Reliability and validity of MRI-based automated volumetry software relative to manual measurement of subcortical structures in HIV-infected patients from a multisite prospective study

J. Dewey[1], G. Han[2], T. Russell[1], J. Price[1], D. McCaffrey[1], E. Sem[1], J. C. Anyanwu[1], C. Guttmann[1], B. Navia[3], R. Cohen[2], and D. Tate[1]

[1]Center for Neurological Imaging, Brigham and Women's Hospital, Boston, MA, United States, [2]Warren Alpert School of Medicine at Brown University, Providence, RI, United States, [3]Tufts New England Medical Center, Boston, MA, United States

Introduction: MRI-based brain volumetry is a valuable technique for investigating the pathology of neurodegenerative disease, as well as the effects of normal ageing. Though held as a "gold standard" of accuracy, volumetric measurements obtained from manual tracings are time consuming and difficult to produce. Multiple automated methods have been developed to reduce tracing time while still obtaining reliable data. Previous comparisons of competing automated methods have shown notable differences despite examining only a limited number of structures. The purpose of this study was to examine the neuroimaging output of several clinically relevant subcortical structures from a large multisite consortium study of HIV infection and compare the accuracy and consistency of volumetric results obtained using three methods: manual tracing, Freesurfer (Martinos Center, Boston, MA), and SPM (Wellcome Trust Centre for Neuroimaging, UK).

Methods: As part of the multisite NIH funded MRS HIV neuroimaging consortium study, T1-weighted SPGR MRI images from 52 HIV-infected patients were acquired using the following sequence parameters: 1.5T, TR=20, TE=5, 1X1X1.3 mm, Flip angle=0. Using these images, manual tracings of the hippocampus, amygdala, caudate, and putamen were completed by trained and reliable raters. T1 weighted images were then processed independently by both the Freesurfer (v4.0.3) and SPM (v5) software packages. The results from these automated processing pipelines were analyzed for accuracy relative to manually obtained volumes (percentage volume difference and paired t-test) as well as consistency (standard deviation of percentage change distributions and Pearson correlation coefficients relative to manual tracings).

Results: Tables 1 and 2 summarize the results of these comparisons. On average, Freesurfer output required less volume change in five of eight structures and exhibited less variable deviation from manually traced volumes in all structures. Of the calculated t-values in paired tests, all measurements from both methods were significantly different from those obtained through manual tracings, with the exception of left and right caudate volumes determined by Freesurfer. These results were confirmed by a Friedman's two-way ANOVA, which showed a significant difference between all three methods in every structure of interest except the left caudate. Pearson correlation coefficients showed significant correlation between SPM and manual volumes in six of eight structures; correlation coefficients between Freesurfer and manual results were significant in all structures and significantly larger than those of SPM in all cases (z-score range: 1.96-3.92).

Discussion: The volume measurements returned by Freesurfer were shown to be both more accurate and more consistent than those of SPM in the majority of the structures examined. However, both of these methods remain too inaccurate and variable in the majority of structures to warrant the use of a purely automated method. These results suggest researchers visually inspect automated Freesurfer and SPM volumetric output to ensure reliable data, especially when examining structures that are more difficult to visualize, such as the amygdala and hippocampus. In order to maximize the efficacy of these automated tools, future work must focus on increasing their accuracy and consistency in order to minimize the amount of manual corrections required and ultimately obviate the need for manual intervention altogether.

Table 1. Manual vs. Freesurfer

	LC	RC	LP	RP	LA	RA	LH	RH
ΔV	5.86% (±3.87%)	6.39% (±4.81%)	17.44% (±7.56%)	12.98% (±7.40%)	17.16% (±8.94%)	17.89% (±9.28%)	21.08% (±9.02%)	24.54% (±7.58%)
t-value (p)	-1.284 (0.205)	-0.845 (0.402)	-13.669 (<0.001)	-11.604 (<0.001)	-8.530 (<0.001)	-4.696 (<0.001)	-12.969 (<0.001)	-18.458 (<0.001)
r-value (p)	0.831 (<0.001)	0.822 (<0.001)	0.686 (<0.001)	0.781 (<0.001)	0.625 (<0.001)	0.617 (<0.001)	0.743 (<0.001)	0.753 (<0.001)

Table 2. Manual vs. SPM

	LC	RC	LP	RP	LA	RA	LH	RH
ΔV	24.63% (±49.28%)	12.74% (±13.39%)	37.68% (±26.29%)	16.87% (±13.40%)	25.54% (±17.62%)	17.38% (±13.31%)	20.68% (±10.46%)	14.95% (±9.70%)
t-value (p)	5.285 (<0.001)	3.581 (0.001)	8.876 (<0.001)	3.068 (0.003)	8.127 (<0.001)	2.770 (0.008)	-10.495 (<0.001)	-3.032 (0.004)
r-value (p)	0.380 (<0.001)	0.582 (<0.001)	0.155 (0.272)	0.373 (0.007)	0.325 (0.019)	0.294 (0.035)	0.384 (0.005)	0.251 (.073)

Tables 1 and 2: Average absolute volume difference (ΔV) between Freesurfer (Table 1) or SPM (Table 2) and manual measurements (expressed as the percentage of the Freesurfer or SPM volume) as well as results of a paired t-test (t-value) and Pearson correlation coefficients (r-value) listed with their associated alphas (p). LC/RC – left/right caudate; LP/RP – left/right putamen; LA/RA – left/right amygdala; LH/RH – left/right hippocampus.