Reproducibility and Reliability of Computing Models in Segmentation and Volumetric Measurement of Brain

Annals of Neurosciences
30(4) 224–229, 2023

DOI: 10.1177/09727531231159959
journals.sagepub.com/home/aon
Sage

Mahender Kumar Singh[1,2]

Abstract

Background: Segmentation and morphometric measurement of brain tissue and regions from non-invasive magnetic resonance images have clinical and research applications. Several software tools and models have been developed by different research groups which are increasingly used for segmentation and morphometric measurements. Variability in results has been observed in the imaging data processed with different neuroimaging pipelines which have increased the focus on standardization.

Purpose: The availability of several tools and models for brain morphometry poses challenges as an analysis done on the same set of data using different sets of tools and pipelines may result in different results and interpretations and there is a need for understanding the reliability and accuracy of such models.

Methods: T1-weighted (T1-w) brain volumes from the publicly available OASIS3 dataset have been analysed using recent versions of FreeSurfer, FSL-FAST, CAT12, and ANTs pipelines. grey matter (GM), white matter (WM), and estimated total intracranial volume (eTIV) have been extracted and compared for inter-method variability and accuracy.

Results: All four methods are consistent and strongly reproducible in their measurement across subjects however there is a significant degree of variability between these methods.

Conclusion: CAT12 and FreeSurfer methods have the highest degree of agreement in tissue class segmentation and are most reproducible compared to others.

Keywords

Brain segmentation, morphometry, FreeSurfer, FSL-FAST, CAT12, ANTs

Received 13 December 2022; **accepted** 10 February 2023

Introduction

Advancements in neuro-imaging have greatly increased our impetus for the study of brain structure and function.[1] Magnetic Resonance Imaging (MRI) techniques provide good contrast between grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF) and it is extensively used for structural and functional (fMRI) imaging of the brain.[2]

Segmentation of 3D brain volumes into tissue types and between different cortical and subcortical regions has applications in a wide range of biomedical research problems.[3,4] Manual segmentation is considered the "gold standard" but suffers from serious drawbacks like being time-consuming, inter-rater variability and not being suitable for large-scale evaluation.[5]

Automatic, semi-automatic segmentation methods including machine learning models have been developed to address these shortcomings and are used for tissue-class segmentation, cortical and sub-cortical segmentation, and are being increasingly used for neuroimaging analysis.[6] There are several tools in the public domain developed by different research groups for neuroimaging analysis[6] and each of them implements its own approach for bias-field correction, brain extraction, segmentation, and so on.

Several past studies have compared some of the segmentation methods applied either on the whole brain or to specific regions but mostly carried on much smaller datasets ($n < 100$).[7–10] The main goal of this study is to compare the

[1] National Brain Research Centre, Manesar, Gurugram, Haryana, India
[2] Starex University, Binola, Gurugram, Haryana, India

Corresponding Author:
Mahender Kumar Singh, Information Scientist, Data Science Laboratory, National Brain Research Centre (NBRC), Manesar, Gurugram, Haryana 122050, India.
E-mail: mks.nbrc@gov.in

Table 1. Demographic Profile of Subjects.

	Subjects (MR Sessions)			Mean Age in Years (standard deviation)
	Male	Female	Total	
Normal − (Group A)	302 (646)	442 (977)	744 (1623)	68.42 (8.47)
Cognitive decline − (Group B)	299 (487)	287 (470)	586 (957)	75.04 (7.56)
Combined Group (A+B)	601 (1133)	729 (1447)	1330 (2580)	70.88 (8.75)

Note: Subjects appearing in CogNorm-Cognitively_Normal_Cohorts table are included in Group A and the rest in Group B.

Table 2. List of Scanner Models Used for 3T MR Acquisition.

Scanner Model	No. of MR Sessions (T1-w)
Siemens TrioTim (3T)	1421
Siemens Magnetom_Vida (3T)	328
Siemens Biograph_mMR (3T)	830
Siemens Prisma_fit (3T)	1

reliability and reproducibility of automatic brain tissue segmentation using FreeSurfer, FSL-FAST, CAT12, and ANTs on a large dataset and compare agreement between them.

Methods

Dataset

The publicly available OASIS3[11] dataset (2022 Release) has been downloaded from http://central.xnat.org in terms of the accepted data use agreement. Briefly, the OASIS3 dataset is a retrospective compilation of longitudinal data of 1379 subjects in the age group of 42−96 years and includes normal subjects as well as those at various stages of cognitive decline. MR images that have failed during any of the automated image processing pipelines have been excluded from the study. The demographic profile of the selected subjects is detailed in Table 1.

MR Images

T1-weighted (T1-w) MR images in NIFTI format for each of the subjects scanned at 3-Tesla scanners at different time points have been extracted from the OASIS3 dataset. The subjects are scanned on different Siemens 3-Tesla MRI Scanners as detailed in Table 2. The OASIS3 imaging methods, scanning protocols, and data dictionary are available on the OASIS website (https://www.oasis-brains.org/).

Image Processing

Each of the MR volumes has been processed through publicly available versions of FreeSurfer, FMRIB Software Library (FSL), Advanced Normalization Tools (ANTs), and Computational Anatomy Toolbox (CAT12) pipelines. The segmentation volumes for GM, WM, and estimated total intracranial volume (eTIV) of the brain have been extracted from the segmentation output. The brief detail of each of the pipelines is detailed as under:

FreeSurfer

FreeSurfer[12] (https://surfer.nmr.mgh.harvard.edu/) is a software package developed by the Laboratory of Computational Neuroimaging at Athinoula A. Martinos Center for Biomedical Imaging and is among the most standardized software packages for processing and analyzing neuroimaging data. The recon-all pipeline from FreeSurfer v7.3 has been used for the automatic segmentation of each T1-w image volume. The WM volume has been computed by summing "cerebral and cerebellum WM volumes," and the GM and eTIV have been directly computed from "Total grey-matter volume" and "Estimated Intracranial Volume" variables respectively.

FMRIB Software Library (FSL)

FSL[13] (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki) has been developed at Wellcome Centre for Integrative Neuroimaging, Oxford, the fsl_anat.sh script of FSL v6.0 has been used for pre-processing and segmentation of MR volumes. The GM, WM, and eTIV have been extracted and calculated from partial volume maps for the 3 tissue classes generated by FSL-FAST.[14]

Advanced Normalization Tools (ANTs)

ANTs[15] (https://stnava.github.io/ANTs/) is a suite of programs for biomedical image analysis, the tissue and brain volume has been extracted from the antsbrainvols.csv file generated by the segmentation pipeline.

Computational Anatomy Toolbox (CAT12)

CAT12[16] (https://neuro-jena.github.io/cat/) is a toolbox for SPM12[17,18] (https://www.fil.ion.ucl.ac.uk/spm/) which runs on MATLAB® platform, segment option on the CAT12 is used for segmentation followed by extraction of GM, WM, and eTIV from the generated XML file. The process has been automated using batch scripts. Additionally, the Image Quality Rating (IQR) and the visual report generated by CAT12 are used for quality assessment.

Table 3. Mean and (Standard Deviation) of Volumetric Measurement in cm^3 Using Different Methods.

	FSL-FAST (a)	CAT12 (b)	FreeSurfer (c)	ANTs (d)
GM (Group A)	539.09 (50.86)	608.92 (58.36)	577.40 (55.36)	454.44 (48.26)
GM (Group B)	526.41 (54.85)	587.33 (60.79)	562.46 (56.82)	443.10 (49.83)
GM (A+B)	534.38 (52.73)	601.06 (60.15)	571.85 (56.37)	450.23 (49.15)
WM (Group A)	503.69 (60.28)	477.79 (62.56)	456.52 (59.26)	409.51 (47.88)
WM (Group B)	499.17 (60.43)	463.96 (61.59)	443.83 (57.81)	417.21 (48.11)
WM (A+B)	502.01 (60.37)	472.66 (62.58)	451.81 (59.04)	412.37 (48.11)
eTIV (Group A)	1344.78 (139.76)	1417.72 (150.31)	1455.58 (148.47)	1375.28 (136.14)
eTIV (Group B)	1374.61 (151.48)	1449.64 (156.96)	1494.19 (170.03)	1387.99 (141.03)
eTIV (A+B)	1355.85 (144.94)	1429.56 (153.58)	1469.90 (157.92)	1379.99 (138.11)

Table 4. Test−Retest Reproducibility for Normal Subjects Re-scanned in Same Scanner Model within 1 Year [n = 56 (26 M, 30 F), Mean Gap Between Scans = 0.39 Y, Mean Age at the Time of Rescan 69.36 Y].

	FSL-FAST	CAT12	FreeSurfer	ANTs
GM	545.08 (+0.50%)	621.80 (+0.07%)	585.71 (+0.11%)	463.21 (+0.28%)
WM	501.14 (–0.47%)	475.82 (+0.07%)	456.09 (+0.04%)	410.50 (+0.04%)
eTIV	1348.37 (+0.10%)	1427.83 (+0.19%)	1459.17 (–0.19%)	1391.86 (–0.16%)

Quality Control

MR images that have failed during any of the automated image processing pipelines have been excluded, similarly, those images for which IQR was estimated to be below 75% have been visually inspected before further analysis.

Statistical Analysis

The mean and standard deviation of GM, WM, and eTIV (in cm^3) from each of the methods across different age bands in normal as well as cognitively declining population has been measured and compared. The agreement between the methods has been analysed with Bland-Altman Plots[19] for each pair of methods, the X-axis in the plot represents the mean measurement of the two methods and the Y-axis denotes the difference between the methods. The lower and upper line of agreements corresponding to a 95% confidence interval is also plotted parallel to the X-axis in the BA plot. Reproducibility of measurement has been evaluated in a subset of the dataset (n = 56) of the normal population having scan-rescan performed within 1 year on the same MRI scanner model.

Results

The mean age of the selected subjects was 70.88 years, the normal group (A) was 6.62 years younger than the cognitive decline group (B). The e-ICV of group B was higher than group A but the GM and WM volumes were consistently reported lower by each of the methods in group B (Table 3) pointing to age-related brain atrophy[20] which may have further been accelerated in the cognitive decline group.[21]

The reproducibility of the four methods tested on a smaller test−retest dataset of n = 56 has a variation of less than 1% pointing to a high degree of reproducibility for each of the four methods. CAT12 with a mean change of 0.07% for both GM and WM was most reproducible for tissue class segmentation followed by FreeSurfer and ANTs (Table 4).

The Bland-Altman plots (Figure 1) for comparing agreement among the methods showed that most of the observations were falling within the 95% confidence interval. The bias between GM and WM measurement (Table 5) was lowest among FreeSurfer and CAT12 methods and the 95% confidence interval was also narrowest.

Discussion

Each of the discussed neuroimaging pipelines used for segmentation has significant technical differences in their implementation and how they define different tissue classes. FSL-FAST and CAT12 pipelines perform partial volume estimation of tissue classes using Markov random field model with expectation-maximization[13] and Adaptive Maximum A Posterior (AMAP) technique,[16] respectively, FreeSurfer uses image intensity along with probabilistic atlas[12] in its segmentation model resulting in fine grain segmentation from which the tissue volumes are computed, ANTs relies on probabilistic tissue segmentation[22] along with machine learning models trained on labelled data in its segmentation approach. The difference in the various methods was observable in GM and WM measurements. The GM and WM tissue observations from the CAT12 and FreeSurfer methods have a higher degree of agreement whereas FSL-FAST has

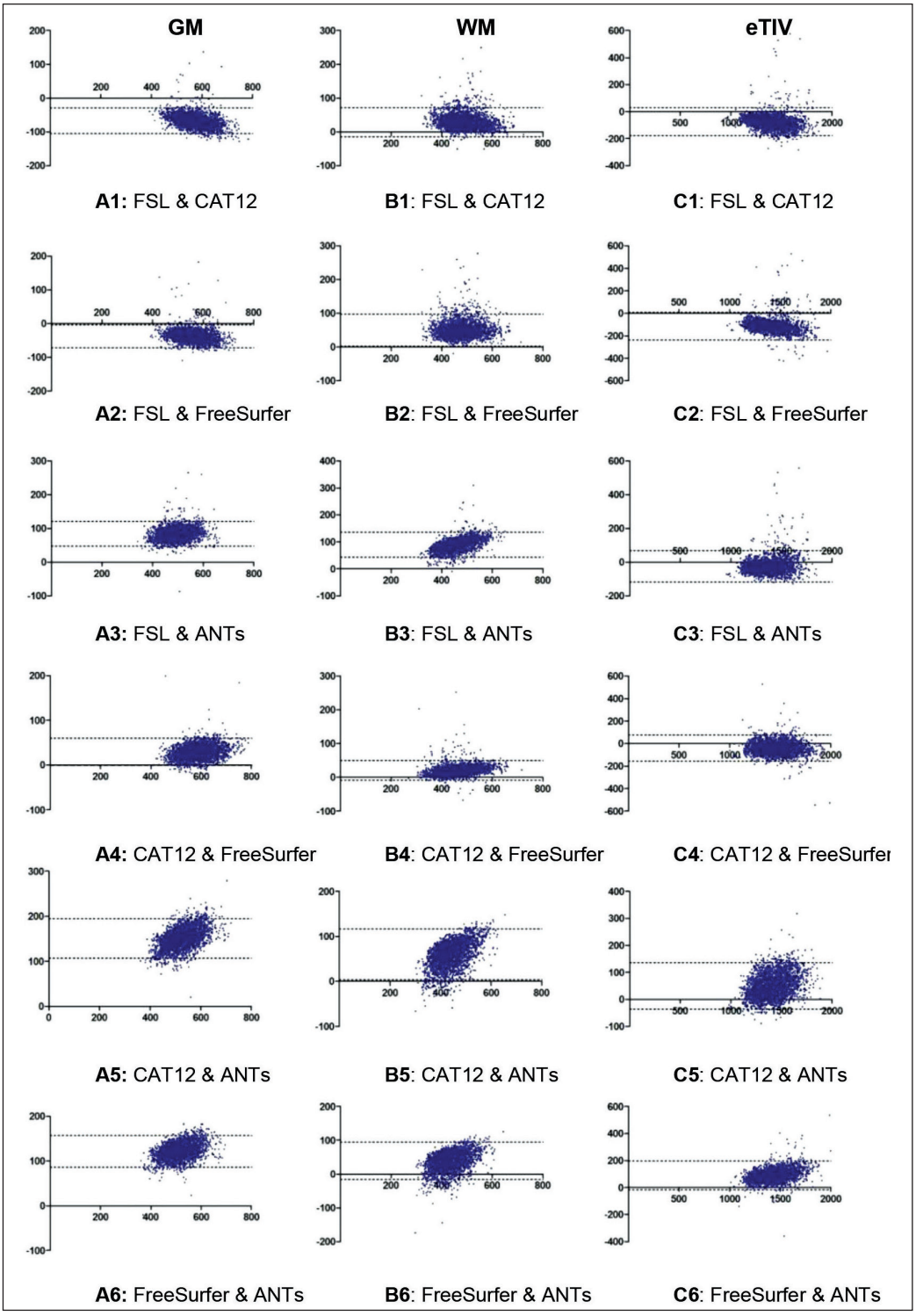

Figure 1. Bland-Altman Plots for Comparing Between Each Pair of Methods for GM, WM, and eTIV.

reported lower GM but higher WM as compared to CAT12 and FreeSurfer. On the other hand, ANTs have consistently underreported GM and WM compared to all other methods and this may be improved by labelled training data. Despite these differences between methods, the observations across subjects for each of the methods were consistent and reproducible as also evident from test-retest reproducibility of less than 1% in the smaller dataset.

However, among the methods CAT12 and FreeSurfer have performed better than others, FreeSurfer performs fine grain

Table 5. Bland-Altman Analysis for Agreement Between Methods.

			95% Limits of Agreement	
N = 1330 (2580 MR sessions)	Bias	SD of bias	From	To
GM				
FSL-FAST and CAT12	−67.00	19.00	−105.00	−29.00
FSL-FAST and FreeSurfer	−37.47	17.39	−71.55	−3.39
FSL-FAST and ANTs	84.15	18.68	47.54	120.80
CAT12 and FreeSurfer	29.00	16.00	−1.20	60.00
CAT12 and ANTs	150.80	22.54	106.60	195.00
FreeSurfer and ANTs	122.00	18.00	87.00	157.00
White Matter (WM)				
FSL-FAST and CAT12	29.00	22.00	−14.00	73.00
FSL-FAST and FreeSurfer	50.00	24.00	2.20	98.00
FSL-FAST and ANTs	90.00	24.00	43.00	136.00
CAT12 and FreeSurfer	20.85	14.83	−8.22	49.92
CAT12 and ANTs	60.00	29.00	3.60	117.00
FreeSurfer and ANTs	39.00	28.00	−16.00	95.00
Estimated Intracranial Volume (eTIV)				
FSL-FAST and CAT12	−74.00	53.00	−178.00	30.00
FSL-FAST and FreeSurfer	−114.10	63.11	−237.70	9.63
FSL-FAST and ANTs	−24.15	47.84	−117.90	69.62
CAT12 and FreeSurfer	−40.34	59.76	−157.50	76.79
CAT12 and ANTs	49.57	43.91	−36.50	135.60
FreeSurfer and ANTs	89.91	53.97	−15.88	195.70

segmentation of cortical and subcortical regions using the Destrieux atlas and the Desikan-Killiany atlas and is suitable for the region of interest studies.

One of the potential drawbacks of the current study is that the selected dataset is primarily composed of the elderly population (mean age = 70.88 years) and the findings may not be representative of younger age groups.

Conclusion

CAT12 and FreeSurfer methods have the highest degree of agreement in tissue class segmentation and are most reproducible compared to others.

Acknowledgements

Data were provided by OASIS, OASIS-3: Longitudinal Multimodal Neuroimaging: Principal Investigators: T. Benzinger, D. Marcus, J. Morris; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. The author would like to thank Dr. Richa Chaturvedi, Professor, School of Computer Science, Starex University for her guidance and support.

Statement of Ethics

Not applicable.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The data analysis has been carried out on computational resources funded by the Department of Biotechnology, Government of India under the erstwhile DIC project.

ORCID iD

Mahender Kumar Singh https://orcid.org/0000-0001-9617-6112

References

1. Bandettini PA. What's new in neuroimaging methods? *Annals of the New York Academy of Sciences* 2009; *1156*(1): 260–293.
2. Lalitanantpong D and Lalitanantpong S. Magnetic resonace imaging study in major psychiatric disorders. *J Med Assoc Thai* 2004; *87*(Suppl 2): S300–S308.
3. Giorgio A and De Stefano N. Clinical use of brain volumetry. *Journal of Magnetic Resonance Imaging*. 2013; *37*(1): 1–14.
4. Raghuprasad MS and Manivannan M. Volumetric and morphometric analysis of pineal and pituitary glands of an Indian inedial subject. *Annals of Neurosciences* 2018; *25*(4): 279–288.
5. Shen L, Firpi HA, Saykin AJ, et al. Parametric surface modeling and registration for comparison of manual and automated segmentation of the hippocampus. *Hippocampus*. 2009; *19*(6): 588–595.
6. Singh MK and Singh KK. A Review of Publicly Available Automatic Brain Segmentation Methodologies, Machine Learning Models, Recent Advancements, and Their Comparison. *Annals of Neurosciences* 2021; *28*(1-2): 82–93.
7. Fellhauer I, Zollner FG, Schroder J, et al. Comparison of automated brain segmentation using a brain phantom and patients with early Alzheimer's dementia or mild cognitive impairment. *Psychiatry Res* 2015; *233*(3): 299–305.
8. Velasco-Annis C, Akhondi-Asl A, Stamm A, et al. Reproducibility of brain MRI segmentation algorithms: Empirical comparison of local MAP PSTAPLE, FreeSurfer, and FSL-FIRST. *Journal of Neuroimaging* 2018; *28*(2): 162–172.
9. Bartel F, Vrenken H, Van Herk M, et al. FAst Segmentation Through SURface Fairing (FASTSURF): A novel semi-automatic hippocampus segmentation method. *PLOS One* 2019; *14*(1): e0210641.
10. Palumbo L, Bosco P, Fantacci ME, et al. Evaluation of the intra- and inter-method agreement of brain MRI segmentation software packages: A comparison between SPM12 and FreeSurfer v6.0. *Physica Medica* 2019; *64*: 261–272.
11. Lamontagne PJ, Benzinger TL, Morris JC, et al. OASIS-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and Alzheimer Disease. 2019.
12. Fischl B. FreeSurfer. *NeuroImage* 2012; *62*(2): 774–781.
13. Jenkinson M, Beckmann CF, Behrens TE, et al. Fsl. *Neuroimage* 2012; *62*(2): 782–790.
14. Zhang Y, Brady M, and Smith S. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Trans Med Imaging* 2001; *20*(1): 45–57.
15. Tustison NJ, Cook PA, Holbrook AJ, et al. The ANTsX ecosystem for quantitative biological and medical imaging. *Scientific Reports*. 2021; *11*(1).
16. Gaser C, Dahnke R, Thompson PM, et al. CAT – A computational anatomy toolbox for the analysis of structural MRI data. 2022.
17. Ashburner J. Computational anatomy with the SPM software. *Magn Reson Imaging* 2009; *27*(8): 1163–1174.
18. Ashburner J. SPM: a history. *Neuroimage* 2012; *62*(2): 791–800.
19. Altman DG and Bland JM. Measurement in medicine: The analysis of method comparison studies. *Journal of the Royal Statistical Society Series D (The Statistician)* 1983; *32*(3): 307–317.
20. Peters R. Ageing and the brain. *Postgraduate Medical Journal* 2006; *82*(964): 84–88.
21. Cole JH, Ritchie SJ, Bastin ME, et al. Brain age predicts mortality. *Mol Psychiatry* 2018; *23*(5): 1385–1392.
22. Avants BB, Tustison NJ, Wu J, et al. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics* 2011; *9*(4): 381–400.