Benchmarking the Reproducibility of Brain Tissue Segmentation Across MRI Scanners

Ekaterina Kondrateva[1,2] EKATERINA.KONDRATEVA@MAASTRICHTUNIVERSITY.NL
Abdalla Z Mohamed[3]
Sandzhi Barg[4]
Florian Kofler[5,6,7,8]

[1] *Department of Radiation Oncology (Maastro), GROW Research Institute for Oncology and Reproduction, Maastricht University Medical Centre+, Maastricht University, Maastricht, The Netherlands*
[2] *Atelic AI, UAE*
[3] *Department of Cognitive Sciences, College of Humanities and Social Sciences, United Arab Emirates University, UAE*
[4] *Higher School of Economics (HSE), Russia*
[5] *Hertie Institute for AI in Brain Health, University of Tübingen, Tübingen, Germany*
[6] *Department of Quantitative Biomedicine, University of Zurich, Zurich, Switzerland*
[7] *Helmholtz AI, Helmholtz Munich, Neuherberg, Germany*
[8] *AI for Image-Guided Diagnosis and Therapy, TUM School of Medicine and Health, Technical University of Munich, Munich, Germany*

Editors: Accepted for publication at MIDL 2026

Abstract

Accurate and reproducible brain morphometry from structural magnetic resonance imaging is critical for monitoring neuroanatomical changes across time and imaging domains. Although deep learning has accelerated segmentation workflows, scanner-induced variability and limited reproducibility remain major obstacles, particularly in longitudinal and multi-site studies. In this study, we benchmark two state-of-the-art segmentation pipelines, *FastSurfer* and *SynthSeg*, integrated into *FreeSurfer*, one of the most widely adopted neuroimaging tools. Using two complementary datasets—a 17-year single-subject longitudinal cohort and a nine-site test–retest cohort—we quantify between-scan segmentation variability with region-wise overlap and distance measures, including the Dice similarity coefficient, surface Dice, the 95th percentile of the Hausdorff distance, and the mean absolute percentage error in regional volumes.

Our results reveal up to 7–8% variation in the volumes of small subcortical structures such as the amygdala and ventral diencephalon, even under controlled test–retest conditions. This level of noise raises a critical question: can we reliably detect subtle longitudinal changes of 5–10% in small brain regions with volumes below 2 milliliters, given the magnitude of scanner- and site-induced morphometric variability? We further analyze how registration choices and interpolation modes contribute additional, although smaller, biases, and we show that surface-based quality filtering can remove outlier segmentations while preserving most scans and maintaining morphometric stability. This work provides a reproducible benchmark of modern FreeSurfer-based segmentation pipelines and highlights the need for harmonization and quality-control strategies to enable robust morphometry in real-world neuroimaging studies.

The code is publicly available on https://github.com/kondratevakate/brain-mri-segmentation

Keywords: Machine Learning, Brain Morphometry, MRI, Multi-Scanner Variability, Dice, FreeSurfer, SynthSeg, Segmentation, Statistics, Test-Retest, Domain Shift

1. Introduction

Advances in AI-driven medical imaging have revolutionized pathology detection, yet reproducible morphometric analysis of healthy brains—especially across scanners and over time—remains a challenge. This gap limits our ability to monitor individual brain health trajectories and detect early pathological changes. While artificial intelligence (AI) has significantly advanced medical imaging, particularly in pathology segmentation tasks such as tumor identification in the BraTS challenge (Menze et al., 2015), there remains a notable gap in applying these advancements to morphometric analyses of healthy brains across varied domains. This underexplored area presents opportunities for developing robust, generalizable AI models that can accurately capture subtle anatomical variations, thereby deepening insight into brain aging and development.

Traditional tools like FreeSurfer (Fischl, 2012b) have been instrumental in providing detailed morphometric analyses of brain structures' volume, cortical surface area, and thickness. Recent integrations, such as SynthSeg (Billot et al., 2023c), offer contrast-agnostic segmentation capabilities trained on synthetic data, aiming to improve generalizability across different imaging protocols. In parallel, FastSurfer (Henschel et al., 2020b) emerged as a deep learning-based alternative to the traditional FreeSurfer pipeline, utilizing convolutional neural networks trained on real FreeSurfer-labeled data to achieve significantly faster processing times while maintaining comparable accuracy. Unlike SynthSeg's domain-randomization approach with synthetic training data, FastSurfer leverages supervised learning on anatomically labeled T1-weighted images, making it well-suited for standard clinical protocols but potentially less robust to imaging variations. Despite these advancements, challenges persist in ensuring reproducibility of volumetric estimates under real-world conditions, particularly when dealing with data from multiple scanners and protocols.

This study aims to assess the consistency of brain volume measurements using FastSurfer and FreeSurfer 8 with integrated SynthSeg across longitudinal MRI scans from a single individual. By quantifying inter-scan variability using metrics like absolute volume difference, Dice, and Surface Dice, we seek to highlight the limitations of current segmentation pipelines in personalized brain health monitoring and early detection of neurodegenerative conditions.

2. Related Works

Deep learning has significantly advanced individual-level brain morphometry from structural MRI. Traditional pipelines such as *FreeSurfer* (Fischl, 2012b) have long served as a gold standard, producing cortical and subcortical morphometric features (e.g., thickness, volume, surface area). However, these methods are computationally intensive and sensitive to scanner variability, limiting their scalability in large-scale or multisite studies.

Recent versions of FreeSurfer integrate *SynthSeg* (Billot et al., 2023c), a contrast-agnostic segmentation model trained on synthetic data. *SynthSeg* provides robust volumetric estimates across diverse contrasts, resolutions, and scanners. Its compatibility with

standard atlases (e.g., Desikan-Killiany, MUSE) makes it suitable for harmonized morphometry across heterogeneous datasets.

To address runtime bottlenecks, *FastSurfer* (Henschel et al., 2020b) provides a FreeSurfer-compatible alternative using a voxel-size independent convolutional neural network (FastSurferVINN) (Henschel et al., 2022), enabling accurate whole-brain segmentation and optional surface-based cortical analysis within minutes. Tools such as BrainChop (Tudosiu et al., 2023) prioritize clinical scalability, though often at the cost of generalization to unseen protocols.

Other high-performing segmentation models include *nnU-Net* (Isensee et al., 2021) and *nnFormer* ((Zhou et al., 2023)), which yield excellent accuracy in controlled benchmarks but often require dataset-specific finetuning to generalize effectively in clinical or real-world settings.

2.1. Longitudinal Modeling and Individualized Morphometry

The reproducibility of brain segmentation directly impacts downstream applications that rely on longitudinal morphometric stability, including disease progression modeling, normative deviation detection, and biological age estimation. Without consistent volumetric measurements across scans, even sophisticated longitudinal models risk misinterpreting noise as biological change, undermining clinical utility in personalized monitoring.

Normative modeling frameworks enable the estimation of z-score deviations from large-scale population references. This approach is particularly effective in identifying early deviations in psychiatric populations and supports both clinical and subclinical applications (Marquand et al., 2016).

Another widely adopted line of work focuses on brain age prediction. *BrainAGE* (Franke and Gaser, 2012) models estimate biological aging based on MRI-derived morphometric features, frequently using *FreeSurfer* outputs. These models have demonstrated strong longitudinal reliability and clinical interpretability.

Emerging tools like *Neurofind* (Vieira et al., 2025) offer user-friendly platforms that integrate normative modeling and brain age estimation, providing individualized reports based on high-resolution structural MRI images.

Despite these advances, challenges remain in achieving sulcal-level surface precision, quantifying uncertainty, and ensuring reproducibility in real-world multisite studies. Although morphometry has clear clinical applications, including epilepsy-focused MRI analysis and dementia-oriented volumetry (Aliev et al., 2021; Khadhraoui et al., 2024), rigorous longitudinal reproducibility benchmarks remain scarce.

2.2. Brain Morphometry as a Biomarker

Longitudinal MRI studies have greatly expanded our understanding of how brain morphometry changes over time, particularly in response to aging, disease, and stress. A growing body of work highlights structural biomarkers in specific brain regions—especially the hippocampus, anterior cingulate, and prefrontal cortex—that reflect vulnerability or resilience to neuropsychiatric conditions.

In healthy populations (Papagni et al., 2011) demonstrated gray matter volume (GMV) reductions in the anterior cingulate cortex (ACC), hippocampus, and medial prefrontal

cortex (mPFC) in individuals exposed to stress. Similar findings were confirmed in large-scale aging studies, including (Schaefer et al., 2018), who reported consistent hippocampal atrophy associated with aging. MacDonald and Pike (2021) provide a broader review of region-specific atrophy across the lifespan. Structural biomarkers also inform psychiatric research. Cardoner et al. (2024) review evidence of stress-induced degeneration in the ACC and dorsolateral prefrontal cortex (dlPFC), while Carnevali and Sgoifo (2018) identify preserved amygdala volumes as potential resilience markers. UK Biobank analyses further support longitudinal volume reductions in fronto-limbic circuits among individuals with high stress exposure (Statsenko et al., 2022). Importantly, several studies have examined structural changes within individuals undergoing therapy. Gryglewski et al. (2019) found hippocampal and amygdalar volume increases after electroconvulsive therapy (ECT) in treatment-resistant depression. Furtado et al. (2012) reported volumetric growth in the dlPFC after rTMS. Frodl et al. (2008) showed that psychotherapy attenuated gray matter loss over three years in depression. Together, these findings suggest that MRI-based brain morphometry, especially when assessed longitudinally, provides meaningful biomarkers for brain health across both normative and pathological aging.

3. Methods

We study the reproducibility of brain MRI segmentation pipelines across longitudinal and multi-site datasets. We employ two publicly available datasets: SIMON (Single Individual volunteer for Multiple Observations across Networks) and SRPBS (Strategic Research Program for Brain Sciences), spanning a wide range of scanners and protocols. We compare segmentation outputs from FreeSurfer, FastSurfer, and SynthSeg, using FreeSurfer's `recon-all` pipeline as a reference. Segmentation reproducibility is evaluated using a targeted subset of cortical and subcortical ROIs most relevant for neuroimaging biomarkers. For surface-based comparisons, we apply rigid registration using ANTs (Avants et al., 2011) and assess the effect of different interpolation modes and reference spaces. Quantitative evaluation is performed using Dice coefficient (DSC), Surface Dice, 95th percentile Hausdorff distance (HD95), and mean absolute percentage error (MAPE) of regional brain volumes. These metrics can be efficiently computed using tools such as panoptica (Kofler et al., 2024).

3.1. Datasets

SIMON Dataset (Duchesne et al., 2019): This dataset comprises 73 T1-weighted MRI scans of a single healthy male subject, collected over 17 years across multiple sites and 1.5T scanners (Duchesne et al., 2019), multiple sites with 35 distinct scanner settings.

SRPBS Traveling Subject Dataset (Tanaka et al., 2021): This dataset includes 411 T1-weighted MRI scans from 9 healthy subjects, each scanned at 9 different sites using 3T MRI scanners in constitutive days. The data is organized following the BIDS format and includes accompanying metadata such as participant demographics and scanner parameters (Tanaka et al., 2021).

Figure 1 provides a representative qualitative example from the SRPBS Traveling Subject dataset, illustrating scanner/protocol-induced domain shift at the image level and its downstream impact on segmentation consistency.

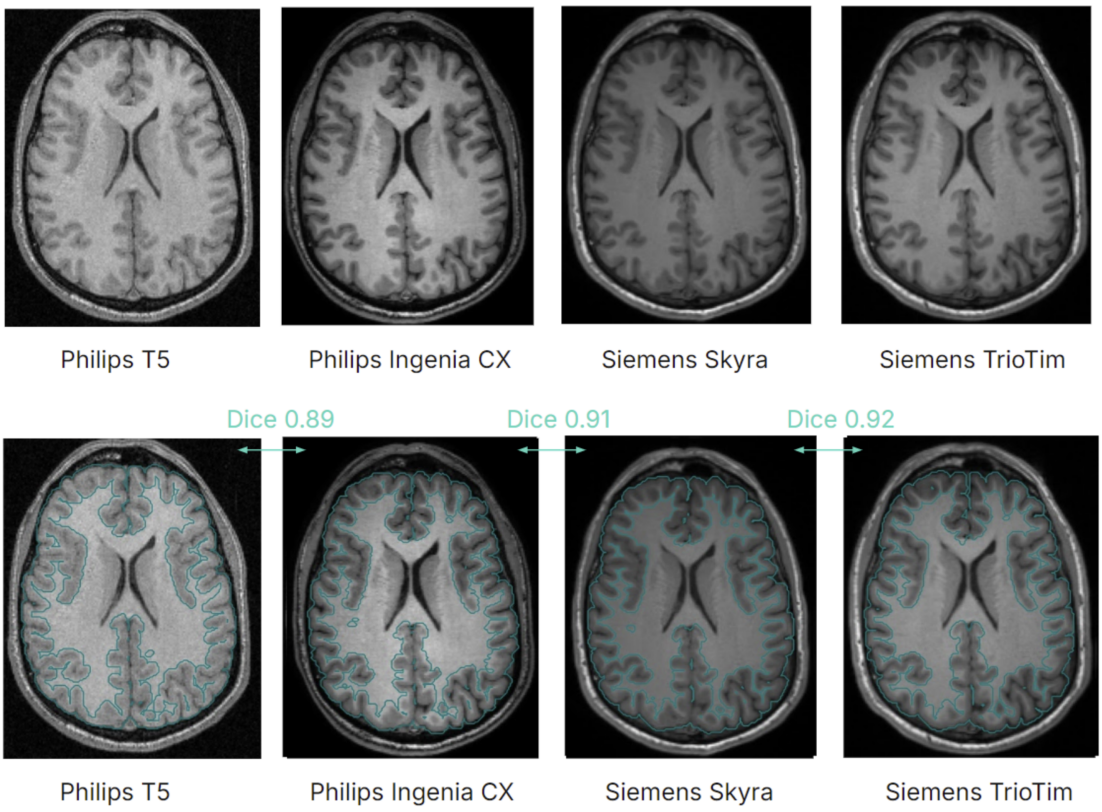

Figure 1: Representative qualitative example from the SRPBS Traveling Subject dataset (Subject 1; four sites/scanners; consecutive-day acquisitions). **Top:** T1-weighted scans of the same subject acquired on different scanners/sites, illustrating scanner/protocol-induced appearance differences (same anatomical slice; consistent display settings). All scans were rigidly registered to the first acquisition, and FreeSurfer `recon-all` label maps were co-registered accordingly; after `recon-all` conforming, images share the same voxel grid (size and spacing). **Bottom:** FreeSurfer outputs overlaid as contours derived from a binarized full-parcellation mask. Dice overlap values (range $[0, 1]$) are shown between adjacent acquisitions to summarize cross-scanner segmentation consistency. **Take-home message:** even when anatomy is held constant and scans are aligned to a common space, scanner/protocol changes visibly alter image appearance and reduce segmentation agreement.

Because both datasets consist of healthy participants and include repeat acquisitions with short inter-scan intervals (SRPBS: consecutive-day traveling-subject scans; SIMON: repeated acquisitions of the same subject), we follow common practice in structural MRI test–retest studies and treat true anatomical change over such intervals as negligible compared to acquisition- and pipeline-induced variability. Therefore, in the absence of voxel-wise man-

ual annotations, we interpret agreement metrics as measures of reproducibility/consistency rather than absolute accuracy (Iscan et al., 2015; Han et al., 2006; Jovicich et al., 2009; Reuter et al., 2012).

3.2. Segmentation pipelines

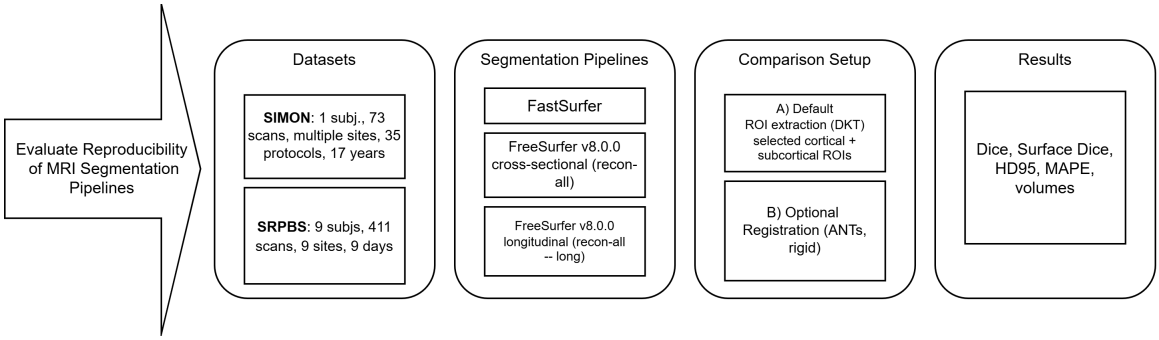

Figure 2: High-level overview of the benchmarking pipeline. The workflow evaluates the reproducibility of FreeSurfer and FastSurfer using the longitudinal SIMON dataset and the multi-site SRPBS traveling-subject dataset. All outputs are mapped to a common ROI space and compared with spatial and volumetric metrics while varying registration and interpolation choices. **Take-home message:** the benchmark is designed to isolate pipeline-, scanner-, and post-processing-induced variability from true biological change.

We employed FreeSurfer 8.0.0 for cortical surface reconstruction and anatomical segmentation using the `recon-all` pipeline. To evaluate segmentation performance, we compared two state-of-the-art deep learning-based methods: FastSurfer (Henschel et al., 2020b) and SynthSeg (Billot et al., 2023c). FastSurfer offers rapid and accurate whole-brain segmentation, replicating FreeSurfer's anatomical outputs, while SynthSeg provides robust segmentation across varying MRI contrasts and resolutions without the need for retraining. For consistency and comprehensive analysis, we selected FreeSurfer's `recon-all` outputs as the common baseline and assessed the Desikan-Killiany-Tourville (DKT) atlas parcellations, encompassing 100 cortical and subcortical regions.

Preprocessing and input standardization. To maximize reproducibility and avoid introducing study-specific choices, we intentionally did not apply additional external preprocessing (e.g., custom intensity normalization) beyond each pipeline's default processing. In prior work on tumor segmentation, the choice of segmentation pipeline was reported to have a negligible effect on segmentation accuracy (Kondrateva et al., 2024). FreeSurfer recon-all includes its canonical internal conforming and intensity processing as part of the standard workflow (Fischl, 2012b). SynthSeg is trained via domain randomization to be robust to wide variations in contrast, resolution, noise and bias fields, and is designed to operate without requiring dedicated preprocessing (Billot et al., 2023c).

For surface-based metrics, we applied rigid-body registration using ANTs (Avants et al., 2011), computing transforms from the original T1-weighted images. We evaluated two interpolation modes: linear and nearest neighbor. Two registration strategies were compared: (1) intra-subject co-registration to the subject's first session for longitudinal alignment, and (2) spatial normalization to an asymmetric MNI atlas. This approach aimed to assess the impact of interpolation schemes and reference space choice on the consistency of surface-derived measurements.

ROI Analysis. We focused our analysis on 9 cortical and 8 subcortical bilateral regions of interest (ROIs), selected based on their relevance as biomarkers in neuroimaging studies. The complete list of analyzed ROIs is provided in Table 4. Differences observed across successive MRI sessions were interpreted as domain variations.

3.3. Metrics.

To evaluate segmentation reproducibility, we report absolute volume differences, as well as spatial similarity metrics: Dice coefficient, Surface Dice, and 95th percentile Hausdorff Distance (HD95). Each metric captures a different aspect of agreement between two segmentations: volumetric overlap, boundary proximity, and outlier misalignment. These are computed for each region of interest (ROI) and aggregated across sessions. To compare volumes across repeated scans, we use the mean absolute percentage error between segmentation volumes.

Dice Coefficient (DSC). DSC (Dice, 1945b) measures the voxel-level overlap between two binary masks A and B (e.g., predicted and reference segmentation):

$$\mathrm{DSC} = \frac{2|A \cap B|}{|A| + |B|} \tag{1}$$

Here, $|A|$ and $|B|$ are the number of voxels in each mask, and $|A \cap B|$ is the number of voxels they share. Dice is widely used due to its simplicity, but can be insensitive to boundary errors.

Surface Dice (S-DSC). Surface Dice (Nikolov et al., 2018) quantifies the proportion of surface points that lie within a distance τ between the two segmentation boundaries ∂A and ∂B:

$$\text{S-DSC} = \frac{|\{x \in \partial A : d(x, \partial B) \leq \tau\}| + |\{y \in \partial B : d(y, \partial A) \leq \tau\}|}{|\partial A| + |\partial B|}. \tag{2}$$

Here, $d(x, \partial B)$ denotes the minimum Euclidean distance from a point x on the surface of A to the surface of B, and τ is the distance tolerance (set to 1 mm in our experiments). This metric captures small surface deviations and is well-suited for assessing perceptual segmentation accuracy.

95th Percentile Hausdorff Distance (HD95). HD95 (Huttenlocher et al., 1993) captures the worst-case boundary discrepancy, ignoring extreme outliers by focusing on the 95th percentile of all boundary distances:

$$\mathrm{HD}_{95}(A, B) = \max \begin{Bmatrix} \mathrm{P}_{95}\big(\{d(x, \partial B) : x \in \partial A\}\big), \\ \mathrm{P}_{95}\big(\{d(y, \partial A) : y \in \partial B\}\big) \end{Bmatrix}. \tag{3}$$

Where P_{95} denotes the 95th percentile, and $d(x, \partial B)$ is the shortest distance from point x to the other surface. HD95 is useful for identifying large local deviations in shape or topology.

Mean Absolute Percentage Error (MAPE): To compare volumes across repeated scans, we use the mean absolute percentage error between segmentation volumes:

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^{n} \left| \frac{V_i^{\text{pred}} - V_i^{\text{ref}}}{V_i^{\text{ref}}} \right| \tag{4}$$

Where V_i^{pred} and V_i^{ref} are the predicted and reference volumes for region i, and n is the number of ROIs. MAPE is intuitive for assessing how much segmentations deviate from expected anatomical volumes.

3.4. Computation Environment.

We used a single compute environment for all benchmark experiments, reflecting practical deployment constraints and emphasizing reproducibility.

- **CPU morphometry:** FreeSurfer 8.0.0 `recon-all` (cross-sectional and longitudinal) was executed on a Google Cloud Platform (GCP) instance equipped with *64* vCPUs and *512* GB of RAM. FreeSurfer was run using a single CPU core per subject, with an average processing time of approximately ∼2 hours/subject (longitudinal runs typically longer due to within-subject template construction and additional processing steps). Attempts to utilize GPU acceleration for FreeSurfer 8.0.0 were unsuccessful due to driver/library compatibility issues; therefore, all FreeSurfer processing was performed on the CPU.
- **CNN-based segmentation:** FastSurfer and SynthSeg were executed within the same benchmark framework. While these models can benefit substantially from GPU acceleration, we report GPU runtimes as reference from the respective projects: FastSurfer reports whole-brain inference on the order of minutes on GPU and substantially reduced end-to-end runtime compared to classical FreeSurfer pipelines (FastSurfer Developers, 2026).

Model parameters: SynthSeg (Billot, 2023) employs a 3D U-Net with five levels, starting from 24 feature maps and doubling at each downsampling layer, corresponding to ∼20–40 million trainable parameters (computed directly from the official Keras/TensorFlow model definition) (Billot et al., 2023c); FastSurferCNN and FastSurferVINN use lightweight U-Net–style fully convolutional networks with only ∼1.8–1.85 million parameters (Henschel et al., 2022); for comparison, browser-based BrainChop leverages a MeshNet-style 3D CNN with dilated convolutions, achieving interactive inference on consumer hardware with a compressed TensorFlow.js model of only tens of megabytes (∼1–5 million parameters) (Tudosiu et al., 2023).

4. Results

We first report short-interval test–retest findings from the SRPBS traveling-subject dataset to isolate scanner- and site-related variability under minimal expected biological change. We

then compare longitudinal behavior in the SIMON dataset and conclude with cross-dataset analyses of overlap, volume reproducibility, and the effects of registration and interpolation choices.

4.1. SRPBS Test–Retest: FastSurfer

We analyzed 15 sessions from the SRPBS Traveling Subject dataset (Tanaka et al., 2021) using FastSurfer . As shown in Figure 3, the first five sessions were acquired on the same scanner across five consecutive days, while the remaining sessions involved different scanners and sites.

For both hippocampus and amygdala, volume estimates during the same-scanner phase were highly consistent. For example, left hippocampus volumes ranged narrowly between 4.42–4.44 cm^3 (SD = 0.01), and right amygdala volumes ranged from 1.73–1.75 cm^3 (SD = 0.008). In contrast, sessions from different scanners showed noticeable variability: left hippocampus ranged from 4.16–4.53 cm^3 (SD = 0.10), and right amygdala from 1.50–1.85 cm^3 (SD = 0.11).

This highlights that even in a highly controlled test-retest design, inter-scanner variability introduces morphometric noise of up to 10%, especially in small structures like the amygdala. Reliable quantification in longitudinal or multisite settings requires either harmonization or robust outlier filtering.

4.2. SIMON Longitudinal: FastSurfer vs. FreeSurfer

We evaluated segmentation reproducibility across 73 sessions over 17 years using FastSurfer and FreeSurfer.

FastSurfer. FastSurfer `recon-all` failed on 3 sessions and 8 runs. For valid outputs, subcortical volumes were stable: Left/Right Amygdala: 1.93 ± 0.17 / 2.10 ± 0.12 cm^3 Left/Right Hippocampus: 4.54 ± 0.19 / 4.82 ± 0.16 cm^3 Volume trajectories showed small upward trends ($R^2 = 0.12$–0.26).

FreeSurfer. Subcortical variation averaged 3.1%, peaking at 15–20%. Cortical parcellations varied by 5% on average, with outliers exceeding 40–90%. Volumes were consistently higher: Amygdala: 2.13 ± 0.07 / 2.22 ± 0.07 cm^3 Hippocampus: 5.10 ± 0.11 / 5.18 ± 0.12 cm^3.

Volume comparisons show that FreeSurfer consistently estimates larger volumes than FastSurfer. For example, the left hippocampus volume averaged 5.10 ± 0.11 cm^3 in FreeSurfer versus 4.54 ± 0.19 cm^3 in FastSurfer.

Table 1 compares FreeSurfer and FastSurfer across eight representative cortical structures. FastSurfer yielded consistently higher Dice scores (e.g., 0.861 vs. 0.793 for Insula, 0.816 vs. 0.728 for Fusiform), suggesting improved anatomical overlap. Surface Dice values remained comparable, with minimal variation between methods. Volume differences were notably smaller in FastSurfer (e.g., 2.0 mm^3 for Insula, compared to 31.6 mm^3 in FreeSurfer), reflecting reduced bias. Interestingly, FreeSurfer produced lower Hausdorff distances in some regions (e.g., Superior Frontal Cortex: 1.21 mm vs. 1.74 mm), but at the cost of greater volume deviation. Overall, FastSurfer offers more consistent cortical segmentation while maintaining competitive boundary accuracy.

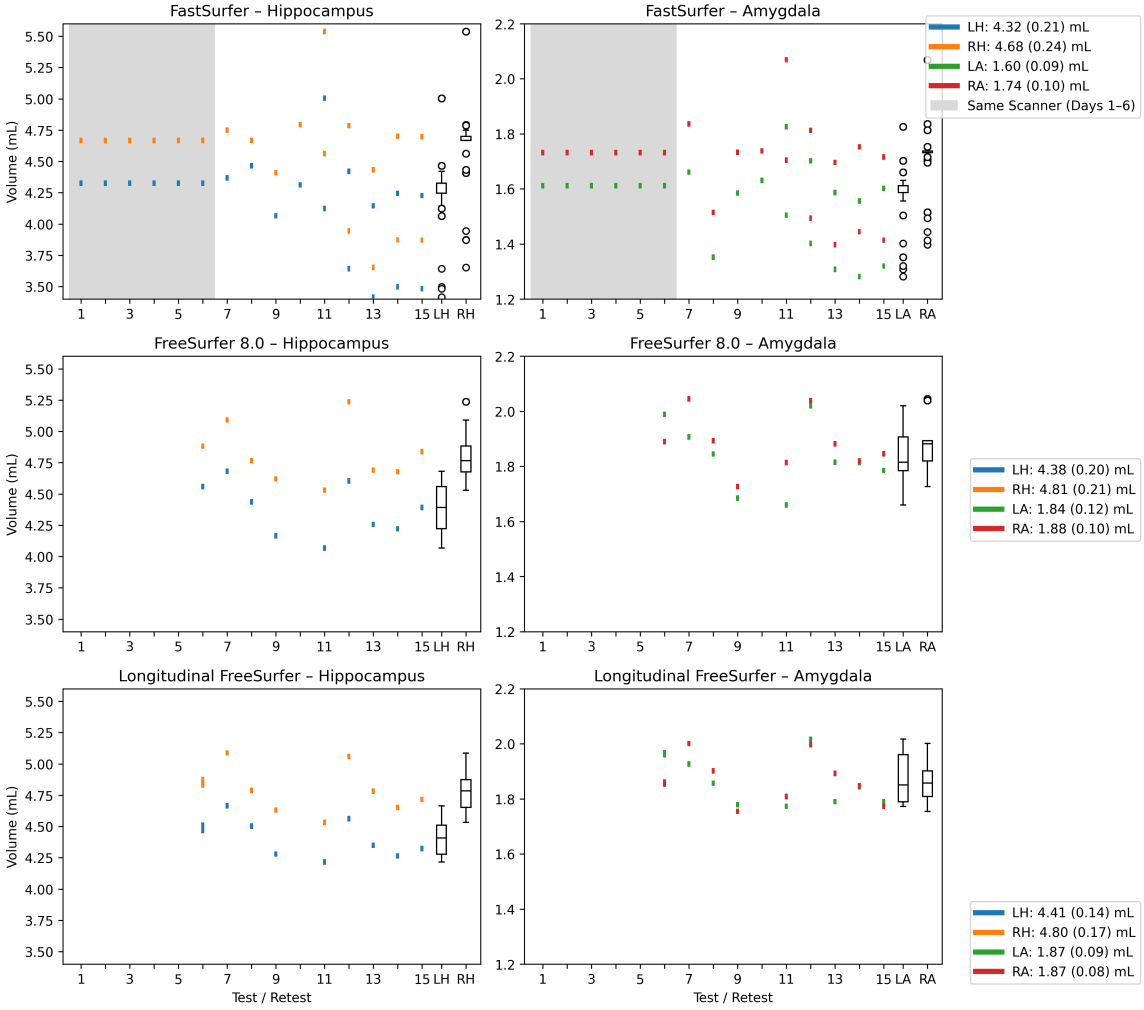

Figure 3: Volume estimates for left and right hippocampus and amygdala across 15 scans of the same traveling subject in the SRPBS dataset (sub-01), processed with FastSurfer, FreeSurfer 8, and longitudinal FreeSurfer 8 with creation of the unbiased template. Each point corresponds to one session; the shaded region marks the first five scans acquired on a single scanner (days 1–6), whereas the remaining sessions were collected at different sites. **Take-home message:** within-scanner repeat scans are tightly clustered, but cross-site acquisitions introduce much larger spread, especially in the amygdala, showing that scanner effects can approach the size of subtle longitudinal change.

4.3. Comparison of Distance-based Metrics Across Datasets

Volume differences (in cm^3) were consistently higher in SRPBS. In contrast, SIMON—being a single-subject longitudinal dataset—showed lower volume deviations across repeated scans. Dice and Surface Dice scores were uniformly higher in SIMON, indicating improved

Table 1: Comparison of FreeSurfer 8 (FS) and FastSurfer (Fast) segmentation performance across subcortical structures. Volume differences are in mm^3, Dice and Surface Dice are unitless, HD95 is in mm.

Metric	**Accumbens**		**Amygdala**		**Caudate**		**Hippocampus**		**Pallidum**		**Putamen**		**Thalamus**		**Ventral DC**	
	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast
Volume Diff (mm^3)	5.20	-0.56	0.22	-2.23	14.18	1.36	12.46	-0.17	11.99	1.94	19.99	-5.04	2.27	8.30	12.32	1.06
Dice	0.803	0.827	0.858	0.862	0.868	0.874	0.850	0.868	0.850	0.859	0.897	0.902	0.909	0.917	0.858	0.873
Surface Dice	0.965	0.955	0.961	0.944	0.972	0.957	0.964	0.963	0.958	0.927	0.969	0.956	0.947	0.948	0.959	0.950
HD95 (mm)	1.23	1.60	1.26	1.50	1.20	1.56	1.23	1.34	1.27	1.64	1.21	1.58	1.33	1.45	1.23	1.43

Table 2: Comparison of FreeSurfer 8 (FS) and FastSurfer segmentation performance across selected cortical structures. Volume difference is in mm^3, Dice and Surface Dice are unitless, HD95 is in mm.

Metric	**Caudal Ant. Cingulate**		**Entorhinal Cortex**		**Fusiform Gyrus**		**Inferior Parietal**		**Insula**		**Lat. Orbitofrontal**		**Med. Orbitofrontal**		**Superior Frontal**		**Superior Temporal**	
	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast	FS	Fast
Volume Diff	21.62	2.90	7.75	-4.78	58.67	-2.95	113.35	3.51	31.60	2.00	64.48	7.46	35.29	5.04	216.32	42.99	115.34	15.80
Dice	0.746	0.820	0.709	0.794	0.728	0.816	0.726	0.807	0.793	0.861	0.712	0.796	0.663	0.780	0.733	0.807	0.759	0.817
Surface Dice	0.965	0.958	0.922	0.922	0.959	0.964	0.973	0.963	0.970	0.971	0.952	0.948	0.938	0.949	0.970	0.966	0.964	0.958
HD95	1.24	1.64	1.72	1.72	1.28	1.35	1.19	1.47	1.35	1.51	1.34	1.85	1.46	1.84	1.21	1.74	1.26	1.47

overlap and surface-level agreement. For example, mean Dice scores for the caudate and putamen reached 0.868 and 0.897 in SIMON, compared to 0.802 and 0.848 in SRPBS. HD95 distances also decreased in SIMON (e.g., 1.234 mm for hippocampus vs. 1.830 mm in SRPBS). These results support the utility of repeated intra-subject data for evaluating segmentation consistency.

Statistical comparison using Mann-Whitney U tests with Benjamini-Hochberg FDR correction confirmed significant differences between SIMON and SRPBS conditions for 89% of ROI-metric combinations (64/72, $p_{\text{adj}} < 0.05$). Effect sizes were predominantly large (Cliff's $\delta > 0.474$ in 56/72 comparisons), with the most pronounced differences observed for Dice and Surface Dice in cortical regions (e.g., inferior parietal: $\delta = 0.81$–0.84; fusiform gyrus: $\delta = 0.74$–0.80). These results provide quantitative evidence that cross-scanner variability introduces systematic and substantial degradation in segmentation consistency.

Contrary to expectations based on study design and field strength, the observed segmentation variability did not align with either nominal time interval or magnet strength: although SIMON (a longitudinal 1.5T dataset) could plausibly exhibit larger true volumetric changes over time and might be expected to be less robust than SRPBS (a consecutive-day 3T traveling-subject dataset), we instead observed generally higher consistency in SIMON

Metric	Accumbens		Amygdala		Caudate		Hippocampus		Pallidum		Putamen		Thalamus		Ventral DC	
	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON	SRPBS	SIMON
Volume diff (cm^3)	0.046	0.030	0.102	0.076	0.119	0.098	0.207	0.125	0.102	0.095	0.206	0.136	0.450	0.374	0.219	0.141
Dice	0.677	0.803	0.790	0.858	0.802	0.868	0.782	0.850	0.789	0.850	0.848	0.897	0.868	0.909	0.806	0.858
Surface Dice	0.849	0.965	0.840	0.961	0.868	0.972	0.845	0.964	0.843	0.958	0.870	0.969	0.820	0.947	0.873	0.959
HD95 (mm)	1.735	1.228	1.697	1.263	1.584	1.200	1.830	1.234	1.675	1.271	1.582	1.210	1.828	1.327	1.620	1.233

Table 3: Comparison of segmentation metrics between the SRPBS and SIMON datasets across subcortical structures. Volume difference is shown in cm^3, Dice and Surface Dice are unitless similarity scores, and HD95 represents the 95th percentile Hausdorff distance in millimeters.

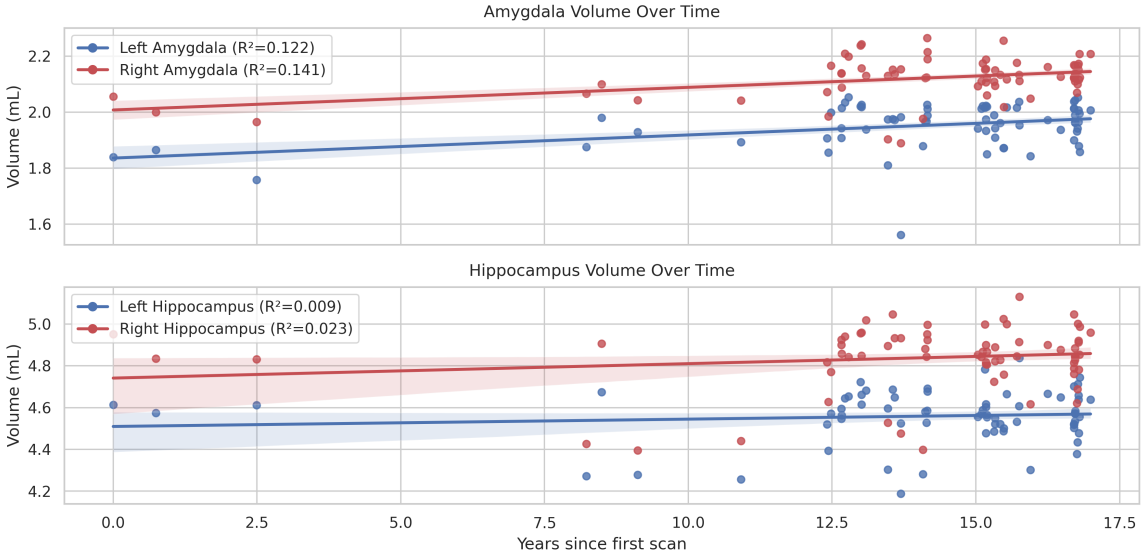

Figure 4: SIMON dataset: volume trajectories of the amygdala and hippocampus over time for 73 MRI scans acquired over 17 years in one healthy individual using FastSurfer. Confidence intervals and regression trends are shown. **Take-home message:** FastSurfer yields comparatively smooth long-term trajectories with modest drift, suggesting that most within-subject variation is small relative to the overall volume scale.

and lower consistency in SRPBS. This suggests that acquisition-related domain shift, driven by scanner- and protocol-specific factors and their interaction with each pipeline, dominates the variability in these datasets, outweighing both short-term biological change and the presumed robustness advantage of 3T over 1.5T.

HD95 Stability Analysis. Hausdorff distance at the 95th percentile (HD95) revealed striking differences in boundary consistency between datasets. For SIMON, HD95 values were remarkably stable: the median was 1.0 mm across most cortical structures, with IQR $= 0$ for 16/18 regions. Values of exactly 1.0, $\sqrt{2} \approx 1.41$, and $\sqrt{3} \approx 1.73$ mm correspond to 1-, 2D-diagonal, and 3D-diagonal voxel distances at 1 mm isotropic resolution, indicating that segmentation boundaries differ by at most 1–2 voxels between consecutive sessions.

In contrast, SRPBS exhibited substantially higher HD95 variability (median 1.4–2.2 mm, IQR up to 1.9 mm). Notably, certain site combinations produced outlier values reaching 7–9 mm (e.g., siteHKH vs. siteHUH), while others remained at 1.0 mm. This site-dependent variability suggests that specific scanner or protocol combinations introduce systematic boundary inconsistencies, even when comparing the same subject.

Subcortical Filtering Based on Segmentation Quality. To assess the impact of quality-based filtering, we evaluated the proportion of subcortical structures removed using various thresholds on Dice and Surface Dice metrics. Applying a strict Surface Dice threshold of 0.92 filtered out only 5% of regions, while retaining a low mean absolute percentage

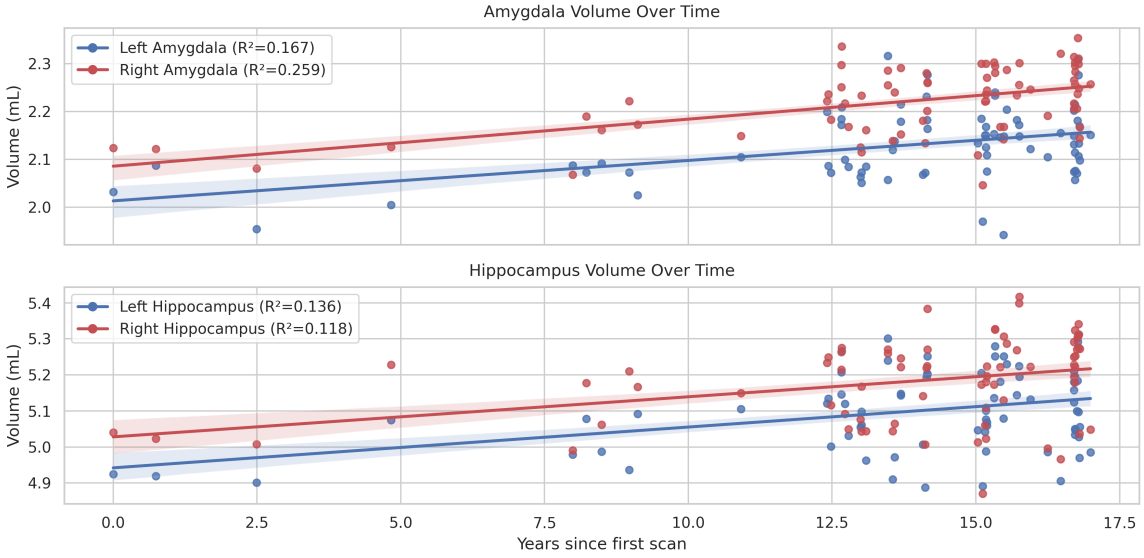

Figure 5: SIMON dataset: volume trajectories of the amygdala and hippocampus over time for 73 MRI scans acquired over 17 years in one healthy individual using FreeSurfer 8. Confidence intervals and regression trends are shown. **Take-home message:** the inferred longitudinal trend depends on the segmentation software, not only the anatomy: FreeSurfer 8 produces systematically larger estimates and visually broader dispersion than the corresponding FastSurfer trajectories.

error (MAPE) across the remaining structures (2.8% at 75th percentile, 8.6% at 95th). Relaxing the threshold to 0.90 slightly reduced filtering (3.8%) without degrading MAPE. In contrast, filtering with a traditional Dice threshold of 0.80 excluded more than half of all structures (52.8%), yet retained comparable or worse error profiles. This supports the use of Surface Dice as a more efficient and precise filtering criterion for detecting outliers in automated segmentation pipelines.

Table 4: Percentage of subcortical regions filtered out using Dice and Surface Dice thresholds, with 75th and 95th percentile MAPE values across retained regions.

Filtering Metric	Threshold	Structures	% Filtered	75th (% MAPE)	95th (%)
Surface Dice	0.92	Subcortical	5.0	2.8	8.6
Surface Dice	0.90	Subcortical	3.8	2.8	8.8
Dice	0.80	Subcortical	52.8	2.2	5.8

4.4. Volume Reproducibility (ICC Analysis)

To quantify the reproducibility of volumetric estimates, we computed intraclass correlation coefficients (ICC(3,1), two-way mixed effects, single measurement, consistency) for

cortical ROIs across repeated measurements. For the SIMON dataset, ICC values were predominantly poor (mean ICC = 0.14, range: −0.04 to 0.32), reflecting high inter-scanner variability in volume estimates despite reasonable segmentation overlap. This finding highlights that even when segmentation boundaries are consistent (Dice ∼0.80), absolute volume estimates can vary substantially across different scanners and protocols in time.

In contrast, the SRPBS dataset exhibited moderate ICC values (mean ICC = 0.68, range: 0.42–0.87), with 8/18 structures achieving good reliability (ICC > 0.75). The higher ICC in SRPBS, despite lower Dice scores, suggests that the traveling-subject protocol with controlled acquisition parameters produces more consistent volume estimates across sites than the heterogeneous SIMON acquisitions spanning 17 years and 35 different protocols.

These findings underscore a critical distinction: segmentation *overlap* (Dice) and volume *reproducibility* (ICC) capture different aspects of reliability. Cross-scanner variability primarily affects absolute volume quantification rather than anatomical boundary delineation.

4.5. Registration Strategies and Interpolation Effects

To compare surface-based metrics, rigid-body registration was applied using ANTs (Avants et al., 2011). We tested two interpolation strategies, `nearestNeighbor` and `genericLabel`, and two reference spaces: subject-native (first session) and standard MNI atlas. Interpolation mode affected mean volume estimates by up to 1.72%, while template choice accounted for a smaller 0.07% deviation.

5. Conclusion and Discussion

This study shows that even state-of-the-art segmentation tools such as FastSurfer and FreeSurfer remain sensitive to scanner and protocol variability, particularly in multi-site and longitudinal settings. Despite their wide adoption and strong benchmark performance, we observed non-negligible instability in repeated measurements, especially for small subcortical structures such as the amygdala and pallidum.

On the SRPBS Traveling Subject dataset, both tools achieved excellent within-scanner consistency over five consecutive days, with volume deviations below 1%. However, cross-site sessions for the same individual and nominal protocol produced fluctuations up to 10%, directly constraining the ability to detect subtle longitudinal changes in small structures. In the 17-year longitudinal SIMON dataset, both FastSurfer and FreeSurfer exhibited increasing volume trends over time, but differed in the magnitude and smoothness of these trajectories, indicating method-dependent biases in long-term morphometric estimates.

FreeSurfer systematically yielded larger subcortical volumes than FastSurfer (e.g., left hippocampus: $5.10\,\mathrm{cm}^3$ vs. $4.58\,\mathrm{cm}^3$) and showed greater inter-scan variation in cortical regions. Such systematic offsets and differential variability imply that absolute volumes and longitudinal slopes are not directly interchangeable between methods. In practical terms, this reinforces the need for harmonization strategies, method-specific calibration, or stringent quality-control filters in real-world neuroimaging pipelines.

We quantified reliability using a set of complementary overlap and distance-based metrics rather than relying on a single indicator. This multi-metric approach captures distinct aspects of segmentation behaviour, such as boundary stability, volumetric agreement, and

sensitivity to outlier scans, and provides a more informative picture of robustness than Dice alone.

While recent work in brain MRI segmentation has focused on speed, automation, and ease of deployment, our results suggest that robustness to scanner and protocol variation remains a primary bottleneck for individualized applications. We release a lightweight, fully reproducible evaluation pipeline on longitudinal and multi-scanner datasets, with the aim of encouraging more transparent and method-agnostic benchmarking of segmentation tools under conditions that approximate real clinical and research use, in line with recent decentralized benchmarking efforts in healthcare AI such as the FeTS challenge (Zenk et al., 2025).

5.1. Limitations

Reference Annotations. Both SRPBS and SIMON datasets lack manual annotations, preventing true accuracy assessment (Kofler et al., 2023). We evaluated reproducibility under the assumption that anatomical structures remain stable in healthy subjects. We use FreeSurfer `recon-all` as a *reference label space* (not as ground truth) to enable like-for-like ROI definitions and surface-based morphometry across pipelines. In the absence of manual voxel-wise annotations, our evaluation targets *reproducibility/consistency* rather than absolute accuracy, and conclusions are based on relative variability patterns across ROIs and sites.

A further limitation is that the available datasets are not ideal for isolating scanner effects: SRPBS provides a traveling-subject design but includes only nine subjects, and SIMON is a single-subject longitudinal dataset. These small sample sizes limit statistical power and may reduce the generalizability of conclusions about inter-scanner and longitudinal variability patterns.

Preprocessing and Augmentation. We processed raw data without denoising, intensity normalization, or augmentation to isolate the effect of domain shift. Although this reflects practical variability, it limits reproducibility. It has been shown that classical preprocessing techniques, such as intensity normalization and histogram matching, do not consistently improve brain tumor segmentation performance across different domains. This limitation underscores the challenges posed by domain shifts in medical imaging. However, recent advancements in generative methods, including those utilizing generative adversarial networks (GANs), offer promising avenues to address these challenges. For instance, methods like M-GenSeg employ semi-supervised generative training strategies for cross-modality tumor segmentation, demonstrating improved generalization across diverse imaging modalities (Alefsen de Boisredon d'Assier et al., 2022). Related brain-MRI domain-adaptation work has also explored Fader networks on ABIDE-II fMRI, reinforcing that representation-level adaptation may help when acquisition domains differ (Pominova et al., 2021).

Software and model choice. We did not compare against an established FreeSurfer baseline such as v7.4 (the benchmark uses FreeSurfer 8.0.0), and we did not evaluate FastSurfer in longitudinal mode (`--long`) alongside its cross-sectional pipeline; additionally, although both T1 and T2 were available for many (but not all) sessions, we did not assess multi-contrast segmentation/parcellation or contrast-synthesis to enable consistent multimodal inputs across all sessions, even though modality-ablation studies in multimodal brain

MRI segmentation suggest that the contribution of each sequence can materially affect baseline performance and transferability (Druzhinina et al., 2022); nor did we include additional state-of-the-art segmentation families (e.g., nnU-Net/nnFormer) or foundation/VLM-style approaches (e.g., MedSAM), which may limit generalization of our conclusions beyond the selected pipelines and input setting.

5.2. Future Directions

1. Software, model, and configuration combinations for robustness. Systematically evaluate robustness-oriented "pipelines of pipelines" rather than single tools: FreeSurfer or FastSurfer and more combined with alternative skull-stripping, intensity standardization, and surface/cortical post-processing choices, and quantify which combinations reduce variability without sacrificing anatomical plausibility. In particular, test inference-time strategies such as test-time augmentation, uncertainty-based filtering, and consensus or ensemble labeling to stabilize ROI estimates across scanners and sessions.

2. Extend the benchmark to include low-field scanner acquisitions, where SNR, bias fields, motion, and resolution differ substantially from standard research-grade protocols. This would enable stress-testing segmentation robustness under realistic clinical constraints and help identify which methods and QC thresholds remain reliable when acquisition quality is degraded or highly variable.

3. While we focus on T1-weighted brain MRI segmentation, the proposed benchmark design is broadly applicable to other anatomical segmentation settings (e.g., different organs, other MRI contrasts, or CT), whenever repeated-measurement data are available (traveling-subject, scan–rescan, or longitudinal acquisitions). Beyond segmentation, the same repeated-measurement framework can be used as a practical QC instrument to monitor scanner drift and protocol changes over time, and to quantify robustness of downstream modules such as registration, interpolation, and ROI extraction under domain shift. Finally, our evaluation protocol can serve as a consistent testbed for comparing harmonization strategies (e.g., feature-level harmonization such as ComBat (Fortin et al., 2018), image-level normalization, or generative image translation) using the same reproducibility metrics. Adapting the benchmark to a new application primarily requires replacing the ROI/label definitions (atlas) and, where appropriate, adjusting metric tolerances (e.g., surface-distance thresholds) to reflect region size and target anatomy; the core idea of quantifying variability under repeated measurements remains unchanged.

CRediT authorship contribution statement

Ekaterina Kondrateva: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing – original draft, Writing – review and editing.
Sandzhi Barg: Software, Data curation, Validation, Visualization, Writing – original draft, Writing – review and editing.
Florian Kofler: Methodology, Validation, Writing – review and editing (reviewed methodological design and manuscript drafts).
Abdalla Z. Mohamed: Methodology, Validation, Writing – review and editing (reviewed methodological design and manuscript drafts; provided critical assessment of missing methodological components, including the longitudinal FreeSurfer analysis).

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT to improve the readability of this paper. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Acknowledgments

This work was supported by Anna Valentina Lioba Eleonora Claire Javid Mamasani and the Gemeinnützige Hertie Stiftung.

We thank Martin Reuter (DZNE) and Thomas Kirk (Quantified Imaging) for their invaluable comments on the preprint version.

We thank Mikhail Vasiliev for help with data visualization (see Appendix A).

This material is based upon work supported by the Google Cloud Research Credits program under award GCP19980904.

References

Malo Alefsen de Boisredon d'Assier, Eugene Vorontsov, and Samuel Kadoury. M-genseg: Domain adaptation for target modality tumor segmentation with annotation-efficient supervision. *arXiv preprint arXiv:2212.07276*, 2022.

Ruslan Aliev, Ekaterina Kondrateva, Maxim Sharaev, Oleg Bronov, Anna Marinets, Sergey Subbotin, Alexander Bernstein, and Evgeny Burnaev. Convolutional neural networks for automatic detection of focal cortical dysplasia. In *Advances in Cognitive Research, Artificial Intelligence and Neuroinformatics*, volume 1358 of *Advances in Intelligent Systems and Computing*, pages 582–588, 2021. doi: 10.1007/978-3-030-71637-0_67.

Brian B. Avants, Nicholas J. Tustison, and Gang Song. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011.

Alexander Bernstein, Renat Akzhigitov, Ekaterina Kondrateva, Svetlana Sushchinskaya, Irina Samotaeva, and Vladislav Gaskin. Mri brain imagery processing software in data analysis. *Transactions on Mass-Data Analysis of Images and Signals*, 9(1):3–17, 2018.

B. Billot, D. N. Greve, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining, 2021.

B. Billot, M. Colin, Y. Cheng, S. E. Arnold, S. Das, and J. E. Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain MRI datasets. *Proceedings of the National Academy of Sciences (PNAS)*, 120(8):e2216399120, 2023a. doi: 10.1073/pnas.2216399120.

B. Billot, D. N. Greve, K. Van Leemput, B. Fischl, A. V. Dalca, and J. E. Iglesias. SynthSeg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Analysis*, 89:102878, 2023b. doi: 10.1016/j.media.2023.102878.

Benjamin Billot. Synthseg github repository. https://github.com/BBillot/SynthSeg, 2023. Accessed: 2026-01-24. Runtime documentation in README.md.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, and Juan Eugenio Iglesias. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical Image Analysis*, 86:102789, 2023c. doi: 10.1016/j.media.2023.102789.

BraTS Challenge Organizers. Brain Tumor Segmentation (BraTS) Challenge 2024, 2024a. URL https://www.brats.org/. 750+ meningioma cases; multi-institutional dataset.

BraTS Challenge Organizers. Analysis of the 2024 BraTS Meningioma Radiotherapy Challenge, 2024b. v3 published August 2024.

N Cardoner, R Andero, and M Cano. Impact of stress on brain morphology: Insights into structural biomarkers of stress-related disorders. *Current Neuropharmacology*, 2024.

L Carnevali and A Sgoifo. Resilience and vulnerability: neurobiological perspectives. *Current Opinion in Behavioral Sciences*, 14:85–92, 2018.

A. Casamitjana, C. Falcon, A. Tucholka, G. Operto, T. Tourdias, X. Franceries, V. Montal, J. González-Ortiz, G. Petrí, D. Altomare, et al. NextBrain: A probabilistic histological atlas of the human brain for MRI analysis. *Nature*, 625(7975):109–116, 2025. doi: 10.1038/s41586-025-09708-2.

Cortechs.ai. NeuroQuant® - FDA 510(k) Cleared Brain Volumetry Software, 2007. URL https://www.cortechs.ai/. FDA 510(k) clearance September 2007.

A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis. I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179–194, 1999. doi: 10.1006/nimg.1998.0395. URL https://pubmed.ncbi.nlm.nih.gov/9931268/.

L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945a. doi: 10.2307/1932409.

Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26 (3):297–302, 1945b.

J. Doshi, G. Erus, M. Habes, and C. Davatzikos. MUSE: MUlti-atlas region Segmentation utilizing Ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage*, 134:431–446, 2016. doi: 10.1016/j.neuroimage.2015.11.073.

Polina Druzhinina, Ekaterina Kondrateva, Arseny Bozhenko, Vyacheslav Yarkin, Maxim Sharaev, et al. Brats2021: Exploring each sequence in multi-modal input for baseline u-net performance. In *Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries*, volume 12962 of *Lecture Notes in Computer Science*, pages 194–203, 2022. doi: 10.1007/978-3-031-08999-2_15.

Simon Duchesne, Isabelle Chouinard, Olivier Potvin, Vladimir S Fonov, April Khademi, Robert Bartha, Pierre Bellec, D Louis Collins, Maxime Descoteaux, Rick Hoge, et al. The canadian dementia imaging protocol: harmonizing national cohorts. *Journal of Magnetic Resonance Imaging*, 49(2):456–465, 2019.

L. D. Eggert, S. Sommer, A. Jansen, T. Kircher, and C. Konrad. Accuracy and reliability of automated gray matter segmentation on brains with large lesions. *PLoS ONE*, 7(12): e45081, 2012. doi: 10.1371/journal.pone.0045081.

Santiago Estrada, David Kügler, Emad Bahrami, Peng Xu, Dilshad Mousa, Monique M. B. Breteler, N. Ahmad Aziz, and Martin Reuter. Fastsurfer-hypvinn: Automated sub-segmentation of the hypothalamus and adjacent structures on high-resolutional brain mri. *Imaging Neuroscience*, 1:1–32, 2023. doi: 10.1162/imag_a_00034. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC11576934/.

FastSurfer Developers. Fastsurfer. GitHub repository, 2026. URL https://github.com/Deep-MI/FastSurfer. Accessed: 2026-01-21.

B. Fischl. FreeSurfer. *NeuroImage*, 62(2):774–781, 2012a. doi: 10.1016/j.neuroimage.2012.01.021. URL https://www.sciencedirect.com/science/article/abs/pii/S1053811912000389.

B. Fischl and A. M. Dale. Measuring the thickness of the cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97(20):11050–11055, 2000. doi: 10.1073/pnas.200033797.

B. Fischl, M. I. Sereno, and A. M. Dale. Cortical surface-based analysis. II. Inflation, flattening, and a surface-based coordinate system. *NeuroImage*, 9(2):195–207, 1999. doi: 10.1006/nimg.1998.0396. URL https://pubmed.ncbi.nlm.nih.gov/9931269/.

Bruce Fischl. Freesurfer. *NeuroImage*, 62(2):774–781, 2012b. doi: 10.1016/j.neuroimage.2011.09.015.

FMRIB Analysis Group. FSL – FMRIB Software Library, 2000. URL https://fsl.fmrib.ox.ac.uk/fsl/. First release 2000.

Jean-Philippe Fortin, Nathaniel Cullen, Yvette I. Sheline, et al. Harmonization of cortical thickness measurements across scanners and sites. *NeuroImage*, 167:104–120, 2018. doi: 10.1016/j.neuroimage.2017.11.024.

Katja Franke and Christian Gaser. Estimating the age of healthy subjects from t1-weighted mri scans using kernel methods: Exploring the influence of various parameters. *NeuroImage*, 50(3):883–892, 2012. doi: 10.1016/j.neuroimage.2010.01.005.

FreeSurfer Development Team. FreeSurfer v8.0.0 release notes: Major update with SynthSeg, SynthStrip, and SynthMorph integration, 2024. URL https://surfer.nmr.mgh.harvard.edu/fswiki/ReleaseNotes. Beta 8.0.0 released November 5, 2024; Stable 8.0.0 released February 27, 2025.

TS Frodl, N Koutsouleris, and R Bottlender. Depression-related variation in brain morphology over 3 years: effects of stress? *Archives of General Psychiatry*, 2008.

CP Furtado, KE Hoy, JJ Maller, and G Savage. Cognitive and volumetric predictors of response to repetitive transcranial magnetic stimulation (rtms)—a prospective follow-up study. *Journal of Affective Disorders*, 2012.

C. Gaser and R. Dahnke. CAT - A Computational Anatomy Toolbox for the Analysis of Structural MRI Data. In *Proceedings of the 22nd Annual Meeting of the Organization for Human Brain Mapping*, Geneva, Switzerland, 2016.

C. Gaser, R. Dahnke, P. M. Thompson, K. Kurth, E. Lüders, and Alzheimer's Disease Neuroimaging Initiative. CAT: a computational anatomy toolbox for the analysis of structural MRI data. *GigaScience*, 13:giae049, 2024. doi: 10.1093/gigascience/giae049.

C. D. Good, I. S. Johnsrude, J. Ashburner, R. N. Henson, K. J. Friston, and R. S. Frackowiak. A voxel-based morphometric study of ageing in 465 normal adult human brains. *NeuroImage*, 14(1):21–36, 2001. doi: 10.1006/nimg.2001.0786.

G Gryglewski, P Baldinger-Melich, and R Seiger. Structural changes in amygdala nuclei, hippocampal subfields and cortical thickness following electroconvulsive therapy in treatment-resistant depression: longitudinal analysis. *The British Journal of Psychiatry*, 2019. doi: 10.1192/bjp.2018.224.

A. Hammers, R. Allom, M. J. Koepp, S. L. Free, R. Myers, E. K. Louis, W. Harkness, J. S. Duncan, and L. Lemieux. On brain atlas choice and automatic segmentation methods. *Scientific Reports*, 10:2775, 2020. doi: 10.1038/s41598-020-57951-6.

Xiao Han, Jorge Jovicich, David Salat, André van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, Paul Maguire, Diana Rosas, Nikos Makris, Anders Dale, Bradford C. Dickerson, and Bruce Fischl. Reliability of mri-derived measurements of human cerebral cortical thickness: The effects of field strength, scanner upgrade and manufacturer. *NeuroImage*, 32(1):180–194, 2006. doi: 10.1016/j.neuroimage.2006.02.051.

K. M. Hasan, J. M. Soares, and J. M. Pereira. A survey of methods for brain tumor segmentation-based MRI and a step forward. *Journal of Clinical and Experimental Data*, 10(1):266–287, 2023. doi: 10.1016/j.jcede.2023.01.010.

L. Henschel, S. Conjeti, S. Estrada, C. Diehl, and M. Reuter. FastSurfer - A fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020a. doi: 10.1016/j.neuroimage.2020.116793.

Leonie Henschel, Sailesh Conjeti, Sergio Estrada, Kerstin Diers, Bruce Fischl, and Martin Reuter. Fastsurfer – a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020b. doi: 10.1016/j.neuroimage.2020.117012.

Leonie Henschel, David Kügler, and Martin Reuter. Fastsurfervinn: Building resolution-independence into deep learning segmentation methods—a solution for highres brain mri. *NeuroImage*, 251:118933, 2022. doi: 10.1016/j.neuroimage.2022.118933.

Daniel P Huttenlocher, Gregory A Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.

J. E. Iglesias and M. R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205–219, 2015. doi: 10.1016/j.media.2015.06.012.

Zafer Iscan, Tony B. Jin, Alexandria Kendrick, Bryan Szeglin, Hanzhang Lu, Madhukar Trivedi, Maurizio Fava, Patrick J. McGrath, Myrna Weissman, Benji T. Kurian, Phillip Adams, Sarah Weyandt, Marisa Toups, Thomas Carmody, Melvin McInnis, Cristina Cusin, Crystal Cooper, Maria A. Oquendo, Ramin V. Parsey, and Christine DeLorenzo. Test–retest reliability of freesurfer measurements within and between sites: Effects of visual approval process. *Human Brain Mapping*, 36(9):3472–3485, 2015. doi: 10.1002/hbm.22856.

Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18(2):203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Jorge Jovicich, Mattia Marizzoni, Beatriz Bosch, David Bartrés-Faz, Jennifer Arnold, Jens Benninghoff, Jens Wiltfang, Cristina Rocchi, Agnese Picco, Flavio Nobili, Gianluigi Forloni, Michela Pievani, Silvia Morbelli, Alessandro Padovani, Claudio Babiloni, Simone Rossi, Pierre Payoux, and Giovanni B. Frisoni. Mri-derived measurements of human subcortical, ventricular and intracranial brain volumes: Reliability effects of scan sessions, acquisition sequences, data analyses, scanner upgrade, scanner vendors and field strengths. *NeuroImage*, 46(2):398–408, 2009. doi: 10.1016/j.neuroimage.2009.02.010.

E. Khadhraoui, T. Nickl-Jockschat, H. Henkes, D. Behme, and S. J. Müller. Automated brain segmentation and volumetry in dementia diagnostics: a narrative review with emphasis on FreeSurfer. *Frontiers in Aging Neuroscience*, 16:1459652, 2024. doi: 10.3389/fnagi.2024.1459652.

Florian Kofler, Johannes Wahle, Ivan Ezhov, Sophia Wagner, Rami Al-Maskari, Emilia Gryska, Mihail Todorov, Christina Bukas, Felix Meissen, Tingying Peng, Ali Ertürk, Daniel Rueckert, Rolf Heckemann, Jan Kirschke, Claus Zimmer, Benedikt Wiestler, Bjoern Menze, and Marie Piraud. Approaching peak ground truth. *arXiv preprint arXiv:2301.00243*, 2023. doi: 10.48550/arXiv.2301.00243. URL https://arxiv.org/abs/2301.00243.

Florian Kofler et al. Brainlesion suite: A flexible and user-friendly framework for modular brain lesion image analysis. *arXiv preprint arXiv:2507.09036*, 2024. Software available at: https://github.com/BrainLesion/panoptica.

Ekaterina Kondrateva, Polina Druzhinina, Alexandra Dalechina, Svetlana Zolotova, Andrey Golanov, Boris Shirokikh, Mikhail Belyaev, and Anvar Kurmukov. Negligible effect of brain mri data preprocessing for tumor segmentation. *Biomedical Signal Processing and Control*, 96:106599, 2024.

Ekaterina A. Kondrateva, Polina Belozerova, Maxim G. Sharaev, Evgeny V. Burnaev, Alexander V. Bernstein, and Irina S. Samotaeva. Machine learning models reproducibility and validation for mr images recognition. In *Twelfth International Conference on Machine Vision*, volume 11433, page 114330Z. SPIE, 2020.

A. Lavoie, E. Buc, F. Giraldo, A. Carrillo, K. Hejazi, A. Cocciadiferro, U. Sinha, K. Strickland, A. Kone, A. Levin, et al. BraTS Toolkit: Translating BraTS Brain Tumor Segmentation Challenge Methods into Clinical Practice. *Frontiers in Neuroscience*, 14:125, 2020. doi: 10.3389/fnins.2020.00125.

ME MacDonald and GB Pike. Mri of healthy brain aging: A review. *NMR in Biomedicine*, 34(11):e4564, 2021.

Andre F. Marquand, Iead Rezek, Jan Buitelaar, and Christian F. Beckmann. Understanding heterogeneity in clinical cohorts using normative models: Beyond case-control studies. *Biological Psychiatry*, 80(7):552–561, 2016. doi: 10.1016/j.biopsych.2015.12.023.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34(10):1993–2024, 2015.

Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cian Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

SA Papagni, S Benetti, and S Arulanantham. Effects of stressful life events on human brain structure: a longitudinal voxel-based morphometry study. *Stress*, 14(3):227–232, 2011.

J. M. Peixoto, M. Petrides, and L. Bonilha. Morphometry of medial temporal lobe subregions using high-field T2 MRI in temporal lobe epilepsy. *Epilepsia*, 65(9):e155–e166, 2024. doi: 10.1111/epi.19034.

D. L. Pham, C. Xu, and J. L. Prince. A review of publicly available automatic brain segmentation and tissue classification methods. *Frontiers in Neuroinformatics*, 15:617029, 2021. doi: 10.3389/fninf.2021.617029.

Marina Pominova, Ekaterina Kondrateva, Maxim Sharaev, Alexander Bernstein, and Evgeny Burnaev. Fader networks for domain adaptation on fmri: Abide-ii study. In *Thirteenth International Conference on Machine Vision*, volume 11605, page 116051Z. SPIE, 2021. doi: 10.1117/12.2587348.

Martin Reuter, Nikolaus J. Schmansky, H. Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *NeuroImage*, 61(4):1402–1418, 2012. doi: 10.1016/j.neuroimage.2012.02.084.

A Schaefer, R Kong, and EM Gordon. Longitudinal atrophy patterns in early and late onset alzheimer's disease. *Neurobiology of Aging*, 64:68–76, 2018.

M. I. Sereno, A. M. Dale, A. Y. Liu, and R. B. Tootell. A Surface-based Coordinate System for a Canonical Cortex. *NeuroImage*, 3(3):S184, 1996. doi: 10.1016/S1053-8119(96)80147-3.

P. E. Shrout and J. L. Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428, 1979. doi: 10.1037/0033-2909.86.2.420.

D. Srinivasan, J. Doshi, G. Erus, M. Habes, I. M. Nasrallah, and C. Davatzikos. A comparison of Freesurfer and multi-atlas MUSE for brain anatomy segmentation: Findings about size and age bias, and inter-scanner stability in multi-site aging studies. *NeuroImage*, 223:117248, 2020. doi: 10.1016/j.neuroimage.2020.117248.

Y Statsenko, T Habuza, and D Smetanina. Brain morphometry and cognitive performance in normal brain aging: age-and sex-related structural and functional changes. *Frontiers in Aging Neuroscience*, 13:713680, 2022.

Saori C Tanaka, Ayumu Yamashita, Noriaki Yahata, Takashi Itahashi, Giuseppe Lisi, Takashi Yamada, Naho Ichikawa, Masahiro Takamura, Yujiro Yoshihara, Akira Kunimatsu, et al. A multi-site, multi-disorder resting-state magnetic resonance image database. *Scientific Data*, 8(1):227, 2021. doi: 10.1038/s41597-021-01004-8.

K. S. Tran, D. P. Schmitz, A. Sahin, and K. H. Taber. Test-retest reliability of FreeSurfer-derived volume, area and cortical thickness from MPRAGE and MP2RAGE brain MRI images. *NeuroImage: Reports*, 2:100089, 2022. doi: 10.1016/j.ynirp.2022.100089.

Petru-David Tudosiu, Abhir Bhalerao, Kunthearith Keo, Paul McCarthy, Nils Forkert, Herve Lombaert, and Andriy Fedorov. Brainchop: In-browser mri volumetric segmentation and rendering. *Journal of Open Source Software*, 8(83):5098, 2023. doi: 10.21105/joss.05098.

Sandra Vieira, Lea Baecker, Walter Lopez Pinaya, Rafael Garcia Dias, Cristina Scarpazza, Vince Calhoun, and Andrea Mechelli. Neurofind: Using deep learning to make individualised inferences in brain-based disorders. *Translational Psychiatry*, 15(1):45, 2025. doi: 10.1038/s41398-025-03290-x.

Wellcome Centre for Human Neuroimaging. SPM12 - Statistical Parametric Mapping, 2014. URL https://www.fil.ion.ucl.ac.uk/spm/software/spm12/. Software release October 2014.

G. P. Winston, M. J. Cardoso, J. L. Ledoux, R. J. Dolan, J. Ashburner, and J. S. Duncan. Automated hippocampal segmentation on routine clinical MRI: Validation in temporal lobe epilepsy. *Neurology*, 90(11):e911–e917, 2018. doi: 10.1212/WNL.0000000000005161.

M. Zenk, U. Baid, S. Pati, et al. Towards fair decentralized benchmarking of healthcare ai algorithms with the federated tumor segmentation (fets) challenge. *Nature Communications*, 16(1):6274, 2025. doi: 10.1038/s41467-025-60466-1.

Zongwei Zhou, Vatsal Sodha, Jun Pang, Chen Chen, and Lin Yang. nnformer: Volumetric medical image segmentation via a 3d transformer. *Computerized Medical Imaging and Graphics*, 102:102183, 2023. doi: 10.1016/j.compmedimag.2022.102183.

Appendix A. Related Work

A.1. Software Usage for Test–Retest Brain MRI Morphometry and Segmentation

Test–retest designs are commonly used to quantify the *measurement reliability* of structural brain MRI morphometry and segmentation pipelines, isolating biological effects from variability induced by acquisition, preprocessing, and algorithmic choices. Importantly, reliability (repeatability across repeated scans) is distinct from accuracy (agreement with a reference delineation): stable measurements may still be inaccurate, especially in the presence of pathology or large lesions (Eggert et al., 2012). Reliability is typically reported using intraclass correlation coefficients (ICC) (Shrout and Fleiss, 1979) (and related consistency metrics), whereas overlap-based agreement with a reference labeling is often summarized using the Dice coefficient (Dice, 1945a).

A substantial share of the research literature relies on the `FreeSurfer recon-all` pipeline as a *de-facto baseline* for surface-based morphometry and whole-brain parcellation, largely due to its early introduction, broad adoption, and comparability with prior work. The core surface-based framework was established through cortical surface reconstruction and coordinate-system formulations (Dale et al., 1999; Fischl et al., 1999; Sereno et al., 1996) and consolidated in the widely used `FreeSurfer` software description (Fischl, 2012a), including cortical thickness estimation (Fischl and Dale, 2000). Test–retest studies have examined `FreeSurfer` reliability within-site and between-site, showing that visual QC and approval criteria can materially affect reproducibility estimates (Iscan et al., 2015). More recent work also evaluated reliability across acquisition sequences (e.g., MPRAGE vs. MP2RAGE), highlighting protocol-dependent effects on derived volumes, areas, and thickness (Tran et al., 2022).

Beyond surface-based processing, voxel-based morphometry (VBM) remains a widely used alternative, supported by `SPM12` (Wellcome Centre for Human Neuroimaging, 2014) and `CAT` (Gaser and Dahnke, 2016; Gaser et al., 2024), with classic large-scale VBM studies demonstrating sensitivity to aging effects (Good et al., 2001). The literature also includes multi-atlas segmentation approaches, grounded in extensive methodological work (Iglesias and Sabuncu, 2015) and implemented in toolchains such as `MUSE` (Doshi et al., 2016). In multi-site aging contexts, comparative analyses have reported differences between `FreeSurfer` and multi-atlas methods in terms of size/age bias and inter-scanner stability (Srinivasan et al., 2020). More recently, deep learning pipelines have been adopted to improve throughput and robustness: `FastSurfer` provides a high-throughput deep learning alternative (Henschel et al., 2020a), with subsequent work addressing resolution-independence (Henschel et al., 2022) and specialized sub-segmentation tasks (Estrada et al., 2023). In parallel, `SynthSeg` was proposed to enable segmentation across heterogeneous clinical MRI contrasts and resolutions without retraining (Billot et al., 2021, 2023b), and later scaled for large heterogeneous datasets (Billot et al., 2023a). Tool and atlas choices can substantially affect downstream morphometric conclusions, motivating explicit reporting of atlas/tool decisions (Hammers et al., 2020) and broader syntheses of publicly available segmentation methods (Pham et al., 2021).

Recent `FreeSurfer` releases have also modernized the ecosystem by distributing and integrating deep learning modules (e.g., `SynthStrip`, `SynthSeg`, `SynthMorph`) within the

broader `recon-all` workflow (FreeSurfer Development Team, 2024). Given that `SynthSeg` targets robustness to domain shift and acquisition heterogeneity (Billot et al., 2023b,a), these updates are expected to further increase uptake in research settings where multi-site and real-world clinical variability is a primary concern.

Finally, tool usage diverges in pathology-specific contexts. Brain tumor segmentation typically relies on dedicated lesion segmentation methods and challenge-driven toolkits (e.g., BraTS and clinical translation efforts) rather than general morphometry pipelines (Lavoie et al., 2020; BraTS Challenge Organizers, 2024a,b; Hasan et al., 2023). In temporal lobe epilepsy (TLE), hippocampal/MTL segmentation is commonly validated in clinical cohorts and may leverage specialized imaging contrasts and protocols beyond standard T1-based morphometry (Winston et al., 2018; Peixoto et al., 2024). In dementia-oriented volumetry, narrative reviews continue to report substantial reliance on `FreeSurfer`, reflecting its enduring role as a research baseline (Khadhraoui et al., 2024), while newer deep learning and atlas resources expand the methodological landscape (Casamitjana et al., 2025).

A.2. Usability in Research Practice: Operational Factors Shaping Tool Choice

In research practice, the "usability" of morphometry and segmentation software is determined not only by nominal accuracy or reported reliability, but also by operational burden and reproducibility at scale. Typical workflows are batch pipelines (preprocessing → segmentation/surface reconstruction → feature extraction → statistical analysis), where usability depends on version stability, automation of consistent parameterization, and robustness to heterogeneous data. Related work outside morphometry-specific benchmarking has likewise emphasized the importance of reproducible validation pipelines (Kondrateva et al., 2020) and practical software ecosystems for MRI data analysis (Bernstein et al., 2018). For surface-based pipelines, visual QC is often essential and can meaningfully affect test–retest outcomes, motivating explicit reporting of QC/approval procedures (Iscan et al., 2015).

Compute cost and runtime are additional practical constraints. Computationally intensive classical pipelines have motivated accelerated alternatives and deep learning-based replacements in high-throughput studies (Henschel et al., 2020a). Likewise, robustness to heterogeneous clinical data (variable contrast, resolution, and quality) is increasingly treated as a usability requirement; methods explicitly designed for such heterogeneity (e.g., `SynthSeg`) are therefore increasingly used as pragmatic components in research pipelines (Billot et al., 2023b,a). These operational factors—QC burden, heterogeneity sensitivity, compute cost, and version stability—are especially consequential in test–retest and multi-site settings.

A.3. Practical Guidelines for Researchers

Based on our limited analysis of segmentation stability, we offer the following guidelines for researchers using automated brain parcellation as per Jan 2026:

Scanner manufacturer effects. Cross-manufacturer comparisons (e.g., Siemens vs. GE) produce substantially degraded metrics. In our SRPBS dataset, comparing scans from HKH (Siemens Spectra) and HUH (GE Signa HDxt) yielded HD95 values 7–8× higher than same-manufacturer comparisons (e.g., 7.81 mm vs. 1.0 mm for caudal anterior cingulate). For

multi-site studies, we recommend stratifying analyses by scanner manufacturer or applying harmonization methods before pooling data.

Software selection. FastSurfer demonstrated superior volume consistency (bias reduced 5–10×) and higher Dice scores for cortical structures (+8–12% improvement over FreeSurfer). FreeSurfer showed marginally better boundary precision (HD95 approximately 0.3 mm lower). For volumetric studies, FastSurfer is preferred; for cortical thickness or atrophy studies where boundary precision is critical, FreeSurfer remains a valid choice despite longer processing time.

Structure-specific reliability. Cortical structures showed substantial variability in reproducibility. The most stable regions were superior temporal (Dice = 0.849) and superior frontal cortex (Dice = 0.823), while the least stable were lateral orbitofrontal (Dice = 0.734) and entorhinal cortex (Dice = 0.757). We recommend applying stricter QC thresholds (Surface Dice > 0.90) for entorhinal and orbitofrontal regions, and validating findings involving these structures with manual review.

QC Recommendations. Our analysis reveals that universal Surface Dice thresholds may be inappropriate for quality control, as structure-specific variability ranges substantially—from 5th percentile values of 0.50 (caudal anterior cingulate) to 0.93 (ventral diencephalon). We recommend structure-specific thresholds based on the 10th percentile of the reference distribution: segmentations falling below this threshold should be flagged for manual review. The complete table of empirically derived failure and acceptability thresholds for all analyzed structures is available in our code repository.

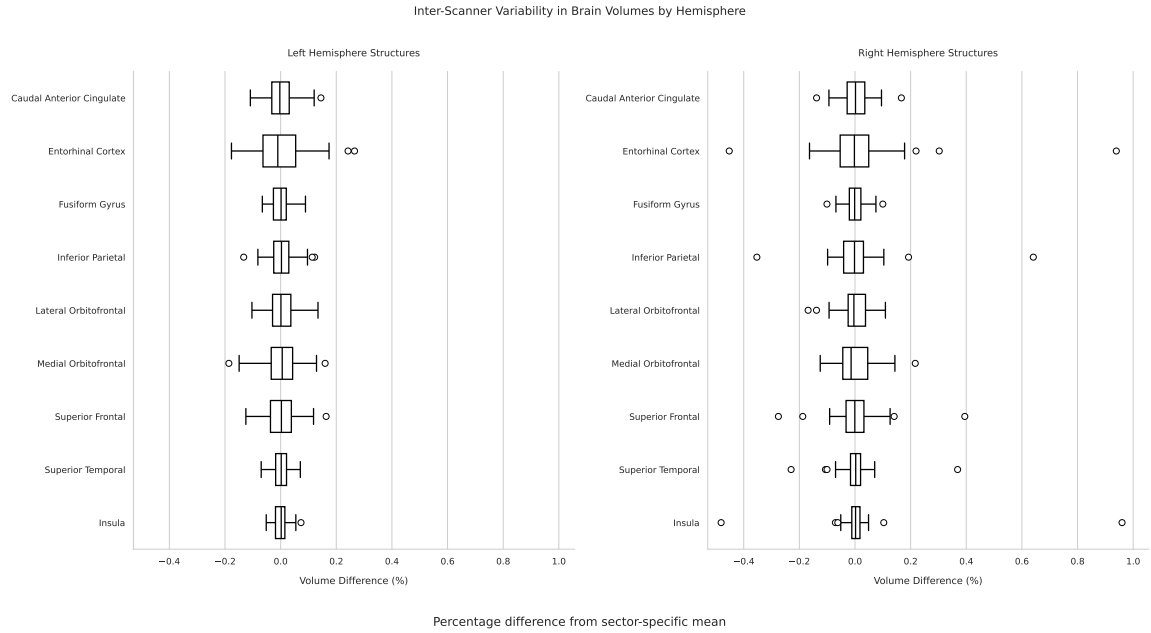

Figure 6: Inter-scanner variability of cortical volumes in the SIMON dataset. Boxplots show the percentage difference from the structure-specific mean across repeated sessions, grouped by hemisphere. **Take-home message:** cortical volume reproducibility is strongly region-dependent, with some regions showing much wider spread and outliers than others, so a single global QC threshold is unlikely to be sufficient.

Tool / Family	Year	Key reference	Typical research usage (test–retest / morphometry)
FreeSurfer (`recon-all`)	1999–2012; 2025 update	Dale et al. (1999); Fischl et al. (1999); Fischl (2012a); FreeSurfer Development Team (2024)	De-facto baseline for surface-based morphometry and parcellation; test–retest reliability depends on QC/approval (Iscan et al., 2015; Tran et al., 2022). **v8 adds Synth* DL modules (incl. SynthSeg)** (FreeSurfer Development Team, 2024), likely increasing adoption for domain-robust processing (Billot et al., 2023a).
SPM12 (VBM)	2014	Wellcome Centre for Human Neuroimaging (2014)	VBM workflows for group-level morphometry; common alternative to surface-based processing (Good et al., 2001).
CAT (VBM toolbox)	2016–2024	Gaser and Dahnke (2016); Gaser et al. (2024)	VBM-oriented toolbox; widely used for streamlined preprocessing and morphometric inference.
FSL (structural suite)	2000	FMRIB Analysis Group (2000)	Structural preprocessing/segmentation components used in broader workflows; sometimes included in reproducibility comparisons.
MUSE (multi-atlas)	2016	Doshi et al. (2016)	Multi-atlas ROI segmentation; compared to FreeSurfer in multi-site aging with reports on stability and bias (Srinivasan et al., 2020).
FastSurfer (DL pipeline)	2020–2021	Henschel et al. (2020a, 2022)	High-throughput DL alternative producing FreeSurfer-like outputs; used when runtime and scalability dominate.
SynthSeg (robust DL)	2021–2023	Billot et al. (2021, 2023b,a)	Designed for heterogeneous clinical MRI without retraining; adopted for domain robustness and large-scale analyses.
NeuroQuant (clinical volumetry)	2007	Cortechs.ai (2007)	Clinical-facing volumetry; sometimes used in research for clinical comparability.
BraTS ecosystem (tumor)	2020–2024	Lavoie et al. (2020); BraTS Challenge Organizers (2024a,b)	Dedicated lesion segmentation workflows for neuro-oncology; distinct from general morphometry pipelines (Hasan et al., 2023).
TLE hippocampal / MTL	2018–2024	Winston et al. (2018); Peixoto et al. (2024)	Epilepsy-focused hippocampal/MTL morphometry with clinical validation and specialized protocols/contrasts.
NextBrain (atlas)	2025	Casamitjana et al. (2025)	Histology-informed probabilistic atlas; emerging use for fine-grained MRI analysis.

Table 5: Publication timeline and typical research usage patterns for common brain MRI morphometry and segmentation toolchains in test–retest and related study designs.

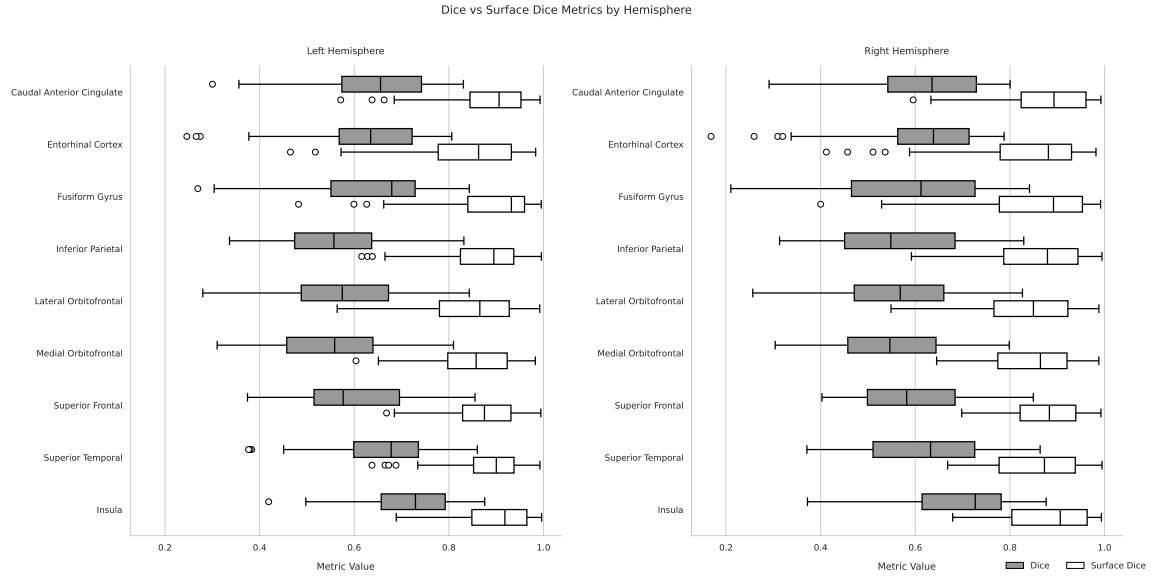

Figure 7: Inter-scanner variability of cortical overlap metrics in the SIMON dataset. Boxplots show Dice and Surface Dice between repeated scans, grouped by hemisphere. **Take-home message:** overlap remains generally high, but reliability is not uniform across cortical regions or hemispheres, which motivates region-aware QC rather than assuming all structures are equally stable.

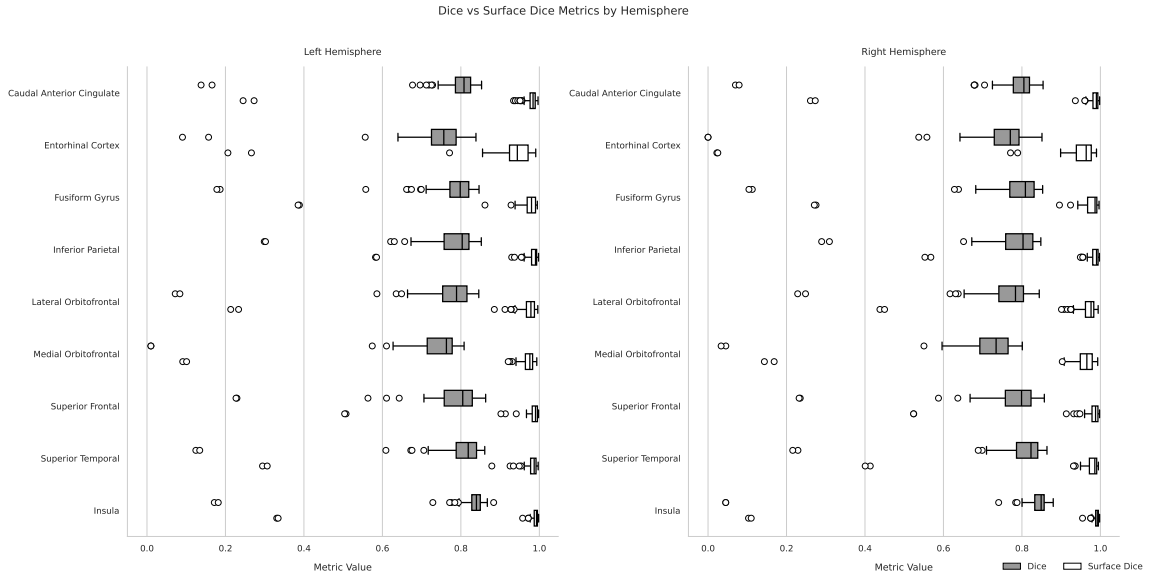

Figure 8: Dice and Surface Dice distributions across cortical regions in the left and right hemispheres of the SIMON dataset using FreeSurfer. Each structure is evaluated over multiple longitudinal scans from the same individual. Surface Dice (white boxes) consistently exceeds traditional Dice (gray boxes), especially in regions with complex geometry such as the entorhinal cortex and insula. **Take-home message:** a surface-aware criterion is less punitive than standard Dice and better separates minor boundary jitter from clear segmentation failures, making it a more informative QC signal for cortical morphometry.