

Online Resource 1

Supplementary text, tables, figures and algorithms for

How reliable is deep learning segmentation of the infant brain? Variability and failure analysis of segmentation pipelines under a common protocol in 51 infants aged 1 to 9 months

Submitted to *Pediatric Radiology*. Section, table, figure and algorithm numbers prefixed with “S” refer to this document; unprefixed numbers refer to the main text. References cited in this document are listed at its end.

S1 Rationale for the study questions

RQ1: magnitude and interpretation of pipeline differences. Segmentation benchmarks commonly rank methods by mean Dice, and small numerical margins are often sufficient for a method to be reported as superior. In infant brain MRI, UinTSeg was reported to achieve the highest overall performance among several competing methods although its Dice improvement over nnU-Net across infants aged 0–24 months was 0.64 percentage points for grey matter and 0.51 percentage points for white matter, with a larger 1.12-point difference for cerebrospinal fluid [1]; on the iSeg-2017 benchmark, MAGNET reported a Dice improvement from 0.835 to 0.840 over a unimodal UNETR for T1-weighted segmentation [2]. Margins of approximately 0.5–1.5 Dice percentage points over strong baselines are likewise described as performance gains in recent brain-tumour segmentation work [3, 4], whereas the original iSeg-2017 analysis found that no method showed strongly statistically significant superiority over all competing methods across the evaluated tissues and metrics [5]. Convolutional encoder–decoder networks have been the mainstay of volumetric brain segmentation since multi-scale 3D architectures were shown to work at scale [6]. RQ1 therefore asks how large the observed differences are when data, subject-disjoint folds, evaluation cases and metrics are held common, answered descriptively and by the effect-size and cluster-robust procedures of Algorithm S2 (Step 1); RQ1.1 asks whether differences among the three 3D pipelines exceed observed retraining variation and RQ1.2 whether rankings are stable across cross-validation partitions (Algorithm S2, Step 2).

RQ2: aggregation and failure sensitivity. In a large-scale reanalysis of biomedical image-analysis competitions, changing the aggregation strategy substantially altered method rankings and for some tasks changed the reported winner, and leave-one-case-out analyses showed that individual test cases could influence the leaderboard [7]. This matters when performance distributions contain a few severe failures, because the arithmetic mean and pooled error summaries are disproportionately affected whereas median or rank-based summaries reflect a different aspect of behaviour; current validation guidance asks that aggregation strategy and the treatment of failed or empty predictions be made explicit [8, 9]. RQ2.1 asks whether interpretation changes with mean, median or subject-level ordering; RQ2.2 whether summaries are driven by a few severe failures.

RQ3: metric and structure behaviour. Larger hippocampal subfields achieve substantially higher Dice than smaller ones while boundary distance is less size-dependent [10]; Dice is particularly sensitive to small spatial shifts in thin hypothalamic subunits [11]; and regions with high surface-to-volume ratio show rapid Dice reductions even when the underlying disagreement is limited [12]. RQ3.1 asks how structure size and surface-to-volume ratio affect overlap and surface metrics across the evaluated infant structures.

RQ4: robustness via consensus. Combining predictions is widely used to exploit complementary errors, from the cross-validation ensembling built into nnU-Net [13] to the majority voting of AssemblyNet across overlapping brain regions [14] and a consensus of five lesion-segmentation algorithms in multiple sclerosis, where the gain depended on metric and lesion burden [15]. An ensemble should not be assumed to provide a meaningful gain over its strongest constituent. RQ4.1 asks whether a simple cross-fitted plurality consensus improves robustness relative to the individual pipelines (Algorithm S1).

Exploratory questions. EQ1 asks whether performance is associated with chronological age or with a prespecified joint contrast index, given the marked contrast changes across the isointense period. EQ2 asks what longitudinal behaviour is observed in the repeated-visit subset; with only 17 infants contributing intervals it is descriptive rather than inferential.

Table S1: Comparison of typical prior infant-segmentation benchmark practice and the reporting adopted in this work.

Aspect	Typical prior practice	This work
Cross-validation	Single train/test split, or subject overlap across folds not verified	Subject-disjoint five-fold; evaluation case list verified against the split file
Pipeline training status	Adult-trained pipelines applied off the shelf to infant data (e.g. SynthSeg, SamSeg; Alexander et al. 2026, main-text reference list)	All four pipelines trained on this cohort under matched subject-disjoint folds
Retraining variability	Rarely repeated; single run per configuration	Repeated runs characterise an empirical noise floor (main text, Table 4a)
Fold-level variability	Not reported separately from the pooled mean	Fold-specific paired differences reported; sign changes shown (main text, Table 4b)
Aggregation choice	Single cohort mean	Mean, median, win count, and equal-weight vs. pooled-surface aggregation compared
Rare failures	Averaged into the cohort mean	Explicitly counted and cross-referenced; sensitivity analysis excluding influential scans
Metric implementation	Often unstated	Independently validated against MONAI and a SciPy cKDTree implementation
Empty predictions	Frequently dropped silently	Explicit predefined policy; valid denominators reported per metric

S2 Age metadata and fold balancing

The `ses-*mo` session identifiers denote nominal scheduled visits. True age at scan exceeded the nominal label by +0.59 months on average (range −0.8 to +1.8). Age at scan ranged from 1.0 to 9.4 months (mean 5.34, SD 2.31 across the 71 visits; mean 5.33, SD 2.02 across the 51 infants), which reconciles the 1–8 month span of the session labels with the 1–9 month range described for the published resource: a single visit (`sub-198202_ses-8mo`) had a true age of 9.4 months. Forty-three of 71 visits were of female infants (31 of 51 infants; 60.8%). Folds were assigned before the age metadata were retrieved and were balanced on nominal labels; nominal label and true age were rank-concordant (Spearman $\rho = 0.984$), and repeating the age analysis and the variance model with true age left every headline result unchanged (main text, EQ1).

S3 Consensus vote composition

Over 51,985,811 foreground voxels, votes were 68.672% unanimous, 26.975% three-to-one majority, 0.863% unique plurality, 3.483% two-to-two tie and 0.008% four-way split. The cross-fitted tie priority placed 3D TransUNet first for all 71 visits, with the second and third positions exchanged for the 14 visits of one fold, confirming that cross-fitting was active.

S4 Relation to the Baby Open Brains measurement-error study

Alexander and colleagues applied four adult-trained pipelines off the shelf without retraining to the same 71 visits from the same 51 participants (SynthSeg [16], SamSeg [17], FastSurfer [18], and a pretrained transformer model), excluded the latter two after anatomically implausible segmentations, and reported that SynthSeg reached Dice above 0.8 for most structures while SamSeg systematically overestimated ventricular and whole-brain volume by roughly 130–190%, that segmentation quality improved with age, and that segmentation-related limits of agreement for volume and fractal dimension frequently exceeded published between-group developmental effects [19]. The estimands differ from ours (downstream measurement error versus stability of the pipeline comparison), the training status differs (adult-trained and unadapted versus trained on this cohort under subject-disjoint folds), and their comparison was not modality matched (SynthSeg used T1-weighted input alone whereas SamSeg used both contrasts, a confound they note), whereas all pipelines in the present study used the same paired T1- and T2-weighted acquisitions. Absolute values are therefore not compared between the studies.

S5 Supplementary tables and figures

Table S2: Boundary-error localisation under two aggregation schemes.

Pipeline	Equal-weight structures		Pooled surfaces	
	≤ 1 mm	≤ 2 mm	≤ 1 mm	≤ 2 mm
3D TransUNet	84.1%	93.9%	88.7%	94.6%
Swin UNETR	82.9%	94.0%	90.3%	97.1%
nnU-Net	82.4%	92.4%	74.9%	80.6%
STSNNet (2.5D)	72.2%	90.3%	63.9%	87.4%

Equal-weight is the mean over case-structures of the per-item fraction; pooled is summed boundary errors divided by summed errors. The schemes disagree on ordering at 1 mm. Excluding the two influential scans raises the nnU-Net pooled 1 mm fraction from 74.9% to 82.8% (post-hoc sensitivity analysis). Percentages should therefore be interpreted with reference to the aggregation scheme.

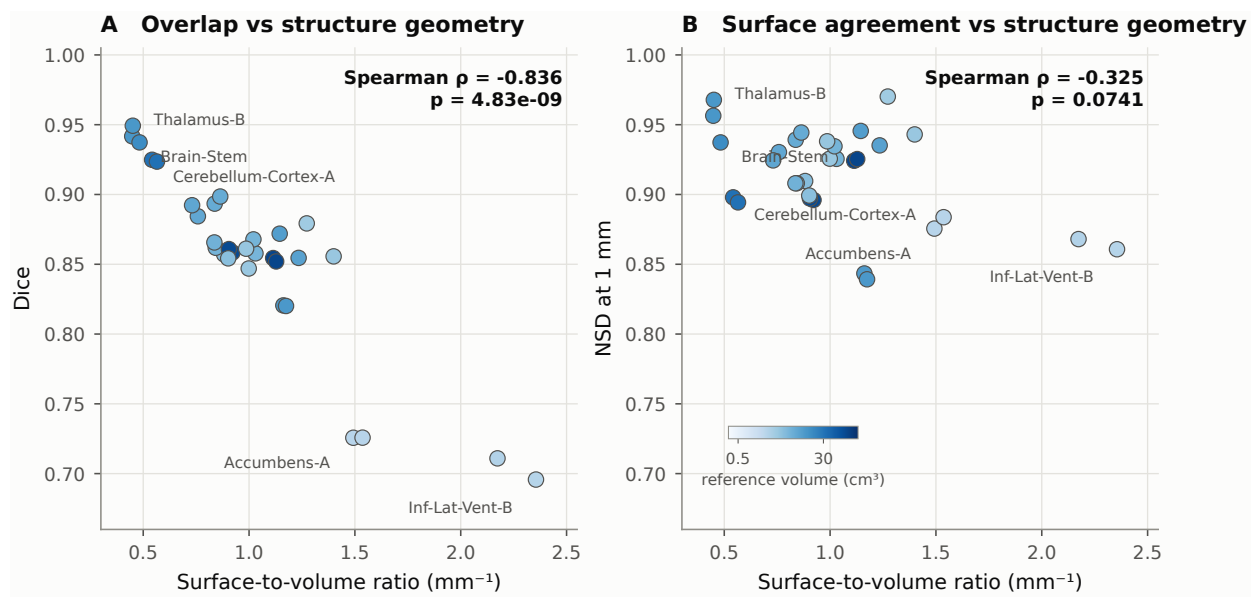


Figure S1: Structure geometry and metric behaviour. (A) Dice and (B) normalised surface Dice at 1 mm against surface-to-volume ratio for the 31 evaluated structures, pooled across pipelines; point shade encodes reference volume. The panels share a vertical range so the difference in slope is directly comparable. Dice is strongly associated with surface-to-volume ratio (Spearman $\rho = -0.836$) whereas NSD is much more weakly associated ($\rho = -0.325$) and unassociated with volume ($\rho = +0.046$, $p = 0.806$). Extremes are labelled. These are descriptive structure-level associations; no causal claim is made, and the data do not establish that NSD is independent of structure geometry.

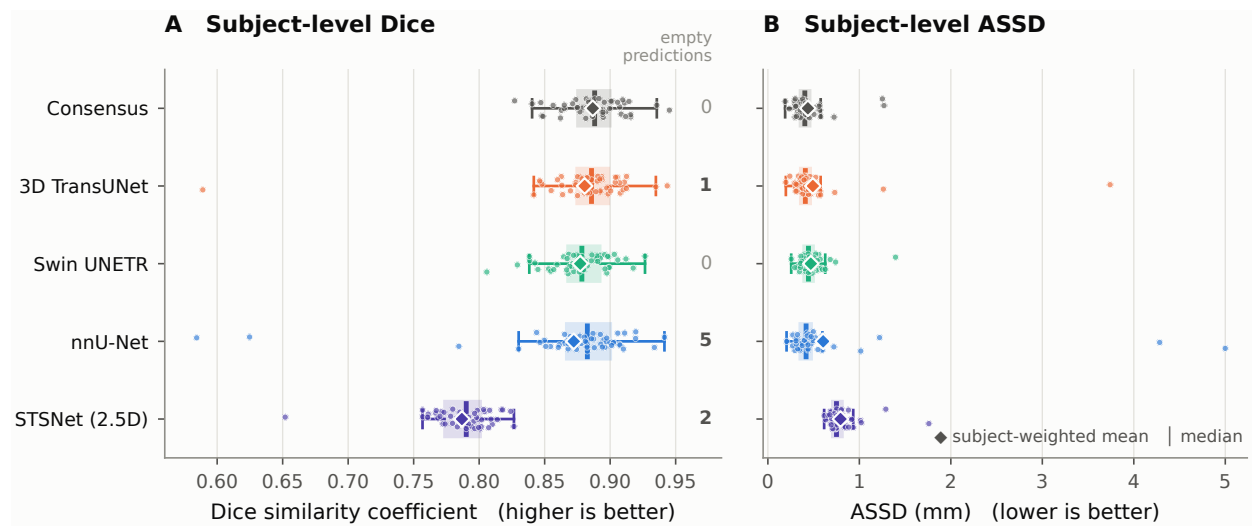


Figure S2: Subject-level distributions by pipeline. (A) Dice and (B) ASSD summarised at the level of the 51 infants. Boxes show the interquartile range and median, whiskers the $1.5 \times \text{IQR}$ range, points individual infants, and the diamond the subject-weighted mean of Table 2 in the main text. The right-hand column of panel A gives empty case-structure predictions per pipeline. The nnU-Net low-Dice and high-ASSD tail appears as isolated points rather than a shift of the whole distribution.

S6 Statistical procedures

The two procedures below formalise the statistical analysis described in the main text.

Algorithm S1 Subject-weighted paired pipeline comparison with multiplicity correction

Require: Per-case–structure metric values for pipelines A, B ; infant identifiers ($n = 51$)

Ensure: Mean/median paired difference, win count, Holm-adjusted Wilcoxon p , standardised effect size

- 1: Average each metric within infant i across all case–structure observations $\rightarrow x_i^A, x_i^B$
 - 2: $d_i \leftarrow x_i^A - x_i^B$ for each infant $i = 1, \dots, 51$
 - 3: Compute $\bar{d} \leftarrow \text{mean}(d)$ with its 95% CI from the t_{50} distribution, and $\text{median}(d)$
 - 4: $w \leftarrow |\{i : d_i \text{ favours } A\}|$ ▷ infant-level win count
 - 5: $(T, p) \leftarrow \text{WILCOXONSIGNEDRANK}(d)$
 - 6: $d_{\text{Cohen}} \leftarrow \bar{d}/s_d$ ▷ standardised effect size, $s_d = \text{SD}(d)$
 - 7: Repeat for every pipeline pair and metric; collect all raw p -values into P
 - 8: $P_{\text{Holm}} \leftarrow \text{HOLMCORRECTION}(P)$ ▷ applied within each metric, across pipeline pairs
 - 9: **return** $\bar{d}$, CI, $\text{median}(d)$, w , P_{Holm} , d_{Cohen} for every pair
-

Algorithm S2 Pipeline-effect variance decomposition, retraining/fold characterisation, and influential-scan flagging

Step 1: Pipeline-effect variance decomposition

- 1: Fit metric $\sim \log_{10}(\text{volume}) + \text{surface/volume} + \text{age} + C(\text{pipeline})$ over all case–structure observations
- 2: Compute the type-II sum-of-squares share attributable to the pipeline term
- 3: Test the pipeline term under two-way cluster-robust inference (clusters: infant, structure)

Step 2: Retraining and fold-level variance characterisation

- 4: For each configuration with ≥ 2 repeated runs on an identical fold, compute the SD of the subject-weighted fold-level mean metric across runs
- 5: For each held-out fold $f = 1, \dots, 5$, compute the subject-weighted paired pipeline difference restricted to fold f
- 6: Report the SD and sign-change count of the five fold-specific differences

Step 3: Influential-scan flagging

- 7: For every scan, compute its case-level paired difference for each pipeline pair
 - 8: Rank scans by |difference| and flag the largest as influential
 - 9: Cross-reference flagged scans against the empty-prediction list
 - 10: Recompute the boundary-fraction summary excluding flagged scans as a sensitivity check
-

References cited in Online Resource 1

- [1] Liu J, Liu F, Sun K, Sun Y, Huang J, Jiang C, Rekik I, Shen D (2024) UinTSeg: unified infant brain tissue segmentation with anatomy delineation. In: Medical Image Computing and Computer Assisted Intervention, MICCAI 2024. Lecture Notes in Computer Science. Springer, Cham, pp 487–497. https://doi.org/10.1007/978-3-031-72069-7_46
- [2] He Q, Dong M, Summerfield N, Glide-Hurst C (2023) MAGNET: a modality-agnostic network for 3D medical image segmentation. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI), pp 1–5. <https://doi.org/10.1109/ISBI53787.2023.10230587>
- [3] Prakash PS, Rao PK, Pasha MJ, Algarni A, Ayadi M, Cho Y, Nam Y (2025) CausalX-Net: a causality-guided explainable segmentation network for brain tumors. Front Med 12:1693603. <https://doi.org/10.3389/fmed.2025.1693603>
- [4] Nie Z, Yang J, Li C, Wang Y, Tang J (2025) DiffBTS: A Lightweight Diffusion Model for 3D Multimodal Brain Tumor Segmentation. Sensors 25:2985. <https://doi.org/10.3390/s25102985>

- [5] Wang L, Nie D, Li G, Puybureau E, Dolz J, Zhang Q et al (2019) Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Trans Med Imaging* 38:2219–2230. <https://doi.org/10.1109/TMI.2019.2901712>
- [6] Kamnitsas K, Ledig C, Newcombe VFJ, Simpson JP, Kane AD, Menon DK, Rueckert D, Glocker B (2017) Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation. *Med Image Anal* 36:61–78. <https://doi.org/10.1016/j.media.2016.10.004>
- [7] Maier-Hein L, Eisenmann M, Reinke A, Onogur S, Stankovic M, Scholz P et al (2018) Why rankings of biomedical image analysis competitions should be interpreted with care. *Nat Commun* 9:5217. <https://doi.org/10.1038/s41467-018-07619-7>
- [8] Maier-Hein L, Reinke A, Godau P, Tizabi MD, Buettner F, Christodoulou E et al (2024) Metrics reloaded: recommendations for image analysis validation. *Nat Methods* 21:195–212. <https://doi.org/10.1038/s41592-023-02151-z>
- [9] Reinke A, Tizabi MD, Baumgartner M, Eisenmann M, Heckmann-Nötzel D, Kavur AE et al (2024) Understanding metric-related pitfalls in image analysis validation. *Nat Methods* 21:182–194. <https://doi.org/10.1038/s41592-023-02150-0>
- [10] Van Leemput K, Bakkour A, Benner T, Wiggins G, Wald LL, Augustinack J, Dickerson BC, Golland P, Fischl B (2009) Automated segmentation of hippocampal subfields from ultra-high resolution in vivo MRI. *Hippocampus* 19:549–557. <https://doi.org/10.1002/hipo.20615>
- [11] Billot B, Bocchetta M, Todd E, Dalca AV, Rohrer JD, Iglesias JE (2020) Automated segmentation of the hypothalamus and associated subunits in brain MRI. *NeuroImage* 223:117287. <https://doi.org/10.1016/j.neuroimage.2020.117287>
- [12] Poirot MG, Caan MWA, Ruhe HG, Bjørnerud A, Groote I, Reneman L, Marquering HA (2022) Robustness of radiomics to variations in segmentation methods in multimodal brain MRI. *Sci Rep* 12:16712. <https://doi.org/10.1038/s41598-022-20703-9>
- [13] Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH (2021) nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods* 18:203–211. <https://doi.org/10.1038/s41592-020-01008-z>
- [14] Coupé P, Mansencal B, Clément M, Giraud R, Denis de Senneville B, Ta VT, Lepetit V, Manjon JV (2020) AssemblyNet: A large ensemble of CNNs for 3D whole brain MRI segmentation. *NeuroImage* 219:117026. <https://doi.org/10.1016/j.neuroimage.2020.117026>
- [15] De Rosa AP, Benedetto M, Tagliaferri S, Bardozzo F, D’Ambrosio A, Bisecco A, Gallo A, Cirillo M, Tagliaferri R, Esposito F (2024) Consensus of algorithms for lesion segmentation in brain MRI studies of multiple sclerosis. *Sci Rep* 14:21348. <https://doi.org/10.1038/s41598-024-72649-9>
- [16] Billot B, Greve DN, Puonti O, Thielscher A, Van Leemput K, Fischl B, Dalca AV, Iglesias JE (2023) SynthSeg: segmentation of brain MRI scans of any contrast and resolution without retraining. *Med Image Anal* 86:102789. <https://doi.org/10.1016/j.media.2023.102789>
- [17] Puonti O, Iglesias JE, Van Leemput K (2016) Fast and sequence-adaptive whole-brain segmentation using parametric Bayesian modeling. *NeuroImage* 143:235–249. <https://doi.org/10.1016/j.neuroimage.2016.09.011>
- [18] Henschel L, Conjeti S, Estrada S, Diers K, Fischl B, Reuter M (2020) FastSurfer: a fast and accurate deep learning based neuroimaging pipeline. *NeuroImage* 219:117012. <https://doi.org/10.1016/j.neuroimage.2020.117012>
- [19] Alexander N, Gucciardi A, Michelucci U (2026) Quantifying segmentation-driven measurement error in infant brain MRI and its impact on volume and fractal dimension reliability. *Eur J Radiol Artif Intell* 6:100087. <https://doi.org/10.1016/j.ejrai.2026.100087>