Mind the Shift: Variability in Brain Segmentation Across MRI Scanners

Ekaterina Kondrateva EKATERINA.KONDRATEVA@MAASTRICHTUNIVERSITY.NL
Maastricht University
Sandzhi Barg SANDZHI.BARG@GMAIL.COM

Editors: Under Review for MIDL 2025

Abstract

Accurate measurement of brain morphometry is essential not only for detecting abnormalities, but also for tracking subtle structural changes in healthy individuals—such as those linked to stress or neuroplasticity. While segmentation of pathological regions (e.g., BraTS challenge) has benefited from well-established benchmarks and metrics, the reproducibility of morphometric estimates in healthy brains, particularly across MRI scanners and protocols, remains underexplored.

In this study, we assess the consistency of brain volume measurements using FreeSurfer 8 with integrated SynthSeg across 73 longitudinal MRI scans from a single individual (SIMON dataset). We quantify inter-scan variability on segmentation stability using absolute volume difference, Dice, and Surface Dice metrics.

Our results show that even state-of-the-art tools exhibit sensitivity to domain shift, with average variability in regional brain volumes reaching 3.1% and maximum of 19% for individual subcortical regions. This highlights a critical limitation for using current segmentation pipelines in personalized brain health monitoring or early detection of conditions such as Alzheimer's disease.

Keywords: Brain Morphometry, MRI, Multi-Scanner Variability, Dice, FreeSurfer, SynthSeg, Segmentation, ANTs

1. Introduction

MRI-accurate brain morphometry is essential for studying aging, neurodegeneration, and tracking structural changes. Although most of the AI work in medical imaging focused on pathology segmentation (e.g., tumors in BraTS), morphometric analysis in healthy brains, especially across domains, remains underexplored.

FreeSurfer (Fischl, 2012) has long been a standard tool, with recent versions that integrate SynthSeg (Billot et al., 2023a), a contrast-agnostic model trained on synthetic data. FastSurfer and Brainchop offer faster alternatives, but SynthSeg remains a benchmark for healthy morphometry.

Despite clinical relevance (e.g., epilepsy, dementia[1]), few studies evaluate the reproducibility of volumetric estimates under real-world conditions.

We evaluated segmentation reproducibility using the SIMON dataset: 73 longitudinal scans over 17 years. We evaluated nine cortical and eight and eight subcortical regions using volume difference, surface Dice, and propose outlier filtering based on segmentation stability.

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

2. Methods

Dataset. The SIMON dataset (Duchesne et al., 2019) includes 73 T1-weighted scans of a healthy male (age 29–46) across multiple sites and scanners over 17 years. Mean interval: 86.2 days; min: 0; max: 1154 days.

Segmentation. T1 scans were segmented using FreeSurfer 8.0.0 Release (Feb 27 2025) with SynthSeg(Billot et al., 2023b,a) and FastSurfer(Henschel et al., 2020). No pre-alignment was applied to preserve raw T1 contrast.

ROI Analysis. We evaluated 9 cortical and 8 subcortical bilateral structures. Differences in scan-to-scan for subsecient magnetic resonance imaging sessions were treated as domain variation.

Registration. For surface metrics, ANTs (Avants et al., 2011) rigid registration was applied using transforms computed from the original T1s. We compared ANTs interpolation modes "Multilabel", and "NearestNeighbor". And registration to the first session or to MRI space asymmetric atlas.

Metrics. We report absolute volume differences (native space), and Dice/Surface Dice scores[2] post-registration.

Computations. Experiments ran on GCP (64 vCPUs, 512 GB RAM). FreeSurfer 8 used a single core (2h/subject). GPU acceleration for SynthSeg failed due to driver issues.[3]

3. Results

Volumetric changes: FastSurfer `recon-all` failed on 8 sessions thus we report only results from FreeSurfer 8 with SynthSeg. Average scan-to-scan variation in subcortical volumes was 3.1%, with individual deviations up to 15% (Thalamus) and 20% (Pallidum). For cortical parcellations average difference constituted 5% and outliers differed by over 40 (Superior Frontal, Inferior Temporal) to 90% (Insula). **Registration:** Chose of a registration template was respondible for maximum 0.07% change from mean values and chose of interpolation strategy - for 1.72% in mean volume changes from original segmentation.

Table 1: 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

Filtering segmentation outliers: We applied quality-based filtering to subcortical segmentations using thresholds on Surface Dice and Dice metrics. A Surface Dice threshold of 0.92 led to the exclusion of 5.0% of regions as low-quality outliers. In contrast, Dice-based filtering at 0.80 removed over 50% of regions, showing its higher sensitivity to shape differences. Mean absolute percentage error (MAPE) across the retained regions is summarized

2. https://github.com/google-deepmind/surface-distance
3. Code: https://github.com/kondratevakate/brain-mri-segmentation

below.

Conclusion: Even state-of-the-art tools like FreeSurfer 8 with SynthSeg are sensitive to scanner variability even for typical 1.5 T machines (no low dose, no brain pathologies). This work highlights the need for robust preprocessing pipelines for longitudinal and multi-scanner studies.

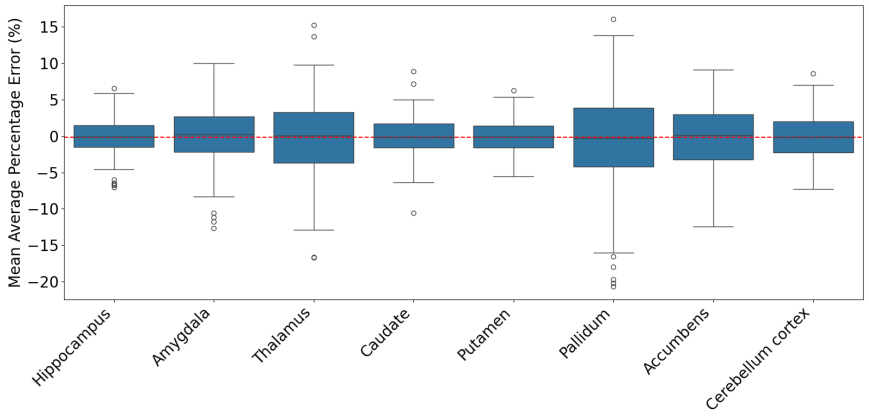

Figure 1: Volume deviation from scan-specific means for subcortical structures. Each boxplot shows segmentation variability across 72 comparisons (for 73 scans) in MAPE, for both right and left paired structures.

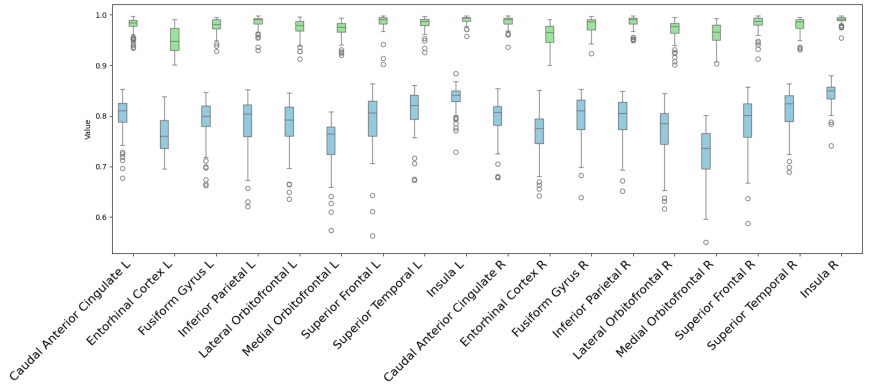

Figure 2: Distribution of Dice (blue) and Surface Dice (green) scores across 73 MRI sessions for 16 bilateral subcortical regions. Surface Dice scores are consistently higher and more stable, highlighting sensitivity of classic Dice to slight boundary misalignments.

References

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.

Benjamin Billot, Douglas N Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V Dalca, Juan Eugenio Iglesias, et al. Synthseg: Segmentation of brain mri scans of any contrast and resolution without retraining. *Medical image analysis*, 86:102789, 2023a.

Benjamin Billot, Colin Magdamo, You Cheng, Steven E Arnold, Sudeshna Das, and Juan Eugenio Iglesias. Robust machine learning segmentation for large-scale analysis of heterogeneous clinical brain mri datasets. *Proceedings of the National Academy of Sciences*, 120(9):e2216399120, 2023b.

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.

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.