

# Leveraging Clinical Imaging Archives for Radiomics: Reliability of Automated Methods for Brain Volume Measurement<sup>1</sup>

Viraj R. Adduru, MS  
Andrew M. Michael, PhD  
Maria Helguera, PhD  
Stefi A. Baum, PhD<sup>2</sup>  
Gregory J. Moore, MD, PhD

<sup>1</sup> From the Institute for Advanced Application (V.R.A., A.M.M., G.J.M.), Autism and Developmental Medicine Institute (A.M.M.), and Department of Radiology (G.J.M.), Geisinger Health System, 100 N Academy Ave, Danville, PA 17822; and Chester F. Carlson Center for Imaging Science, Rochester Institute of Technology, Rochester, NY (V.R.A., A.M.M., M.H., S.A.B.). Received September 8, 2016; revision requested November 21; revision received December 23; accepted January 13, 2017; final version accepted February 9. Address correspondence to A.M.M. (e-mail: [ammichael@geisinger.edu](mailto:ammichael@geisinger.edu)).

Study supported by Geisinger Health System Foundation. This study is supported, in part, by the Pennsylvania Department of Health (#SAP 4100070267). The Pennsylvania Department of Health specifically disclaims responsibility for any analyses, interpretations or conclusions.

## Current address:

<sup>2</sup> Geisinger Health System, Danville, Pa.

V.R.A. and A.M.M. contributed equally to this work.

© RSNA, 2017

## Purpose:

To validate the use of thick-section clinically acquired magnetic resonance (MR) imaging data for estimating total brain volume (TBV), gray matter (GM) volume (GMV), and white matter (WM) volume (WMV) by using three widely used automated toolboxes: SPM ([www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/)), FreeSurfer ([surfer.nmr.mgh.harvard.edu](http://surfer.nmr.mgh.harvard.edu)), and FSL (FMRIB software library; Oxford Centre for Functional MR Imaging of the Brain, Oxford, England, <https://fsl.fmrib.ox.ac.uk/fsl>).

## Materials and Methods:

MR images from a clinical archive were used and data were deidentified. The three methods were applied to estimate brain volumes from thin-section research-quality brain MR images and routine thick-section clinical MR images acquired from the same 38 patients (age range, 1–71 years; mean age, 22 years; 11 women). By using these automated methods, TBV, GMV, and WMV were estimated. Thin- versus thick-section volume comparisons were made for each method by using intraclass correlation coefficients (ICCs).

## Results:

SPM exhibited excellent ICCs (0.97, 0.85, and 0.83 for TBV, GMV, and WMV, respectively). FSL exhibited ICCs of 0.69, 0.51, and 0.60 for TBV, GMV, and WMV, respectively, but they were lower than with SPM. FreeSurfer exhibited excellent ICC of 0.63 only for TBV. Application of SPM's voxel-based morphometry on the modulated images of thin-section images and interpolated thick-section images showed fair to excellent ICCs (0.37–0.98) for the majority of brain regions (88.47% [306924 of 346916 voxels] of WM and 80.35% [377282 of 469502 voxels] of GM).

## Conclusion:

Thick-section clinical-quality MR images can be reliably used for computing quantitative brain metrics such as TBV, GMV, and WMV by using SPM.

© RSNA, 2017

Online supplemental material is available for this article.

**C**urrent radiology practice primarily relies on treating images as pictures, and interpretations are made on the basis of visual inspection of the images (1). According to Gillies et al, radiomics refers to the “extraction of quantitative features that result in the conversion of images into mineable data and the subsequent analysis of these data for decision support” (1). Brain volume estimation, referred to as brain volumetry, is often used in radiomics to estimate brain tissue volumes. Automated brain volumetry is increasingly used on structural magnetic resonance (MR) images in both research and clinical domains to help diagnose disease, track disease progression, and monitor effects of treatment (2). In the clinical setting, routine head MR images are commonly acquired with lower spatial resolution and higher section thicknesses to enable faster acquisition times. It is unclear, however, if low-resolution clinical brain MR images are limited in their use with automated segmentation tools that have traditionally been used with high-resolution images (ie, usually with a section thickness of <2 mm). Specifically, we do not know how brain

volume metrics derived from thick-section images compare with those derived from thin-section images. This is important because establishing the reliability of clinical MR imaging data for research-driven volumetric analyses will allow for the use of vast archives of previously unused clinical images.

Brain volumetry typically involves segmentation and quantification of various tissue types of the brain such as gray matter (GM), white matter (WM), and cerebrospinal fluid. Other metrics include total intracranial volume and total brain volume (TBV). Automated segmentation tools are widely applied because of the labor-intensive nature of manual segmentation. A variety of automated methods currently exist, of which SPM ([www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/)) (3), FreeSurfer ([surfer.nmr.mgh.harvard.edu](http://surfer.nmr.mgh.harvard.edu)) (4), and FSL (FMRIB software library; Oxford Centre for Functional MR Imaging of the Brain, Oxford, England, <https://fsl.fmrib.ox.ac.uk/fsl>) (5) are among the most popular. A study by Helms (6) provided an extensive review of these and other available brain segmentation methods.

To date, few studies performed brain volumetry by using thick-section images. Smith et al (5,7) used FSL on MR images of varying section thicknesses (1–6 mm) from the same patients and found that FSL estimates of TBV did not vary with section thickness. Eritaia et al (8) examined the effect of sparse sampling of image sections and showed that reliable estimates of total intracranial volume can be achieved up to a sampling density of one in 25 sections. These results were confirmed in a recent study (9). Klauschen et al (10) compared the performances of SPM version 5, FSL, and FreeSurfer in calculating GM volume (GMV), WM volume (WMV), and TBV by using thin-section images. This study found that volumetric accuracy of SPM version 5 and FSL

was better than that of FreeSurfer. A more recent study (11) showed that SPM version 12 performed better than FreeSurfer for calculation of total intracranial volume. However, the reliability of applying automated methods on clinical-quality MR images has not been well established in the literature. In this study, we aim to validate the use of thick-section clinically acquired MR images for estimating GMV, WMV, and TBV by using three widely used automated methods (SPM, FreeSurfer, and FSL).

## Materials and Methods

### Patients

This study was reviewed and approved by Geisinger Health System Foundation institutional review board. The data used in this study were not identifiable and no protected health information was collected, accessed, used, or distributed. This study was part of a larger research initiative that addressed the question of leveraging clinical imaging archives for research studies. As part of that initiative, we deidentified 2500 randomly selected patients with head MR images from

### Advances in Knowledge

- Automated tools applied to thick-section MR images reliably estimated brain volumes compared with thin-section images from the same patients with excellent intraclass correlation coefficient (ICC) (ICC, 0.83–0.97).
- Of the three automated toolboxes evaluated, the statistical parametric mapping (SPM; [www.fil.ion.ucl.ac.uk/spm/](http://www.fil.ion.ucl.ac.uk/spm/)) toolbox performed most reliably for all three brain volumes; total brain volume (ICC, 0.97), gray matter volume (ICC, 0.85), and white matter volume (ICC, 0.83).
- In voxel-based morphometry of SPM, tissue density for the majority of brain voxels (>80%) could also be reliably estimated by using thick-section MR images.

### Implication for Patient Care

- Automated methods can be used for brain volume estimations on standard-of-care clinical-quality brain MR images.

<https://doi.org/10.1148/radiol.2017161928>

Content code: 

Radiology 2017; 284:862–869

#### Abbreviations:

GM = gray matter  
GMV = GM volume  
ICC = intraclass correlation coefficient  
TBV = total brain volume  
WM = white matter  
WMV = WM volume

#### Author contributions:

Guarantors of integrity of entire study, V.R.A., A.M.M.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, V.R.A., M.H., S.A.B., G.J.M.; experimental studies, V.R.A., A.M.M.; statistical analysis, V.R.A., A.M.M., M.H., S.A.B.; and manuscript editing, all authors

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

Table 1

Imager Parameters				
Parameter	Thin-Section Images		Thick-Section Images	
	GE	Philips	GE	Philips
MR imaging pulse sequence	T1 SPGR	T1 SPGR	T1 SE	T1 SE
Acquisition plane	Coronal	Coronal	Axial	Axial
No. of patients	16	22	16	22
Section thickness (mm)	0.8–1.6	1.10	5.5–6	4.4–5.5
Voxel width (mm)	0.35–0.43	0.43–0.83	0.39–0.47	0.63–0.9

Note.—SPGR = spoiled gradient recalled, SE = spin echo.

our clinical picture archiving and communication system; all images were acquired between March and November of 2014. Of these patients, 44 had both thick- and thin-section images with complete head coverage acquired from the same imager in the same imaging session. Of the 44 patients, 38 were free of intracranial abnormalities on the basis of a neuroradiologist's clinical review (G.J.M., with 6 years of experience). The MR images of these 38 patients (mean age, 22 years; age range, 1–71 years; 11 female patients) were used as the final dataset of this study. A retrospective inspection of the deidentified reports indicated that these 38 images were acquired for clinical purposes as part of our institution's routine clinical imaging protocol for evaluating patients with seizures or reported headaches.

### Image Acquisition

Twenty-two of the patients were imaged by using an Achieva 1.5-T imager (Philips, Best, the Netherlands) and 16 patients were imaged by using a Signa HDxt 1.5-T imager (GE Medical Systems, Milwaukee, Wis). Further information on image parameters is available in Table 1.

### Brain Volume Calculations

To estimate brain volumes by using SPM, we applied the unified segmentation algorithm (3), provided as the Segment tool in SPM (version 12). GMV and WMV were calculated according to Approach 2 outlined in Malone et al (11) by using the native

space probabilistic tissue maps produced during segmentation. FreeSurfer volumes were obtained by using the recon-all-all pipeline of FreeSurfer (version 6 beta). The total GMV and cerebral WMV values found in the aseg.stats output file of FreeSurfer were used for GMV and WMV, respectively. Further information on the FreeSurfer segmentation process can be found in Fischl et al (12). In FSL (version 5.0.8) segmentation, volumes were obtained by using SIENAX (5). We used the nonnormalized GMV and WMV volumes produced by SIENAX. All images were processed by using the default parameters of the toolboxes.

### Voxel-based Morphometry

Voxel-based morphometry in SPM uses the modulated GM and WM maps produced by the unified segmentation pipeline (13). Modulated images allow users to compare regional tissue density in a standard space and absolute volume differences across patients (14). To examine if thick-section clinical images can be reliably used for voxel-based morphometry, we obtained modulated images for GM and WM for both thick- and thin-section MR images by using the modulated option for warped tissue selection in the segment tool in SPM. Modulated images produced from thick-section images had aliasing artifacts because of low resolution. To remove this artifact, we resectioned the thick-section images to 1 mm by using nearest-neighbor interpolation before running the unified segmentation pipeline.

### Statistical Methods

GMV, WMV, and TBV (GMV + WMV) were obtained for thick- and thin-section images by using SPM, FreeSurfer, and FSL. The reliability and level of agreement between thick- and thin-section image volumes were evaluated by using intraclass correlation coefficient (ICC) for all three volumes. ICC was computed by using one-way random effects model (case-1, as defined in McGraw et al [15]) with patients (ie, row effects) as random effects assuming a normal distribution. Before applying ICC, volumes estimated by the automated methods were tested for normal distribution by using Kolmogorov-Smirnov test (16). Reliability was classified on the basis of ICCs with the following scale: poor, 0–0.36; fair, 0.37–0.47; good, 0.48–0.55; and excellent, 0.56–1.0. For our sample size of 38 images, fair, good, and excellent corresponded to *P* values of less than .01, less than .001, and less than .0001, respectively. The difference in volumes between thick- and thin-section images was compared by using the percentage difference and Bland-Altman plots (17). Percentage differences between the thick- and thin-section estimates were calculated as a percentage of the thin-section estimate. Figure 1 outlines the intermethod comparison methodologic analysis.

In addition to the reliability tests and after establishing the most reliable method, analyses were performed by using estimates from the most reliable method to study the effect of age and imager heterogeneity. The effect of age on GM and WM tissue intensity contrast in structural MR images was previously demonstrated by several studies (18,19). To evaluate the effect of age on the reliability of thick-section brain volume estimates, the following experiment was performed: ICCs between thick- and thin-section volume estimates were iteratively calculated starting with the 10 youngest patients (age, 1–4 years) and then sequentially adding the next-older patient to the group.

The structure and signal intensities of infant MR images are different from

those of adults. To verify whether the infants influenced the reliability, ICCs were calculated between thick- and

thin-section volumes for the cohort after excluding the infants (age, <2 years;  $n = 5$ ).

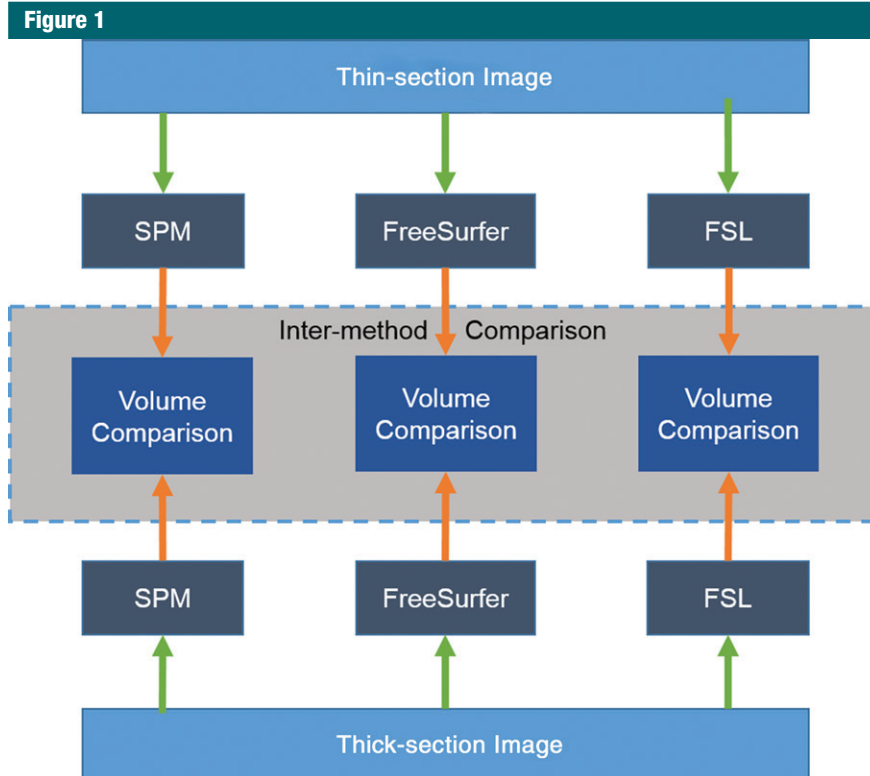
MR image signal intensity range, noise, and tissue contrast are different for images acquired by using different imagers. The effect of imager heterogeneity on reliability is verified by computing ICCs for each imager separately (Achieva,  $n = 22$ ; Signa,  $n = 16$ ).

For voxel-based morphometry analysis, voxel-by-voxel ICC was performed between the voxels of the modulated images obtained from thick- and thin-section images for GM and WM maps. The voxel-by-voxel analysis created a stereotaxic map of voxels where the tissue concentrations were reliably reproduced. As in previous studies (20,21), voxels with 10% or greater probability of belonging to GM or WM were selected for analysis. The tissue probability was provided by the tissue probability map distributed with the SPM package. The tissue probability map is defined in the Montreal Neurologic Institute (22) space. All statistical analyses were performed by using statistical software (Matlab version 8.6.0; Mathworks, Natick, Mass).

## Results

### Brain Volumetry

Among the 38 patients included in the study, FreeSurfer processing failed for three patients. FreeSurfer failed to automatically register images of two



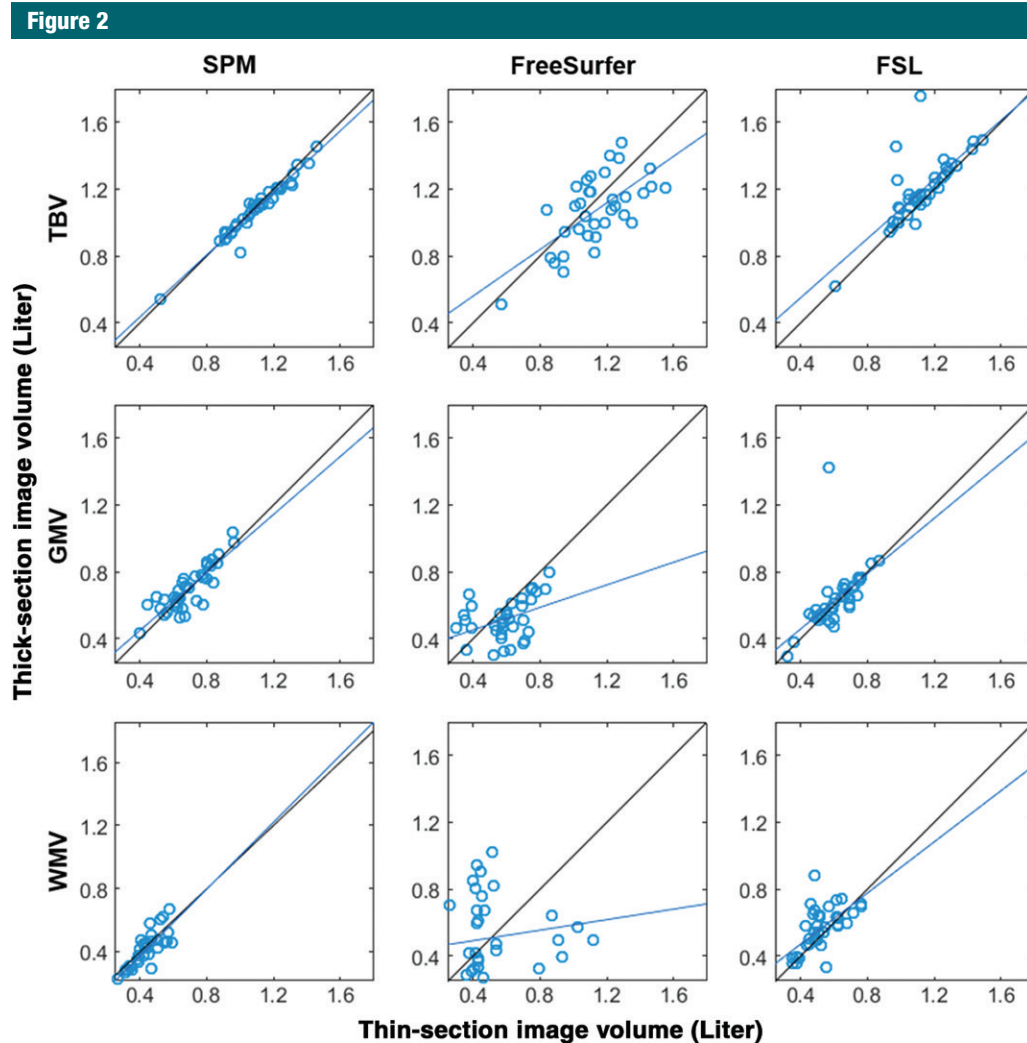
**Figure 1:** Flowchart of the intermethod brain volume comparison between thick- and thin-section MR images. Green arrows represent the raw image input to three different automated volume estimation methods. Orange arrows represent estimated brain volumes. The volume comparison box represents performance of statistical analyses to compare thick- and thin-section image volumes. The intermethod comparison box (gray box) represents the comparison of performance between the three methods.

**Table 2**

### Summary of Statistical Analysis on Thick- and Thin-Section Volume Estimations

Parameter	Thin Section (L)	Thick Section (L)	Percentage Difference (%)	ICC	P Value
<b>TBV</b>					
SPM	1.10 ± 0.18	1.08 ± 0.17	−1.19 ± 3.95	0.97	<.0001
FreeSurfer	1.13 ± 0.20	1.07 ± 0.21	−4.87 ± 14.80	0.63	<.0001
FSL	1.13 ± 0.17	1.19 ± 0.20	6.25 ± 12.52	0.69	<.0001
<b>GMV</b>					
SPM	0.69 ± 0.13	0.70 ± 0.13	2.29 ± 11.88	0.85	<.0001
FreeSurfer	0.59 ± 0.15	0.52 ± 0.12	−7.92 ± 30.98	0.30	.0349
FSL	0.61 ± 0.12	0.63 ± 0.18	4.43 ± 25.96	0.51	.0004
<b>WMV</b>					
SPM	0.41 ± 0.11	0.38 ± 0.13	−7.35 ± 16.29	0.83	<.0001
FreeSurfer	0.49 ± 0.22	0.51 ± 0.24	12.84 ± 59.70	0.16	.1692
FSL	0.52 ± 0.11	0.56 ± 0.13	9.53 ± 21.05	0.60	<.0001

Note.—Data are mean ± standard deviation. There were 38 patients for SPM, 35 patients for FreeSurfer, and 38 patients for FSL.



**Figure 2:** Scatterplots between thick-section (y-axis) and thin-section estimates (x-axis) of TBV, GMV, and WMV as estimated by three different automated methods. Blue lines represent the trend lines fitted to the scatter points. Black lines represent the  $y = x$  reference lines.

patients to the atlas. The processing of the third patient had to be manually terminated because the processing time exceeded 36 hours. These three patients were excluded from the FreeSurfer analysis.

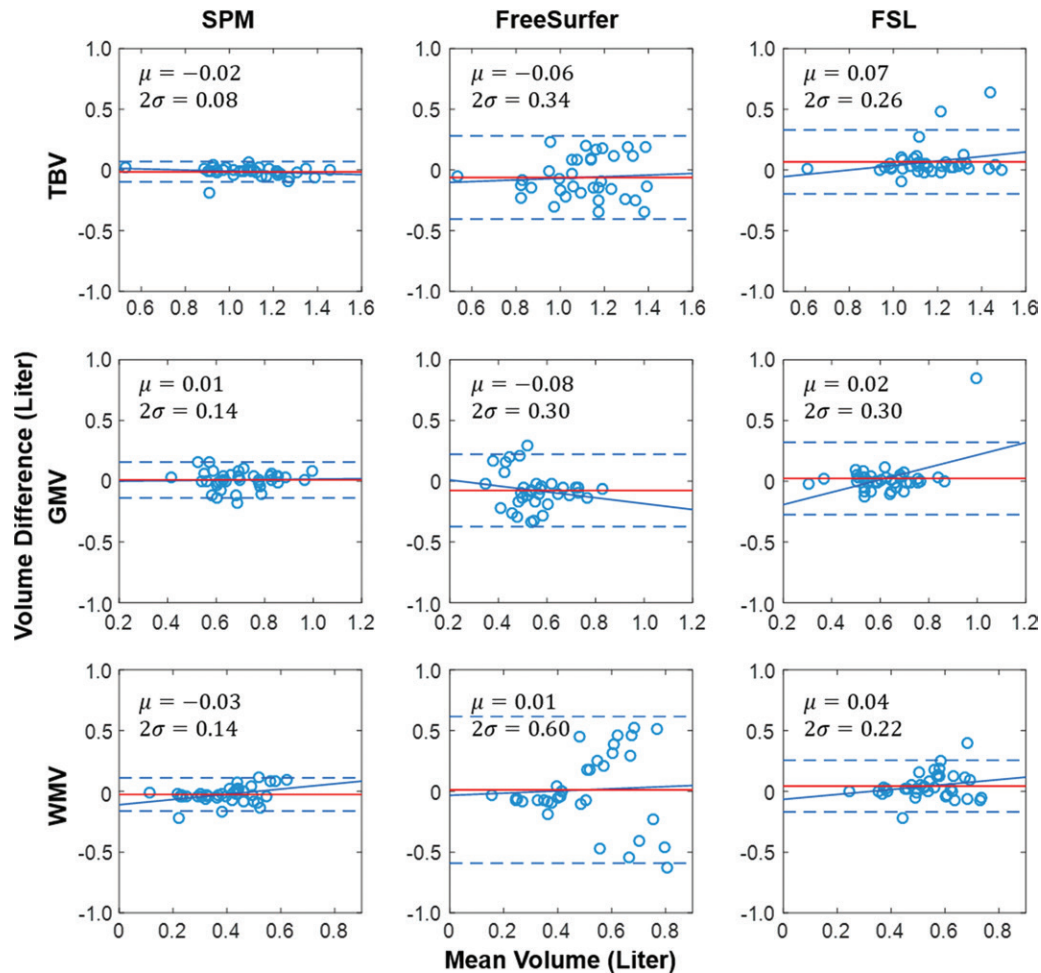
The performance of the three automated methods between thick-section and thin-section image volumes is presented in Table 2. All volumes estimated by the methods satisfied the criterion for normal distribution as determined by the Kolmogorov-Smirnov test. SPM showed excellent reliability between thick- and thin-sectioned image volumes

for TBV, GMV, and WMV (ICCs were 0.97, 0.85, and 0.83, respectively). FSL exhibited excellent reliability for TBV and WMV (ICCs were 0.69 and 0.60, respectively) and good reliability for GMV (ICC, 0.51), but ICC values were lower than those of SPM. FreeSurfer showed the lowest reliability among the methods for all the volumes and had excellent reliability only for TBV (ICC, 0.63) and poor reliability for GMV and WMV (ICCs were 0.30 and 0.16, respectively). Average GMV estimates indicated that SPM was the largest (0.70 L) and FreeSurfer was

the lowest (0.52 L). In WMV, however, SPM had the lowest estimate for WMV (0.38 L), whereas FSL exhibited the highest (0.56 L). One outlier (Fig 2) was observed in FSL, which exhibited a thick-section GMV of 1.423 L, which was more than 4 standard deviations away from the mean thick-section GMV of FSL (0.63 L). The GMVs from the same patient's thick-section image for SPM and FreeSurfer were 0.58 L and 0.43 L, respectively, which were within 1 standard deviation from their respective means. Figure 2 illustrates volumes derived from thick-section images plotted



Figure 3



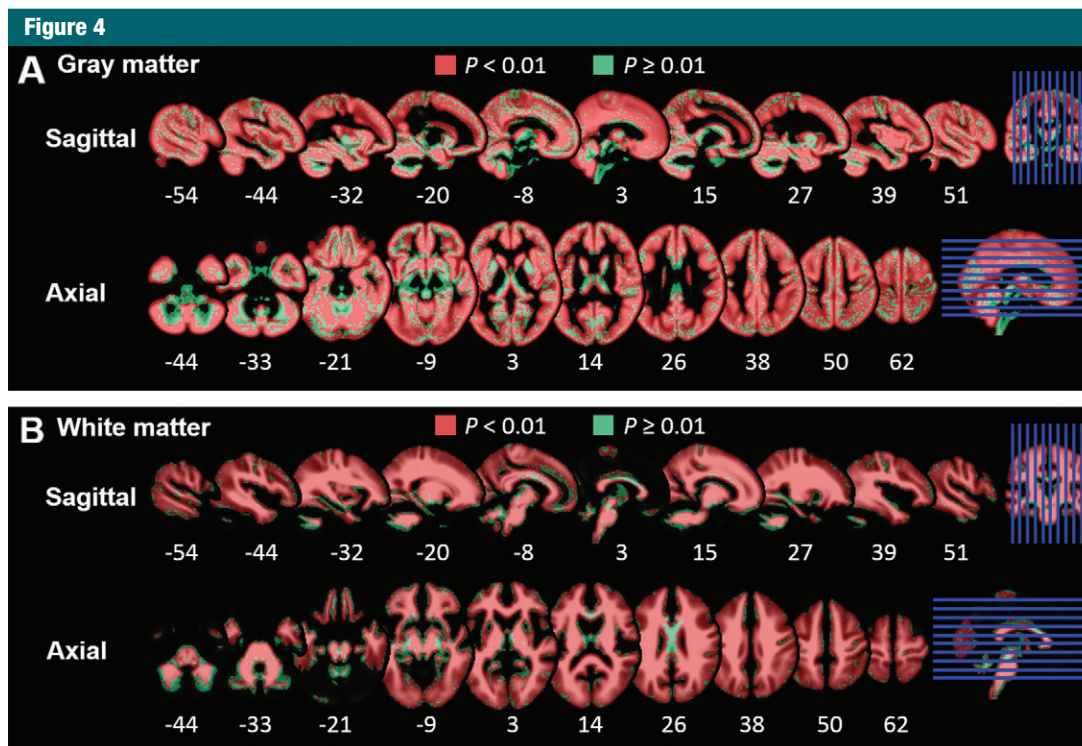
**Figure 3:** Bland-Altman plots showing thick minus thin volume difference (y-axis) plotted against the respective mean value (x-axis) of thick and thin volumes for each patient for TBV, GMV, and WMV estimated by three automated volume estimation. Solid blue line represents the trend lines. Numerical values of the mean difference (red line) and  $\pm 2$  standard deviations (dashed blue line) are also shown.

against thin-section images for TBV, GMV, and WMV for all three methods with trend lines and reference lines. Points with identical estimates from thick-section and thin-section volumes should fall at the reference line. SPM showed the best linear trend among the three methods for all the volumes followed by FSL. FreeSurfer showed large deviations from the trend line for all the volumes. We also compared the three methods after excluding from SPM and FSL the three patients for whom FreeSurfer failed, and we observed similar results (Table E1 [online]).

SPM exhibited the lowest mean and standard deviation of the percentage difference for all three volumes (Table 2). The difference between thick-section and thin-section volumes was plotted against their average in Bland-Altman plots (Figure 3). SPM showed the lowest standard deviation of the volume error (thick - thin volume) for the three brain volumes. The trend line fitted between volume difference versus mean volume of the thick- and thin-section pair showed no particular pattern for all three methods. The mean of the difference between thick- and thin-section

volumes represents the bias introduced by the methods.

The effect of age on SPM reliability is provided in Figure E1 (online). Results indicate that the reliability of TBV is stable with increasing age, but the reliability of GMV and WMV declined marginally with increasing age. However, the reliability for all three volumes remained excellent at all ages. When ICCs were calculated for SPM estimates after excluding the infants (age, <2 years;  $n = 5$ ) from the cohort, the volumes showed excellent agreement. After the removal of infants, the ICC did not change for TBV (ICC,



**Figure 4:** Sagittal and axial views of ICC between thick- and thin-section MR images in SPM voxel-based morphometry for, A, GM and, B, WM. All red regions represent fair to excellent agreement (ICC,  $>0.37$ ;  $P < .01$ ;  $n = 38$ ) and all green regions represent insignificant ICC. The section coordinates are in Montreal Neurologic Institute space.

0.97), ICC marginally improved from 0.85 to 0.86 for GMV, and ICC marginally declined from 0.83 to 0.78 for WMV.

When the effect of imager heterogeneity was assessed by using SPM estimates, the reliability was robust for each imager separately despite lower sample size. For the Philips imager ( $n = 22$ ) reliability was excellent for TBV, GMV, and WMV (ICC was 0.99, 0.90, and 0.85, respectively;  $P < .0001$ ). The reliability of the GE imager ( $n = 16$ ) was excellent for TBV (ICC, 0.90;  $P < .0001$ ), but reliability decreased but remained statistically significant for WMV (ICC, 0.74;  $P < .001$ ) and GMV (ICC, 0.57;  $P < .01$ ).

#### Voxel-based Morphometry

ICC for GM and WM tissue regions between thick- and thin-section images are presented in Figure 4. The sagittal and axial maps of the GM (Fig 4a) and WM (Fig 4b) are shown; red regions represent significantly correlated voxels ( $n = 38$ ;  $P < .01$ ; fair to

excellent reliability; ICC,  $>0.37$ ) and green regions show voxels with ICC of 0.36 or less. Of the GM and WM voxels, 80.35% (377 282 of 469 502 voxels; ICC, 0.37–0.97) and 88.47% (306 924 of 346 916 voxels; ICC, 0.37–0.98), respectively, show agreement between thick- and thin-section images. GM voxels above the spinal cord and GM–WM boundary in the cerebellum and near the ventricles showed poor reliability. In WM, mismatches were observed in voxels near the ventricles and GM–WM boundary in the cerebellum.

#### Discussion

In this study we compared brain volume estimates from thick-section ( $>3$  mm) MR images to those obtained from thin-section ( $<2$  mm) MR images from routine clinical imaging. We evaluated the reliability of three separate automated toolboxes (SPM, FSL, and FreeSurfer) in estimating TBV, GMV, and WMV by using thick-section brain images. The

best method is characterized by excellent reliability and lowest bias. With this criterion, SPM was the superior overall performer for all three volumes examined. FSL exhibited good to excellent agreement, albeit much lower than SPM. However, FreeSurfer exhibited excellent agreement for TBV only and poor agreement for GMV and WMV. In voxel-based morphometry with SPM, our results indicated fair to excellent reliability between thick- and thin-section images for GM and WM maps for the majority of brain voxels. Our findings suggest that thick-section MR images can be reliably used for estimating TBV and brain tissue volumes by using SPM, a widely applied automated method. Additionally, clinical quality images may be used for voxel-based morphometry in SPM.

Previous work has not directly attempted to compare thick- and thin-section MR image reliability for computing brain volumes, although several studies (8,9) examined the effect of sparse sampling

of MR image sections for manually computing total intracranial volume. These studies indicated that sparse sampling can provide accurate total intracranial volume estimates. Similarly, our findings indicate that thick-section images can provide reliable brain volume estimates by using automated methods. Previous studies (10,11) indicated that SPM gives the most accurate estimates for brain volumes compared with volumes calculated from manual segmentation. Our study built on these findings, and we report that SPM gives the most reliable estimates for TBV, GMV, and WMV among the three methods when thick-section images were used.

In addition, we demonstrated that the reliability of automated methods for thin-versus-thick estimation for TBV was stable with increasing age, but declined marginally for GMV and WMV. Although there was a marginal decline of ICC for GMV and WMV, all ICC values remained excellent. The decline in GMV and WMV reliability can be attributed to decreasing GM and WM tissue intensity contrast with age (18,19).

There are two important limitations to address regarding our findings. First, images were acquired from different imagers and had varying acquisition planes and acquisition sequences. In spite of this heterogeneity, consistent results emerged. Therefore, we do not expect that differences in imagers and imager parameters influenced the results. The second limitation is that our study has a small sample size and our findings need to be further validated with a large dataset.

Despite the presence of large archives of clinical-quality MR images, to our knowledge this source of valuable data has not been used to its full extent in research studies. Our study demonstrates the potential to use large existing clinical brain imaging archives for radiomics, which opens a vast new data resource to researchers and clinicians

for a variety of studies. These include clinical outcome and association studies, studies of pharmacologic efficacy, and phenome- and genome-wide association studies.

**Disclosures of Conflicts of Interest:** V.R.A. disclosed no relevant relationships. A.M.M. disclosed no relevant relationships. M.H. disclosed no relevant relationships. S.A.B. disclosed no relevant relationships. G.J.M. disclosed no relevant relationships.

## References

1. Gillies RJ, Kinahan PE, Hricak H. Radiomics: images are more than pictures, they are data. *Radiology* 2016;278(2):563–577.
2. Giorgio A, DeStefano N. Clinical use of brain volumetry. *J Magn Reson Imaging* 2013;37(1):1–14.
3. Ashburner J, Friston KJ. Unified segmentation. *Neuroimage* 2005;26(3):839–851.
4. Dale AM, Fischl B, Sereno MI. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *Neuroimage* 1999;9(2):179–194.
5. 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.
6. Helms G. Segmentation of human brain using structural MRI. *MAGMA* 2016;29(2):111–124.
7. Smith SM. Fast robust automated brain extraction. *Hum Brain Mapp* 2002;17(3):143–155.
8. Eritaja J, Wood SJ, Stuart GW, et al. An optimized method for estimating intracranial volume from magnetic resonance images. *Magn Reson Med* 2000;44(6):973–977.
9. Sargolzaei S, Sargolzaei A, Cabrerizo M, et al. A practical guideline for intracranial volume estimation in patients with Alzheimer's disease. *BMC Bioinformatics* 2015;16(Suppl 7):S8.
10. Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A, Lundervold A. Evaluation of automated brain MR image segmentation and volumetry methods. *Hum Brain Mapp* 2009;30(4):1310–1327.
11. Malone IB, Leung KK, Clegg S, et al. Accurate automatic estimation of total intracranial volume: a nuisance variable with less nuisance. *Neuroimage* 2015;104:366–372.
12. Fischl B, Salat DH, Busa E, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron* 2002;33(3):341–355.
13. Mechelli A, Price CJ, Friston KJ, Ashburner J. Voxel-based morphometry of the human brain: Methods and applications. *Curr Med Imaging Rev* 2005;1(2):105–113.
14. Good CD, Johnsrude IS, Ashburner J, Henson RN, Friston KJ, Frackowiak RS. A voxel-based morphometric study of ageing in 465 normal adult human brains. *Neuroimage* 2001;14(1 Pt 1):21–36.
15. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods* 1996;1(1):30.
16. Massey FJ. The Kolmogorov-Smirnov Test for Goodness of Fit. *J Am Stat Assoc* 1951;46(253):68–78.
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet* 1986;1(8476):307–310.
18. Kim DM, Xanthakos SA, Tupler LA, et al. MR signal intensity of gray matter/white matter contrast and intracranial fat: effects of age and sex. *Psychiatry Res* 2002;114(3):149–161.
19. Knight MJ, McCann B, Tsivos D, Couthard E, Kauppinen RA. Quantitative T1 and T2 MRI signal characteristics in the human brain: different patterns of MR contrasts in normal ageing. *MAGMA* 2016;29(6):833–842.
20. Peelle JE, Cusack R, Henson RNA. Adjusting for global effects in voxel-based morphometry: gray matter decline in normal aging. *Neuroimage* 2012;60(2):1503–1516.
21. Lorio S, Lutti A, Kherif F, et al. Disentangling in vivo the effects of iron content and atrophy on the ageing human brain. *Neuroimage* 2014;103:280–289.
22. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 2002;15(1):273–289.