the assessment of brain atrophy. Neurology. 2005; 64(3):548–550. DOI: 10.1212/01.WNL.0000150542.16969.CC [PubMed: 15699394]
43. Sampat MP, Healy BC, Meier DS, Dell'Oglio E, Liguori M, Guttmann CRG. Disease modeling in multiple sclerosis: assessment and quantification of sources of variability in brain parenchymal fraction measurements. NeuroImage. 2010; 52(4):1367–1373. DOI: 10.1016/j.neuroimage.2010.03.075 [PubMed: 20362675]
44. Nakamura K, Brown RA, Araujo D, Narayanan S, Arnold DL. Correlation between brain volume change and T2 relaxation time induced by dehydration and rehydration: implications for monitoring atrophy in clinical studies. NeuroImage Clin. 2014; 6:166–170. DOI: 10.1016/j.nicl.2014.08.014 [PubMed: 25379428]
45. Nakamura K, Brown RA, Narayanan S, Collins DL, Arnold DL. Alzheimer's Disease Neuroimaging Initiative. Diurnal fluctuations in brain volume: Statistical analyses of MRI from large populations. NeuroImage. 2015; 118:126–132. DOI: 10.1016/j.neuroimage.2015.05.077 [PubMed: 26049148]
46. Jovicich J, Czanner S, Han X, et al. MRI-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. NeuroImage. 2009; 46(1):177–192. DOI: 10.1016/j.neuroimage.2009.02.010 [PubMed: 19233293]
47. Kruggel F, Turner J, Muftuler LT. Alzheimer's Disease Neuroimaging Initiative. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage. 2010; 49(3):2123–2133. DOI: 10.1016/j.neuroimage.2009.11.006 [PubMed: 19913626]
48. Filippi M, Wolinsky JS, Comi G. Effects of oral glatiramer acetate on clinical and MRI-monitored disease activity in patients with relapsing multiple sclerosis: a multicentre, double-blind, randomised, placebo-controlled study. Lancet Neurol. 2006; 5(3):213–220. DOI: 10.1016/S1474-4422(06)70327-1 [PubMed: 16488376]
49. Shinohara RT, Sweeney EM, Goldsmith J, et al. Statistical normalization techniques for magnetic resonance imaging. NeuroImage Clin. 2014; 6:9–19. DOI: 10.1016/j.nicl.2014.08.008 [PubMed: 25379412]
50. Nyul LGL, Udupa JJK, Zhang X. New variants of a method of MRI scale standardization. Med Imaging IEEE Trans On. 2000; 19(2):143–150.
51. Ghassemi R, Brown R, Narayanan S, Banwell B, Nakamura K, Arnold DL. Normalization of white matter intensity on T1-weighted images of patients with acquired central nervous system

demyelination. J Neuroimaging Off J Am Soc Neuroimaging. 2015; 25(2):184–190. DOI: 10.1111/jon.12129

52. Fortin J-P, Sweeney EM, Muschelli J, Crainiceanu CM, Shinohara RT. Alzheimer's Disease Neuroimaging Initiative. Removing inter-subject technical variability in magnetic resonance imaging studies. NeuroImage. 2016; 132:198–212. DOI: 10.1016/j.neuroimage.2016.02.036 [PubMed: 26923370]
53. Madabhushi A, Udupa JK. New methods of MR image intensity standardization via generalized scale. Med Phys. 2006; 33(9):3426–3426. DOI: 10.1118/1.2335487 [PubMed: 17022239]
54. Nyúl LG, Udupa JK. On standardizing the MR image intensity scale. Magn Reson Med Off J Soc Magn Reson Med Soc Magn Reson Med. 1999; 42(6):1072–1081.
55. Chua AS, Egorova S, Anderson MC, et al. Handling changes in MRI acquisition parameters in modeling whole brain lesion volume and atrophy data in multiple sclerosis subjects: Comparison of linear mixed-effect models. NeuroImage Clin. 2015; 8:606–610. DOI: 10.1016/j.nicl.2015.06.009 [PubMed: 26199872]
56. Leek JT, Scharpf RB, Bravo HC, et al. Tackling the widespread and critical impact of batch effects in high-throughput data. Nat Rev Genet. 2010; 11(10):733–739. DOI: 10.1038/nrg2825 [PubMed: 20838408]
57. Fennema-Notestine C, Gamst AC, Quinn BT, et al. Feasibility of multi-site clinical structural neuroimaging studies of aging using legacy data. Neuroinformatics. 2007; 5(4):235–245. DOI: 10.1007/s12021-007-9003-9 [PubMed: 17999200]

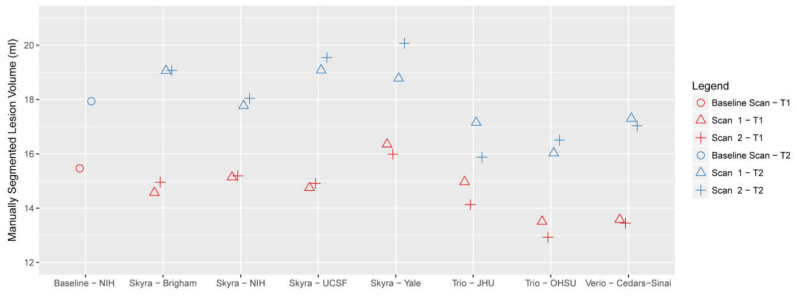

FIG 1.

Manually measured T1 (red) and T2 (blue) lesion volumes for scan-rescan pairs at each of seven NAIMS sites. Results from the baseline scan, acquired on the same Skyra scanner and subsequent imaging acquired at NIH, are shown using circles. Points have been slightly offset relative to one another for ease of visualization.

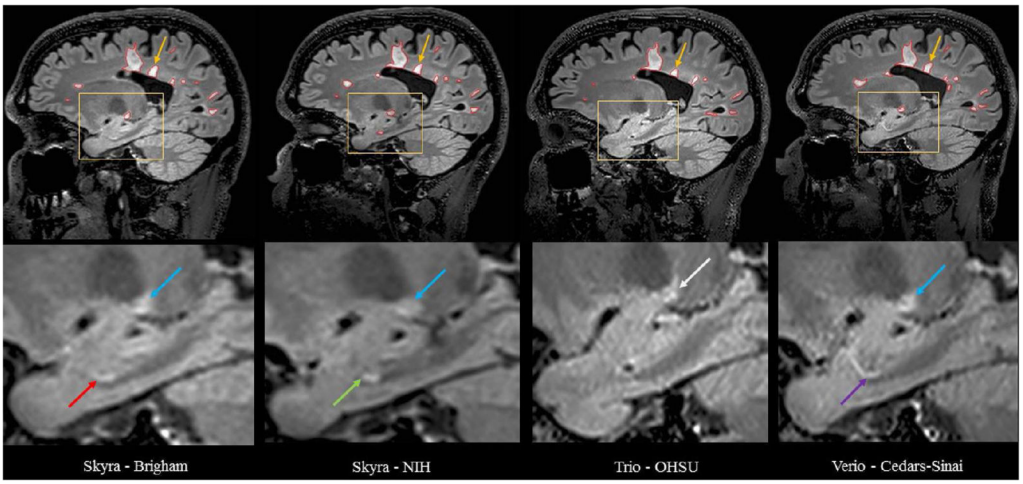

FIG 2.

Comparison of manual segmentation of cerebral T2 hyperintense lesions at four NAIMS sites. 3T MRI scans on Siemens scanners from a single subject with multiple sclerosis showing T2 hyperintense lesions from sagittal fluid-attenuated inversion recovery (FLAIR) sequences from 4 different North American Imaging in Multiple Sclerosis (NAIMS) sites and scanner models: Brigham and Women's Hospital, Skyra; National Institutes of Health (NIH), Skyra; Oregon Health & Sciences University (OHSU), Trio; Cedars-Sinai, Verio. The upper panel shows the native images. The lower panel shows zoomed and cropped images to illustrate the key findings. The green arrow (lower panel) shows a possible lesion detected and traced on the NIH scan; the red arrow shows the same lesion not detected by the expert procedure on Brigham scan; the purple arrow shows a similar tubular area that was interpreted as a blood vessel on the Cedars-Sinai scan, which was not selected as a lesion by the expert tracing; no lesion was detected on the OHSU scan in this area on this slice or any of the adjacent slices (not shown). The blue arrow shows a different lesion detected and traced on the Brigham, NIH, and Cedars-Sinai scans but not detected by the expert review on the OHSU scan, appearing hazy/subtle (white arrow). The yellow arrow (upper panel) shows a lesion on all scans; however, when adding the tracing of all slices showing the lesion, the 3D volume of the lesion differed among sites: Brigham = 0.059 ml, NIH = 0.053 ml, OHSU = 0.033 ml, Cedars-Sinai = 0.053 ml.

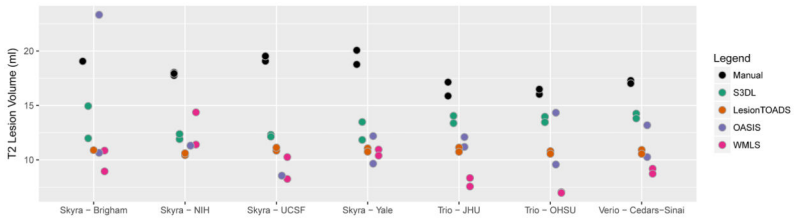

FIG 3.

Comparison of manual and automated methods for measuring lesional volume. Scan-rescan imaging is shown using multiple dots for each site and algorithm.

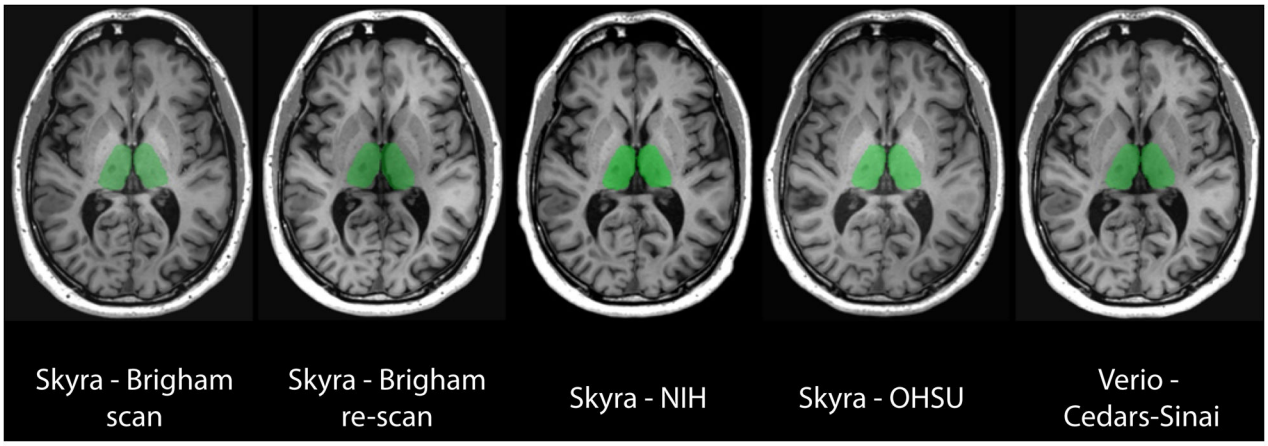

FIG 4.

FSL-FIRST Automated Segmentation Results: Thalamus. Representative anatomic slice showing segmentation of the thalamus (green) in the single subject. The segmentation maps are overlaid to the original raw 3D T1-weighted images after re-orientation to the axial plane. Segmentation was performed by the fully automated FSL/FIRST pipeline ([FMRIB (Oxford Centre for Functional MRI of the Brain) Software Library Integrated Registration and Segmentation Tool]). The scan site and 3T Siemens model are shown for each image. The first two scans are from the scan/re-scan at Brigham and Women's Hospital. OHSU= Oregon Health & Sciences University; NIH = National Institutes of Health

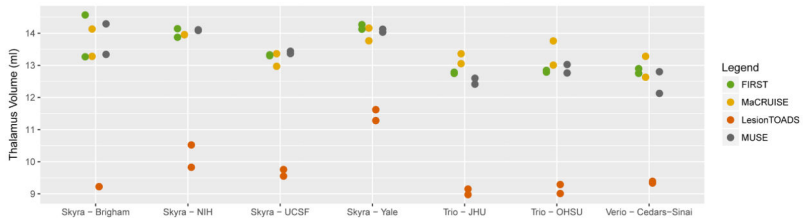

FIG 5.

Comparison of automated methods for measuring thalamic volume. Scan-rescan imaging is shown using multiple dots for each site and algorithm.

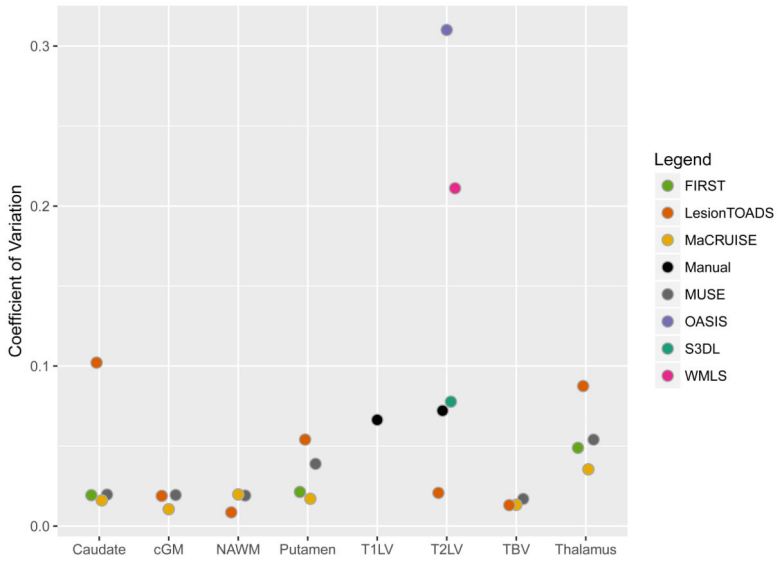

FIG 6.

Estimated across-site coefficient of variation for each structure using various methods for volumetric measurement. cGM = cortical gray matter; NAWM = normal-appearing white matter; T1LV = T1 lesion volume; T2LV = T2 lesion volume; TBV = total brain volume.

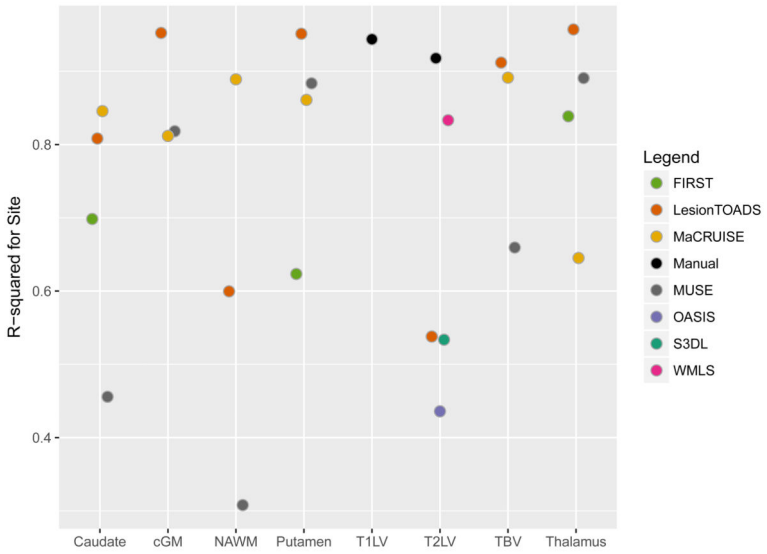

FIG 7.

Estimated proportion of variation explained by site for using various segmentation methods for different structures in the brain. cGM = cortical gray matter; NAWM = normal-appearing white matter; T1LV = T1 lesion volume; T2LV = T2 lesion volume; TBV = total brain volume.

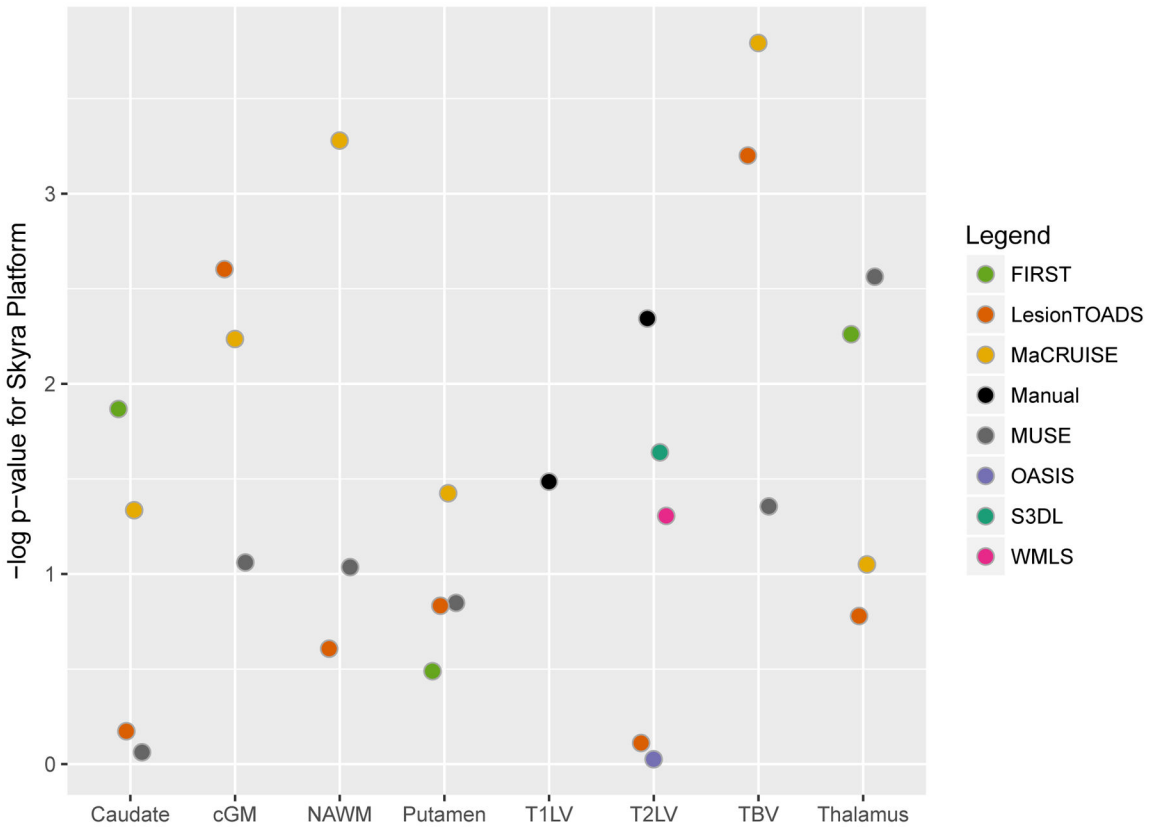

FIG 8.

Negative logarithm (base 10) p-value from t-test describing the difference in average volume between Skyra vs non-Skyra platforms explained by site for using various segmentation methods for different structures in the brain. cGM = cortical gray matter; NAWM = normal-appearing white matter; T1LV = T1 lesion volume; T2LV = T2 lesion volume; TBV = total brain volume.

Table 1

3T brain MRI anatomical acquisition protocols

Each of the 7 sites used one of 3 different Siemens scanner models (Skyra, Verio and TrioTim), necessitating 3 model-specific protocols for each of the 3 pulse sequences.

	3D T2 FLAIR			3D T1 MPRAGE			2D PD T2		
Scanner manufacturer	**Siemens Skyra**	**Siemens Verio**	**Siemens TrioTim**	**Siemens Skyra**	**Siemens Verio**	**Siemens TrioTim**	**Siemens Skyra**	**Siemens Verio**	**Siemens TrioTim**
Operation system version	Syngo MR D13	Syngo MR B17	Syngo MR B17	Syngo MR D13	Syngo MR B17	Syngo MR B17	Syngo MR D13	Syngo MR B17	Syngo MR B17
Coil	20, 32, or 64 Channel*	32 Channel	32 Channel	20, 32, or 64 Channel*	32 Channel	32 Channel	20, 32, or 64 Channel*	32 Channel	32 Channel
Acceleration factor for parallel imaging	2	2	2	2	2	2	3	3	3
Orientation	Sagittal	Sagittal	Sagittal	Sagittal	Sagittal	Sagittal	Axial	Axial	Axial
Field of view (cm)	25.6×25.6	25.6×25.6	25.6×25.6	25.6×25.6	25.0×25.0	25.0×25.0	21.7×24.0	21.7×24.0	21.7×24.0
Matrix size	512×512	512×512	512×512	256×256	256×256	256×256	232×256	232×256	232×256
Number of slices	176	176	176	176	176	176	108	108	108
Repetition time (msec)	4800	4800	4800	1900	1900	1900	3000	3000	3000
Echo time (msec)	353	354	355	2.52	2.52	2.52	11/101	11/101	11/101
Flip angle (degrees)	120	120	120	9	9	9	150	150	150
Voxel size (mm)	0.5×0.5×1.0	0.5×0.5×1.0	0.5×0.5×1.0	1.0×1.0×1.0	0.977×0.977×1.0	0.977×0.977×1.0	0.9375×0.9375×3.0	0.9375×0.9375×3.0	0.9375×0.9375×3.0
Scan time (mins)	6:53	7:00	7:00	4:15	4:16	4:16	1:57	2:15	2:15
Number of signal averages	1	1	1	1	1	1	1	1	1

Key: FLAIR = fluid-attenuated inversion recovery, MPRAGE = magnetization-prepared rapid acquisition gradient echo, PD = proton density.

* Brigham and Women's Hospital = 20-channel, University of California San Francisco = 64 channel (the other two Skyra sites = 32 channel).