frontiers in Neuroscience

ORIGINAL RESEARCH
published: 09 November 2016
doi: 10.3389/fnins.2016.00503

Reproducibility and Bias in Healthy Brain Segmentation: Comparison of Two Popular Neuroimaging Platforms

Dana L. Tudorascu[1, 2, 3], Helmet T. Karim[4], Jacob M. Maronge[5], Lea Alhilali[6], Saeed Fakhran[7], Howard J. Aizenstein[3, 4], John Muschelli[8] and Ciprian M. Crainiceanu[8]*

[1] Department of Internal Medicine, University of Pittsburgh, Pittsburgh, PA, USA, [2] Department of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA, [3] Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA, [4] Department of Biomedical Engineering, University of Pittsburgh, Pittsburgh, PA, USA, [5] Biostatistics Program, Louisiana State University Health Sciences Center, New Orleans, LA, USA, [6] Department of Neuroradiology, Barrow Neurological Institute, Phoenix, AZ, USA, [7] Department of Radiology, Banner Health and Hospital Systems, Mesa, AZ, USA, [8] Department of Biostatistics, Bloomberg School of Public Health, John Hopkins University, Baltimore, MD, USA

We evaluated and compared the performance of two popular neuroimaging processing platforms: Statistical Parametric Mapping (SPM) and FMRIB Software Library (FSL). We focused on comparing brain segmentations using Kirby21, a magnetic resonance imaging (MRI) replication study with 21 subjects and two scans per subject conducted only a few hours apart. We tested within- and between-platform segmentation reliability both at the whole brain and in 10 regions of interest (ROIs). For a range of fixed probability thresholds we found no differences between-scans within-platform, but large differences between-platforms. We have also found very large differences between- and within-platforms when probability thresholds were changed. A randomized blinded reader study indicated that: (1) SPM and FSL performed well in terms of gray matter segmentation; (2) SPM and FSL performed poorly in terms of white matter segmentation; and (3) FSL slightly outperformed SPM in terms of CSF segmentation. We also found that tissue class probability thresholds can have profound effects on segmentation results. We conclude that the reproducibility of neuroimaging studies depends on the neuroimaging software-processing platform and tissue probability thresholds. Our results suggest that probability thresholds may not be comparable across platforms and consistency of results may be improved by estimating a probability threshold correspondence function between SPM and FSL.

Keywords: MRI reproducibility, segmentation bias, healthy brain segmentation

OPEN ACCESS

Edited by:
John Ashburner,
UCL Institute of Neurology, UK

Reviewed by:
Matthew Brett,
University of Cambridge, UK
Théodore Papadopoulo,
INRIA, France

****Correspondence:***
Dana L. Tudorascu
tudorascudl@upmc.edu

Specialty section:
This article was submitted to Brain Imaging Methods, a section of the journal Frontiers in Neuroscience

Received: *14 May 2016*
Accepted: *21 October 2016*
Published: *09 November 2016*

Citation:
Tudorascu DL, Karim HT, Maronge JM, Alhilali L, Fakhran S, Aizenstein HJ, Muschelli J and Crainiceanu CM (2016) Reproducibility and Bias in Healthy Brain Segmentation: Comparison of Two Popular Neuroimaging Platforms. Front. Neurosci. 10:503. doi: 10.3389/fnins.2016.00503

INTRODUCTION

Magnetic Resonance Imaging (MRI) is widely used in clinical practice and research. While MRI acquisition techniques have standard protocols within institutions across the world, the population level analysis of MRI obtained from heterogeneous sources is still under intense methodological development. The current state-of-the-art for pre-processing MRI data is to use standard software packages and develop research-group-specific processing pipelines. In practice, the choice of processing steps and associated parameters can substantially affect brain measurements and the conclusions of the study. We focus on studying the reproducibility and bias

of brain MRI segmentation software. We consider two popular neuroimaging software platforms, Statistical Parametric Mapping (SPM, http://www.fil.ion.ucl.ac.uk/spm/software/spm12) and FMRIB Software Library (FSL, http://www.fmrib.ox.ac.uk/fsl/index.html), and compare results on the Kirby21 dataset (Landman et al., 2011). Kirby21 is a publicly available dataset containing scan-rescan imaging sessions on 21 healthy volunteers with no history of neurological disorders. Multiple imaging modalities were acquired on these volunteers including a three-dimensional, T1-weighted, gradient-echo sequence (MPRAGE), fluid attenuated inversion recovery (FLAIR), diffusion tensor imaging (DTI), resting state functional magnetic resonance imaging (fMRI), B0, and B1 field maps. For the purpose of this paper we use only the MPRAGE structural images.

Tsang et al. (2008) have investigated segmentation methods using SPM5 and FSL. They compared the performance of the methods on a phantom dataset as well as on 32 healthy volunteers and showed that SPM5 was more accurate than FSL in terms of gray matter (GM)/white matter (WM) segmentation. A similar investigation (Kazemi and Noorizadeh, 2014) was performed with newer versions of SPM and FSL using both a simulated and a real dataset. Kazemi and Noorizadeh's investigation found that SPM performed better in terms of accuracy of segmentation than FSL on both real and simulated data with varying level of noise and intensity inhomogeneity. In addition they also investigated Brainsuite, which performed worse in terms of accuracy than both, SPM and FSL. Klauschen et al. (2009) compared the segmentation performance of SPM, FSL, and FreeSurfer (Fischl, 2012). Specifically, they found that SPM had higher sensitivity and that SPM/FSL performed similarly in calculation of volumes, however in terms of gray matter FreeSurfer, SPM, and FSL performed differently (from best to worst). The results of Kazemi and Noorizadeh and Klauschen and colleagues support the result of Tsang and colleagues that SPM performs better in terms of segmentation. Eggert et al. (2012) compared reliability and accuracy of gray matter tissue segmentation using SPM, VBM8 (http://www.neuro.uni-jena.de/vbm/; Ashburner and Friston, 2000), FSL, and FreeSurfer. In addition to differences in gray matter mean segmented volumes between segmentation algorithms, Eggert and colleagues observed that the segmentation is highly sensitive to the skull-stripping technique applied.

Using manual segmentations of the gray matter, white matter, and CSF, Mendrik et al. compared three well-known neuroimaging methods (FSL, SPM, and FreeSurfer), as well as several other custom methods (Mendrik et al., 2015). They found that SPM, FSL, and FreeSurfer performed, in this order, from best to worst. The authors proposed that FreeSurfer's poor performance might have been due to the low resolution of the structural scan used. However, in gray and white matter FreeSurfer outperformed FSL (Mendrik et al., 2015). These results seem to further support previous findings (when comparing the common neuroimaging methods).

Our approach adds to the literature in at least three novel ways. First, we compare tissue segmentation at different probability thresholds and in each subject's native space (compared to a standard neuroanatomical space). Second, we characterize the scan/rescan reproducibility using a repeated measures model that includes a factor for scan. The scan factor plays a very important role since it can be used to test whether, on average, there is a statistically significantly difference between scan and rescan regardless of the segmentation method. This work on scan/rescan reproducibility builds upon our previous work on studying reproducibility of resting state fMRI, fractional anisotropy, and brain morphology (Shou et al., 2013). Third, we use a blinded randomized reader study to compare segmentation results of SPM and FSL. This provides valuable clinical information about the accuracy of the segmented tissues.

Proper classification of brain tissue plays a crucial role in the statistical analysis of neuroimaging data. Thus, there is an urgent need to understand and quantify the reproducibility of brain segmentation results across software platforms and studies. Our results indicate that: (1) SPM and FSL provide results that exhibit moderate to large differences indicating differences between the two software platforms; (2) there is no statistically significant scan effect; and (3) there is a statistically significant segmentation method effect: significant differences were detected between the two segmentation methods for gray matter and white matter at all probability thresholds considered and for cerebrospinal fluid at two of the three thresholds.

MATERIALS AND METHODS

Subjects

The dataset is named Kirby21 (Landman et al., 2011) and is publicly available online at https://www.nitrc.org/projects/multimodal/. Twenty-one healthy volunteers (average age 31.8, $sd = 9.5$, 10 Females), were scanned using multiple imaging techniques including MRI. Local institutional review board approval and written informed consent were obtained prior to examination. Two MRI scans were collected, taken approximately 3 h apart. The sequence parameters for the MPRAGE scans in the Kirby21 dataset (Landman et al., 2011) were as follows; "A 3D inversion recovery sequence was used (TR/TE/TI = 6.7/3.1/842 ms) with a $1.0 \times 1.0 \times 1.2$ mm^3 resolution over an FOV of $240 \times 204 \times 256$ mm acquired in the sagittal plane. The SENSE acceleration factor was 2 in the right-left direction. Multi-shot fast gradient echo (TFE factor = 240) was used with a 3 s shot interval and the turbo direction being in the slice direction (right-left). The flip angle was 8°. No fat saturation was employed. The total scan time was 5 min 56 s."

Image Segmentation

Image segmentation for both neuroimaging software tools, as well as for regions of interest, was performed for both MRI scans using identical approaches. Images were processed using two standard neuroimaging packages: Statistical Parametric Mapping version 12 (Penny et al., 2011) implemented in MatLab (MathWorks) and FMRIB Software Library v5.0 (Jenkinson et al., 2012). The FSL package was used via the statistical package R through the FSLR library, a wrapper implemented by John Muschelli (https://cran.r-project.org/web/packages/fslr/index.html). Images were segmented into gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). A detailed

description of segmentation approaches is provided below. Links with our code and generated datasets are provided in the Supplemental Material.

FSL Segmentation

For FSL, images were first bias-field corrected using the N4 algorithm (Tustison et al., 2010) to remove low frequency intensity variations. Images were then skull-stripped using FSL Brain Extraction Tool (BET) (Smith, 2002) with the default parameters. The FAST (FMRIB's Automated Segmentation Tool) algorithm in FSL was used on the N4-normalized skull-stripped images to generate a tissue probability map for three tissue classes: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). The result of the FAST algorithm (Zhang et al., 2001) is the relative probability that every voxel is GM, WM, or CSF. Segmentations were performed in the native space (subject space) and not in a standard anatomical space.

To calculate the volume for each tissue class, probability maps were threshold at three levels 0.5, 0.8, and 0.9, each generating a different binary mask. The volume for each tissue class and threshold pair was obtained by multiplying the number of voxels with an assigned probability above a given threshold with the dimension of the voxel. The volume was expressed in mL by dividing this number by 1000. In addition to the volumes computed at different thresholds, a weighted volume, presented in the Supplemental Material, was also calculated by summing the voxel-specific probabilities and multiplying this sum by the dimension of the voxel within each tissue class and dividing by 1000.

SPM Segmentation

Segmentation (via SPM12) combines bias-field correction with segmentation and registration to a standard anatomical space. Unlike FSL, SPM12 does not perform skull stripping before segmentation and uses tissue probability maps as priors for the segmentation (Ashburner and Friston, 2005). SPM segmentation provides classification into six tissue classes (GM, WM, CSF, skull, soft-tissue, and air). In addition to probability maps in the standard anatomical space, SPM provides an inverse deformation field that can be used to generate tissue probability maps in the native space. To calculate the volume for each tissue class, probability maps were threshold at three levels 0.5, 0.8, and 0.9, each generating a different binary mask. Using the same technique described for FSL, the volume for GM, WM, and CSF can be computed.

Region of Interest Selection and Extraction

To quantify differences between SPM and FSL we compared estimated volumes of regions of interest. Ten random gray matter regions of interest (ROI) were selected from the Automated Anatomical Labeling Atlas (Tzourio-Mazoyer et al., 2002), available in SPM through the WFU pickatlas toolbox (Maldjian et al., 2003). The ROIs considered were: anterior cingulate cortex, middle frontal gyrus, superior frontal gyrus, paracentral lobule, parietal inferior, parietal superior, postcentral, precentral, superior motor, temporal superior, all bilateral. **Figure 1** displays one of these ROIs (parietal inferior). All ROIs were extracted in the MNI template space. ROIs were then mapped to the native space using the registration approach described in ROI Extraction.

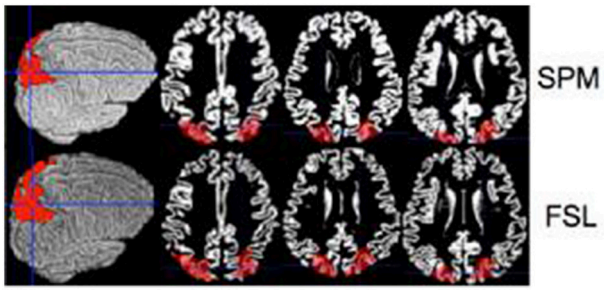

FIGURE 1 | Inferior parietal ROI is shown on a single brain (native space) for the SPM (top row) and the FSL (bottom row).

ROI Extraction

SPM ROI Extraction

Importantly, ROI's were first coregistered to native space and then the volume was calculated. During segmentation in SPM12, for each subject an inverse deformation field is generated that registers every component of the brain (i.e., ROI) in the standard space (MNI) to the native space. This deformation field was applied to all ten ROIs using a nearest-neighbor interpolation. To extract the volume of the ROIs in the native space, probability maps were threshold at the same three probability thresholds used for SPM and the extracted volume was expressed in milliliters.

FSL ROI Extraction

Similar to SPM, ROI's were first coregistered to native space and then the volumes were calculated. To generate a similar deformation field for FSL, we used FSL to register the ROIs to the native space. The FSL ROI extraction involved the following steps: (1) use an affine registration between the structural MPRAGE and the MNI template, using the function FLIRT (FMRIB's Linear Image Registration Tool) in FSL; (2) use the parameters from FLIRT to conduct a non-linear registration between the structural and the MNI template, using the function FNIRT (FMRIB's Non-linear Registration; Jenkinson and Smith, 2001; Jenkinson et al., 2002) in FSL; and (3) use the FNIRT warp file to coregister the ROI from the template space into each subject's native space, using the applywarp FSL function. Volumes were expressed in mL.

Neuroradiology Ratings

Each GM, WM, and CSF was rated independently by two neuroradiologists on a scale from 1 to 4 with 1 being poor (incorrect classification of all or a significant portion of an anatomic structure) and 4 being excellent. Results of our ratings study are described in Neuroradiology Ratings. Neuroradiologists were blinded to the image segmentation method. The neuroradiologists have never used SPM or FSL for segmentation and they confirmed that they could not tell which

image was from SPM or from FSL. They have used FSL in their previous work on Diffusion Tensor Imaging (DTI).

Statistical Methods

Descriptive statistics (means, standard deviations) were calculated for all measurements. To assess the differences in the brain volume measurements between the two methods and two scans, a two way repeated measures analysis was performed with two fixed factors (scan and method) and a random subject effect to account for within-subject correlation. The interaction between scan and method was tested but it was not found to be statistically significant in any of the models. Thus, results for the main model include only main effects. The Kenward-Rogers method (Kenward and Roger, 1997) was used for computing the number of degrees of freedom. The following statistical model was fit for each subject's brain tissue type volume (y):

$$y_{ij} = \beta_0 + \beta_1 M_{ij} + \beta_2 S_{ij} + b_{i0} + \varepsilon_{ij},$$

where, β_0 represents the intercept, M_{ij} is the segmentation method factor ($M = 0$ for FSL, $M = 1$ for SPM) for the j^{th} observation on the i^{th} subject ($i = 1, 2, ..., 21, j = 1, 2, ..., 4$ ($=n_i$) since there are $n_i = 4$ observations per subject), S_{ij} is the scan factor ($S = 0$ for scan 1, $S = 1$ for scan 2), b_{i0} is the subject-specific random effect $[b_{0i} \sim N(0, \sigma_0^2)]$, and $\varepsilon_{ij} \sim N(0, \sigma_e^2)$ is the random error term. The term β_1 represents the difference in the tissue volume between methods (SPM vs. FSL) when the scan index is fixed and β_2 represents the difference in the tissue volume between scan 1 and scan 2 when the method is fixed.

The same statistical model presented above was used for the analysis of each of the calculated tissue type volume, at each threshold as well as for the weighted sum volume and for each ROI (at all thresholds). The repeated measures analysis was performed in SAS 9.3 (SAS Institute, Cary, NC) while the descriptive analysis and plots were performed in R (R Core Team, 2013; http://www.R-project.org/). All statistical tests were two sided and test results were considered significant if the associated $p < 0.05$. No multiple comparison correction was performed.

RESULTS

Descriptive Statistics: Gray Matter, White Matter, and CSF

The average brain volume for each tissue type across each scan was almost identical within each method, but different between methods (**Table 1** in the paper, Figure 1 in the Supplemental Material). The gray matter volume computed using FSL was lower on average than the gray matter volume computed from SPM at all thresholds considered.

The white matter volume computed using FSL was lower than that computed with SPM at all thresholds, except for the 0.5 probability threshold. The CSF volume computed using FSL was lower at probability thresholds of 0.5 and 0.8 than the SPM volume, but higher at the 0.95 probability threshold. Weighted sum probability volumes descriptive statistics and parameter estimates from the repeated measures model are presented in the Supplemental Material (Tables 2, 3).

In addition, we have also computed intra-class correlation coefficients (ICC) using a one-way random effects model between scan 1 and scan 2 for each tissue type volume within each segmentation method and each threshold (Table 7 in Supplemental Material; Shrout and Fleiss, 1979; McGraw and Wong, 1996). The random effects model that was used for the ICC is not for binary data but for continuous data (e.g., GM volume), for each subject and each scan. The ICC was used to quantify the within-method correlation for each tissue type and probability threshold. A high degree of reliability (all ICC's above 0.8) was found between scan 1 and 2 for each tissue type at each threshold within each method.

Gray Matter

Parameter estimates for each method and individual parameter test results are presented in **Table 2**. There was no statistically significant effect of scan at any threshold ($p > 0.05$). There was a statistically significant effect of segmentation method for the gray matter volume at all probability thresholds ($p < 0.0001$), with higher volumes on average for the SPM segmentation.

White Matter

Parameter estimates for each method and individual parameter test results are presented in **Table 3**. There was no statistically significant effect of scan at any threshold ($p > 0.05$). There was a statistically significant effect of segmentation method for the white matter volume at all probability thresholds ($p < 0.0001$), with higher volumes on average for the SPM segmentation at 0.8 and 0.95 probability thresholds and lower at the 0.5 threshold.

Regions of Interest

Parameter estimates for all ROIs, each method, and individual parameter test results are presented in the Supplemental Material (4–6).

Regions of interest (ROI) analysis results followed the same pattern with results for the gray matter volume. A statistically significant result of scan effect was identified for only 1 out of 10 ROIs at the 0.95 probability threshold ($p = 0.046$), which could very well be due to chance. A statistically significant effect of segmentation method was detected for all ROI's at the probability thresholds 0.8 and 0.95. For the 0.5 probability threshold a statistically significant effect of segmentation method was identified for 6 out of the 10 ROI's. There was no statistically significant effect of scan for any of the ROI's at the 0.8 and 0.5 probability thresholds. The descriptive statistics for the 0.5 probability threshold are provided in **Figure 2**.

Neuroradiology Ratings

Two neuroradiologists provided good and close performance ratings for gray matter segmentation for the two methods and scans (40 out of 42 were rated excellent for both SPM and FSL by each rater) (Figure 2 in Supplemental Material). In contrast, for white matter segmentation both methods were rated poorly (Figure 3 in Supplemental Material). FSL had a higher percentage of being rated poorly on both scans (42 white matter segmentations out of 42 were rated poorly for FSL compared to 9

TABLE 1 | Descriptive statistics for threshold volumes for each tissue type.

Tissue type	Scan1 mean (*sd*)	Scan2 mean (*sd*)	Threshold	Scan1 mean (*sd*)	Scan2 mean (*sd*)
	SPM			***FSL***	
GM	983.59 (100.05)	984.14 (96.86)		829.25 (72.81)	826.21 (72.71)
WM	595.54 (62.17)	595.81 (58.37)	***Threshold 0***	575.63 (59.10)	575.77 (58.68)
CSF	769.16 (146.98)	775.11 (132.12)		462.19 (35.79)	465.09 (37.19)
GM	711.57 (66.21)	711.35 (66.42)		566.94 (53.67)	561.90 (54.64)
WM	456.28 (49.72)	454.90 (47.28)	***Threshold 0.5***	507.95 (55.74)	507.84 (56.06)
CSF	274.60 (73.02)	276.50 (74.33)		270.41 (23.73)	274.76 (27.71)
GM	626 (55.82)	624.55 (58.16)		327.50 (34.42)	323.32 (35.05)
WM	424.56 (47.42)	423.19 (45.01)	***Threshold 0.8***	384.54 (46.90)	384.19 (47.24)
CSF	190.26 (60.58)	190.65 (61.06)		174.55 (17.03)	176.77 (22.12)
GM	507.99 (45.13)	504.54 (49.10)		312.36 (32.23)	308.38 (33.12)
WM	384.89 (44.51)	383.54 (42.23)	***Threshold 0.95***	358.95 (42.93)	358.88 (43.03)
CSF	116.50 (43.70)	115.72 (43.24)		161.87 (17.17)	164.24 (22.63)

TABLE 2 | Repeated measures analysis results for gray matter by threshold.

Threshold	Effect	β (*SE*)	*t*(*df*)	*p*-value	95% CI for β
GRAY MATTER (GM)					
0	Intercept	828.36 (18.71)	44.28 (21.6)	<0.0001	(789.92, 867.19)
	Method (FSL = ref)	156.13 (5.17)	30.20 (61)	<0.0001	(145.80, 166.47)
	Scan (Scan1 = ref)	−1.24 (5.17)	−0.24 (61)	0.81	(−11.58, 9.09)
0.5	Intercept	565.73 (13.10)	43.19 (21.4)	<0.0001	(538.53, 592.93)
	Method (FSL = ref)	147.04 (3.42)	42.89 (61)	<0.0001	(140.19, 153.90)
	Scan (Scan1 = ref)	−2.62 (3.42)	−0.77 (61)	0.44	(−9.47, 4.23)
0.8	Intercept	326.82 (10.09)	32.40 (23.6)	<0.0001	(305.98, 347.66)
	Method (FSL = ref)	299.86 (4.07)	73.66 (61)	<0.0001	(291.72, 308)
	Scan (Scan1 = ref)	−2.81 (4.07)	−0.69 (61)	0.49	(−10.96, 5.32)
0.95	Intercept	312.23 (8.63)	36.20 (24.7)	<0.0001	(294.45, 330.00)
	Method (FSL = ref)	195.89 (3.89)	50.28 (61)	<0.0001	(188.1, 203.68)
	Scan (Scan1 = ref)	−3.71 (3.89)	−0.95 (61)	0.34	(−11.50, 4.07)

β coefficient for the method represents the difference in mean estimates between SPM and FSL when scan is fixed; β_0 from the model equation corresponds to intercept row, β_1 corresponds to method row, β_2 corresponds to scan row.

for SPM by rater 1; 42 out of 42 rated poorly for FSL compared to 10 out of 42 for SPM by rater 2).

Neuroradiology Case Study

Figure 3 provides an illustration of neuroradiology ratings of gray matter tissue segmentation that was rated excellent vs. one that was rated poorly. Red arrows indicate areas where the tissue was not properly classified. **Figure 4** provides a similar contrast for white matter. The original MPRAGE is displayed on the last row of **Figure 4**.

Additional examples with segmentation issues are presented and discussed in the Supplemental Material (Figures 5–10 in Supplemental Material). These case studies seem to indicate that BET could sometimes affect the follow-up segmentation algorithm and that FSL may have more problems differentiating white from gray matter in sub-cortical regions even when BET performs skull stripping well.

DISCUSSION

Artifacts and partial volume effects can affect brain tissue segmentation. In this paper, we compared SPM and FSL segmentation methods and we focused on: (1) differences between the segmentation methods; (2) reliability of the segmentations across two scans taken a few hours apart; and (3) randomized reader studies to compare the perceived quality of segmentation by clinical neuroradiologists.

TABLE 3 | Repeated measures analysis results for white matter by threshold.

Threshold	Effect	β (*SE*)	*t(df)*	*p*-value	95% CI for β
WHITE MATTER (WM)					
0	Intercept	575.60 (12.89)	44.65 (21.4)	<0.0001	(548.81, 602.38)
	Method (FSL = ref)	19.97 (3.34)	5.98 (61)	<0.0001	(13.30, 26.65)
	Scan (Scan1 = ref)	0.21 (3.33)	0.06 (61)	0.95	(−6.47, 6.88)
0.5	Intercept	508.27 (11.36)	44.71 (20.7)	<0.0001	(484.61, 531.93)
	Method (FSL = ref)	−52.31 (2.14)	−24.44 (61)	<0.0001	(−56.58, −48.03)
	Scan (Scan1 = ref)	−0.74 (2.14)	−0.35 (61)	0.73	(-5.02, 3.54)
0.8	Intercept	384.80 (10.13)	37.99 (20.8)	<0.0001	(363.73, 405.88)
	Method (FSL = ref)	39.50 (1.99)	19.77 (61)	<0.0001	(35.51,43.50)
	Scan (Scan1 = ref)	−0.87 (1.99)	−0.43 (61)	0.67	(−4.86, 3.13)
0.95	Intercept	359.27 (9.38)	38.30 (20.7)	<0.0001	(339.74, 378.79)
	Method (FSL = ref)	25.30 (1.75)	14.43 (61)	<0.0001	(21.80, 28.81)
	Scan (Scan1 = ref)	−0.71 (1.75)	−0.40 (61)	0.69	(−4.21, 2.80)

β coefficient for the method represents the difference in mean estimates between SPM and FSL when scan is fixed.; β_0 from the model equation corresponds to intercept row, β_1 corresponds to method row, β_2 corresponds to scan row.

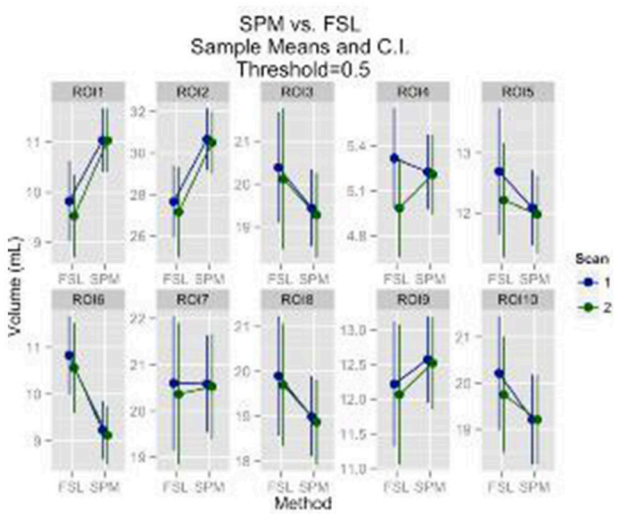

FIGURE 2 | GM ROI's mean volumes and 95% CI are plotted. The ROI's are: anterior cingulate cortex, middle frontal gyrus, superior frontal gyrus, paracentral lobule, parietal inferior, parietal superior, postcentral, precentral, superior motor, temporal superior, respectively.

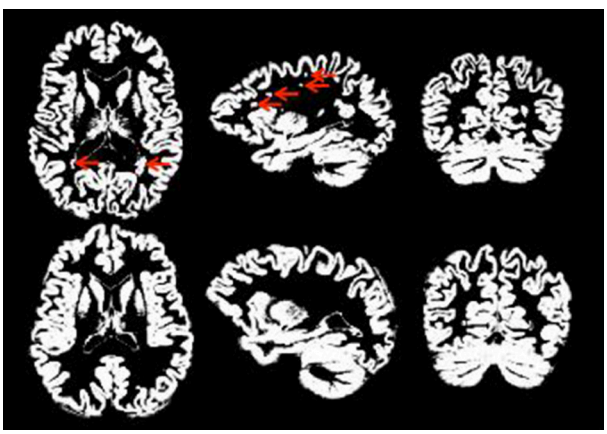

FIGURE 3 | Good (bottom row) vs. poor (top row) GM segmentation. Red arrows indicate regions that are problematic/incorrectly classified as gray matter.

We found that there were moderate to large differences between segmentation platforms and a strong within-subject, within-platform reliability. We have also found that clinical neuroradiologists agree that: SPM and FSL perform well in gray matter segmentation; and SPM and FSL both perform poorly in white matter segmentation.

MRBrainS is an imaging computational challenge started at Landman et al. (2012) dedicated to comparing the performance of segmentation of gray matter, white matter and cerebrospinal fluid on multi-sequence (T1-weighted, T1-weighted inversion recovery and FLAIR) 3 Tesla MRI scans of the brain (http://mrbrains13.isi.uu.nl/). MRBrainS compared various methods among themselves, and relative to a ranking system (Mendrik et al., 2015). Using manual segmentations as the ground truth, results indicate that SPM and FSL performed worse than other algorithms, and that SPM seems to outperform FSL in overall ranking as well as gray/white matter segmentation (Mendrik et al., 2015). Several other previous studies have also compared SPM and FSL segmentations. Tsang et al. found that SPM5's segmentation performed more accurately than FSL segmentation (Tsang et al., 2008). Kazemi and Noorizadeh (2014) reported that SPM8's segmentation performed better compared to FSL in the presence of noise. Klauschen et al. (2009) reported significant differences in gray matter volume and white matter volume between SPM and FSL. Our results complement these studies; they indicate that there are differences between the methods;

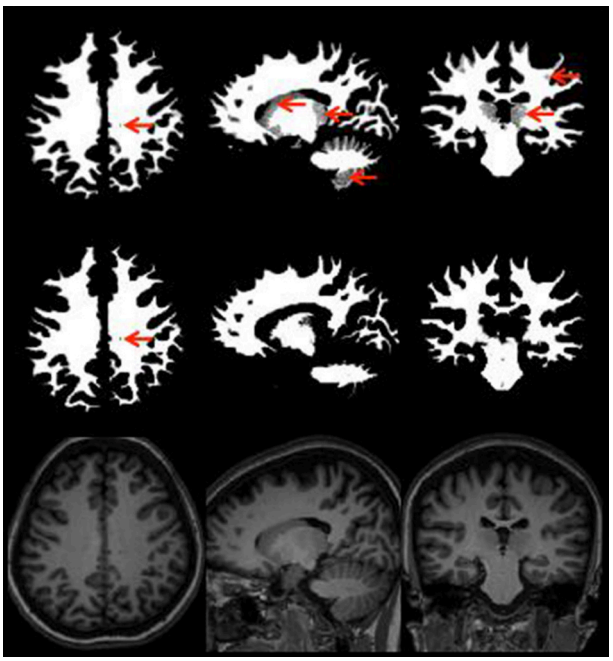

FIGURE 4 | White matter segmentation with the areas poorly rated (top row), rated as good (middle row), and MPRAGE of the same subject (bottom row) showing the same slice white matter without any potential problems. Red arrows point to areas of potential issues.

that these differences are not due to software reliability, that results can differ dramatically with the probability threshold used, and that SPM and FSL perform quite differently for different tissue classes in terms of perceived clinical accuracy. Our results add that SPM and FSL perform similarly within subject across two scans indicating that these methods are robust to between scan factors. Similar to previous studies (Mendrik et al., 2015), we found that the perceived differences between FSL and SPM (see Supplemental Material from neuroradiologists in Neuroradiology Case Study) were due to FSL's inability to accurately distinguish between deep cortical gray matter and white matter.

Results indicate that there are significant differences in gray matter, white matter, and CSF segmentations between SPM and FSL. They indicate that differences may be due to the segmentation approach, but the choice of probability thresholds may have a much larger impact on results. While sensitivity in the probability thresholds is expected, the large effect of these thresholds has been under-reported.

A study of threshold comparison between SPM and FSL may reduce observed differences between results. Investigating if such equivalent probability thresholds exist, one would need to study if the relationships are preserved across subjects, tissue classes, and regions of interest.

Results indicate that the reported effects are both global and local. Indeed, ROI volumes results mirrored whole brain results for different regions of the brain. In general, differences between scans are negligible when compared to differences between platforms and probability thresholds.

We also identified strong within-subject reliability of segmentation, though reliability is only a part of the story. Indeed, it is actually worse to produce reliably poor results than to produce unreliably good results. As there are large differences in results between platforms and probability thresholds, we conclude that either or both methods are biased.

Our study indicates that the differences found between SPM and FSL tissue volumes computed from the segmentations depend in a complex way on the various tuning parameters associated with individual segmentation steps of each algorithm. FSL relies only on image intensity to conduct segmentation, which may be more prone to inaccurately segment parts of the gray matter as white matter. This is likely due to heterogeneity across images as well as overlap between gray and white matter intensities in certain images. We have noticed that this occurs in several subcortical structures (e.g., caudate/thalamus), which seems to support the hypothesis that such substructures are more likely to exhibit white-gray matter intensity overlap. Moreover, the scatter plots of white and gray matter intensities do not indicate perfect separation in intensities. Thus, irrespective to the performance of the clustering algorithm used by FSL, it remains more difficult to segment gray matter from white matter. In contrast, SPM uses both image intensity and spatial prior information, which may be the reason for improved segmentation. For example, in situations where registration to a template is decent, SPM's spatial priors provide information about where the caudate is, which helps segmentation. This suggests that: (1) at least for healthy brains that register relatively well to the SPM template, the spatial priors contain additional information; (2) in un-healthy brains that have sizeable pathology and deformation SPM may actually induce bias and perform worse than FSL or other methods; and (3) improved spatial registration, such as multi-atlas label fusion, and population-specific templates may improve performance of segmentation algorithms. One approach to test this is to perform manual segmentations of individual ROIs, such as the caudate, and compare them to white/gray matter segmentation algorithms in these sub-structures. Another possible explanation could be that skull-stripping using BET in FSL may have an effect on segmentation. One approach to testing whether BET reliably segments the skull is to perform BET on multiple subjects with hand segmentation of the brain. As we perform manual segmentation on many images acquired in our lab, such a study could be performed on a relatively large population of older individuals. This could give insight as to how accurate/inaccurate BET is, but it could also reveal where in the brain BET is inaccurate.

Our study has several limitations. The sample size of the study was small and the archival study has multiple scans taken a few hours apart, but no ground truth segmentations were available. The randomized reader study provides additional insight into when SPM or FSL perform better and are more useful.

Kirby21 is an archival dataset that collected high-resolution structural MRI images over a short period of time (within the same day). Previous studies (Tsang et al., 2008; Klauschen et al., 2009; Kazemi and Noorizadeh, 2014) have looked at the differences between these methods, however few (Morey et al.,

2010) have investigated the reliability of a single segmentation within the same subject for scans taken only hours apart. Importantly, there seems to be a high within platform reliability: segmentations of scans that were only hours apart yielded very similar segmentations and volumes. This may indicate that probability thresholds may have a much bigger effect than previously reported. This may suggest that the interpretation of probabilities or their calculation may be different across platforms. This suggests that estimating a universal correspondence function between the SPM and FSL probability thresholds may reduce the discrepancy between results.

AUTHOR CONTRIBUTIONS

Analyzed the data: DT, HK, JMM. Contributed analysis tools: JM. Wrote the paper: DT, HK, CC. Discussed the analysis and results: DT, HK, HA, CC. Segmentation ratings: LA, SF.

ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health/NHLBI grant number R01 HL123407, and National Institute of Neurological Disorders and Stroke, grant number, R01 NS060910. We would also like to acknowledge the Statistical and Applied Mathematical Sciences Institutes (SAMSI) program (part of our work was conducted while CC was leading the Clinical Brain Imaging group at SAMSI). No conflicts of interest to report from any of the authors except HA, who has received research support from Novartis Pharmaceuticals.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://journal.frontiersin.org/article/10.3389/fnins.2016.00503/full#supplementary-material

REFERENCES

Ashburner, J., and Friston, K. J. (2000). Voxel-based morphometry—the methods. *Neuroimage* 11, 805–821. doi: 10.1006/nimg.2000.0582

Ashburner, J., and Friston, K. J. (2005). Unified segmentation. *Neuroimage* 26, 839–851. doi: 10.1016/j.neuroimage.2005.02.018

Eggert, L. D., Sommer, J., Jansen, A., Kircher, T., and Konrad, C. (2012). Accuracy and reliability of automated gray matter segmentation pathways on real and simulated structural magnetic resonance images of the human brain. *PLoS ONE* 7:e45081. doi: 10.1371/journal.pone.0045081

Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Jenkinson, M., Bannister, P., Brady, M., and Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage* 17, 825–841. doi: 10.1006/nimg.2002.1132

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., and Smith, S. M. (2012). Fsl. *Neuroimage* 62, 782–790. doi: 10.1016/j.neuroimage.2011.09.015

Jenkinson, M., and Smith, S. (2001). A global optimisation method for robust affine registration of brain images. *Med. Image Anal.* 5, 143–156.

Kazemi, K., and Noorizadeh, N. (2014). Quantitative comparison of SPM, FSL, and brainsuite for brain mr image segmentation. *J. Biomed. Phys. Eng.* 4, 13–26. doi: 10.1002/hbm.20216

Kenward, M. G., and Roger, J. H. (1997). Small sample inference for fixed effects from restricted maximum likelihood. *Biometrics* 53, 983–997.

Klauschen, F., Goldman, A., Barra, V., Meyer-Lindenberg, A., and Lundervold, A. (2009). Evaluation of automated brain MR image segmentation and volumetry methods. *Hum. Brain Mapp.* 30, 1310–1327. doi: 10.1002/hbm.20599

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A., Farrell, J. A., et al. (2011). Multi-parametric neuroimaging reproducibility: a 3-T resource study. *Neuroimage* 54, 2854–2866. doi: 10.1016/j.neuroimage.2010.11.047

Landman, B., and Warfield, S. (2012). "MICCAI 2012 workshop on multi-atlas labeling", in *Medical Image Computing and Computer Assisted Intervention Conference 2012: MICCAI 2012 Grand Challenge andWorkshop on Multi-Atlas Labeling Challenge Results* (Athens).

Maldjian, J. A., Laurienti, P. J., Kraft, R. A., and Burdette, J. H. (2003). An automated method for neuroanatomic and cytoarchitectonic atlas-based interrogation of fMRI data sets. *Neuroimage* 19, 1233–1239. doi: 10.1016/S1053-8119(03)00169-1

McGraw, K. O., and Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychol. Methods* 1, 30.

Mendrik, A. M., Vincken, K. L., Kuijf, H. J., Breeuwer, M., Bouvy, W. H., de Bresser, J., et al. (2015). MRBrainS challenge: online evaluation framework for brain image segmentation in 3T MRI scans. *Comput. Intell. Neurosci.* 2015:813696. doi: 10.1155/2015/813696

Morey, R. A., Selgrade, E. S., Wagner, H. R. II., Huettel, S. A., Wang, L., and McCarthy, G. (2010). Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Hum. Brain Mapp.* 31, 1751–1762. doi: 10.1002/hbm.20973

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., and Nichols, T. E. (2011). *Statistical Parametric Mapping: the Analysis of Functional Brain Images: The Analysis of Functional Brain Images*. Cambridge, MA: Academic Press.

R Core Team (2013). *R: A Language and Environment for Statistical Computing.*

Shou, H., Eloyan, A., Lee, S., Zipunnikov, V., Crainiceanu, A. N., Nebel, N. B., et al. (2013). Quantifying the reliability of image replication studies: the image intraclass correlation coefficient (I2C2). *Cogn. Affect. Behav. Neurosci.* 13, 714–724. doi: 10.3758/s13415-013-0196-0

Shrout, P. E., and Fleiss, J. L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychol. Bull.* 86, 420.

Smith, S. M. (2002). Fast robust automated brain extraction. *Hum. Brain Mapp.* 17, 143–155. doi: 10.1002/hbm.10062

Tsang, O., Gholipour, A., Kehtarnavaz, N., Gopinath, K., Briggs, R., and Panahi, I. (2008). Comparison of tissue segmentation algorithms in neuroimage analysis software tools. *Conf. Proc. IEEE Eng. Med. Biol. Soc.* 2008, 3924–3928. doi: 10.1109/IEMBS.2008.4650068

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: improved N3 bias correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi: 10.1109/TMI.2010.2046908

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., et al. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15, 273–289. doi: 10.1006/nimg.2001.0978

Zhang, Y., Brady, M., and Smith, S. (2001). Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans. Med. Imaging* 20, 45–57. doi: 10.1109/42.906424

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.