Journal Pre-proof

Brain volumes quantification from MRI in healthy controls: Assessing correlation, agreement and robustness of a convolutional neural network-based software against FreeSurfer, CAT12 and FSL

Hernán Chaves, Francisco Dorr, Martín Elías Costa, María Mercedes Serra, Diego Fernández Slezak, Mauricio F. Farez, Gustavo Sevlever, Paulina Yañez, Claudia Cejas

PII: S0150-9861(20)30280-7

DOI: https://doi.org/10.1016/j.neurad.2020.10.001

Reference: NEURAD 943

To appear in: *Journal of Neuroradiology*

Accepted Date: 19 October 2020

Please cite this article as: { doi: https://doi.org/

Brain volumes quantification from MRI in healthy controls: Assessing correlation, agreement and robustness of a convolutional neural network-based software against FreeSurfer, CAT12 and FSL.

Hernán Chaves, MD[a,b,*]; Francisco Dorr, BsC[b]; Martín Elías Costa, PhD[b]; María Mercedes Serra, MD[a,b]; Diego Fernández Slezak, PhD[b,c,d]; Mauricio F. Farez MD MPH[b,e,f,g]; Gustavo Sevlever, MD PhD[e]; Paulina Yañez, MD[a]; Claudia Cejas, MD[a]

[a] Diagnostic Imaging Department, Fleni, Buenos Aires, Argentina

[b] Entelai, Buenos Aires, Argentina

[c] Universidad de Buenos Aires, Facultad de Ciencias Exactas y Naturales. Departamento de Computación. Buenos Aires, Argentina.

[d] CONICET-Universidad de Buenos Aires. Instituto de Investigación en Ciencias de la

[e] Neurology Department, Fleni, Buenos Aires, Argentina

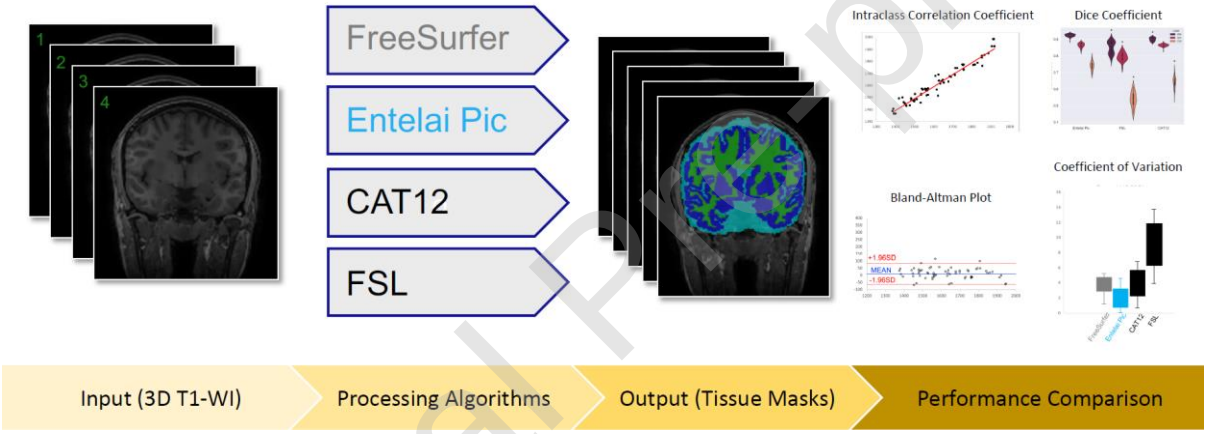
[f] Center for Research on Neuroimmunological Diseases (CIEN), Fleni, Buenos Aires, Argentina.

[g] Center for Biostatistics, Epidemiology and Public Health (CEBES), Fleni, Buenos Aires, Argentina.

***Corresponding author:** Hernán Chaves. E-mail: hchaves@fleni.org.ar. Address: Montañeses 2325 (C1428AQK), Ciudad de Buenos Aires, Argentina. Phone: (+54) 11-5777-3200 (ext.: 2901). ORCID: 0000-0001-8649-6374

WORD COUNT: 3215

Graphical abstract



HIGHLIGHTS

- Evaluated a novel CNN-based model (Entelai Pic) for brain volume estimation.
- Entelai Pic had excellent correlation and agreement with FreeSurfer.
- Entelai Pic provided robust segmentations of brain volumes.
- Post-processing time is 480 minutes for FreeSurfer and 5 minutes for Entelai Pic.
- This novel CNN-based model is suitable for brain volumetry on clinical practice.

ABSTRACT

Background and purpose: There are instances in which an estimate of the brain volume should be obtained from MRI in clinical practice. Our objective is to calculate cross-sectional robustness of a convolutional neural network (CNN) based software (Entelai Pic) for brain volume estimation and compare it to traditional software such as FreeSurfer, CAT12 and FSL in healthy controls (HC).

Materials and Methods: Sixteen HC were scanned four times, two different days on two different MRI scanners (1.5T and 3T). Volumetric T1-weighted images were acquired and post-processed with FreeSurfer v6.0.0, Entelai Pic v2, CAT12 v12.5 and FSL v5.0.9. Whole-brain, grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) volumes were calculated. Correlation and agreement between methods was assessed using intraclass correlation coefficient (ICC) and Bland Altman plots. Robustness was assessed using the coefficient of variation (CV).

Results: Whole-brain volume estimation had better correlation between FreeSurfer and Entelai Pic (ICC (95% CI) 0.96 (0.94-0.97)) than FreeSurfer and CAT12 (0.92 (0.88-0.96)) and FSL (0.87 (0.79-0.91)). WM, GM and CSF showed a similar trend. Compared to FreeSurfer, Entelai Pic provided similarly robust segmentations of brain volumes both on same-scanner (mean CV 1.07, range 0.20–3.13% vs. mean CV 1.05, range 0.21–3.20%, p=0.86) and on different-scanner variables (mean CV 3.84, range 2.49–5.91% vs. mean CV 3.84, range 2.62–5.13%, p=0.96). Mean post-processing times were 480, 5, 40 and 5 minutes for FreeSurfer, Entelai Pic, CAT12 and FSL respectively.

Conclusion: Based on robustness and processing times, our CNN-based model is suitable for cross-sectional volumetry on clinical practice.

KEYWORDS

Magnetic Resonance Imaging; Brain; Deep Learning; Segmentation; Freesurfer.

ABBREVIATIONS

ANTs, Advanced normalization tools

BET, Brain extraction tool

CAT, computational anatomy toolbox

CNN, convolutional neural networks

CSF, cerebrospinal fluid

CV, coefficient of variation

DC, Dice coefficient

DDDS, different-day different-scanner

DDSS, different-day same-scanner

FAST, FMRIB's automated segmentation tool

FSL, FMRIB Software Library

GM, grey matter

HC, healthy controls

ICC, intraclass correlation coefficients

MRI, magnetic resonance images

SDDS, same-day different-scanner

SDSS, same-day same-scanner

SPM, Statistical parametrical mapping

WM, white matter

MANUSCRIPT

Introduction

There are many instances in which an estimate of the brain volume should be obtained from magnetic resonance images (MRI) in clinical practice (e.g.: cognitive impairment, developmental delay, multiple sclerosis, etc.). Even though several visual rating scales have been developed to provide semi quantitative measures of the degree of atrophy, their domain is limited to specific brain regions, subjective –prone to intra and interrater variations– and cumbersome [1].

To be easily integrated in the clinical practice workflow of a radiology department, brain volume quantification methods should be fast, reliable and robust. This is why automated brain volume estimation may be the best method available to be incorporated in clinical practice. There are several available tools for automated brain volume estimation, and these have been tested both in healthy subjects and patients with neurological conditions. These tools can be broadly divided in atlas-based, deformable, region-based and learning-driven methods [2]. The most commonly used software includes FreeSurfer [3], Computational Anatomy Toolbox (CAT12) [4], and the FMRIB Software Library (FSL) [5].

In recent years, learning-based methods –and more specifically convolutional neural networks (CNN)– have grown exponentially outperforming traditional methods and human-level performance [6,7].

Our purpose is to calculate cross-sectional correlation and robustness of a novel CNN-based software (Entelai Pic) for brain volume estimation and compare it to traditional software such as FreeSurfer, CAT12 and FSL in healthy controls (HC).

Material and methods

Subjects and MRI acquisition

We recruited 20 HC for this study. Three subjects did not show on the day of the study. All acquired images were visually inspected by a neuroradiologist. One subject was excluded because of an incidental finding on MRI (white matter lesions and an extra-axial lesion). Sixteen subjects were finally included in the study. The study was approved by the institutional review board and subjects gave informed consent.

All subjects were scanned four times, two different days on two MRI scanners (Philips Achieva 1.5T and GE Discovery 750 3T). Group A included 9 subjects who were scanned twice on the same scanner on the first day and twice on the other scanner on the second day. Group B included 7 subjects who were scanned twice on different scanners on each day. Same-day scans were separated by 30-60 minutes, different-day scans were separated by 1-3 weeks (Fig. 1). On same-day scans, subjects were allowed to drink water and/or use the restrooms, but they were asked not to leave the MRI facilities.

We acquired ~~a 3D FLAIR and A~~ 3D T1-weighted images without contrast administration on each scan session. MRI sequences parameters are summarized in Table 1. This study was approved by an ethics committee and was performed in accordance with the ethical standards laid down in the 1964 Declaration of Helsinki and its later amendments. All persons gave their informed consent prior to their inclusion in the study.

Post-processing

3D T1-weighted images were post-processed with FreeSurfer v6.0.0, Entelai Pic v2, CAT12 v12.5 and FSL v5.0.9. The whole-brain, grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF) volumes were calculated (Fig. 2). ~~3D FLAIR images were not used in any post-processing pipeline.~~ Average post-processing times were calculated for each segmentation software.

FreeSurfer

Volumetric segmentation was performed with the Freesurfer image analysis suite, which is documented and freely available for download online (http://surfer.nmr.mgh.harvard.edu/). Briefly, this processing includes motion correction, removal of non-brain tissue, automated Talairach transformation and intensity normalization for surface and intensity-based segmentation of the cortex, subcortical white matter and deep gray matter volumetric structures [8]. Procedures for the measurement of cortical thickness have been validated against histological analysis [9,10], and manual measurements [11,12]. Segmentation was carried out using the standard 'recon-all' command, which performs all cortical and subcortical reconstruction processes. Using Python, SimpleITK and NumPy, FreeSurfer segmentation output was converted into GM, WM and CSF masks.

Entelai Pic

MRI images are first preprocessed to reduce bias, normalize the brightness distribution (to zero mean and unitary SD) and ensure a homogeneous resolution (1x1x1 mm voxel size). Bias reduction is carried out using the N4BiasFieldCorrection routine from Advanced Normalization Tools (ANTs). All remaining preprocessing steps are performed using Python libraries and SimpleITK. Once this step is completed, images are fed to a series of deep convolutional networks that first separate the brain from the rest of the skull and later produce labeled images for both cortical and subcortical structures. The volume for every structure is calculated from these labeled images. The architecture selected for these deep learning networks is a 20-layer 3D convolutional network with residual connections. This type of architecture has been shown to be especially well suited for 3D parcellation of MRI images into a large number of classes (>100) [13]. The models were trained with over 1,500 visually inspected and corrected FreeSurfer masks. Training was done using Niftynet [14], a deep learning framework for medical images. Optimization was carried out with the Adam algorithm + L2 regularization. Data augmentation transformations included: scalings, rotations, flips and quadratic bias field additions.

CAT12

CAT12 toolbox is a free extension to Statistical Parametrical Mapping 12 (SPM12) to provide computational anatomy (http://www.neuro.uni-jena.de/). 3D T1-weighted images are interpolated, normalized using an affine followed by non-linear registration, denoised, corrected for bias field inhomogeneities, and then segmented into GM, WM, and CSF components. The segmentation approach is based on an

AMAP (Adaptive Maximum A Posterior) technique without the need for a priori information on the tissue probabilities, and Partial Volume Estimation with a simplified mixed model of a maximum of two tissue types [15]. Cross-sectional segmentation was carried out using default settings, activating the option to output GM, WM and CSF masks on patients' native space.

FSL

Brain tissue volume was estimated with FMRIB's Automated Segmentation Tool (FAST), part of FSL toolbox (https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Before running FAST an image of the brain should first be extracted, using Brain Extraction Tool (BET) from FSL toolbox. The resulting brain-only image can then be fed into FAST. FAST starts by extracting brain and skull images from the single whole-head input data. Next, tissue-type segmentation with partial volume estimation is carried out with FAST in order to calculate total volume of brain tissue (including separate estimates of volumes of GM, WM and CSF), whilst also correcting for spatial intensity variations (i.e.: bias field). The underlying method is based on a hidden Markov random field model and an associated Expectation-Maximization algorithm. The whole process is fully automated. We used default processing parameters, activating the option to output binary segmentation images of 3 tissue classes.

~~Brain tissue volume, normalized for subject head size, was estimated with SIENAX, part of FSL [16]. SIENAX starts by extracting brain and skull images from the single whole-head input data. The brain image is then affine-registered to MNI152 (using the skull image to determine the registration scaling); this is primarily in order to~~

~~obtain the volumetric scaling factor, to be used as normalization for head size. Next, tissue-type segmentation with partial volume estimation is carried out in order to calculate total volume of brain tissue (including separate estimates of volumes of GM, WM and CSF).~~

Statistical analysis

To obtain a concurrent estimate of consistency and agreement between volumes derived from the different segmentation techniques, we computed intraclass correlation coefficients (ICC) [16]. An ICC value of 1 indicates a perfect reproducibility between two (or more) raters and a value of 0 or less, a reproducibility that is lower than what is expected on the basis of chance alone. A strong correlation would confirm a good consistency between techniques. ICCs were computed automatically specifying a two-way mixed-effect model.

To ~~further~~ investigate agreement between volumes derived from different techniques, we computed Bland–Altman plots. This graphical method is used to illustrate differences in estimation between two techniques or raters [17]. Bland–Altman plots are created using the mean of the two studied techniques as the estimation of reference.

To assess spatial agreement, we used Dice coefficient (DC) between the output segmentations generated by the different software, using FreeSurfer as a gold standard [18]. FreeSurfer segmentations were visually checked and manually corrected or excluded by a neuroradiologist with experience in segmentation. A DC of 1 indicates a perfect spatial agreement between 2 segmentations.

The robustness (repeatability and reproducibility) of repeated measures was assessed using the within-subject coefficient of variation (CV). CV may be defined as the ratio of the standard deviation of a number of measurements to the arithmetic mean [19]. A software is considered to be robust if its output is consistently accurate even if one or more of the input variables are changed.

All the statistical analyses were performed using STATA version 14. Group comparisons between ~~the four~~ software were tested using the Kruskal-Wallis rank test, and in case of significant differences among the software, *post hoc* paired analysis were performed using the Wilcoxon rank-sum test. A $p<0.05$ was considered statistically significant.

Results

Subjects

Sixteen healthy subjects were included in the study (8 females and 8 males, age range: 25 to 37 years, mean age = 30.4 ± 2.9 years). Subjects reported no history of neurological or psychiatric disease.

For CV estimation, four variables were defined (Fig. 3): same-day same-scanner (SDSS), different-day same-scanner (DDSS), same-day different-scanner (SDDS) and different-day different-scanner (DDDS).

Correlation and agreement of brain volumetry

There were differences in the numeric brain tissue segmentation output from FreeSurfer, Entelai Pic, CAT12 and FSL as detailed in Table 2.

Whole brain volume estimation had better correlation between FreeSurfer and Entelai Pic (ICC (95% CI) 0.96 (0.94-0.97)) than FreeSurfer compared to CAT12 (0.92 (0.88-0.96)) and FSL (0.87 (0.79-0.91)). WM, GM and CSF volume estimation showed a similar trend (Table 3).

Bland-Altman plots show better agreement between whole brain, WM, GM and CSF volume measures obtained by FreeSurfer and Entelai Pic, compared to the agreement between FreeSurfer and CAT12 or FSL (Fig. 4).

Entelai Pic tissue segmentation masks had also better spatial agreement with FreeSurfer when compared to FSL and CAT12. WM mask had the highest (mean DC (range) 0.89 (0.77-0.94) and CSF had the lowest (0.64 (0.45-0.80)) spatial agreement in all three methods (Table 4 and Fig. 5). To further analyze the difference observed in the CSF mask, we averaged the differences in CSF segmentations for the three evaluated methods and our gold standard (Fig. 6). Each method has a unique error pattern, convexal subarachnoid space CSF segmentation seems clearly problematic for all three. Ventricles segmentation was fairly similar among the three methods when compared to the gold standard. GM masks had a good spatial agreement for all three methods (0.84 (0.72-0.88)). We analyzed the GM masks in the same fashion as CSF masks. Differences were higher in the striatum, thalami and mesencephalon for FSL and CAT12 compared to the gold standard (Fig. 7).

Robustness of brain volumetry

Mean CV for the whole brain, GM, WM and CSF from FreeSurfer, Entelai Pic, CAT 12 and FSL are summarized in Table 5. Whole brain segmentation had the lowest CV among all methods, except for Entelai Pic in which GM segmentation had the lowest CV. On the contrary CSF segmentation was the tissue class that had the highest CV among all methods.

FreeSurfer and Entelai Pic mean CV were very similar among all tissue classes. CAT12, achieved the highest robustness of all in whole brain segmentation (mean CV 0.57) and the lowest robustness of all in CSF segmentation (mean CV 9.68). CAT12 GM and WM segmentation robustness were slightly worse compared to FreeSurfer and Entelai Pic. Mean FSL CV was higher than the other three methods for all tissue classes, except for CSF when compared to CAT12. All methods CV were lowest among same-scanner (SDSS and DDSS) compared to different-scanner (SDDD and DDDS) variables (CV range 0.31–2.70% vs. 2.75–5.49%, $p<0.0001$) (Fig. 8).

Compared to FreeSurfer, Entelai Pic provided similarly robust segmentations of brain volumes both on same-scanner (mean CV 1.07, range 0.20–3.13% vs. mean CV 1.05, range 0.21–3.20%, $p=0.86$) and on different-scanner variables (mean CV 3.84, range 2.49–5.91% vs. mean CV 3.84, range 2.62–5.13%, $p=0.96$). Specifically, whole brain different-scanner CV were statistically significant lower in FreeSurfer compared to Entelai Pic (mean CV 2.58 vs 4.27, $p=0.0058$ for SDDS and 3.42 vs 4.69, $p=0.016$ for DDDS) and, on the other hand GM different-scanner CV were statistically significant higher in FreeSurfer compared to Entelai Pic (mean CV 3.48 vs 1.48, $p=0.001$ for SDDS and 3.64 vs 2.02, $p=0.0013$ for DDDS).

Post-processing times

Mean post-processing times for each 3D T1-WI was 480 ± 10 minutes for FreeSurfer, 5 minutes ± 30 seconds for Entelai Pic, 40 ± 2 minutes for CAT12 and 15 ± 1 minutes for FSL. These times were measured on a g4dn.2xlarge Amazon Web Services instance with GPU. Intervals correspond to estimates of the standard deviation for those times.

Discussion

Deep learning is a subset of machine learning that learns representations of data based on models composed of multiple processing layers [20]. Deep learning algorithms, and specifically CNN models, are starting to be applied to medical image analysis in general [21], gaining an important role in the process of brain segmentation in the field of neuroimaging [22]. Here, we present a CNN-based software for brain tissue segmentation, and we compare it to other well-known traditional brain segmentation software used in the neuroimaging field such as FreeSurfer, CAT12 and FSL. FreeSurfer is considered as one of the most accurate brain segmentation tools that is available and has been used as a gold standard by many authors as it has been proven to have a good agreement with histologic and manual measurements of cortical thickness as it has been previously commented. We have also decided to use FreeSurfer´s segmentations as gold standard to assess correlation, agreement and robustness. One of FreeSurfer´s most mentioned drawbacks is that long processing times are inevitable [23]. In this regard, Entelai Pic reduced processing times in several orders of magnitude.

Multiple CNN-based methods for normal brain segmentation have been published in recent years [24–33]. Most of these papers only report their performance in terms of spatial agreement. As our main goal was to assess the robustness of our model, we opted to add a CV analysis like Guo and colleagues, after verifying for brain volume correlation and agreement [34]. As with other methods, CV was lowest among the same-scanner compared to different-scanner. As it has been previously reported, we found greater variability between 1.5T and 3T when measuring different brain structures [35–37]. This could be due to different contrast-noise ratio and signal-noise ratio among volumetric acquisitions related to the different static magnetic fields on 1.5T and 3T scanners [38]. It is also worth mentioning the fact that the scanners were manufactured by different companies, and that the 3D T1-weighted images parameters were not exactly identical. However, it is difficult to estimate which of these factors had more weight in the observed differences. As it has been previously proved, this variation between scanners could be partially compensated by reporting volumes normalized to intracranial volume [34].

CNN-based models have been built not only for whole brain tissues and structures segmentation. There are also deep learning segmentation algorithms with specific purposes such as subcortical structures segmentation [39–42], striatum segmentation [43], or brain ventricles parcellation [44], to mention a few. Although our paper only analyzed the segmentation performance of our model in the main brain tissue classes (WM, GM and CSF), it also performs cortical and subcortical structures segmentation.

In their paper Moeskops and colleagues, assuming that many patients have WM hyperintensities of presumed vascular origin, have included the segmentation of these lesions with a different tissue class than normal WM, GM and CSF using a

multi-scale CNN with FLAIR and T1-weighted images as input [26]. For this paper, we recruited healthy controls and specifically excluded patients with WM hyperintensities, so this approach was not necessary, albeit this method could be incorporated in future versions of our model to make it more robust among all kind of patients.

As it has been previously mentioned, labelling 3D brain images requires laborious efforts by expert anatomists because of the differences among images in terms of their noise, contrast, or ambiguous boundaries [45]. To overcome this difficulty, Ito and colleagues trained a deep neural network on a small number of annotated images, but also a large number of unlabeled images by leveraging image registration to attach pseudo-labels to images that were originally unlabeled [30]. To elude these difficulties, we opted to train our model based only on a large number of FreeSurfer segmentation masks that were visually checked and manually corrected or excluded by a neuroradiologist with experience in segmentation.

This work has some limitations. First of all, we included only healthy patients to evaluate the performance of our brain tissue segmentation model. There are many issues regarding brain tissue segmentation in patients with WM hyperintensities of presumed vascular origin, multiple sclerosis, brain malformations or tumors that were not contemplated in this work. Second, we only included adult patients. Tissue segmentation is more complex in pediatric patients who have not reached a complete myelination, this problem is particularly difficult to surpass in the isointense stage (approximately 6–8 months of age) were WM and GM exhibit almost the same level of intensity in both T1- and T2-weighted images. Zhang and colleagues used a CNN-based method for segmenting isointense stage brain tissues using multi-modality MR images [25]. They showed that their CNN approach outperforms prior

methods and classical machine learning algorithms like support vector machine and random forest classifiers. Nie and colleagues extended the conventional CNN architectures from 2D to 3D, and integrated coarse and dense feature maps to better model tiny tissue regions [33]. They obtained improved results compared to Zhang and colleagues [25]. Third, we trained our CNN model with neuroradiologists-curated FreeSurfer masks, and we also used FreeSurfer masks as the gold standard. So, it could be expected to find better agreement between the segmentations produced by our model and the defined gold standard.

Conclusions

In this paper, we developed a CNN-based model for automatically segmenting brain tissues from 3D T1-weighted images, named Entelai Pic, and analyzed its performance.

Our results show that consistent use of the same scanner is essential for accurate brain volume estimation with both CNN and traditional brain segmentation software. Based on robustness and processing times, our CNN-based model is particularly suitable for cross-sectional volumetry on clinical practice.

AUTHOR CONTRIBUTIONS

All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by Hernán Chaves, Francisco Dorr, Martín Elías Costa and María Mercedes Serra. The first draft of the manuscript was

written by Hernán Chaves and all authors commented on previous versions of the manuscript. All authors read and approved the final version of the manuscript.

CONFLICT OF INTEREST STATEMENT

Hernán Chaves: has received stipends as a medical advisor from Entelai.

Francisco Dorr: Entelai employee.

Martín Elías Costa: Entelai employee.

María Mercedes Serra: Entelai employee.

Diego Fernández Slezak: is CTO and co-founder of Entelai.

Mauricio Franco Farez: is CEO and co-founder of Entelai.

Gustavo Sevlever: none

Paulina Yañez: none

Claudia Cejas: none

REFERENCES

[1] Harper L, Barkhof F, Fox NC, Schott JM. Using visual rating to diagnose dementia: a critical evaluation of MRI atrophy scales. J Neurol Neurosurg Psychiatry 2015;86:1225–33. https://doi.org/10.1136/jnnp-2014-310090.

[2] González-Villà S, Oliver A, Valverde S, Wang L, Zwiggelaar R, Lladó X. A review on brain structures segmentation in magnetic resonance imaging. Artificial Intelligence in Medicine 2016;73:45–69. https://doi.org/10.1016/j.artmed.2016.09.001.

[3] Fischl B. FreeSurfer. NeuroImage 2012;62:774–81. https://doi.org/10.1016/j.neuroimage.2012.01.021.

[4] Ashburner J, Friston KJ. Unified segmentation. NeuroImage 2005;26:839–51. https://doi.org/10.1016/j.neuroimage.2005.02.018.

[5] Jenkinson M, Beckmann CF, Behrens TEJ, Woolrich MW, Smith SM. FSL. NeuroImage 2012;62:782–90. https://doi.org/10.1016/j.neuroimage.2011.09.015.

[6] Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in Neural Information Processing Systems, 2012.

[7] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. Proceedings of the IEEE international conference on computer vision, 2015, p. 1026–1034.

[8] Fischl B, Salat DH, Busa E, Albert M, Dieterich M, Haselgrove C, et al. Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron 2002;33:341–55.

[9] Cardinale F, Chinnici G, Bramerio M, Mai R, Sartori I, Cossu M, et al. Validation of FreeSurfer-estimated brain cortical thickness: comparison with histologic measurements. Neuroinformatics 2014;12:535–42. https://doi.org/10.1007/s12021-014-9229-2.

[10] Rosas HD, Liu AK, Hersch S, Glessner M, Ferrante RJ, Salat DH, et al. Regional and progressive thinning of the cortical ribbon in Huntington's disease. Neurology 2002;58:695–701. https://doi.org/10.1212/wnl.58.5.695.

[11] Kuperberg GR, Broome MR, McGuire PK, David AS, Eddy M, Ozawa F, et al. Regionally localized thinning of the cerebral cortex in schizophrenia. Arch Gen Psychiatry 2003;60:878–88. https://doi.org/10.1001/archpsyc.60.9.878.

[12] Salat DH, Buckner RL, Snyder AZ, Greve DN, Desikan RSR, Busa E, et al. Thinning of the cerebral cortex in aging. Cereb Cortex 2004;14:721–30. https://doi.org/10.1093/cercor/bhh032.

[13] Li W, Wang G, Fidon L, Ourselin S, Cardoso MJ, Vercauteren T. On the Compactness, Efficiency, and Representation of 3D Convolutional Networks: Brain Parcellation as a Pretext Task. ArXiv:170701992 [Cs] 2017;10265:348–60. https://doi.org/10.1007/978-3-319-59050-9_28.

[14] Gibson E, Li W, Sudre C, Fidon L, Shakir DI, Wang G, et al. NiftyNet: a deep-learning platform for medical imaging. Computer Methods and Programs in Biomedicine 2018;158:113–22. https://doi.org/10.1016/j.cmpb.2018.01.025.

[15] Tohka J, Zijdenbos A, Evans A. Fast and robust parameter estimation for statistical partial volume models in brain MRI. Neuroimage 2004;23:84–97. https://doi.org/10.1016/j.neuroimage.2004.05.007.

[16] Shrout PE, Fleiss JL. Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 1979;86:420–8. https://doi.org/10.1037/0033-2909.86.2.420.

[17] Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;8:307–10.

[18] Dice LR. Measures of the Amount of Ecologic Association Between Species. Ecology 1945;26:297–302. https://doi.org/10.2307/1932409.

[19] Hendricks WA, Robey KW. The Sampling Distribution of the Coefficient of Variation. The Annals of Mathematical Statistics 1936;7:129–32.

[20] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature 2015;521:436–44. https://doi.org/10.1038/nature14539.

[21] Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, et al. A survey on deep learning in medical image analysis. Medical Image Analysis 2017;42:60–88. https://doi.org/10.1016/j.media.2017.07.005.

[22] Akkus Z, Galimzianova A, Hoogi A, Rubin DL, Erickson BJ. Deep Learning for Brain MRI Segmentation: State of the Art and Future Directions. Journal of Digital Imaging 2017;30:449–59. https://doi.org/10.1007/s10278-017-9983-4.

[23] Seiger R, Ganger S, Kranz GS, Hahn A, Lanzenberger R. Cortical Thickness Estimations of FreeSurfer and the CAT12 Toolbox in Patients with Alzheimer's Disease and Healthy Controls. J Neuroimaging 2018;28:515–23. https://doi.org/10.1111/jon.12521.

[24] de Brebisson A, Montana G. Deep neural networks for anatomical brain segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Boston, MA, USA: IEEE; 2015, p. 20–8. https://doi.org/10.1109/CVPRW.2015.7301312.

[25] Zhang W, Li R, Deng H, Wang L, Lin W, Ji S, et al. Deep convolutional neural networks for multi-modality isointense infant brain image segmentation. NeuroImage 2015;108:214–24. https://doi.org/10.1016/j.neuroimage.2014.12.061.

[26] Moeskops P, de Bresser J, Kuijf HJ, Mendrik AM, Biessels GJ, Pluim JPW, et al. Evaluation of a deep learning approach for the segmentation of brain tissues and white matter hyperintensities of presumed vascular origin in MRI. NeuroImage: Clinical 2018;17:251–62. https://doi.org/10.1016/j.nicl.2017.10.007.

[27] Moeskops P, Viergever MA, Mendrik AM, de Vries LS, Benders MJNL, Isgum I. Automatic Segmentation of MR Brain Images With a Convolutional Neural Network. IEEE Trans Med Imaging 2016;35:1252–61. https://doi.org/10.1109/TMI.2016.2548501.

[28] Bao S, Chung ACS. Multi-scale structured CNN with label consistency for brain MR image segmentation. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization 2018;6:113–7. https://doi.org/10.1080/21681163.2016.1182072.

[29] Chen H, Dou Q, Yu L, Qin J, Heng P-A. VoxResNet: Deep voxelwise residual networks for brain segmentation from 3D MR images. NeuroImage 2018;170:446–55. https://doi.org/10.1016/j.neuroimage.2017.04.041.

[30] Ito R, Nakae K, Hata J, Okano H, Ishii S. Semi-supervised deep learning of brain tissue segmentation. Neural Networks 2019;116:25–34. https://doi.org/10.1016/j.neunet.2019.03.014.

[31] Guha Roy A, Conjeti S, Navab N, Wachinger C. QuickNAT: A fully convolutional network for quick and accurate segmentation of neuroanatomy. NeuroImage 2019;186:713–27. https://doi.org/10.1016/j.neuroimage.2018.11.042.

[32] Wachinger C, Reuter M, Klein T. DeepNAT: Deep convolutional neural network for segmenting neuroanatomy. NeuroImage 2018;170:434–45. https://doi.org/10.1016/j.neuroimage.2017.02.035.

[33] Nie D, Wang L, Adeli E, Lao C, Lin W, Shen D. 3-D Fully Convolutional Networks for Multimodal Isointense Infant Brain Image Segmentation. IEEE Transactions on Cybernetics 2019;49:1123–36. https://doi.org/10.1109/TCYB.2018.2797905.

[34] 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. European Radiology 2019;29:1355–64. https://doi.org/10.1007/s00330-018-5710-x.

[35] Chow N, Hwang KS, Hurtz S, Green AE, Somme JH, Thompson PM, et al. Comparing 3T and 1.5T MRI for Mapping Hippocampal Atrophy in the Alzheimer's Disease Neuroimaging Initiative. AJNR Am J Neuroradiol 2015;36:653–60. https://doi.org/10.3174/ajnr.A4228.

[36] Chu R, Tauhid S, Glanz BI, Healy BC, Kim G, Oommen VV, et al. Whole Brain Volume Measured from 1.5T versus 3T MRI in Healthy Subjects and Patients with Multiple Sclerosis. J Neuroimaging 2016;26:62–7. https://doi.org/10.1111/jon.12271.

[37] Chu R, Hurwitz S, Tauhid S, Bakshi R. Automated segmentation of cerebral deep gray matter from MRI scans: effect of field strength on sensitivity and

reliability. BMC Neurol 2017;17:172. https://doi.org/10.1186/s12883-017-0949-4.

[38] Fushimi Y, Miki Y, Urayama S-I, Okada T, Mori N, Hanakawa T, et al. Gray matter-white matter contrast on spin-echo T1-weighted images at 3 T and 1.5 T: a quantitative comparison study. Eur Radiol 2007;17:2921–5. https://doi.org/10.1007/s00330-007-0688-9.

[39] Milletari F, Ahmadi S-A, Kroll C, Plate A, Rozanski V, Maiostre J, et al. Hough-CNN: Deep Learning for Segmentation of Deep Brain Regions in MRI and Ultrasound. ArXiv:160107014 [Cs] 2016.

[40] Shakeri M, Tsogkas S, Ferrante E, Lippe S, Kadoury S, Paragios N, et al. Sub-cortical brain structure segmentation using F-CNN's. ArXiv:160202130 [Cs] 2016.

[41] Kushibar K, Valverde S, González-Villà S, Bernal J, Cabezas M, Oliver A, et al. Automated sub-cortical brain structure segmentation combining spatial and deep convolutional features. Medical Image Analysis 2018;48:177–86. https://doi.org/10.1016/j.media.2018.06.006.

[42] Dolz J, Desrosiers C, Ben Ayed I. 3D fully convolutional networks for subcortical segmentation in MRI: A large-scale study. NeuroImage 2018;170:456–70. https://doi.org/10.1016/j.neuroimage.2017.04.039.

[43] Choi H, Jin KH. Fast and robust segmentation of the striatum using deep convolutional neural networks. J Neurosci Methods 2016;274:146–53. https://doi.org/10.1016/j.jneumeth.2016.10.007.

[44] Shao M, Han S, Carass A, Li X, Blitz AM, Shin J, et al. Brain ventricle parcellation using a deep neural network: Application to patients with

ventriculomegaly. NeuroImage: Clinical 2019;23:101871. https://doi.org/10.1016/j.nicl.2019.101871.

[45] Hanbury A. A survey of methods for image annotation. Journal of Visual Languages & Computing 2008;19:617–27. https://doi.org/10.1016/j.jvlc.2008.01.002.

FIGURES

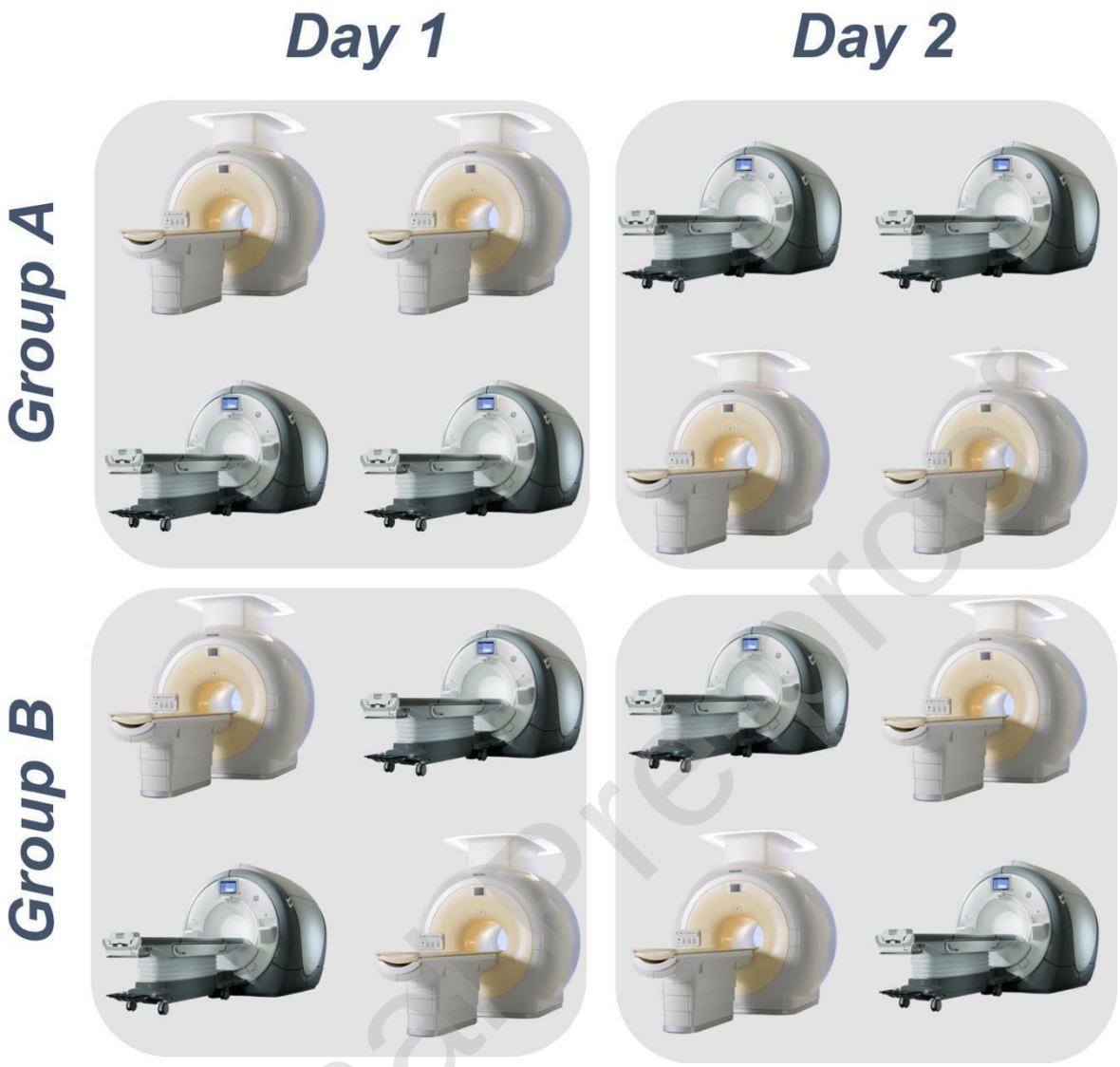


Fig. 1 Subjects were divided in two groups and scanned on two different days. On group A, subjects were scanned on the same scanner on day 1 and on the other scanner on day 2. First day scans were acquired either on the Philips 1.5T scanner (first row) or on the GE 3.0T scanner (second row). On group B, subjects were scanned on different scanners both on day 1 and 2. First day scan were acquired on the Philips 1.5T scanner and GE 3.0T scanner subsequently (third row) or on the GE 3.0T scanner and Philips1.5T scanner subsequently (fourth row)

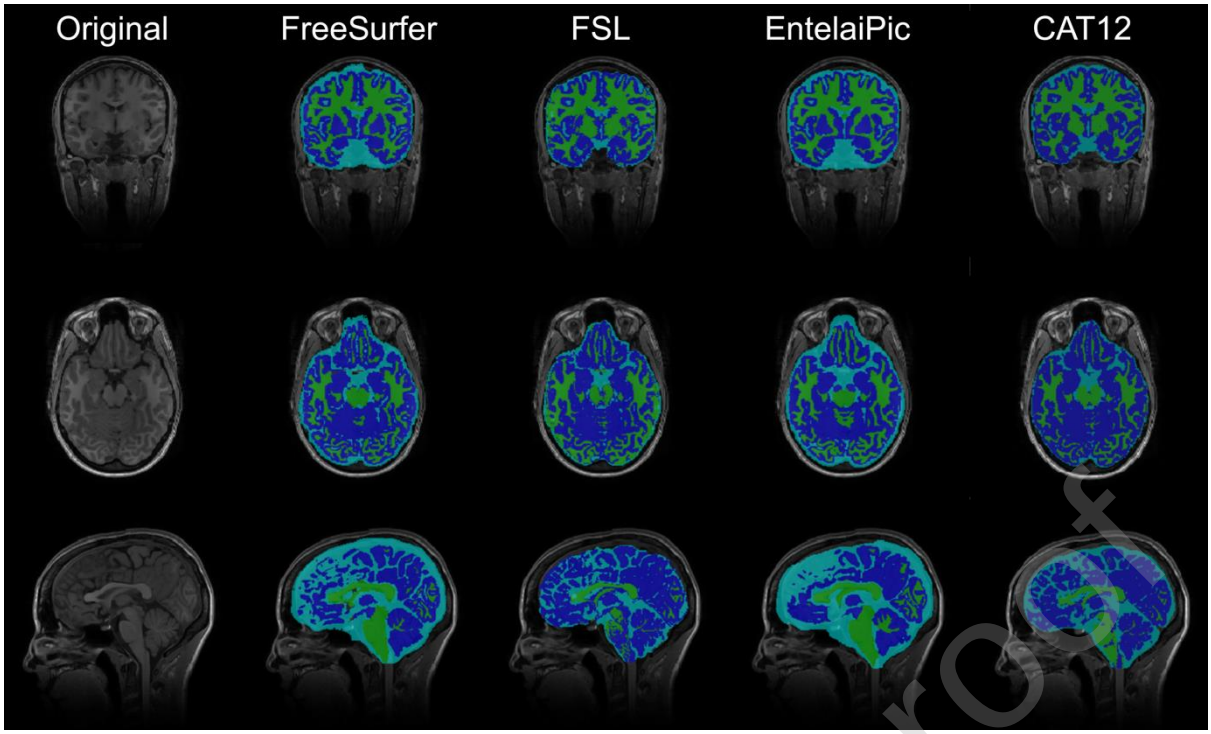


Fig. 2 Original 3D T1-weighted images and segmentation masks from a subject obtained by FreeSurfer, FSL, Entelai Pic and CAT12. Coronal (top row), axial (middle row) and sagittal (bottom row) images are shown. Color coded segmentations shown include WM (green), GM (blue) and CSF (teal)

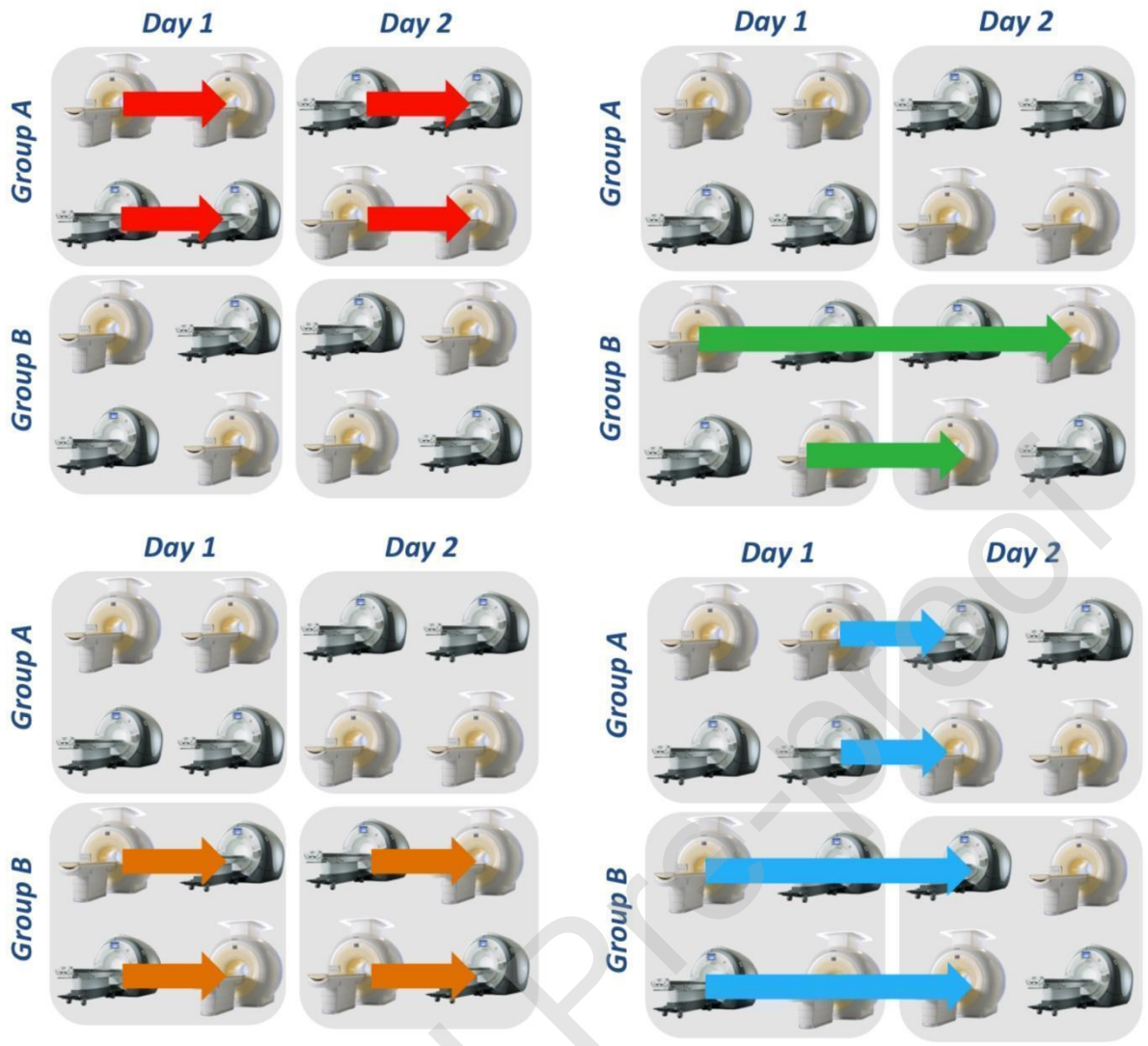


Fig. 3 Graphical explanation of the four variables defined for CV estimation based on the MRI acquisition design: same-day same-scanner, SDSS (red); different-day same-scanner, DDSS (green); same-day different-scanner, SDDS (orange); and different-day different-scanner, DDDS (blue)

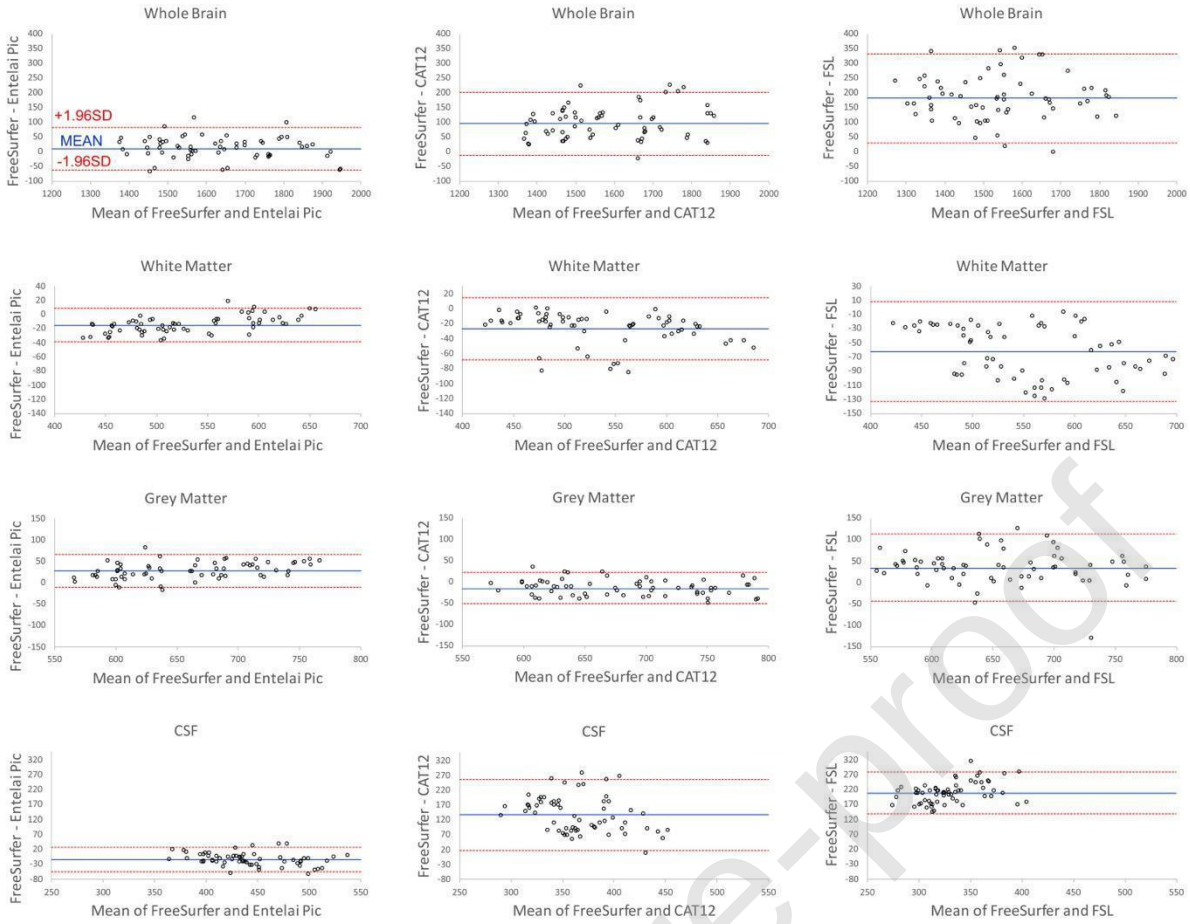


Fig. 4 Bland-Altman plots depicting the agreement between quantitative measurements obtained by Entelai Pic (first column), CAT12 (second column) and FSL (third column) compared to FreeSurfer. Whole brain (first row), WM (second row), GM (third row) and CSF (fourth row) volumes comparisons are shown

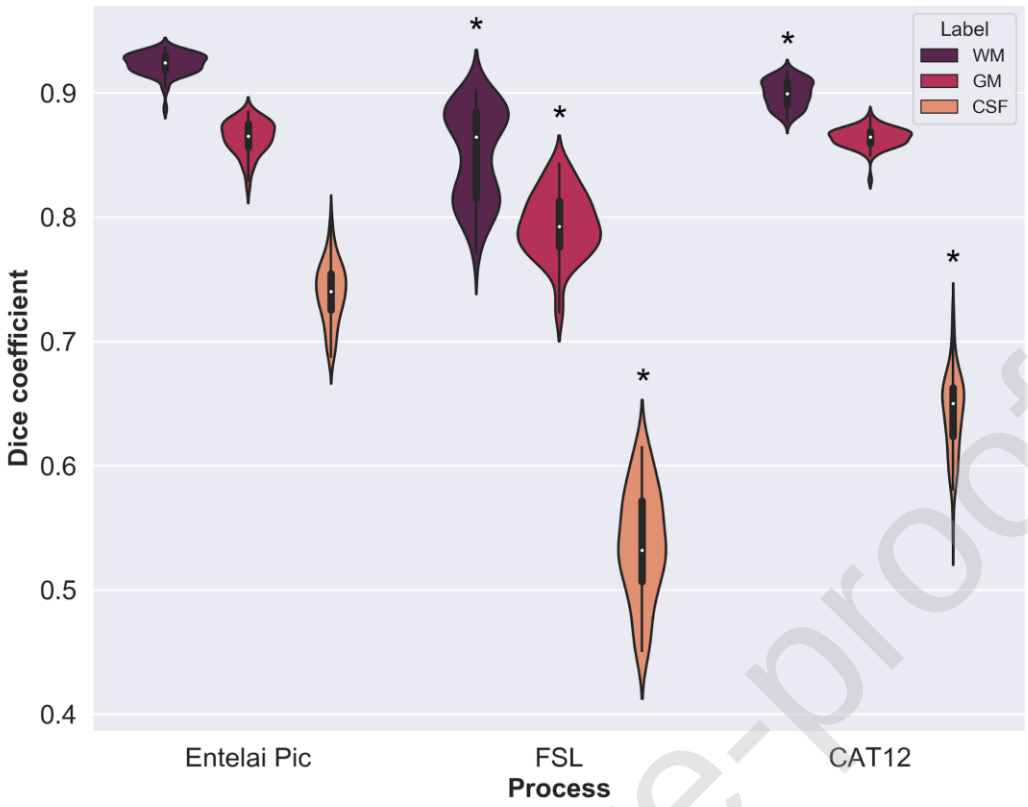


Fig. 5 Distribution of Dice coefficients for each method and tissue type are shown as violin plots. Notice in every method WM Dice > GM Dice > CSF Dice. Statistically significant differences with FreeSurfer ($p<0.05$) are marked with an asterisk (*)

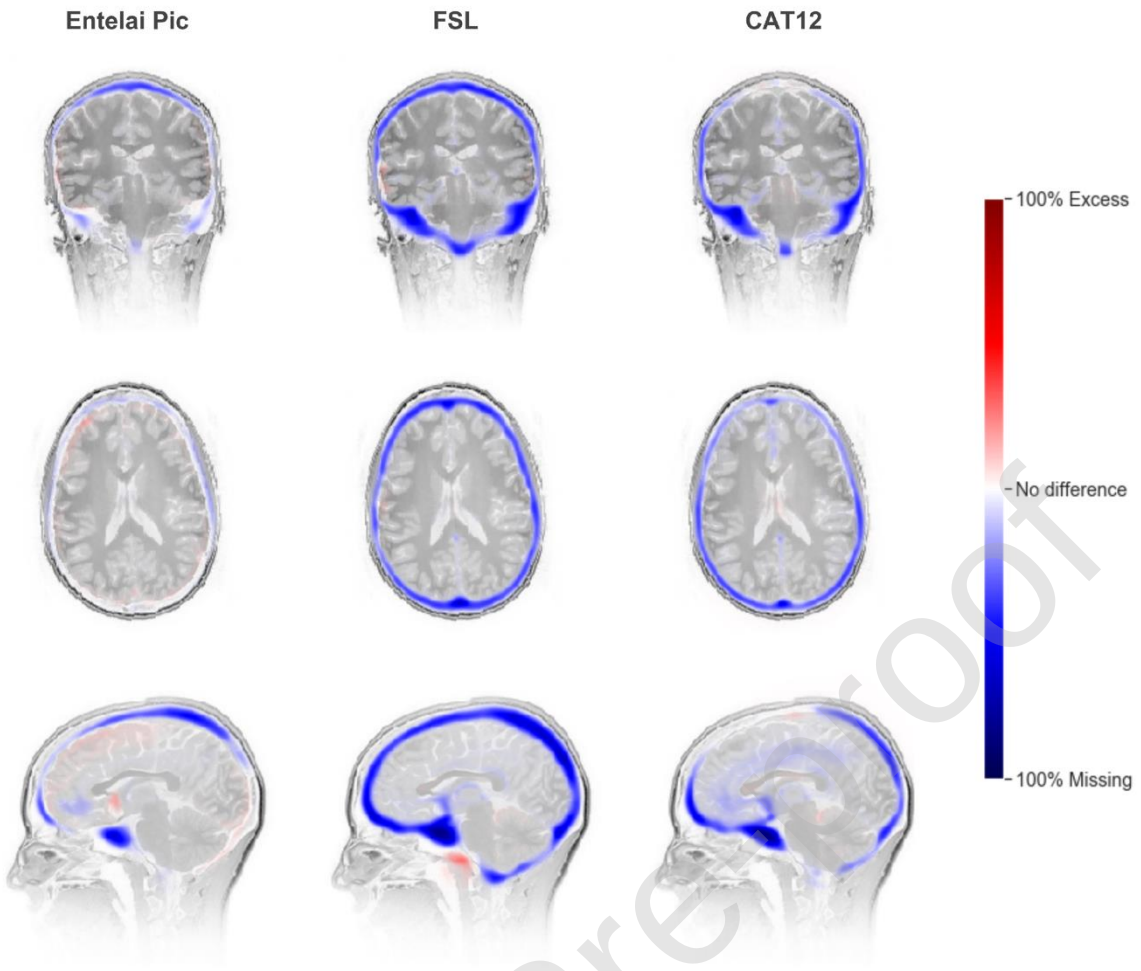


Fig. 6 Average differences in CSF segmentations for the three evaluated methods (viewed in coronal, axial and sagittal slices). Each voxel is colored according to the average discrepancy between the segmentations of each method and our gold standard (FreeSurfer). Voxels where the method tends to undersegment (i.e.: voxels which are not marked as CSF and should be) are painted blue. Conversely, voxels that were segmented as CSF and should have been labeled as something else, are marked in red. Difference maps for all sessions were registered to a reference image (in this case the first session of the first subject) before averaging. Notice each method has a unique error pattern, though some areas are clearly problematic for all three.

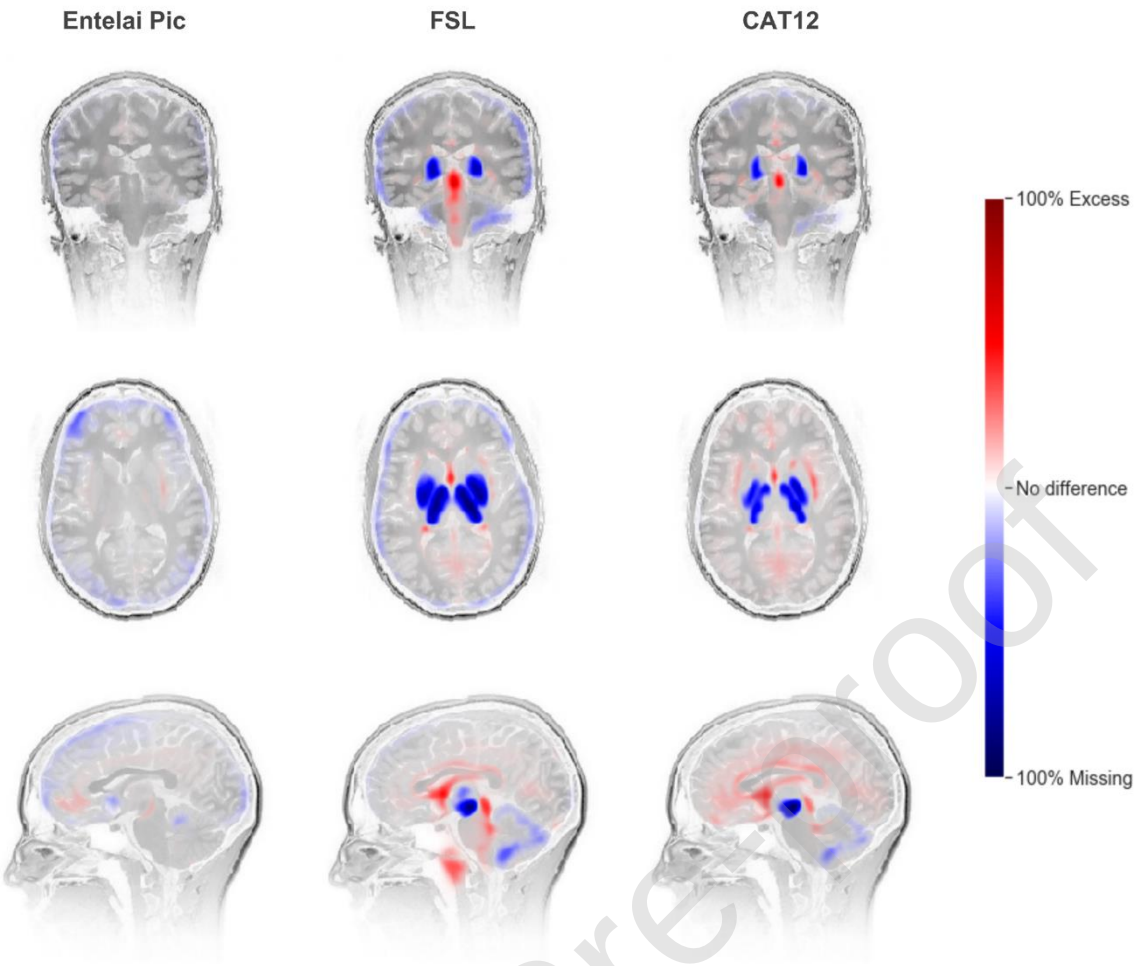


Fig. 7 Average differences in GM segmentations for the three evaluated methods (viewed in coronal, axial and sagittal slices). Each voxel is colored according to the average discrepancy between the segmentations of each method and our gold standard (FreeSurfer). Voxels where the method tends to undersegment (i.e.: voxels which are not marked as GM and should be) are painted blue. Conversely, voxels that were segmented as GM and should have been labeled as something else, are marked in red. Difference maps for all sessions were registered to a reference image (in this case the first session of the first subject) before averaging. Notice some areas are clearly problematic for FSL and CAT12.

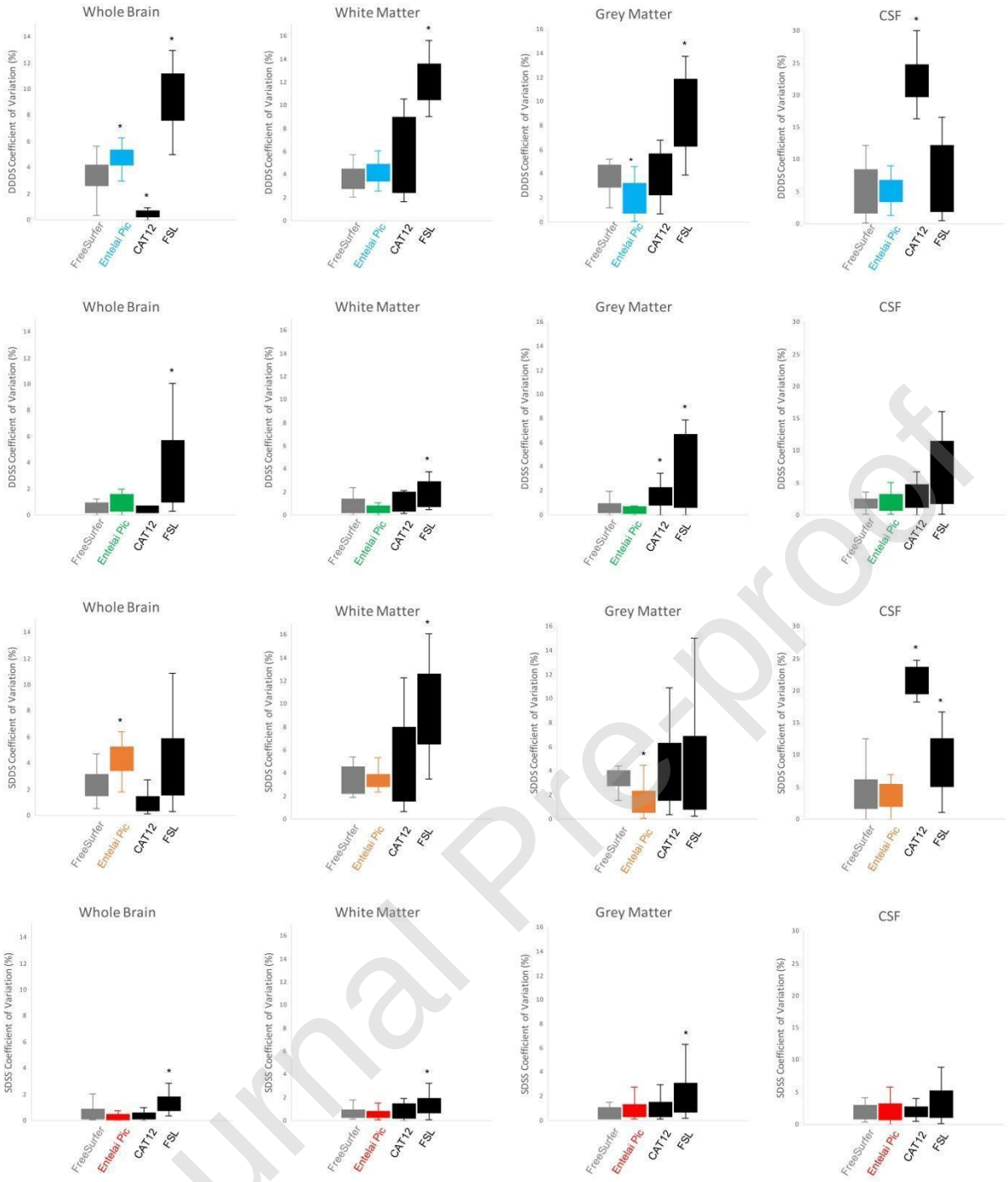


Fig. 8 Box and whisker plots depicting CV estimation: same-day same-scanner, SDSS (first row); different-day same-scanner, DDSS (second row); same-day different-scanner, SDDS (third row); and different-day different-scanner, DDDS (fourth row). CV for the whole brain (first column), WM (second column), GM (third column) and CSF (fourth column) are shown. Statistically significant differences with FreeSurfer ($p<0.05$) are marked with an asterisk (*)

TABLES

Table 1

3D T1-WI acquisition parameters

Scanner	GE Discovery 750	Philips Achieva
Field Strength	3T	1.5T
Field-of-view, mm2	250	256
Number of acquisitions	1	1
Repetition time, ms	8.19	7.14
Inversion time, ms	450	-
Echo time, ms	3.2	3.4
Flip angle, °	12	8
Voxel size, mm	1 x 1 x 1.2	0.7 x 0.7 x 1.2

ms: milliseconds, wi: weighted images

Table 2

Mean ± SD of brain volume estimation (in cm3)

	FreeSurfer	Entelai Pic	CAT12	FSL
Whole brain	1635 ± 154	1624 ± 159	1536 ± 145	1450 ± 155
Grey matter	675 ± 61	647 ± 53	668 ± 65	639 ± 66
White matter	524 ± 67	539 ± 60	550 ± 73	586 ± 80
CSF	436 ± 40	449 ± 46	297 ± 55	225 ± 28

CSF: cerebrospinal fluid

Table 3

ICC (range) of WM, GM and CSF brain volumes compared to FreeSurfer

	Entelai Pic	CAT12	FSL
White matter	0.96 (0.53-0.99)	0.88 (0.24-0.97)	0.65 (-0.07-0.87)
Grey matter	0.84 (0.38-0.94)	0.92 (0.74-0.96)	0.69 (0.23-0.87)
CSF	0.84 (0.61-0.92)	0.04 (-0.04-0.15)	0.02 (-0.01-0.15)

CSF: cerebrospinal fluid

Table 4

Mean Dice coefficient (range) of WM, GM and CSF brain volumes compared to FreeSurfer

	Entelai Pic	CAT12	FSL
White matter	0.92 (0.89-0.94)	0.90 (0.88-0.92)	0.85 (0.77-0.90)
Grey matter	0.86 (0.82-0.88)	0.86 (0.83-0.88)	0.79 (0.72-0.84)
CSF	0.74 (0.69-0.80)	0.64 (0.55-0.72)	0.54 (0.45-0.61)

CSF: cerebrospinal fluid.

Table 5

Mean (range) coefficient of variation (CV)

	FreeSurfer	Entelai Pic	CAT12	FSL
Whole brain	1.50 (0.01-5.66)	1.99 (0.01-6.38)	0.57 (0.00-2.71)	4.14 (0.27-13.85)
Grey matter	1.64 (0.01-6.75)	1.19 (0.00-4.48)	2.63 (0.00-10.89)	4.28 (0.01-17.88)
White matter	1.77 (0.01-5.41)	1.82 (0.00-6.03)	2.60 (0.00-12.10)	4.90 (0.12-16.06)
CSF	3.13 (0.06-12.12)	3.00 (0.01-8.95)	9.68 (0.00-32.26)	5.81 (0.06-16.07)

CSF: cerebrospinal fluid