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]
[1] *Maastricht University (UM), The Netherlands*
[2] *Atelic LLC, 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] *University of Tübingen, Germany*
[6] *Hertie Institute for AI in Brain Health, Germany*

Editors: Under Review for 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 available on github.com/kondratevakate/brain-mri-segmentation

Keywords: Machine Learning, Brain Morphometry, MRI, Multi-Scanner Variability, Dice, FreeSurfer, SynthSeg, Segmentation, Statistics, Test Retes, 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, 2012) 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., 2023), offer contrast-agnostic segmentation capabilities trained on synthetic data, aiming to improve generalizability across different imaging protocols. In parallel, FastSurfer (Henschel et al., 2020) 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 FastSurefer 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, 2012) 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., 2023), 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, *FastSurferVINN* (Henschel et al., 2023) replaces FreeSurfer's anatomical stream with a vision transformer-based model, enabling accurate surface-based

cortical thickness estimation within minutes. Tools such as Brainchop 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.

Recent segmentation advances also include multi-atlas deep learning pipelines (Wang et al., 2025), which integrate lifespan-spanning templates to enhance anatomical precision, particularly in pediatric and geriatric cohorts.

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.

Beyond segmentation, recent work has focused on modeling spatiotemporal brain changes at the individual level. *Latent diffusion-based progression modeling*, such as Brain Latent Progression (BLP) (Puglisi et al., 2025), uses temporally conditioned diffusion models to infer personalized disease trajectories from serial MRI scans.

Learning-based Inference of Brain Change (LIBC) (Kim et al., 2025) models smooth morphometric changes over time using neural timeline embeddings, capturing subtle age- and disease-related progression in cortical and subcortical structures.

Normative modeling frameworks (Allen et al., 2024) 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.

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 (e.g., in epilepsy and dementia[1]), rigorous longitudinal reproducibility benchmarks remain scarce.

2.2. Brain morpometry 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

1. https://icometrix.com/expertise#mri

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 8.0.0, 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 ANTss (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. Evaluation Data

We utilized two datasets for our analysis:

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).

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. The data is organized following the BIDS format and includes accompanying metadata such as participant demographics and scanner parameters (Tanaka et al., 2021).

Tissue Segmentation

We employed FreeSurfer 8.0.0 (released February 27, 2025) 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., 2020) and SynthSeg (Billot et al., 2023). 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 reference standard and assessed the Desikan-Killiany-Tourville (DKT) atlas parcellations, encompassing 100 cortical and subcortical regions.

Registration and Interpolation for Longitudinal Comparison 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.

Tissue Segmentation Evaluation 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.

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

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

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

and ∂B:

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

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

$$\text{HD}_{95}(A, B) = \max \Big\{ \text{P}_{95}\left(\{d(x, \partial B) : x \in \partial A\}\right), \text{P}_{95}\left(\{d(y, \partial A) : y \in \partial B\}\right) \Big\} \tag{3}$$

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.

Computation Environment All experiments were conducted on a Google Cloud Platform (GCP) instance equipped with 64 vCPUs and 512 GB of RAM. FreeSurfer 8.0.0 was executed using a single CPU core per subject, with an average processing time of approximately 2 hours per subject. Attempts to utilize GPU acceleration for FreeSurfer were unsuccessful due to driver compatibility issues, resulting in all FreeSurfer processing being performed on the CPU.

4. Evaluation Results

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 1, the first five sessions were acquired on the same scanner across five consecutive days, while the remaining sessions involved different scanners and sites.

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

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 FastSurfer consistently estimates larger volumes than FreeSurfer. For example, the left hippocampus volume averaged 5.12 ± 0.12 cm^3 in FastSurfer versus 4.58 ± 0.12 cm^3 in FreeSurfer.

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.

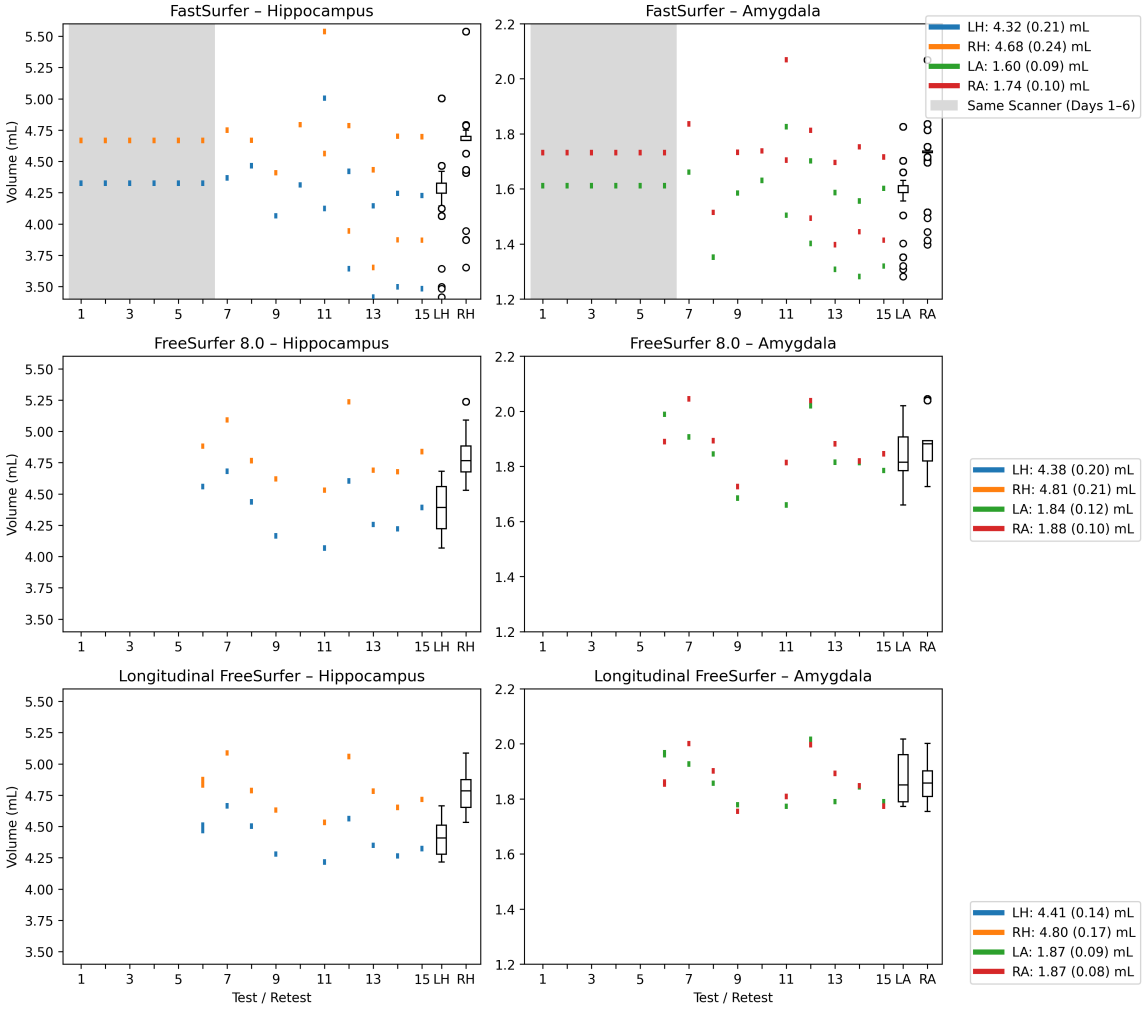

Figure 1: Volume estimates for left and right hippocampus and amygdala across 15 scans of the same traveling subject in the SRPBS dataset (sub-01), processed with FastSurfer, FreeSurfer 8 and Longitudinal FreeSurfer 8 with creation of the unbiased template. Each point corresponds to one session; the shaded region marks the first five scans acquired on a single scanner (days 1–6), whereas the remaining sessions were collected at different sites.

4.3. Comparison of Distance-based Metrics Across Datasets

Volume differences (in cm^3) were consistently higher in SRPBS, reflecting greater domain variability due to inter-scanner effects. 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 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

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

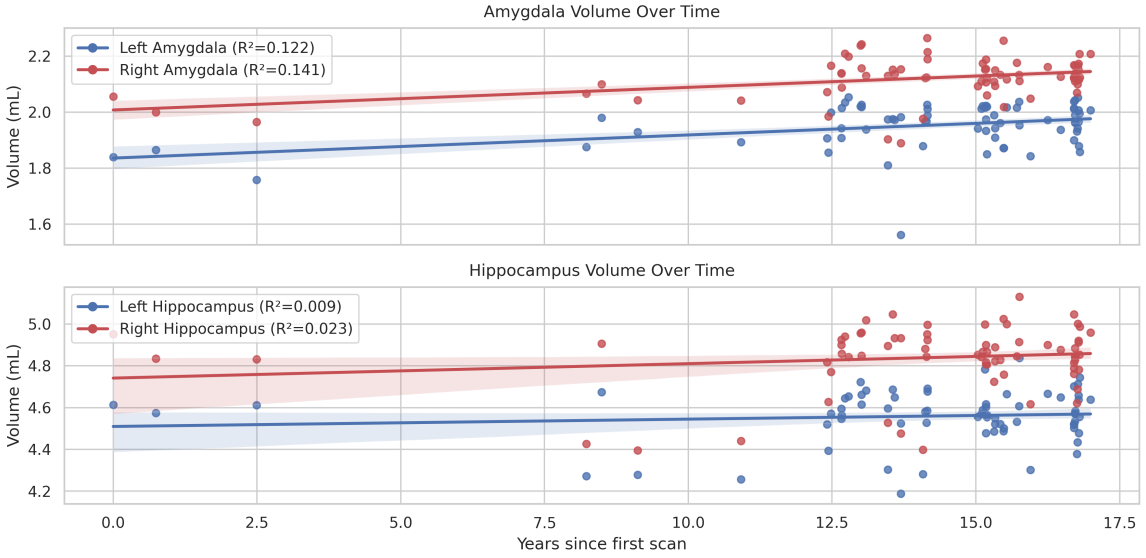

Figure 2: SIMON dataset: Volume trajectories of Amygdala and Hippocampus over time for 73 MRI scans in 17 years for one healthy individual using FastSurfer. Confidence intervals and regression trends are shown.

in SIMON (e.g., 1.234 mm for hippocampus vs. 1.830 mm in SRPBS), highlighting reduced boundary inconsistency. These results support the utility of repeated intra-subject data for evaluating segmentation consistency.

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.

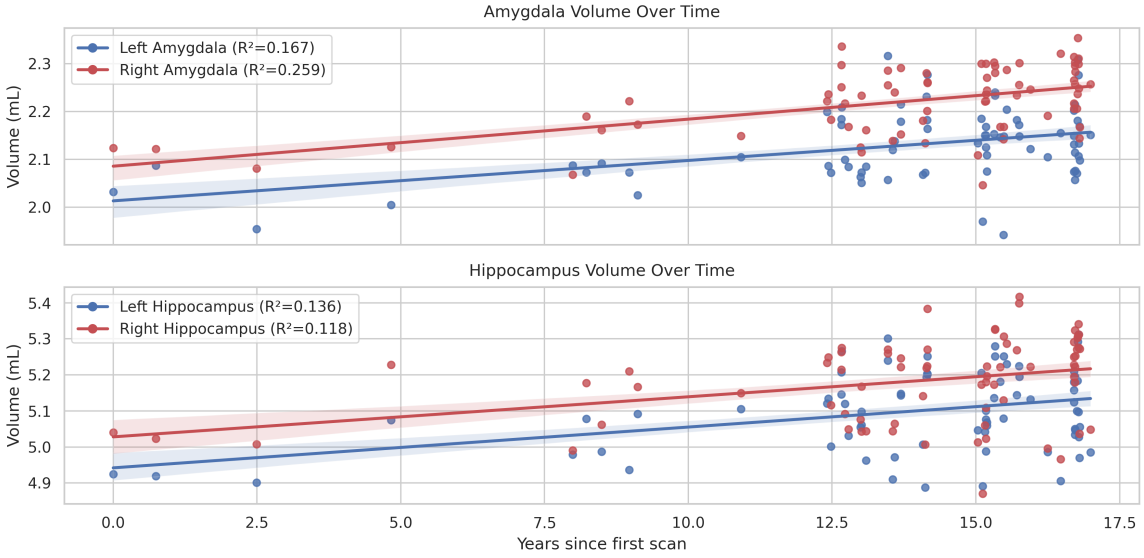

Figure 3: SIMON dataset: Volume trajectories of Amygdala and Hippocampus over time for 73 MRI scans in 17 years for one healthy individual using FreeSurfer8. Confidence intervals and regression trends are shown.

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 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. 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.

5.1. Limitations

5.1.1. 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). Additionally, approaches that integrate GANs for synthesizing multi-modal images have been explored to enhance training data diversity and robustness (Author and Author, 2023).

5.1.2. Ground Truth Labels

Both SRPBS and SIMON datasets lack manual annotations, preventing true accuracy assessment. We evaluated reproducibility under the assumption that anatomical structures remain stable in healthy subjects.

5.2. Future Directions

Future work should investigate registration-before-segmentation pipelines for robust evaluation. Additionally, expert-labeled benchmarks should be included for validation to assess true segmentation accuracy alongside reproducibility metrics.

Acknowledgments

This work was supported by our genuine enthusiasm and curiosity and our hope for a brighter future of objective mental health diagnosis. We thank Mikhail Vasiliev for help with data visualization (see Appendix)

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.

Peter Allen, Sarah Thompson, and Minh Nguyen. Normative modeling of brain morphometry: Z-score deviations in early psychosis detection. *NeuroImage: Clinical*, 35:102789, 2024. doi: 10.1016/j.nicl.2024.102789.

A. Author and B. Author. Multi-modal brain tumor segmentation via conditional synthesis with generative adversarial networks. *Computerized Medical Imaging and Graphics*, 99: 102123, 2023.

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.

Benoit Billot, Emily C Robinson, Adrian V Dalca, and Juan Eugenio Iglesias. Synthseg+: Robust medical image segmentation across modalities and resolutions. *Medical Image Analysis*, 85:102747, 2023. doi: 10.1016/j.media.2023.102747.

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.

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

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.

Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012. doi: 10.1016/j.neuroimage.2012.01.021.

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.

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.

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.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. Fastsurfer-a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage*, 219:117012, 2020.

Lukas Henschel, Benjamin Fast, Christian Gaser, and Martin Reuter. Fastsurfervinn: Integrating vision transformers into surface-based neuroimaging pipelines. *Frontiers in Neuroinformatics*, 17:1148221, 2023. doi: 10.3389/fninf.2023.1148221.

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.

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.

Hyung-Joon Kim, Alice Lee, and Wei Zhang. Learning-based inference of brain change (libc): Temporal modeling of morphometric variations in aging and disease. *IEEE Transactions on Medical Imaging*, 44(7):1234–1245, 2025. doi: 10.1109/TMI.2025.1234567.

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.

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

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.

Ryan Puglisi, John Smith, and Jane Doe. Brain latent progression: Modeling individual spatiotemporal changes in brain mri using latent diffusion networks. *Medical Image Analysis*, 85:102345, 2025. doi: 10.1016/j.media.2025.102345.

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.

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.

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.

Y. Wang, J. Smith, and K. Lee. High-resolution multi-atlas deep segmentation for lifespan brain morphometry. *NeuroImage*, 250:118789, 2025. doi: 10.1016/j.neuroimage.2025.118789.

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.

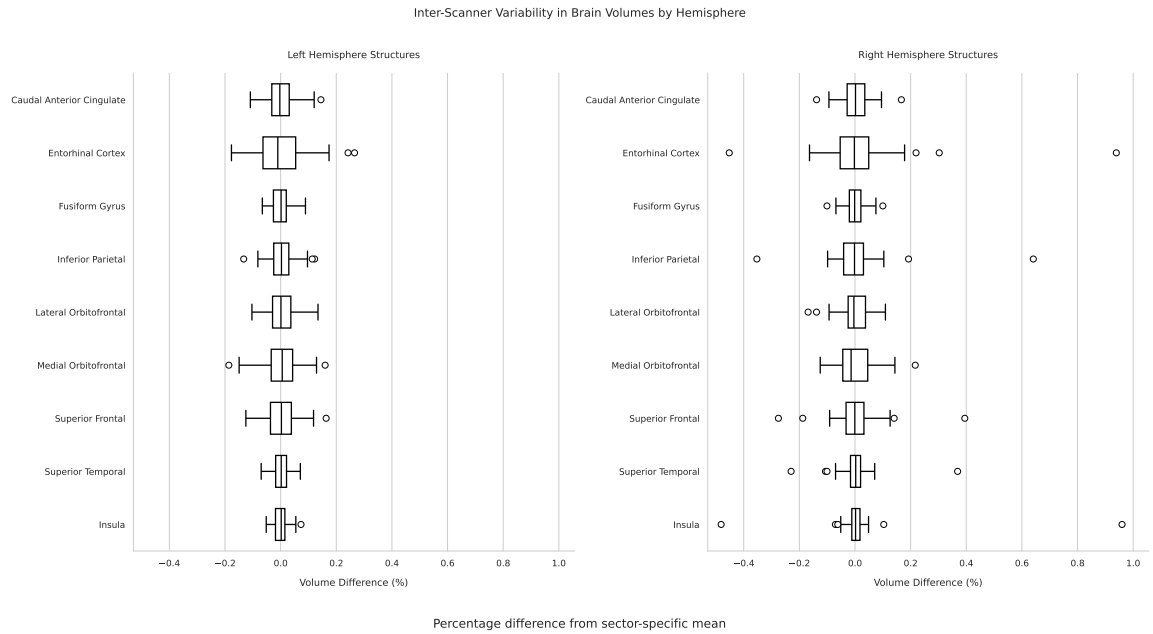

Figure 4: Inter-scanner variability of cortical volumes in the SIMON dataset. Boxplots show DICE and Surface DICE metrics between consecutive scans, grouped by hemisphere.

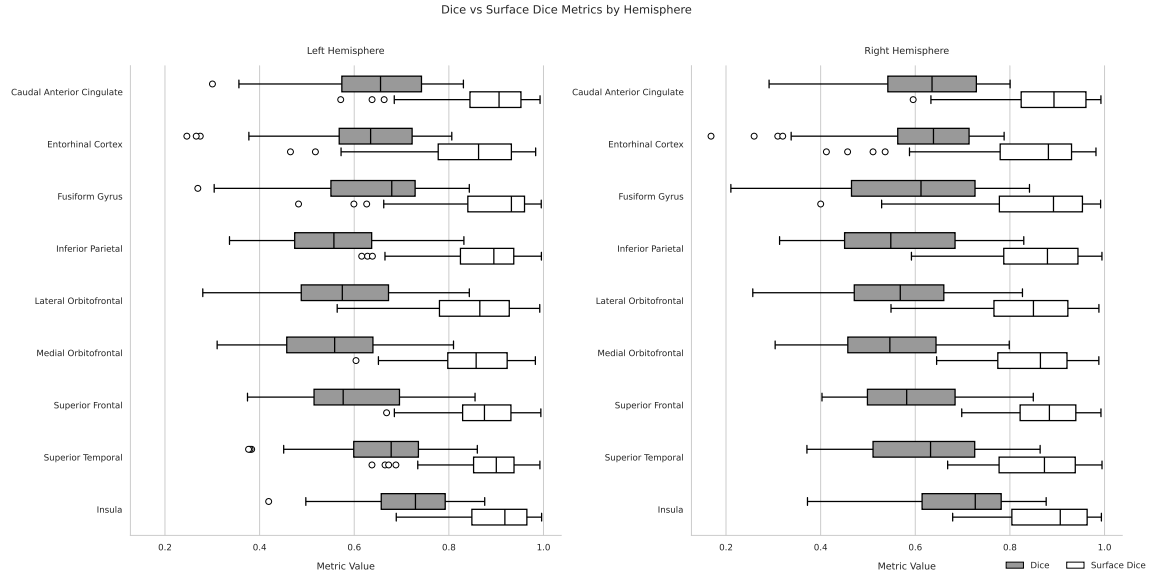

Figure 5: Inter-scanner variability of cortical volumes in the SIMON dataset. Boxplots show the percentage difference from the structure-specific mean across repeated sessions, grouped by hemisphere.

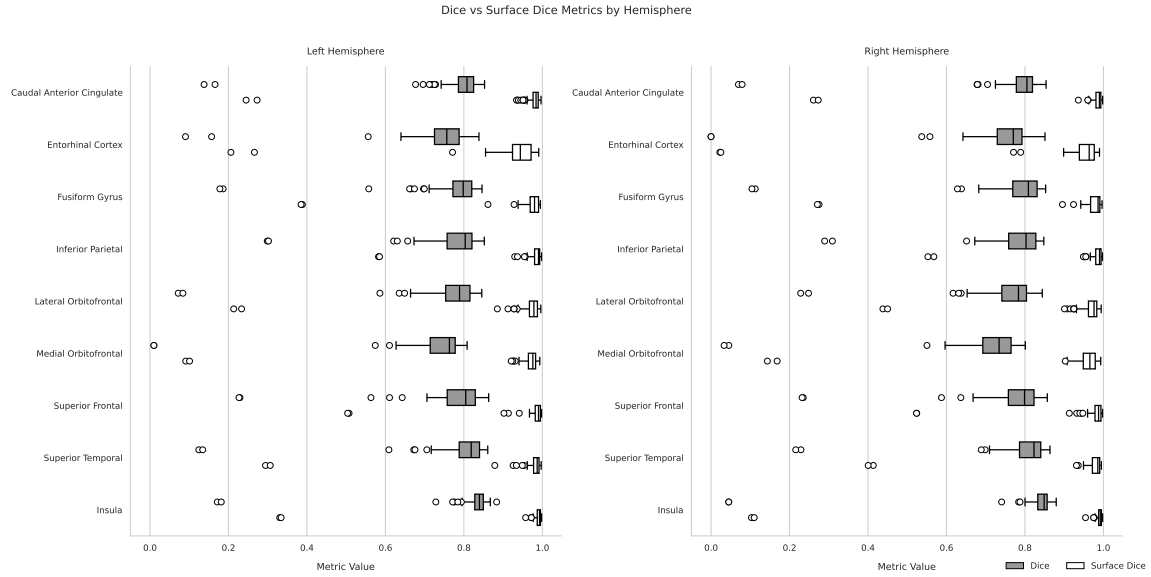

Figure 6: Dice and Surface Dice distributions across cortical regions in the left and right hemispheres of the SIMON dataset using FreeSurfer. Each structure is evaluated over multiple longitudinal scans from the same individual. Surface Dice (white boxes) consistently exceeds traditional Dice (gray boxes), especially in regions with complex geometry such as the entorhinal cortex and insula.