

Comparison of Cortical and Subcortical Measurements in Normal Older Adults across Databases and Software Packages

Swati Rane^{a,*}, Andrew Plassard^b, Bennett A. Landman^b, Daniel O. Claassen^c and Manus J. Donahue^{c,d,e} for the Alzheimer's Disease Neuroimaging Initiative^{1,2}

^a*Department of Radiology, University of Washington Medical Center, Seattle, WA, USA*

^b*Department of Electrical Engineering, School of Engineering, Vanderbilt University, Nashville, TN, USA*

^c*Department of Neurology, Vanderbilt University School of Medicine, Nashville, TN, USA*

^d*Department of Radiology and Radiological Sciences, Vanderbilt University School of Medicine, Nashville, TN, USA*

^e*Department of Psychiatry, Vanderbilt University School of Medicine, Nashville, TN, USA*

Accepted 20 June 2017

Abstract. This work explores the feasibility of combining anatomical MRI data across two public repositories namely, the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Progressive Parkinson's Markers Initiative (PPMI). We compared cortical thickness and subcortical volumes in cognitively normal older adults between datasets with distinct imaging parameters to assess if they would provide equivalent information. Three distinct datasets were identified. Major differences in data were scanner manufacturer and the use of magnetization inversion to enhance tissue contrast. Equivalent datasets, i.e., those providing similar volumetric measurements in cognitively normal controls, were identified in ADNI and PPMI. These were datasets obtained on the Siemens scanner with TI=900 ms. Our secondary goal was to assess the agreement between subcortical volumes that are obtained with different software packages. Three subcortical measurement applications (FSL, FreeSurfer, and a recent multi-atlas approach) were compared. Our results show significant agreement in the measurements of caudate, putamen, pallidum, and hippocampus across the packages and poor agreement between measurements of accumbens and amygdala. This is likely due to their smaller size and lack of gray matter-white matter tissue contrast for accurate segmentation. This work provides a segue to combine imaging data from ADNI and PPMI to increase statistical power as well as to interrogate common mechanisms in disparate pathologies such as Alzheimer's and Parkinson's diseases. It lays the foundation for comparison of anatomical data acquired with disparate imaging parameters and analyzed with disparate software tools. Furthermore, our work partly explains the variability in the results of studies using different software packages.

Keywords: Alzheimer's disease, cortical thickness, magnetic resonance imaging, Parkinson's disease, subcortical

¹Data used in preparation of this article were obtained from the Alzheimer's disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

²Data used in preparation of this article were obtained from the Parkinson's Progression Markers Initiative ([http://www.ppmi-](http://www.ppmi-info.org)

info.org). A complete list of PPMI Study Investigators can be found at <http://www.ppmi-info.org/authorslist>.

*Correspondence to: Swati Rane, PhD, Department of Radiology, University of Washington Medical Center, 1959 NE Pacific Street, Seattle, WA 98195, USA. Tel.: +1 206 543 6159; E-mail: srleven@uw.edu.

INTRODUCTION

The Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Progressive Parkinson's Markers Initiative (PPMI) are public databases with imaging data for Alzheimer's disease (AD) and Parkinson's disease (PD) respectively [1–3]. They provide an excellent opportunity to test hypotheses of pathological and systemic overlap in the two diseases. Of interest, these databases contain large samples of anatomical magnetic resonance imaging (MRI) data that can be harvested to determine degenerative patterns in AD and PD as well as the spectrum of diseases such as Lewy body disease, which show mixed pathologies [4–7]. A critical barrier to conduct such a direct comparative analysis between AD and PD is the difference between acquisition protocols for the imaging sequences.

Our goal was to identify which disparate T1-weighted MR image acquisition protocols provide equivalent volumetric information. Precedence for this study is set by similar comparisons of imaging protocols based on resolution, scanner types, and scanner upgrades. Especially, studies by Han et al. and Reuters et al. support that similar image acquisitions across scanners (Siemens versus GE) may not significantly affect FreeSurfer-based cortical thickness measurements when slice thickness is <1.3 mm even if they have different signal intensity scalings [8–11]. Differences due to imaging resolutions were also tested in 11 individuals by Wonderlick et al. [12]. Differences in cortical thickness with respect to other scanner manufacturers and imaging parameters such as inversion time have not yet been systematically investigated. Recently we showed, that differences in inversion times can significantly affect image segmentation, often requiring a modification to the template or reference image in the software to accommodate the difference in the tissue gray and white matter contrast [13]. We therefore present this work to evaluate the differences in cortical and subcortical measurements due to dissimilar imaging parameters (resolution and inversion times) and scanners (Siemens versus Philips) in three distinct groups identified in ADNI and PPMI.

Our second goal was to compare subcortical volume measurements from different software packages in the same subjects. Different software packages such as FreeSurfer and FSL [14] use different segmentation routines and likely produce different values. In the study, we evaluated how the output measures (subcortical volumes) relate across software

packages and affect study outcomes. Multiple studies show excellent agreement between FSL and FreeSurfer, especially for larger structures. Manual tracing is still the most accurate [15–17]. In children, it has been shown that for smaller structures such as the hippocampus and amygdala, software such as FreeSurfer and FSL did not correlate strongly with manual tracings [18], which are often time-consuming and impractical for large datasets such as ADNI and PPMI. In this study, we compare volumes of subcortical structures using FreeSurfer and FSL in healthy older adults with no symptoms of dementia (AD or PD). Additionally, we applied a recently developed multi-atlas approach [19], which uses multiple atlas templates to segment different brain regions. This approach has the advantage of utilizing multiple (in this case, 15) atlases to delineate brain regions with a statistical fusion algorithm to improve the accuracy of the segmentation.

MATERIALS AND METHODS

Acquisition

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). The ADNI was launched in 2003 as a public – private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial MRI, positron emission tomography, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment and early AD. For up-to-date information, see <http://www.adni-info.org>. Data used in the preparation of this article were also obtained from the Parkinson's Progression Markers Initiative (PPMI) database (<http://www.ppmi-info.org/data>). For up-to-date information on the study, visit <http://www.ppmi-info.org>.

For this study, only older adult control subjects without symptoms of AD or PD, i.e., subjects with no signs of depression (Clinical Dementia Rating [20], CDR=0), mild cognitive impairment or dementia (as assessed by the Mini-Mental State Examination [21] in ADNI or the Montreal Cognitive Assessment [22] in PPMI), no first degree relative with PD, were chosen from the datasets to minimize disease related anatomical differences. Table 1 records the subject demographics used in the comparison. Subjects within ADNI and PPMI are de-identified

Table 1
Demographic and scan information

Group	Sex	Age (years)	TR (ms)	TI (ms)	TE (ms)	Resolution (mm ³)	Manufacturer
ADNI1	5F/10M	68.1 ± 7.2	2300	900	2.98	1.0 × 1.0 × 1.2	Siemens
PPMI1	5F/14M	62.4 ± 13.7	2296	no inversion	3.18	0.9 × 0.9 × 1.2	Philips, GE
PPMI2	4F/15M	62.4 ± 13.7	2300	900	2.98	1.0 × 1.0 × 1.0	Siemens

and corresponding data were accessed in accordance with their data use agreements. Acquisition parameters are also shown in Table 1. Specifically, this study compared three groups separated based on resolution and TR/TE/TI. Most T1 acquisitions in the ADNI database are similar to ADNI1 ($n = 15$). In PPMI, two different acquisitions (PPMI1; $n = 19$ and PPMI2; $n = 19$) are pre-dominant among the data. Few acquisitions with a slice thickness >1.2 mm exist, and were not considered (Siemens or GE or Philips). ADNI1 and PPMI2 differ only in imaging resolution. PPMI1 and PPMI2 differ in resolution, inversion time, TI, and scanner manufacturer. Only 19 datasets with the PPMI1 acquisition were available at the time of this work. PPMI2 and ADNI1 were then age and gender matched to PPMI1 at a group level, thereby defining the sample size. Control PPMI participants are significantly younger (61 ± 10 years) than control ADNI participants (74 ± 6 years) and one-to-one age and gender matching is difficult. Please see Discussion for a *post-hoc* comparison in 137 additional individuals.

Analyses

Cortical thickness measurement

We applied the standard FreeSurfer (v5.1.0) pipeline comprising of intensity-normalization, skull tissue removal, tissue segmentation, resampling to a $1 \times 1 \times 1$ mm³ resolution, and registration to the standard FreeSurfer brain (fsaverage). Briefly, this pipeline includes removal of non-brain tissue using a hybrid watershed/surface deformation procedure [23], automated Talairach transformation, segmentation of the subcortical white matter and deep gray matter volumetric structures (including hippocampus, amygdala, caudate, putamen, ventricles) intensity normalization [24], tessellation of the gray matter/white matter boundary, automated topology correction, and surface deformation following intensity gradients to optimally place the gray/white and gray/cerebrospinal fluid borders at the location where the greatest shift in intensity defines the transition to the other tissue class [25, 26]. Imperfections in tissue segmentations were corrected by manual editing. Manual editing included removing remnants of

skull or adding cortical region in the temporal and occipital lobes that were cut-off by the automated segmentation. We confined thickness measurements to five cortical subdivisions: frontal, parietal, temporal, insular, and occipital. This was achieved by combining the regions from the Desikan-Killiany Atlas as outlined at <https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation> [27]. We compared these five regional measurements between ADNI1, PPMI1, and PPMI2. To ensure that the differences were not normal variations and to overcome the issue of multiple comparisons, we performed permutations testing [28] using 5000 iterations per group and 5000 iteration per cortical region. Significance was set at $p = 0.01$.

Subcortical volume measurement

We then performed subcortical measurements and calculated volumes of the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and accumbens, using FIRST (part of FSL v5.0.8), FreeSurfer (v5.1.0), and multi-atlas segmentation. FIRST: We registered all T1-weighted images to the standard 1 mm MNI template in FSL. FIRST was applied to detect subcortical volumes using prior shape models [14, 29]. We compared the FIRST subcortical volumes between ADNI1, PPMI1, and PPMI2 using a similar permutations-testing approach as for the cortical thickness measurements. FreeSurfer: We used the automated subcortical parcellation in FreeSurfer to detect all the above subcortical structures based on a probabilistic atlas [24, 25]. Multi-atlas segmentations: The multi-atlas segmentation pipeline by Asman et al. uses 15 atlases to resolve voxel assignments to pre-labeled brain regions [19, 30].

To reduce the effect of head-size variability, we normalized all the subcortical volumes for each subject with the respective intra-cranial volume (ICV, calculated using eTIV in FreeSurfer). Similar to the cortical thickness measurements, we used permutations testing to evaluate differences between ADNI1, PPMI1, and PPMI2.

Finally, we compared the subcortical measurements calculated by the three different analytical tools to determine cross-software agreement. We

obtained Bland Altman plots and evaluated intra-class correlation coefficients (ICC) for consistency (also referred to as norm-referenced reliability or consistency among measurements) between the subcortical measurements of each software package [31, 32].

RESULTS

Cortical measurements

Volumetric measurements were performed on three groups of controls subjects with T1-weighted MRI data, which was acquired using disparate imaging parameters. Mean SNR values (mean signal intensity divided by the standard deviation of the signal intensity) of the anatomical images in a uniform normal appearing white matter region (5 mm × 5 mm) for ADNI1, PPMI1, and PPMI2 were 25.34 ± 5.10 , 22.48 ± 5.63 , and 16.95 ± 3.39 respectively. SNR in all cortical gray matter region in ADNI1, PPMI1, and PPMI2, was 5.48 ± 0.35 , 4.90 ± 0.95 , and 5.34 ± 0.34 respectively while in all white matter was 4.71 ± 0.77 , 4.53 ± 0.76 , and 4.88 ± 0.83 , respectively. SNRs were not significantly different between groups. Manual editing was performed on 3 subjects in ADNI1, 5 subjects in PPMI1, and 2 subjects in PPMI2.

Table 2 summarizes the cortical thickness measurements from FreeSurfer. Measurements between ADNI1 and PPMI2 were more similar to each other (slightly larger p -values for most cortical regions) than between ADNI1 and PPMI1. Mean cortical thickness values for all subjects combined in the frontal lobe, cingulate, occipital lobe, parietal lobe, temporal lobe, and insula were 2.44 ± 0.11 , 2.58 ± 0.18 , 1.98 ± 0.12 , 2.71 ± 0.13 , 2.27 ± 0.13 , and 2.92 ± 0.18 mm, respectively. In general, cortical thickness in the frontal, temporal, parietal, and occipital lobes was lower in PPMI1 compared to ADNI1 and PPMI2. In order to ensure that the differences in the lobar thickness with the three

protocols is not normal variability, we performed permutations testing, which showed significant differences in the frontal ($p=0.005$), occipital ($p=0.01$), and parietal ($p<0.0001$) lobe thicknesses between PPMI1 and PPMI2. Frontal lobe was marginally different between ADNI1 and PPMI1 ($p=0.03$) and parietal lobe was marginally different between ADNI1 and PPMI2 ($p=0.02$).

Subcortical measurements

Overall average (\pm standard deviation) volumes of the thalamus, caudate, putamen, pallidum, hippocampus, amygdala, and the accumbens of all 53 subjects were 7813 ± 872 , 3447 ± 458 , 4848 ± 631 , 1815 ± 267 , 3838 ± 497 , 1410 ± 322 , 442 ± 112 mm³ respectively using FIRST. Average ICV estimation from FreeSurfer was 1429986 ± 204533 mm³. Normalizing by ICV, the subcortical volumes were 0.0055 ± 0.0007 , 0.0024 ± 0.0004 , 0.0034 ± 0.0005 , 0.0012 ± 0.0002 , 0.0027 ± 0.0005 , 0.0009 ± 0.0002 , and 0.0003 ± 0.0001 , respectively. Table 3 shows the comparison of subcortical volumes between ADNI1, PPMI1, and PPMI2 using FIRST. After permutations testing, most subcortical volumes were not significantly different between the three groups. Amygdala measurements however were significantly different ($p<0.01$) between groups (ADNI1 versus PPMI1: $p=0.008$, PPMI1 versus PPMI2: $p=0.001$).

Agreement between software packages

Figure 1 shows the segmentation on a representative subject using FIRST, FreeSurfer, and multi-atlas tools. Segmentation of cortical structures in the same slice is shown for one subject. The white arrows point to the discrepant segmentations of the putamen and pallidum between software packages. We used Bland-Altman plots to show whether systematic differences exist between subcortical measurements obtained with different packages. We show comparisons of three structures (i) thalamus (well visible

Table 2
Cortical thickness comparison between ADNI1, PPMI1, and PPMI 2

Cortical thickness (mm)	Frontal lobe	Cingulate	Occipital lobe	Temporal lobe	Parietal lobe	Insula
ADNI1	2.46 ± 0.08	2.55 ± 0.13	2.00 ± 0.13	2.75 ± 0.12	2.26 ± 0.13	2.90 ± 0.17
PPMI1	2.38 ± 0.13	2.62 ± 0.23	1.92 ± 0.12	2.66 ± 0.16	2.21 ± 0.13	2.94 ± 0.19
PPMI2	2.49 ± 0.96	2.56 ± 0.14	2.02 ± 0.08	2.72 ± 0.09	2.35 ± 0.09	2.93 ± 0.17
p (ADNI1 versus PPMI1)	0.03	0.15	0.08	0.06	0.10	0.16
p (PPMI1 versus PPMI2)	0.005	0.14	0.01	0.11	<0.0001	0.16
p (ADNI1 versus PPMI2)	0.15	0.16	0.16	0.15	0.02	0.16

Table 3
Subcortical volumes in ADNI1, PPMI1, and PPMI2, normalized by total intracranial volume using FIRST

Subcortical volume	Thalamus	Caudate	Putamen	Pallidum	Hippocampus	Amygdala	Accumbens
ADNI1	0.0055 ± 0.0008	0.0024 ± 0.0004	0.0035 ± 0.0005	0.0013 ± 0.0002	0.0027 ± 0.0006	0.0010 ± 0.0002	0.0003 ± 0.0001
PPMI1	0.0057 ± 0.0008	0.0026 ± 0.0004	0.0035 ± 0.0006	0.0013 ± 0.0002	0.0030 ± 0.0006	0.0009 ± 0.0002	0.0003 ± 0.0001
PPMI2	0.0054 ± 0.0006	0.0024 ± 0.0003	0.0033 ± 0.0004	0.0012 ± 0.0001	0.0026 ± 0.0003	0.0011 ± 0.0002	0.0003 ± 0.0001
p (ADNI1 versus PPMI1)	0.13	0.14	0.11	0.12	0.14	0.008	0.14
p (PPMI1 versus PPMI2)	0.14	0.14	0.14	0.14	0.07	0.001	0.13
p (ADNI1 versus PPMI2)	0.14	0.14	0.12	0.13	0.13	0.14	0.11

using gray-white matter contrast and large in size), (ii) amygdala (well visible using gray-white matter contrast and small in size), and (iii) accumbens (no visible tissue contrast and small in size) in Fig. 2. Volumes of all large and well-defined structures such as the thalamus, caudate, putamen, hippocampus are well-correlated (ICC values between 0.64–0.85) across different software packages. Bland-Altman Plots show bias between methods (non-zero difference between measurements by two methods) for all structures. In the putamen, the difference between FreeSurfer and FIRST as well as the difference between FreeSurfer and Multi-atlas derived volumes increased with increase in the mean size thereby indicating a proportional difference i.e., as putaminal volume increased, discrepancy between FreeSurfer and Multi-atlas volumes also increased. Figure 2a shows the comparison of all three packages using correlation plots and Bland-Altman plots for a large structure, Thalamus. Figure 2b and c show similar plots for smaller structures such as the amygdala (with visible gray and white matter contrast) and the accumbens (which has no obvious gray and white matter contrast for segmentation purposes). Intra-class correlation coefficient and Fishers-Z ($0.5 \times (\ln(1+r) - \ln(1-r))$, where r = Pearson's correlation coefficient) for all subcortical volume measurements are noted in Table 4. While hippocampal volume measurements were very consistent across software packages, interestingly, consensus was poor with amygdala segmentation. Bias and proportional differences were detected for all small subcortical structures (hippocampus, amygdala, and accumbens).

DISCUSSION

In this work, we have identified data sets from the two databases; ADNI and PPMI, that provide similar volumetric measurements on healthy subjects. We also show that there is consistency between volumetric measurements calculated with FreeSurfer, FIRST, and the multi-atlas approach in most regions.

ADNI and PPMI data sets (ADNI1 and PPMI2) from the same (in this case, Siemens) scanners and identical TI values were similar. PPMI data with different in-plane resolution and manufacturer (Philips) showed differences in volumetric measures in subcortical and posterior cortical structures. While the differences in scanner manufacturers maybe a contributing factor, the observed difference is more likely

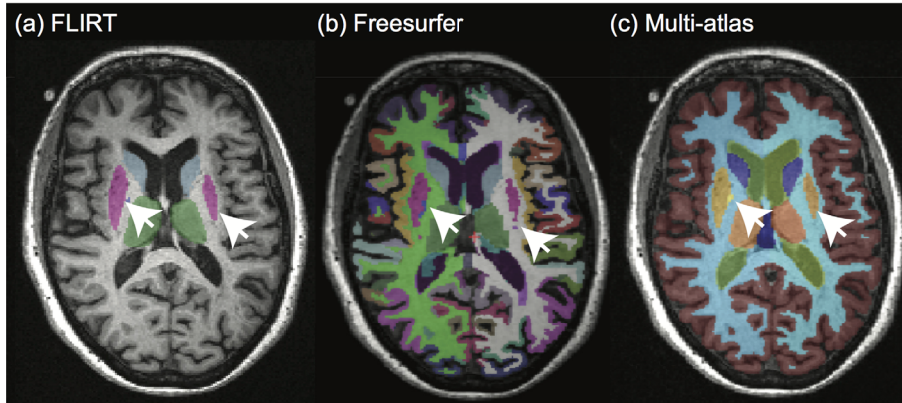


Fig. 1. Subcortical segmentation output using FIRST, FreeSurfer and Multi-atlas segmentation in the same subject and slice location. Differences are evident in the segmentation of the putamen [pink in (a) and (b), dark yellow in (c)] and the pallidum [blue in (a) and (b), bright yellow in (c)].

due to difference in the application of magnetization inversion to enhance gray matter-white matter contrast. Other studies [8, 12] suggest that scanner differences and nominal differences in imaging resolution do not affect cortical and subcortical volume measurements. Our previous work shows that tissue segmentations with different TI values produce significantly different volumetric outcomes [13]. In this study, in the absence of inversion, with T1 of gray matter = 1100 ms, and T1 of white matter = 800 ms at 3T, the ratio of signal intensities following using a simple inversion recovery is 1.07, while at a TI = 900 ms, it is 1.69. The higher contrast between the white and gray matter are likely to result in better segmentation boundaries.

As a *post-hoc* test, we extended the comparison to 107 new controls subjects, with the same acquisition as ADNI1 and 20 new controls subjects with the same acquisition as PPMI2. The two groups had significantly different age (ADNI1 : 75.4 ± 5.7 years, PPMI2 : 69.7 ± 4.8 , $p < 0.001$) and gender distributions (ADNI1 : 53M/54F, PPMI2 : 15M/5F). The thickness values in frontal, cingulate, occipital, temporal, parietal, and insular regions were 2.36 ± 0.09 , 2.36 ± 0.11 , 1.97 ± 0.11 , 2.42 ± 0.09 , 2.15 ± 0.10 , and 2.60 ± 0.11 mm respectively in ADNI1 and 2.37 ± 0.12 , 2.37 ± 0.15 , 2.06 ± 0.11 , 2.43 ± 0.13 , 2.22 ± 0.12 , and 2.60 ± 0.17 mm respectively. After adjusting for age and gender, there were no differences between the frontal ($p = 0.86$), cingulate ($p = 0.96$), occipital ($p = 0.04$), temporal ($p = 0.36$), parietal ($p = 0.10$), and insular thickness ($p = 0.89$) values, using a significance of $p = 0.01$ as before. Note that the comparisons are very similar to the main comparison performed earlier in the Results section.

Similarly, the subcortical comparisons between the new subjects from ADNI1 and PPMI2, after adjusting for age and gender showed no significant differences except in the Accumbens, which was significantly ($p = 0.005$) different in the two groups. Volumes of the Thalamus, Caudate, Putamen, Pallidum, Hippocampus, Amygdala, and Accumbens in ADNI1 were 6328 ± 749 , 3499 ± 506 , 4873 ± 626 , 1380 ± 180 , 3623 ± 462 , 1489 ± 253 , and 468 ± 88 and were 7037 ± 1131 , 3510 ± 873 , 4869 ± 1079 , 1460 ± 268 , 3750 ± 575 , 1535 ± 245 , and 420 ± 123 , respectively, in PPMI2 (p values for Thalamus, Caudate, Putamen, Pallidum, Hippocampus, Amygdala, were 0.11, 0.20, 0.03, 0.95, 0.08, and 0.16 respectively). Therefore, it is likely that the sample size chosen for the initial comparison was not ideal for comparison of smaller structures such as the Accumbens which do not have distinct gray-white matter contrast.

Furthermore, it is likely that volumetric measurements between ADNI1 and PPMI2 are similar only for larger regions and results differ for small regions of interest. Typical volumes of the frontal, temporal, parietal, occipital lobes, cingulate gyrus, and insula are about $60,000 \text{ mm}^3$, $40,000 \text{ mm}^3$, $20,000 \text{ mm}^3$, $10,000 \text{ mm}^3$, and $5,000 \text{ mm}^3$. Therefore, over a wide range of size, ADNI1 and PPMI2 provide similar measurement. In smaller region such as the parahippocampal gyrus (volume $\sim 2500 \text{ mm}^3$) with permutations testing no difference was observed between ADNI1, PPMI1, and PPMI2 (p -values are ADNI1 versus PPMI1 : 0.68, PPMI1 versus PPMI2 : 0.25, ADNI1 versus PPMI2 : 0.56). For the extended comparison between ADNI1 and PPMI2 as well, no statistical difference was observed after adjusting for age and gender ($p = 0.92$).

(2a)

Thalamus segmentation

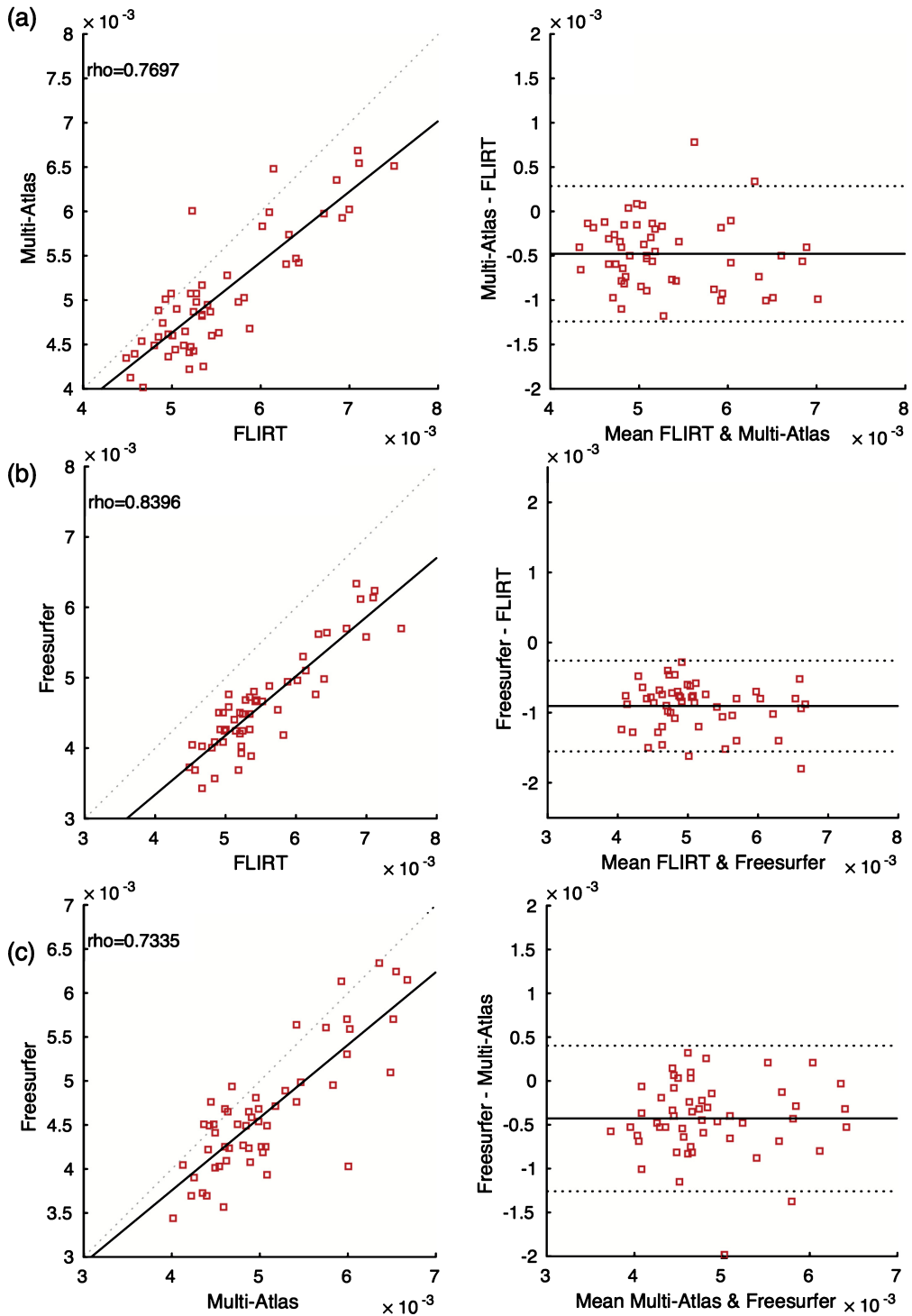


Fig. 2. (Continued)

(2b)

Amygdala segmentation

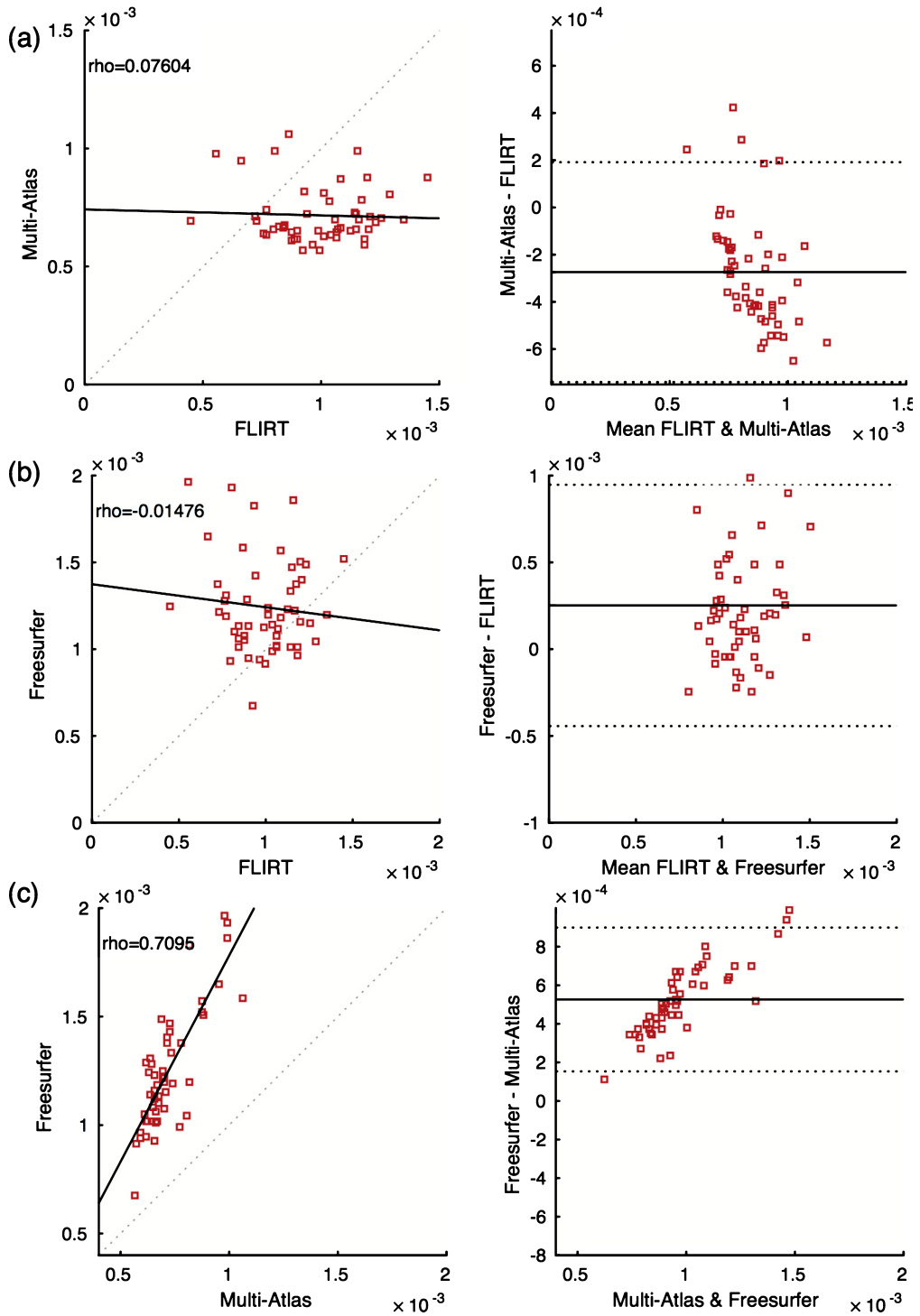


Fig. 2. (Continued)

(2c)

Accumbens segmentation

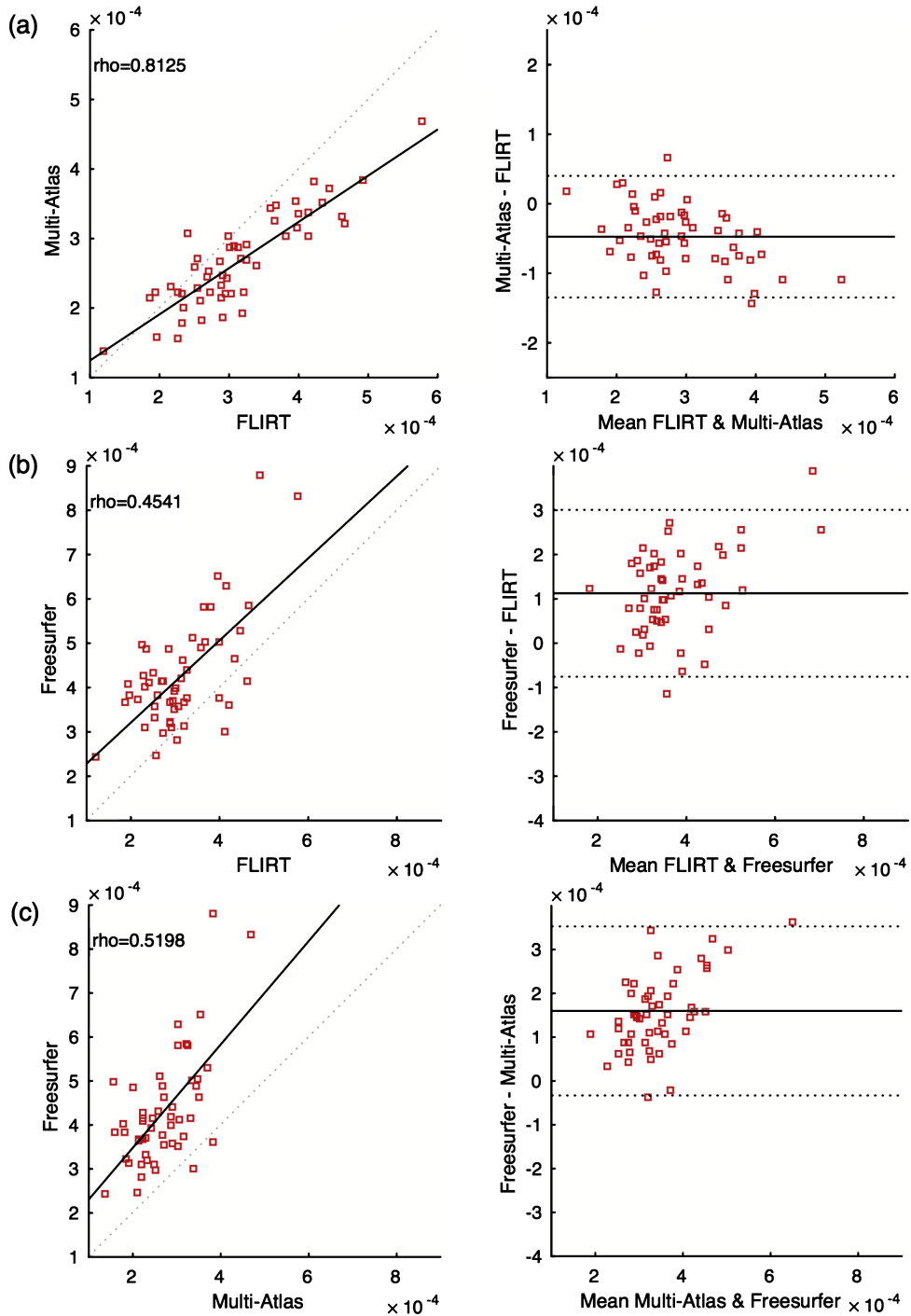


Fig. 2. Agreement between different software packages for subcortical segmentation. While large ($>3000 \text{ mm}^3$) structures with distinct gray matter-white matter contrast such as the thalamus are well-delineated and consistent among software, segmentation of smaller structures such as the amygdala shows considerable variability and bias depending on the software used. The plots for Caudate, Putamen, Hippocampus, Pallidum were similar to Thalamus. However, the putamen showed a proportional error between FreeSurfer and Multi-atlas measurements. Accumbens, which is a small structure and has no discernible gray-white matter contrast, showed surprisingly good agreement between software packages.

Table 4
Subcortical volume measurement using FIRST, FreeSurfer, and Multi-Atlas

Volumes (mm ³)	Thalamus	Caudate	Putamen	Pallidum	Hippocampus	Amygdala	Accumbens
FIRST	7813 ± 872	3448 ± 458	4848 ± 631	1815 ± 267	3838 ± 498	1410 ± 322	442 ± 112
FreeSurfer	6527 ± 780	3576 ± 538	5268 ± 799	1608 ± 244	3908 ± 489	1738 ± 264	594 ± 123
Multi-Atlas	7128 ± 700	3117 ± 472	4211 ± 535	1376 ± 180	3689 ± 445	1007 ± 115	377 ± 93
	Intra-class correlation [95% confidence intervals]						
FIRST & Multi Atlas	0.75 [0.61,0.85]	0.85 [0.76,0.91]	0.78 [0.65,0.87]	0.64 [0.45,0.77]	0.78 [0.66,0.87]	0.06 [-0.21,0.32]	0.83 [0.72,0.90]
FreeSurfer & Multi-Atlas	0.69 [0.52,0.82]	0.86 [0.78,0.92]	0.77 [0.63,0.86]	0.74 [0.59,0.84]	0.79 [0.67,0.87]	0.39 [0.14,0.60]	0.45 [0.21,0.64]
FIRST & FreeSurfer	0.83 [0.73,0.89]	0.85 [0.76,0.91]	0.71 [0.55,0.82]	0.54 [0.32,0.70]	0.82 [0.70,0.89]	-0.17 [0.41,0.10]	0.44 [0.20,0.73]
	Fishers - Z						
FIRST & Multi-Atlas	1.04	1.29	1.09	0.86	1.07	0.1	1.25
FreeSurfer & Multi-Atlas	0.86	1.37	1.18	1.04	1.09	0.6	0.51
FIRST & FreeSurfer	1.2	1.33	0.96	0.61	1.15	-0.17	0.48

We acknowledge that for an ideal comparison, the same subjects should be scanned using the different scanner manufacturers and imaging parameters. We have used three groups of healthy control subjects of similar ages and sex distribution with the assumption that each group will have similar morphometric measurements. Han et al. [8], showed in their study that cortical thickness measurements varied by 0.12 mm across scanner manufacturers (Siemens versus GE) using the same subjects [20]. Difference within the same scanner manufacturers was about 0.03. Within (Siemens) scanner difference in cortical thickness was similar in our study (0.02 ± 0.04 mm). Our results show a difference in cortical thickness between scanners (Siemens versus Philips) of 0.05 ± 0.07 mm.

In the subcortical structures, studies by the same group concluded that different scanner manufacturers did not show significant difference in the amygdala or the thalamus but there was a bias in the hippocampal measurements. In our study, there was a medium-strong correlation between subcortical measurements of most structures except the amygdala. Since the subjects in each group are different for our study, we did not measure bias using Bland-Altman plots.

We, however, assessed correlation and system bias between the different popular software packages as well as a multi-atlas segmentation approach to measure subcortical volumes. We found that there was bias between measurements made with FIRST, Multi-Atlas, as well as FreeSurfer. As in the comparison of scanner manufacturers, the direction of bias was different for different structures in this comparison of software packages. Whether the observed bias is significant or not, could not be determined in this

study. To do so, we would need multiple data sets and comparisons i.e., multiples of this study. This will allow us to determine if the confidence intervals for the line of mean difference, i.e., bias, if the Bland-Altman plots excluded 0.

While comparison of data from different scanner manufacturers must be done with caution, our work suggests that data from the same manufacturers from different databases (ADNI and PPMI, in this case), which is acquired with similar imaging parameters, can be compared. There are three advantages of this finding: 1) It reduces the burden of needing or acquiring more data by allowing the merge of multiple databases; 2) This result provides an opportunity to combine data to increase statistical power to detect subtle changes that may occur in the early stages of disease and may otherwise remain elusive; and 3) Specifically, for the ADNI and PPMI databases, common phenomena such as presence of APOE-ε4 allele and its relative risk, concentrations of amyloid and tau, cardiovascular factors in disparate diseases such as AD and PD can now be investigated.

One limitation of this study is that while the selected subjects were cognitively normal older adults, it is likely adults with APOE-ε4 genetic status and varying degrees of amyloid, tau, and alpha-synuclein pathology were included and could contribute to some of the variability observed in the data. However, it is unlikely, that the variability is more prominent in one group than another.

Lastly, variance in the measurement with each software will determine the statistical power required for hypothesis testing. For instance, based on Table 4 (First 3 rows), the number of subjects required to significantly measure 15% atrophy in hippocampal volume between two groups with 80% power

and false-positive rate (α) of 0.05 would be 13, 12, and 12 with FIRST, FreeSurfer, and Multi-Atlas respectively. However, the same comparison for the amygdala would require 38, 18, and 11 subjects with FIRST, FreeSurfer, and Multi-Atlas respectively. This could likely be one of the reasons why studies using similar number of subjects but different software packages show different outcomes for the same measurement.

Conclusion

We tested the feasibility of combining anatomical imaging data from two distinct data repositories, i.e., ADNI and PPMI. Based on lobar cortical thickness measurements and subcortical volume measurements, we identified imaging datasets in ADNI and PPMI, acquired with magnetization inversion across multiple Siemens scanners that provided equivalent volumetric information. Further, we also compared the subcortical measurements using FSL, FreeSurfer, and Multi-Atlas segmentation. While a general congruence was observed between most measurements, there was significant difference between measurements of smaller brain structures with poor gray matter-white matter tissue contrast.

ACKNOWLEDGMENTS

Data used in preparation of this article were obtained from the Parkinson's Progression Markers Initiative (PPMI: <http://www.ppmi-info.org>) and Alzheimer's Disease Neuroimaging Initiative (ADNI) database (<http://adni.loni.usc.edu>). As such, the investigators within PPMI and ADNI contributed to the design and implementation of both databases, and/or provided data, but did not participate in analysis or writing of this report. A complete list of PPMI Study Investigators can be found at <http://www.ppmi-info.org/authorslist/>. A complete listing of ADNI investigators can be found at: http://www.adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's

Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (<http://www.fnih.org>). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. The Laboratory for Neuro Imaging at the University of Southern California disseminates ADNI data.

PPMI, a public-private partnership, is funded by the Michael J. Fox Foundation for Parkinson's Research and funding partners, including Abbvie, Avid Radiopharmaceuticals, Biogen Idec, Bristol-Myers Squibb, Covance, GE Healthcare, Genentech, GSK GlaxoSmithKline, Lilly, Lundbeck, Merck, MSD Meso Scale Discovery, Pfizer Inc., Piramal, Roche, Servier, and UCB.

Grant support for this work was provided by Biomarkers Across Neurodegenerative Disease from Alzheimer's Association, the Michael J Fox Foundation, and the W. Garfield Weston Foundation awarded to SR and R01NS097783-01 awarded to DC.

CONFLICT OF INTEREST

The authors have no conflict of interest to report.

REFERENCES

- [1] Mueller SG, Weiner MW, Thal LJ, Petersen RC, Jack CR, Jagust W, Trojanowski JQ, Toga AW, Beckett L (2005) Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement* 1, 55-66.
- [2] Jack CR Jr, Bernstein MA, Fox NC, Thompson P, Alexander G, Harvey D, Borowski B, Britson PJ, L Whitwell J, Ward C, Dale AM, Felmlee JP, Gunter JL, Hill DL, Killiany R,

- Schuff N, Fox-Bosetti S, Lin C, Studholme C, DeCarli CS, Krueger G, Ward HA, Metzger GJ, Scott KT, Mallozzi R, Blezek D, Levy J, Debbins JP, Fleisher AS, Albert M, Green R, Bartzokis G, Glover G, Mugler J, Weiner MW (2008) The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *J Magn Reson Imaging* **27**, 685-691.
- [3] Marek K, Jennings D, Lasch S, Siderowf A, Tanner C, Simuni T, Coffey C, Kieburz K, Flagg E, Chowdhury S, Poewe W (2011) The parkinson progression marker initiative (PPMI). *Prog Neurobiol* **95**, 629-635.
- [4] Cuingnet R, Gerardin E, Tessieras J, Auzias G, Lehéricy S, Habert MO, Chupin M, Benali H, Colliot O, Alzheimer's Disease Neuroimaging Initiative (2011) Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *Neuroimage* **56**, 766-781.
- [5] Pereira JB, Svenningsson P, Weintraub D, Brønneck K, Lebedev A, Westman E, Aarsland D (2014) Initial cognitive decline is associated with cortical thinning in early Parkinson disease. *Neurology* **82**, 2017-2025.
- [6] Cairns NJ, Perrin RJ, Franklin EE, Carter D, Vincent B, Xie M, Bateman RJ, Benzinger T, Friedrichsen K, Brooks WS, Halliday GM, McLean C, Ghetti B, Morris JC, Alzheimer Disease Neuroimaging Initiative; Dominantly Inherited Alzheimer Network (2015) Neuropathologic assessment of participants in two multi-center longitudinal observational studies: The Alzheimer Disease Neuroimaging Initiative (ADNI) and the Dominantly Inherited Alzheimer Network (DIAN). *Neuropathology* **35**, 390-400.
- [7] Henchcliffe C, Dodel R, Beal MF (2011) Biomarkers of Parkinson's disease and Dementia with Lewy bodies. *Prog Neurobiol* **95**, 601-613.
- [8] Han X, Jovicich J, Salat D, van der Kouwe A, Quinn B, Czanner S, Busa E, Pacheco J, Albert M, Killiany R, Maguire P, Rosas D, Makris N, Dale A, Dickerson B, Fischl B (2006) Reliability of MRI-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *Neuroimage* **32**, 180-194.
- [9] Reuter M, Rosas HD, Fischl B (2010) Highly accurate inverse consistent registration: A robust approach. *Neuroimage* **53**, 1181-1196.
- [10] Dale AM, Fischl B, Sereno MI (1999) Cortical surface-based analysis: I. Segmentation and surface reconstruction. *Neuroimage* **9**, 179-194.
- [11] Fischl B, Sereno MI, Dale AM (1999) Cortical surface-based analysis: II: Inflation, flattening, and a surface-based coordinate system. *Neuroimage* **9**, 195-207.
- [12] Wonderlick JS, Ziegler DA, Hosseini-Varnamkhasti P, Locascio JJ, Bakkour A, Van Der Kouwe A, Triantafyllou C, Corkin S, Dickerson BC (2009) Reliability of MRI-derived cortical and subcortical morphometric measures: Effects of pulse sequence, voxel geometry, and parallel imaging. *Neuroimage* **44**, 1324-1333.
- [13] Plassard AJ, Harrigan RL, Newton AT, Rane S, Pallavaram S, D'Haese PF, Dawant BM, Claassen DO, Landman BA (2016) On the fallacy of quantitative segmentation for T1-weighted MRI. *Proc SPIE Int Soc Opt Eng* **9784**, 978416.
- [14] Jenkinson M, Beckmann CF, Behrens TE, Woolrich MW, Smith SM (2012) FSL. *Neuroimage* **62**, 782-90.
- [15] Morey RA, Selgrade ES, Wagner HR, Huettel SA, Wang L, McCarthy G (2010) Scan-rescan reliability of subcortical brain volumes derived from automated segmentation. *Hum Brain Mapp* **31**, 1751-1762.
- [16] Dewey J, Hana G, Russell T, Price J, McCaffrey D, Harezlak J, Sem E, Anyanwu JC, Guttmann CR, Navia B, Cohen R, Tate DF, HIV Neuroimaging Consortium (2010) Reliability and validity of MRI-based automated volumetry software relative to auto-assisted manual measurement of subcortical structures in HIV-infected patients from a multisite study. *Neuroimage* **1**, 1334-1344.
- [17] Klauschen F, Goldman A, Barra V, Meyer-Lindenberg A, Lundervold A (2009) Evaluation of automated brain MR image segmentation and volumetry methods. *Hum Brain Mapp* **30**, 1310-1327.
- [18] Schoemaker D, Buss C, Head K, Sandman CA, Davis EP, Chakravarty MM, Gauthier S, Pruessner JC (2016) Hippocampus and amygdala volumes from magnetic resonance images in children: Assessing accuracy of FreeSurfer and FSL against manual segmentation. *Neuroimage* **129**, 1-4.
- [19] Asman AJ, Landman BA (2013), Non-local statistical label fusion for multi-atlas segmentation. *Med Image Anal* **17**, 194-208.
- [20] Morris JC (1993) The Clinical Dementia Rating (CDR): Current version and scoring rules. *Neurology* **43**, 2412-2414.
- [21] Folstein MF, Robins LN, Helzer JE (1983) The mini-mental state examination. *Arch Gen Psychiatry* **40**, 812.
- [22] Nasreddine ZS, Phillips NA, Bédirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H (2005) The Montreal Cognitive Assessment, MoCA: A brief screening tool for mild cognitive impairment. *J Am Geriatr Soc* **53**, 695-699.
- [23] Ségonne F, Dale AM, Busa E, Glessner M, Salat D, Hahn HK, Fischl B (2004) A hybrid approach to the skull stripping problem in MRI. *Neuroimage* **22**, 1060-1075.
- [24] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, Van Der Kouwe A, Killiany R, Kennedy D, Klaveness S, Montillo A (2002) Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron* **33**, 341-355.
- [25] Sled JG, Zijdenbos AP, Evans AC (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Trans Med Imaging* **17**, 87-97.
- [26] Ségonne F, Pacheco J, Fischl B (2007) Geometrically accurate topology-correction of cortical surfaces using non-separating loops. *IEEE Trans Med Imaging* **26**, 518-529.
- [27] Desikan RS, Ségonne F, Fischl B, Quinn BT, Dickerson BC, Blacker D, Buckner RL, Dale AM, Maguire RP, Hyman BT, Albert MS, Killiany RJ (2006) An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *Neuroimage* **31**, 968-980.
- [28] Nichols TE, Holmes AP (2002) Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum Brain Mapp* **15**, 1-25.
- [29] Patenaude B, Smith SM, Kennedy DN, Jenkinson M (2011) A Bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage* **56**, 907-922.
- [30] Asman AJ, Landman BA (2012) Formulating spatially varying performance in the statistical fusion framework. *IEEE Trans Med Imaging* **31**, 1326-1336.
- [31] McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients. *Psychol Methods* **1**, 30.
- [32] Shrout PE, Fleiss JL (1979) Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull* **86**, 420.