JCN Open Access ORIGINAL ARTICLE

pISSN 1738-6586 / eISSN 2005-5013 / J Clin Neurol 2021;17(2):307-316 / https://doi.org/10.3988/jcn.2021.17.2.307

Evaluation of Reproducibility of Brain Volumetry between Commercial Software, Inbrain and Established Research Purpose Method, FreeSurfer

Jungbin Lee[a]
Ji Young Lee[b]
Se Won Oh[c]
Mi Sun Chung[d]
Ji Eun Park[e]
Yeonsil Moon[f]
Hong Jun Jeon[g]
Won-Jin Moon[h]

[a]Department of Radiology, Soonchunghyang University Bucheon Hospital, Bucheon, Korea
[b]Department of Radiology, Hanyang University Medical Center, Seoul, Korea
[c]Department of Radiology, Soonchunhyang University Cheonan Hospital, Cheonan, Korea
[d]Department of Radiology, Chung-Ang University Hospital, Seoul, Korea
[e]Department of Radiology, Asan Medical Center, Seoul, Korea
[f]Departments of Neurology, [g]Psychiatry, and [h]Radiology, Konkuk University Medical Center, Konkuk University School of Medicine, Seoul, Korea

Background and Purpose We aimed to determine the intermethod reproducibility between the commercial software Inbrain (MIDAS IT) and the established research-purpose method FreeSurfer, as well as the effect of MRI resolution and the pathological condition of subjects on their intermethod reproducibility.

Methods This study included 45 healthy volunteers and 85 patients with mild cognitive impairment (MCI). In 43 of the 85 patients with MCI, three-dimensional, T1-weighted MRI data were obtained at an in-plane resolution of 1.2 mm. The data of the remaining 42 patients with MCI and the healthy volunteers were obtained at an in-plane resolution of 1.0 mm. The within-subject coefficient of variation (CoV), intraclass correlation coefficient (ICC), and effect size were calculated, and means were compared using paired *t*-tests. The parameters obtained at 1.0-mm and 1.2-mm resolutions in patients with MCI were compared to evaluate the effect of the in-plane resolution on the intermethod reproducibility. The parameters obtained at a 1.0-mm in-plane resolution in patients with MCI and healthy volunteers were used to analyze the effect of subject condition on intermethod reproducibility.

Results Overall the two methods showed excellent reproducibility across all regions of the brain (CoV=0.5–3.9, ICC=0.93 to >0.99). In the subgroup of healthy volunteers, the intermethod reliability was only good in some regions (frontal, temporal, cingulate, and insular). The intermethod reproducibility was better in the 1.0-mm group than the 1.2-mm group in all regions other than the nucleus accumbens.

Conclusions Inbrain and FreeSurfer showed good-to-excellent intermethod reproducibility for volumetric measurements. Nevertheless, some noticeable differences were found based on subject condition, image resolution, and brain region.

Key Words magnetic resonance imaging; reproducibility of results; mild cognitive impairment; brain, volumetry.

INTRODUCTION

Recent advances in digital image analysis and MRI technology have allowed easy-access, quantitative volumetry software programs to be applied clinically. As a result, the application of brain volumetry can now be helpful in various diseases such as Alzheimer's disease (AD), multiple sclerosis, and epilepsy in which brain atrophy must be evaluated. Brain volume measurements facilitate early diagnosis and allow both disease progress and therapeutic responses to be monitored.[1-4] Brain volumetry can also be used in other diseases such as normal-pressure hydrocephalus.[5] However, it can take hours to interpret the images provided by volumetry software programs used in such research, with considerable computing processing power also being needed.[6] These factors restrict their usefulness in daily

Received July 27, 2020
Revised February 2, 2021
Accepted February 2, 2021

Correspondence
Won-Jin Moon, MD, PhD
Department of Radiology,
Konkuk University Medical Center,
Konkuk University School of Medicine,
120-1 Neungdong-ro, Gwangjin-gu,
Seoul 05030, Korea
Tel +82-2-2030-5544
Fax +82-2-2030-5549
E-mail mdmoonwj@kuh.ac.kr

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (https://creativecommons.org/licenses/by-nc/4.0) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Copyright © 2021 Korean Neurological Association

clinical practice.

Many of the currently available volumetry software programs were inspired by Neuroquant, which is approved by the Food and Drug Administration (FDA) of the United States and has already been applied clinically.[7-10] The commercial software Inbrain (https://www.inbrain.co.kr/) was recently also introduced to the neuroscience community. This is a deep-learning-augmented clinical volumetry software program that has been approved by the FDA of South Korea. Unlike similar clinical volumetry software that preceded it, the Inbrain platform is identical to that of the established research-purpose method FreeSurfer (https://surfer.nmr.mgh.harvard.edu/), but corrects processing errors without manual correction. Both Inbrain and FreeSurfer are based on volumetric- and surface-based segmentation. However, Inbrain, unlike other software, applies a deep-learning segmentation module in intermediate processes such as skull stripping and white-matter segmentation, which are known to require manual correction. In addition, some of the limitations of clinical software programs developed in Western countries may be overcome by incorporating data from Asian populations that were current at the time of its development. However, only a few clinical studies have used Inbrain.[11-13]

Reproducibility has been a prominent topic in recent studies involving neuroimaging, having been somewhat neglected previously.[14] However, this factor has not yet been studied using Inbrain, despite the program having been used in clinical practice and academic research. To interpret volume measurement results obtained in different studies and compare volume measurement results between individuals, it is crucial to ascertain the intermethod reproducibility between FreeSurfer and Inbrain. Hence, in the present study, we aimed to determine the intermethod reproducibility and differences between the commercial software Inbrain and the established research-purpose method FreeSurfer. In addition, we investigated the effect of MRI resolution and the pathological condition of subjects on their intermethod reproducibility.

METHODS

Study design and subjects

The present retrospective study was approved by the Institutional Review Board of Konkuk University Medical Center (IRB number: 2019-08034), and the need for informed consent was waived. As a part of a research initiative by the Korean Society of Neuroradiology to develop clinical practice guidelines, the present study used the images of 85 patients with mild cognitive impairment (MCI) who had visited the memory clinic of the Konkuk University Medical Center between September 2016 and December 2017. All subjects underwent brain MRI, including volumetric T1-weighted imaging with an in-plane resolution of either 1.0 mm or 1.2 mm. Subjects were assigned to either the 1.0-mm or 1.2-mm group, both of which were age-matched. Initially 43 patients were assigned to each group, and 1 patient was subsequently excluded from the 1.2-mm group due to focal encephalomalacia. MCI was diagnosed by either neurologists or psychiatrists based on the Petersen criteria.[15] All patients with MCI were evaluated using the Mini Mental State Examination (MMSE)[16] and the Clinical Dementia Rating (CDR).[17] For comparison, we included 45 healthy controls who had visited the healthcare center and undergone brain MRI that included volumetric T1-weighted imaging with a 1.0-mm in-plane resolution as part of routine medical checkups. The inclusion criteria for the healthy controls were as follows: age over 55 years, and no neurological or psychiatric symptoms observed during an evaluation by a family medicine physician.

A flow chart summarizing the study design is included in Supplementary Fig. 1 (in the online-only Data Supplement). In order to more clearly evaluate the effect of subject condition on reproducibility, 48 AD patients who visited the memory clinic from July 2016 to November 2017 and had undergone brain MRI including volumetric T1-weighted imaging with a 1.0-mm in-plane resolution were recruited and included in the study. The diagnosis of AD was confirmed by the above two physicians based on the criteria of the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association.[18]

Image acquisition

All MRI was performed using a 3-T MRI unit with a 32-channel head coil (DISCOVERY 750, GE Medical Systems, Milwaukee, WI, USA). We used a routine brain MRI protocol with added T1-weighted, volumetric, fast spoiled gradient-recalled echo imaging. Two kinds of resolution parameters were used in the volumetric T1-weighted imaging. Specifically, in the 1.0-mm in-plane-resolution group, source images had a slice thickness of 1.0 mm, a 256×256 matrix, and a 25.6-cm field of view, with a repetition time (TR) of 8.224 ms, echo time (TE) of 3.192 ms, flip angle (FA) of 12°, and acquisition time of 4 min 37 sec. In the 1.2-mm in-plane-resolution group, source images had a slice thickness of 1.2 mm, a 192×192 matrix, and a 24.0-cm field of view, with a TR of 5.692 ms, TE of 2.36 ms, FA of 8°, and acquisition time, of 4 min 31 sec.

Image analysis

All MRI images were reviewed by an experienced neuroradiologist to exclude any major neuropathological conditions other than neurodegeneration. The absence of imaging artifacts that could have limited the evaluation was confirmed. Each volumetric T1-weighted image was processed separately in both FreeSurfer and Inbrain. No manual correction was performed for segmentation error.

FreeSurfer (http://surfer.nmr.mgh.harvard.edu; Harvard University, Boston, MA, USA) was used to analyze the MRI volumetric data according to methods described elsewhere.[6] Inbrain (MIDAS IT, Seoul, Korea) uses both volumetric and surface-based segmentation, as well as a template-driven approach, similar to FreeSurfer's segmentation method.[19,20] The volumetric analysis of Inbrain involves the following procedures: analysis-failure prediction, intensity normalization, brain extraction, registration into the volume and surface atlas, white-matter segmentation, white-matter surface smoothing, topology correction, pial and white-matter surface optimization, and output postprocessing. A deep-learning algorithm is applied to the analysis-failure prediction, brain extraction, and white-matter segmentation.

Statistical analysis

The demographic data were compared using the independent *t*-test for continuous variables and the Mann-Whitney U test for nonnormally distributed variables. Before the volumetric data were analyzed, the FreeSurfer data were divided by 1,000 to match the units. To ensure conciseness, volume data were calculated by summing the values for the right and sides in regions such as the ventricle and cerebellum, while the regional gray-matter (GM) thickness was calculated based on the average of the values on both sides. Although the terms "reproducibility," "reliability," and "agreement" are used as umbrella terms, in this article these terms are used as defined elsewhere.[21-24] Briefly, reproducibility refers to measurements with conditions that vary between replicate measurements,[22] reliability indicates how well one subject in a certain group can be distinguished from others in that group, and agreement refers to the closeness of different measurements.[21] The intermethod reliability of brain volumetric software was assessed using intraclass correlation coefficients (ICCs).[21] We obtained two-way, mixed-model, single-measure ICCs to assess the absolute agreement between the measurements made using the two volumetric software programs. The reliability was classified based on ICC values as follows: <0.50=poor, 0.50–0.75=moderate, 0.75–0.90=good, and >0.90=excellent.[25]

We assessed the agreement between two measurements by calculating the within-subject coefficient of variation (CoV),[21] which was defined as the within-subject standard deviation divided by the mean, as proposed elsewhere (https://www-users.york.ac.uk/~mb55/meas/cv.htm). To assess differences between the methods, we also calculated Cohen's effect size D and Spearman's R^2. We used the following guidelines to interpret effect size D: <0.2=small, 0.2–0.8=moderate, and >0.8=large.[26] If a significant difference in reliability or agreement between the groups was found in the subgroup analysis, scatter plot and Bland-Altman plot analysis were additionally performed. All statistical analyses were performed using MedCalc (version 18.1.1, MedCalc Software, Mariakerke, Belgium).

RESULTS

The 85 patients with MCI were divided into 43 in the 1.0-mm group and 42 in the 1.2-mm group. Table 1 lists the demographic and clinical data of the subjects.

Overall reproducibility comparison between FreeSurfer and Inbrain

Among all subjects (*n*=130), measurements made using the two volumetric software programs showed excellent ICC values across all regions of the brain (ICC > 0.93). The calculated CoV values were close to 0% for all regions of the brain (CoV=0.5–3.9%) (Table 2). In the subgroup analyses consisting of healthy subjects, patients with MCI in the 1.0-mm group, and patients with MCI in the 1.2-mm group, the volumetric software measurements showed good reproducibility, with a good-to-excellent ICC values (0.806 to >0.999) and relatively low CoV values (0.3–5.9%) (Tables 3, 4, and 5).

Intermethod reproducibility based on subject group

When comparing measurements of the regional GM thickness, healthy controls showed lower ICC values (and hence lower reproducibility) than patients with MCI in the 1.0-mm group in all areas: frontal (0.836 vs. 0.964), parietal (0.913 vs. 0.981), temporal (0.806 vs. 0.969), occipital (0.927 vs. 0.966), cingulate (0.850 vs. 0.949), and insular (0.833 vs. 0.859) (Tables 3 and 4). Moreover, CoV values were higher in the healthy

Table 1. Demographics of the study populations

	Healthy controls (1.0 mm)	MCI patients (1.0 mm)	MCI patients (1.2 mm)	*p*
Subjects	45	43	42	
Males	23	15	16	
Age, years	62.8±5.3	71.3±7.3	72.1±6.7	
MMSE score	NA	25.9±3.1	21.6±4.2	<0.05
CDR score	NA	0.5 (0.5–0.5)	0.5 (0.5–1)	<0.05

Data are *n* or mean±standard-deviation values for continuous variables, or median (interquartile range) values for nonnormally distributed variables.

CDR: Clinical Dementia Rating, MCI: mild cognitive impairment, MMSE: Mini Mental State Examination, NA: not applicable.

controls (0.013, 0.011, 0.013, 0.011, 0.019, and 0.016, respectively) than in patients with MCI in the 1.0-mm group (0.009, 0.008, 0.010, 0.009, 0.014, and 0.018, respectively) in all GM areas other than the insular cortex. Effect size D and R^2 values showed similar trends. On scatter plots comparing the FreeSurfer and Inbrain measurements of regional GM thickness, the FreeSurfer measurements were mostly higher than those of Inbrain, and this tendency seemed to be more pronounced in the healthy controls than in patients with MCI in the 1.0-mm group (Fig. 1). Supplementary Fig. 2 (in the on-

Table 2. Reproducibility between FreeSurfer and Inbrain (MIDAS IT) in all subjects (*n*=130)

	ICC		CoV		Effect size		Correlation
	Value	95% CI	Value	95% CI	D	95% CI	R^2
GM vol.	0.985	0.687–0.996	0.016	0.016–0.017	0.144	0.112–0.172	0.993
WM vol.	0.988	0.980–0.993	0.005	0.005–0.005	0.058	0.037–0.084	0.984
Ventricle vol.	>0.999	>0.999	0.005	0.005–0.005	0.001	-0.001–0.002	>0.999
Cerebellum vol.	0.979	0.822–0.993	0.013	0.012–0.013	0.150	0.121–0.185	0.979
Frontal GM thk.	0.970	0.918–0.986	0.013	0.013–0.013	0.132	0.076–0.194	0.980
Parietal GM thk.	0.981	0.919–0.992	0.011	0.011–0.011	0.122	0.085–0.166	0.983
Temporal GM thk.	0.968	0.868–0.987	0.013	0.013–0.014	0.159	0.098–0.226	0.978
Occipital GM thk.	0.979	0.930–0.990	0.011	0.011–0.011	0.120	0.078–0.162	0.975
Cingulate gyrus thk.	0.956	0.825–0.982	0.019	0.019–0.020	0.186	0.124–0.250	0.964
Insular cortex thk.	0.934	0.888–0.959	0.016	0.015–0.017	0.140	0.071–0.223	0.907
Accumbens vol.	0.949	0.929–0.964	0.039	0.035–0.044	0.000	-0.050–0.056	0.901
Amygdala vol.	0.982	0.974–0.987	0.021	0.020–0.023	0.038	0.004–0.068	0.966
Hippocampus vol.	0.991	0.987–0.994	0.011	0.011–0.012	0.021	-0.003–0.043	0.982
Pallidum vol.	0.963	0.947–0.974	0.019	0.018–0.020	-0.058	-0.102–0.009	0.930
Putamen vol.	0.964	0.949–0.975	0.014	0.014–0.015	-0.045	-0.092–0.000	0.934
Caudate nucleus vol.	0.960	0.943–0.971	0.015	0.014–0.016	0.052	0.008–0.101	0.929
Thalamus vol.	0.963	0.947–0.973	0.017	0.016–0.018	0.003	-0.045–0.049	0.926

CI: confidence interval, CoV: within-subject coefficient of variation, GM: gray matter, ICC: intraclass correlation coefficient, thk.: thickness, vol.: volume, WM: white matter.

Table 3. Reproducibility in healthy subjects for 1.0-mm in-plane resolution (*n*=45)

	ICC		CoV		Effect size		Correlation
	Value	95% CI	Value	95% CI	D	95% CI	R^2
GM vol.	0.973	0.228–0.994	0.016	0.016–0.017	0.209	0.164–0.258	0.258
WM vol.	0.998	0.995–0.999	0.005	0.005–0.005	0.030	0.014–0.056	0.056
Ventricle vol.	>0.999	>0.999	0.005	0.004–0.005	0.001	-0.003–0.005	0.005
Cerebellum vol.	0.986	0.800–0.996	0.013	0.012–0.013	0.131	0.097–0.171	0.171
Frontal GM thk.	0.836	-0.043–0.958	0.013	0.013–0.013	0.555	0.447–0.669	0.669
Parietal GM thk.	0.913	0.095–0.977	0.011	0.011–0.011	0.367	0.275–0.461	0.461
Temporal GM thk.	0.806	-0.043–0.952	0.013	0.013–0.014	0.639	0.486–0.794	0.794
Occipital GM thk.	0.927	0.418–0.978	0.011	0.011–0.011	0.297	0.214–0.400	0.400
Cingulate gyrus thk.	0.850	0.030–0.957	0.019	0.019–0.020	0.483	0.363–0.608	0.608
Insular cortex thk.	0.833	0.418–0.934	0.016	0.015–0.017	0.394	0.227–0.575	0.575
Accumbens vol.	0.931	0.878–0.962	0.040	0.035–0.044	-0.014	-0.119–0.094	0.094
Amygdala vol.	0.975	0.956–0.986	0.021	0.019–0.022	0.036	-0.033–0.096	0.096
Hippocampus vol.	0.992	0.985–0.995	0.011	0.010–0.011	0.030	-0.003–0.072	0.072
Pallidum vol.	0.975	0.954–0.986	0.019	0.018–0.020	-0.033	-0.114–0.027	0.027
Putamen vol.	0.978	0.960–0.988	0.016	0.015–0.017	0.020	-0.040–0.088	0.088
Caudate nucleus vol.	0.975	0.955–0.986	0.019	0.017–0.020	0.023	-0.049–0.083	0.083
Thalamus vol.	0.975	0.954–0.986	0.017	0.017–0.018	-0.029	-0.112–0.028	0.028

CI: confidence interval, CoV: within-subject coefficient of variation, GM: gray matter, ICC: intraclass correlation coefficient, thk.: thickness, vol.: volume, WM: white matter.

line-only Data Supplement) shows Bland-Altman plots of the intermethod reproducibility. In other regions, there were no marked differences in reproducibility parameters, although those obtained in the 1.2-mm group were slightly higher than those in the 1.0-mm group. The reproducibility parameters obtained in AD patients showed poor reproducibility in all parameters compared to the MCI patients for 1.0-mm imaging. However, compared with healthy controls, regional GM showed a relatively high ICC and low effect size D, with this trend similar to that in MCI patients (Supplementary

Table 4. Reproducibility in patients with mild cognitive impairment for 1.0-mm in-plane resolution (*n*=43)

	ICC		CoV		Effect size		Correlation
	Value	95% CI	Value	95% CI	D	95% CI	R^2
GM vol.	0.985	0.343–0.997	0.011	0.011–0.011	0.157	0.110–0.210	0.996
WM vol.	0.995	0.989–0.998	0.007	0.007–0.007	0.041	0.015–0.074	0.992
Ventricle vol.	>0.999	>0.999	0.005	0.005–0.005	0.002	-0.001–0.005	>0.999
Cerebellum vol.	0.968	0.722–0.990	0.016	0.015–0.017	0.188	0.129–0.264	0.970
Frontal GM thk.	0.964	0.604–0.990	0.009	0.008–0.009	0.208	0.147–0.277	0.975
Parietal GM thk.	0.981	0.934–0.992	0.008	0.007–0.008	0.112	0.059–0.177	0.976
Temporal GM thk.	0.969	0.306–0.992	0.010	0.010–0.010	0.219	0.143–0.297	0.985
Occipital GM thk.	0.966	0.754–0.989	0.009	0.009–0.009	0.187	0.110–0.271	0.965
Cingulate gyrus thk.	0.949	0.536–0.985	0.014	0.014–0.014	0.247	0.170–0.335	0.959
Insular cortex thk.	0.859	0.676–0.932	0.018	0.017–0.019	0.269	0.119–0.453	0.795
Accumbens vol.	0.953	0.915–0.974	0.050	0.044–0.055	-0.045	-0.143–0.051	0.912
Amygdala vol.	0.981	0.965–0.990	0.027	0.024–0.029	0.057	-0.005–0.114	0.970
Hippocampus vol.	0.988	0.978–0.993	0.014	0.014–0.015	-0.004	-0.057–0.040	0.976
Pallidum vol.	0.917	0.818–0.959	0.029	0.027–0.031	-0.185	-0.301–-0.068	0.868
Putamen vol.	0.928	0.860–0.962	0.025	0.023–0.027	-0.138	-0.224–-0.040	0.880
Caudate nucleus vol.	0.968	0.942–0.983	0.027	0.025–0.030	0.055	-0.015–0.127	0.944
Thalamus vol.	0.950	0.911–0.973	0.021	0.019–0.023	0.056	-0.055–0.143	0.905

CI: confidence interval, CoV: within-subject coefficient of variation, GM: gray matter, ICC: intraclass correlation coefficient, thk.: thickness, vol.: volume, WM: white matter.

Table 5. Reproducibility in patients with mild cognitive impairment for 1.2-mm in-plane resolution (*n*=42)

	ICC		CoV		Effect size		Correlation
	Value	95% CI	Value	95% CI	D	95% CI	R^2
GM vol.	0.989	0.929–0.996	0.011	0.011–0.011	0.101	0.055–0.149	0.991
WM vol.	0.977	0.941–0.989	0.021	0.019–0.021	0.105	0.053–0.170	0.973
Ventricle vol.	>0.999	>0.999	0.003	0.003–0.003	-0.001	-0.005–0.002	>0.999
Cerebellum vol.	0.976	0.855–0.992	0.016	0.015–0.017	0.149	0.082–0.207	0.973
Frontal GM thk.	0.984	0.971–0.992	0.013	0.012–0.014	-0.018	-0.064–0.038	0.990
Parietal GM thk.	0.986	0.970–0.993	0.010	0.010–0.011	0.067	0.012–0.126	0.984
Temporal GM thk.	0.980	0.963–0.989	0.017	0.016–0.018	0.030	-0.042–0.098	0.981
Occipital GM thk.	0.977	0.957–0.988	0.012	0.012–0.013	0.059	-0.009–0.139	0.965
Cingulate gyrus thk.	0.976	0.957–0.987	0.016	0.015–0.017	0.042	-0.030–0.120	0.977
Insular cortex thk.	0.947	0.903–0.971	0.023	0.021–0.024	0.016	-0.076–0.132	0.919
Accumbens vol.	0.821	0.692–0.899	0.059	0.049–0.068	0.104	-0.070–0.289	0.676
Amygdala vol.	0.975	0.955–0.987	0.022	0.020–0.023	0.027	-0.045–0.094	0.951
Hippocampus vol.	0.980	0.963–0.990	0.015	0.014–0.016	0.054	-0.022–0.107	0.963
Pallidum vol.	0.966	0.938–0.982	0.028	0.026–0.030	-0.011	-0.098–0.071	0.932
Putamen vol.	0.958	0.924–0.977	0.031	0.028–0.035	-0.047	-0.144–0.044	0.934
Caudate nucleus vol.	0.930	0.874–0.962	0.037	0.031–0.043	0.078	-0.014–0.204	0.880
Thalamus vol.	0.935	0.883–0.965	0.020	0.019–0.022	-0.011	-0.133–0.098	0.874

CI: confidence interval, CoV: within-subject coefficient of variation, GM: gray matter, ICC: intraclass correlation coefficient, thk.: thickness, vol.: volume, WM: white matter.

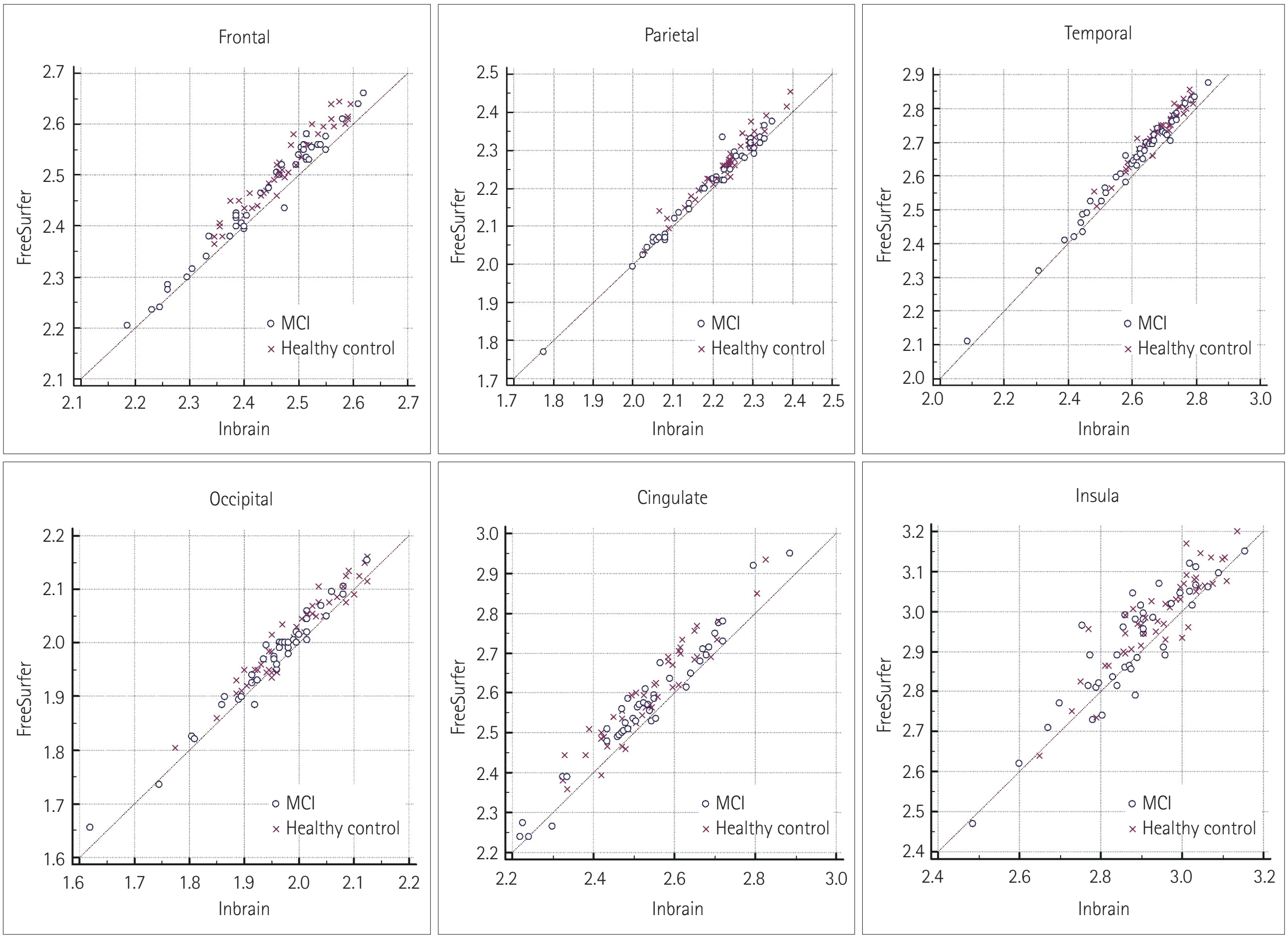

Fig. 1. Evaluation of intermethod reproducibility according to subject condition (healthy controls vs. patients with MCI); scatter plots of regional gray-matter thickness. The open dots represent the measurements from patients with MCI for 1.0-mm in-plane resolution, while the crosses indicate the measurements from healthy controls. On the scatter plots, FreeSurfer measurements (y axis) were mostly larger than those from Inbrain (MIDAS IT) (x axis). This tendency seemed more pronounced in the healthy controls. MCI: mild cognitive impairment.

Table 1 in the online-only Data Supplement).

Intermethod reproducibility based on image resolution

The intermethod reproducibility showed different patterns in different brain regions. The nucleus accumbens, which is the smallest deep GM structure, showed lower reproducibility in the 1.2-mm group (*n*=42) than in the 1.0-mm group (*n*=43), with lower ICC (0.821 vs. 0.953, respectively) and higher CoV (0.059 vs. 0.050). The D and R^2 values were similar (Tables 4 and 5). Supplementary Fig. 3 (in the online-only Data Supplement) shows scatter plots comparing the nucleus accumbens measurements of FreeSurfer and Inbrain, while Supplementary Fig. 4 (in the online-only Data Supplement) shows the corresponding Bland-Altman plots. The reproducibility parameters of regional GM thickness (other than CoV) were slightly better in the 1.2-mm group than in the 1.0-mm group.

DISCUSSION

The present study evaluated the intermethod reproducibility between FreeSurfer, which is one of the most popular freely available brain segmentation tools,[6] and Inbrain, which is a deep-learning-based volumetric software program that has been approved by the Korean FDA. We also investigated factors affecting the intermethod reproducibility. The measured reproducibility parameters showed relatively high correspondence between these two methods, but they appeared to be affected by the pathological condition of the subject, image resolution, and brain region.

Several MRI volumetric software programs have recently been developed, and numerous clinical validation studies have been conducted in various settings.[27-32] However, correspondingly few studies have evaluated the reproducibility of the results obtained using these software programs, and few have evaluated the effect of the subject's pathological condition on intermethod reproducibility.[26] One previous study[27] that compared reproducibility between FreeSurfer and Neuroquant by selecting 20 subjects with AD and 20 normal controls from the Alzheimer's Disease Neuroimaging Initiative (ADNI) data set found that the ICC values for each anatomic location ranged from 0.29 to 0.95. Another study[28] used a similar method to compare reproducibility between the same two software programs by extracting 20 AD and 20 NCs from the ADNI data set and adding 20 patients with mild traumatic brain injury, and found ICC values ranging from 0.13 to 0.98. Both our and previous studies measured the reproducibility of volume measurements in similar anatomic locations such as GM, white matter, and deep GM. However, in previous studies the reproducibility was measured by subdividing the ventricular spaces including into the lateral third and fourth ventricles. Our study found differences in measurements of the regional GM thickness rather than by subdividing the ventricles. It was particularly interesting that the anatomic location with the lowest ICC value in the above two previous studies was the pallidum. However, in our study, the pallidum showed good reproducibility, which supports the assumption that the low reproducibility is due to a different atlas being used by each software program.[33]

To estimate the effect of intermethod difference on a real-world application, we also measured the intermethod CoV values, which ranged from 0.5% to 6.8%. The annual whole-brain atrophy rate is reportedly 0.2–0.7% in healthy subjects but 1–4% in patients with AD.[34] Hippocampal atrophy has been reported in about 4% of patients with AD and 1% of healthy controls.[35] Considering the intermethod CoV values found in the present study, it is likely that the volumetric measurements made using the different software programs deviated markedly from the true value over time because they either missed or exaggerated a small atrophy rate over an interval of 1–2 years. Particularly careful interpretation is needed in the case of the nucleus accumbens, which showed the highest intermethod CoV (4–6.8%). Specifically, because FreeSurfer and Inbrain are technically similar, this result emphasizes that longitudinal studies of volumetric software must be interpreted carefully if they involve a major software upgrade with machine-learning adjustments during the study duration.

The present study found that the intermethod reproducibility of the measured regional GM thickness was more robust in patients with MCI than in healthy subjects. This might be due to a larger brain volume being associated with smaller cerebrospinal fluid (CSF) spaces, which shows clear T1 signal differences between the brain and skull. It follows that FreeSurfer would record more-severe misregistrations than Inbrain, which corrects for skull stripping using machine learning (Fig. 2). The present results support this conjecture, because the regional cortical thicknesses of healthy subjects measured using FreeSurfer were consistently larger than those measured using Inbrain across all brain regions. We also found that the intermethod reproducibility of the regional cortical thickness was better in the 1.2-mm group than in the 1.0-mm group of patients with MCI. Considering the MMSE and CDR scores in the 1.2-mm group, this may represent a difference in brain atrophy between the two groups. Alternatively, the segmentation of thicker image slices might not have been affected by the deep-learning-based optimization of Inbrain. Indeed, thick slices are associated with better signal-to-noise and contrast-to-noise ratios, which are crucial factors for accurate tissue segmentation.[36]

In contrast to volumetry of the regional cortical GM, that of the nucleus accumbens (a small deep GM structure) showed better intermethod reliability in the 1.0-mm group than in the 1.2-mm group. It is unclear why the in-plane resolution had different effects on the volumetric results depending on the GM region. It may be that the lower image resolution exagger-

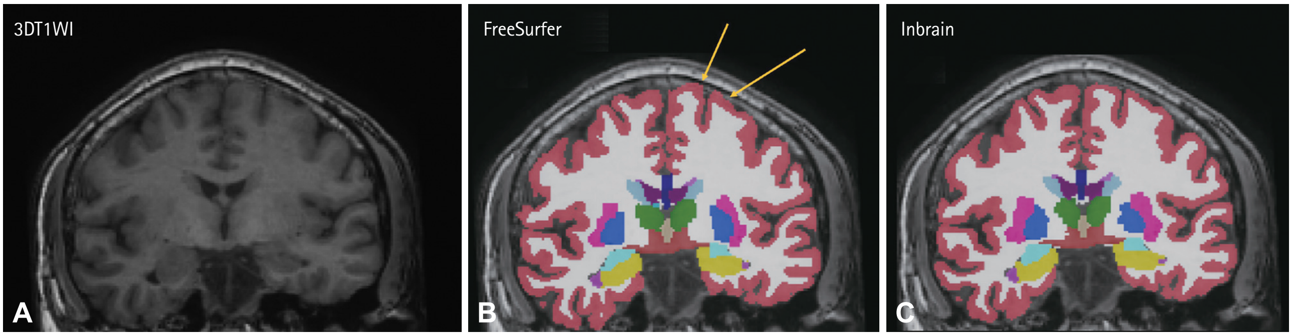

Fig. 2. Representative images of GM thickness measurement differences. Images from a healthy 63-year-old female who visited the healthcare center. A: On three-dimensional T1-weighted imaging, the cerebrospinal fluid space was small between the cortex and dura mater. B: In the segmentation output from FreeSurfer, a misregistration of the dura mater was noted around the cortex. C: In contrast, Inbrain (MIDAS IT) segmented the dura and cortex relatively well. The mean frontal GM thicknesses measured by FreeSurfer and Inbrain were 2.58 mm and 2.49 mm, respectively. GM: gray matter.

ated the difference in segmentation performance between Inbrain and FreeSurfer, despite the compensatory high signal-to-noise ratio. Furthermore, the nucleus accumbens itself is not a clearly definable structure. One previous study using MRI with 1.0-mm imaging and a 256×256 matrix found that T1 hypointensity—which distinguishes the nucleus accumbens from the adjacent striatum (caudate and putamen)—is not consistently evident on MRI images.[37]

The present study had several limitations. Firstly, we only tested the intermethod reproducibility on a single machine in one institution between two software programs, whereas MRI volumetric findings differ between scanners and centers in postmortem and phantom brains.[38] However, the recognition performance of Inbrain is worse than that of other volumetry software. Therefore, validation through head-to-head comparisons with multiple parameters is required before comparing the various software programs. For this reason, the present results should be confirmed in future multicenter studies using various MRI machines and volumetry software programs. Secondly, MMSE and CDR scores differed between the 1.0-mm and 1.2-mm imaging groups because the data were collected in a clinical setting and because we presumed that the results would show similar effects (worse intermethod reproducibility for smaller CSF spaces) on regional GM thickness measurements in the comparison between healthy controls and the MCI patients for 1.0-mm imaging (Supplementary Fig. 5 in the online-only Data Supplement). To address this issue it is necessary to perform back-to-back analysis by obtaining images from one patient with different voxel sizes in a single imaging session. However, this method was not implemented because it could be criticized as wasting medical resources due to the relatively long scan time of three-dimensional T1-weighted MRI. This meant that the effect of image resolution on regional GM thickness was not fully assessed in the present study, and so should be evaluated further in future well-controlled studies. Thirdly, this study did not include patients with diverse conditions. Ideally, diverse various pathological conditions such as epilepsy or AD should be used to evaluate the effect of subject condition on reproducibility between volumetry software. However, the inclusion of diverse disease entities may overly restrict the number of participants due to the heterogeneity of these entities before final diagnoses. Therefore, this study included subjects that were predicted to contain relatively small structural anomalies that could result in dropout. Although this approach it did not achieve the ideal large number of subjects for maximizing the statistical power of reproducibility evaluations, our study did include a total of 130 subjects, with more than 40 subjects in each group, which is the largest clinical study population that has been included in evaluations of the intermethod reproducibility between MRI volumetric software programs. In addition, although ideal control of variables was not achieved, the data of AD patients were added and compared with other data in the study. However, due to the limitations described above, the results of this study should be confirmed in future studies that include larger numbers of subjects with various pathological conditions. Fourthly, instead of using the source data for each of the right and left sides, this study summed the values on both sides for volume and the average value for thickness, and cortical parcellation data were not included in the reproducibility evaluations. This may have led to a dilution of regional errors in the reproducibility evaluations. However, it is also possible that the inclusion of an excessively large number of subgroups can result in overestimations due to a smaller number of outliers. Unfortunately, there is no clear consensus to the level of subgroup evaluation to conduct. We therefore added the reproducibility parameters of the each right and left sides separately obtained from all participants (Supplementary Tables 2 and 3 in the online-only Data Supplement). Most of the parameters obtained from each side were slightly worse than those of the sum of the two sides, but the difference in ICC values was <0.038, which was very small compared to the values for different subject groups and different in-plane resolutions. This meant that the right and left sides could be regarded as a single group sharing similar features. On the other hand, several atlases are available for cortical parcellation (https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation), and since reproducibility varies by location during utility use,[39] further evaluation of this was beyond the scope of the present study. Therefore, since the reproducibility evaluation was applied to relatively large anatomic structures, the results of this study should not be overinterpreted as representing evaluations of all utilities in both software problems, including cortical parcellation.

In conclusion, FreeSurfer and Inbrain exhibited relatively good intermethod reproducibility. However, because there were small measurement differences, the results from longitudinal studies involving both software programs need to be interpreted carefully. Furthermore, intermethod reproducibility appears to be affected by subject condition, image resolution, and brain region.

Supplementary Materials

The online-only Data Supplement is available with this article at https://doi.org/10.3988/jcn.2021.17.2.307.

Author Contributions

Conceptualization: Jungbin Lee, Won-Jin Moon. Data curation: Yeonsil Moon, Hong Jun Jeon. Formal analysis: Jungbin Lee, Yeonsil Moon, Hong Jun Jeon. Funding acquisition: Won-Jin Moon. Investigation: all authors.

Methodology: Jungbin Lee, Won-Jin Moon, Ji Yeong lee, Ji Eun Park. Project administration: Jungbin Lee. Resources: Yeonsil Moon, Hong Jun Jeon. Software: Jungbin Lee. Supervision: Won-jin Moon. Validation: Jungbin Lee, Won-Jin Moon. Writing—original draft: Jungbin Lee, Won-Jin Moon. Writing—review & editing: all authors.

ORCID iDs

Jungbin Lee	https://orcid.org/0000-0002-6240-8277
Ji Young Lee	https://orcid.org/0000-0003-1181-8070
Se Won Oh	https://orcid.org/0000-0003-1336-4498
Mi Sun Chung	https://orcid.org/0000-0003-1141-9555
Ji Eun Park	https://orcid.org/0000-0002-4419-4682
Yeonsil Moon	https://orcid.org/0000-0001-7770-4127
Hong Jun Jeon	https://orcid.org/0000-0002-0260-0494
Won-Jin Moon	https://orcid.org/0000-0002-8925-7376

Conflicts of Interest

The authors have no potential conflicts of interest to disclose.

Acknowledgements

This study was supported by 2017 Clinical Practice Guideline Research Fund grant from the Korean Society of Radiology and a grant from the Korea Health Technology R&D Project through the Korea Health Industry Development Institute (KHIDI), funded by the Ministry of Health & Welfare, Republic of Korea (grant number HI18C1038).

REFERENCES

1. Marciniewicz E, Podgórski P, Sąsiadek M, Bladowska J. The role of MR volumetry in brain atrophy assessment in multiple sclerosis: a review of the literature. *Adv Clin Exp Med* 2019;28:989-999.
2. Galovic M, van Dooren VQH, Postma T, Vos SB, Caciagli L, Borzì G, et al. Progressive cortical thinning in patients with focal epilepsy. *JAMA Neurol* 2019;76:1230-1239.
3. Bosco P, Redolfi A, Bocchetta M, Ferrari C, Mega A, Galluzzi S, et al. The impact of automated hippocampal volumetry on diagnostic confidence in patients with suspected Alzheimer's disease: a European Alzheimer's Disease Consortium study. *Alzheimers Dement* 2017;13:1013-1023.
4. Giorgio A, De Stefano N. Clinical use of brain volumetry. *J Magn Reson Imaging* 2013;37:1-14.
5. Miskin N, Patel H, Franceschi AM, Ades-Aron B, Le A, Damadian BE, et al. Diagnosis of normal-pressure hydrocephalus: use of traditional measures in the era of volumetric MR imaging. *Radiology* 2017;285:197-205.
6. Fischl B. FreeSurfer. *Neuroimage* 2012;62:774-781.
7. Massat MB. Artificial intelligence in radiology: hype or hope? *Appl Radiol* 2018;47:22-26.
8. Ribbens A, Billiet T, Beadnall H, Vaneckova M, Weinstock-Guttman B, Ly L, et al. Assessment of brain atrophy in multiple sclerosis patients in clinical routine: a multi-center comparison study of radiological and quantitative reports (P4.373). *Neurology* 2017;88(16 Suppl): P4.373.
9. Frisoni GB, Jack CR Jr, Bocchetta M, Bauer C, Frederiksen KS, Liu Y, et al. The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: evidence of validity. *Alzheimers Dement* 2015;11:111-125.
10. Kovacevic S, Rafii MS, Brewer JB; Alzheimer's Disease Neuroimaging Initiative. High-throughput, fully automated volumetry for prediction of MMSE and CDR decline in mild cognitive impairment. *Alzheimer Dis Assoc Disord* 2009;23:139-145.
11. Kim YT, Kim WJ, Choi JE, Bae MJ, Jang H, Lee CJ, et al. Cohort profile: Firefighter Research on the Enhancement of Safety and Health (FRESH), a prospective cohort study on Korean firefighters. *Yonsei Med J* 2020;61:103-109.
12. Jung NY, Han JC, Ong YT, Cheung CY, Chen CP, Wong TY, et al. Retinal microvasculature changes in amyloid-negative subcortical vascular cognitive impairment compared to amyloid-positive Alzheimer's disease. *J Neurol Sci* 2019;396:94-101.
13. Lee JS, Kim C, Shin JH, Cho H, Shin DS, Kim N, et al. Machine learning-based individual assessment of cortical atrophy pattern in Alzheimer's disease spectrum: development of the classifier and longitudinal evaluation. *Sci Rep* 2018;8:4161.
14. Zuo XN, Xu T, Milham MP. Harnessing reliability for neuroscience research. *Nat Hum Behav* 2019;3:768-771.
15. Petersen RC, Doody R, Kurz A, Mohs RC, Morris JC, Rabins PV, et al. Current concepts in mild cognitive impairment. *Arch Neurol* 2001;58:1985-1992.
16. Cockrell JR, Folstein MF. Mini-Mental State Examination. In: Copeland JRM, Abou-Saleh MT, Blazer DG, editors. P*rinciples and practice of geriatric psychiatry.* New York, NY: John Wiley & Sons, Inc., 2002; 140-141.
17. Morris JC. Clinical Dementia Rating: a reliable and valid diagnostic and staging measure for dementia of the Alzheimer type. *Int Psychogeriatr* 1997;9 Suppl 1:173-176.
18. McKhann GM, Knopman DS, Chertkow H, Hyman BT, Jack CR Jr, Kawas CH, et al. The diagnosis of dementia due to Alzheimer's disease: recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* 2011;7:263-269.
19. Cho Y, Seong JK, Jeong Y, Shin SY; Alzheimer's Disease Neuroimaging Initiative. Individual subject classification for Alzheimer's disease based on incremental learning using a spatial frequency representation of cortical thickness data. *Neuroimage* 2012;59:2217-2230.
20. Lee JY, Oh SW, Chung MS, Park JE, Moon Y, Jeon HJ, et al. Clinically available software for automatic brain volumetry: comparisons of volume measurements and validation of intermethod reliability. *Korean J Radiol* 2021;22:405-414.
21. Park JE, Han K, Sung YS, Chung MS, Koo HJ, Yoon HM, et al. Selection and reporting of statistical methods to assess reliability of a diagnostic test: conformity to recommended methods in a peer-reviewed journal. *Korean J Radiol* 2017;18:888-897.
22. Sullivan DC, Obuchowski NA, Kessler LG, Raunig DL, Gatsonis C, Huang EP, et al. Metrology standards for quantitative imaging biomarkers. *Radiology* 2015;277:813-825.
23. Hernaez R. Reliability and agreement studies: a guide for clinical investigators. *Gut* 2015;64:1018-1027.
24. de Vet HC, Terwee CB, Knol DL, Bouter LM. When to use agreement versus reliability measures. *J Clin Epidemiol* 2006;59:1033-1039.
25. Portney LG, Watkins MP. *Foundations of clinical research: applications to practice.* Upper Saddle River, NJ: Pearson/Prentice Hall, 2009.
26. Cohen J. *Statistical power analysis for the social sciences.* 2nd ed. Hillsdale, NJ: Lawrence Erlbaum Associates, 1988.
27. Ross DE, Ochs AL, Tate DF, Tokac U, Seabaugh J, Abildskov TJ, et al. High correlations between MRI brain volume measurements based on NeuroQuant® and FreeSurfer. *Psychiatry Res Neuroimaging* 2018;278:69-76.
28. Ochs AL, Ross DE, Zannoni MD, Abildskov TJ, Bigler ED; Alzheimer's Disease Neuroimaging Initiative. Comparison of automated brain volume measures obtained with NeuroQuant and FreeSurfer. *J Neuroimaging* 2015;25:721-727.
29. Guo C, Ferreira D, Fink K, Westman E, Granberg T. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. *Eur Radiol* 2019;29:1355-1364.
30. Velasco-Annis C, Akhondi-Asl A, Stamm A, Warfield SK. Reproducibility of brain MRI segmentation algorithms: empirical comparison of local MAP PSTAPLE, FreeSurfer, and FSL-FIRST. *J Neuroimaging* 2018;28:162-172.
31. Ross DE, Seabaugh J, Cooper L, Seabaugh J. NeuroQuant® and Neu-

roGage® reveal effects of traumatic brain injury on brain volume. *Brain Inj* 2018;32:1437-1441.
32. Storelli L, Rocca MA, Pagani E, Van Hecke W, Horsfield MA, De Stefano N, et al. Measurement of whole-brain and gray matter atrophy in multiple sclerosis: assessment with MR imaging. *Radiology* 2018;288:554-564.
33. Chung J, Kim H, Moon Y, Moon WJ. Comparison of vendor-provided volumetry software and NeuroQuant using 3D T1-weighted images in subjects with cognitive impairment: how large is the intermethod discrepancy? *Investig Magn Reson Imaging* 2020;24:76-84.
34. Sluimer JD, van der Flier WM, Karas GB, Fox NC, Scheltens P, Barkhof F, et al. Whole-brain atrophy rate and cognitive decline: longitudinal MR study of memory clinic patients. *Radiology* 2008;248:590-598.
35. Leung KK, Barnes J, Ridgway GR, Bartlett JW, Clarkson MJ, Macdonald K, et al. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and Alzheimer's disease. *Neuroimage* 2010;51:1345-1359.
36. Wieshmann UC, Free SL, Stevens JM, Shorvon SD. Image contrast and hippocampal volumetric measurements. *Magn Reson Imaging* 1998;16:13-17.
37. Neto LL, Oliveira E, Correia F, Ferreira AG. The human nucleus accumbens: where is it? A stereotactic, anatomical and magnetic resonance imaging study. *Neuromodulation* 2008;11:13-22.
38. Droby A, Lukas C, Schänzer A, Spiwoks-Becker I, Giorgio A, Gold R, et al. A human post-mortem brain model for the standardization of multi-centre MRI studies. *Neuroimage* 2015;110:11-21.
39. Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, et al. An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* 2006;31:968-980.