Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Tasnia Binte Mamun
Bangladesh University of Engineering and Technology
Samiul Based Shuvo
sbshuvo@bme.buet.ac.bd

Bangladesh University of Engineering and Technology

Research Article

Keywords:

Posted Date: October 30th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11019607/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

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

Abstract

Background Deep learning segmentation increasingly supplies the brain volumes used in infant neuroimaging, yet pipelines are usually ranked by a single mean Dice score, which hides variability and rare failures.

Objective To quantify how much of the apparent difference between infant brain MRI segmentation pipelines reflects retraining variability, cross-validation partitioning, aggregation choice and rare failures, and to derive a reporting template for validating such tools.

Materials and methods Four pipelines (nnU-Net, 3D TransUNet, Swin UNETR, 2.5D STSNet) and a cross-fitted consensus were trained and evaluated on the public Baby Open Brains dataset (71 T1- and T2-weighted MRI visits from 51 infants aged 1.0 to 9.4 months, 31 expert-corrected labels) with subject-disjoint five-fold cross-validation, subject-weighted summaries and Holm-corrected paired tests.

Results Mean Dice was 0.8867 (95% confidence interval 0.8802–0.8932) for the consensus, 0.8805 for TransUNet, 0.8772 for Swin UNETR, 0.8722 for nnU-Net and 0.7868 for STSNet. Among the 3D pipelines the pipeline term explained 0.134% of Dice variation ($p = 0.177$), all paired effects were small ($|d| \leq 0.22$) and every mean-difference confidence interval included zero. TransUNet exceeded Swin UNETR in 46 of 51 infants; for nnU-Net versus Swin UNETR the mean (−0.0050) and median (+0.0072) disagreed in sign. Fold-level differences (−0.0490 to +0.0169) exceeded retraining standard deviations (0.0005–0.0016). Two scans drove the reversal and most failures; the consensus produced none.

Conclusion Three competently configured 3D pipelines segmented the infant brain accurately with small mutual differences and no detected age dependence. Choosing among them mattered less than reporting variability and failures; a simple consensus removed all failures.

Keywords: Infant brain MRI · Deep learning · Image segmentation · Artificial intelligence · Reproducibility · Neurodevelopment

1 Introduction

Deep learning segmentation is one of the main ways in which artificial intelligence (AI) enters paediatric neuroimaging: it converts an infant's MRI into labelled brain structures from which regional volumes, cortical morphometry and growth trajectories are derived [1, 2, 3]. The first postnatal year is the period of most rapid brain growth and the period in which structural MRI is hardest to segment. Myelination raises grey/white matter contrast on T1-weighted images while lowering it on T2-weighted images, so that between approximately the fourth and eighth postnatal months an isointense period arises in which the grey/white boundary is poorly expressed on either contrast, and methods developed for older brains degrade [1, 4]. Because morphometry is computed downstream of a segmentation, segmentation reliability propagates directly into developmental inference [2, 5, 3]. For the paediatric radiologist the practical question is therefore not only whether AI can segment the infant brain, but how far its output can be trusted and how that trust should be evidenced.

Infant-specific resources exist. The iSeg challenges established a six-month benchmark and showed how strongly rankings depend on evaluation choices [6]; Infant FreeSurfer [7] and BIBSNet [8] were built for early postnatal data; the Baby Open Brains (BOBs) resource [9], drawn from the Baby Connectome Project [10], released 71 paired T1- and T2-weighted visits from 51 infants with expert-reviewed segmentations that sample

the isointense window densely. The pipelines compared here span the main design families in volumetric segmentation: the self-configuring convolutional nnU-Net [11, 12], the convolution–transformer hybrid 3D TransUNet [13], the shifted-window transformer Swin UNETR [14, 15], and STSNet, a small-target network adapted here to a 2.5D slice-context setting [16]. We treat them as complete pipelines rather than isolated architectures: they differ in input representation, patch configuration, optimisation, augmentation and implementation, so a common dataset and split equalise the data, not the training procedure, and the comparisons describe end-to-end pipelines without isolating individual architectural components.

Ordinary leaderboard reporting is a poor instrument for deciding whether such pipelines differ in a way that matters. Rankings in biomedical image-analysis competitions are unstable under changes of metric, aggregation rule and test-case composition [17, 18]. Overlap metrics penalise small and geometrically complex structures more heavily than large compact ones [19, 20]. Surface metrics are easy to implement incorrectly and are undefined when a prediction is empty, which silently changes denominators unless handled explicitly [21, 22]. Repeated training of an identical configuration produces run-to-run variation [23, 24]; cross-validation folds combine different held-out subjects with independently fitted models, so pairwise ordering can reverse between folds; and a few catastrophic per-scan failures can dominate a cohort mean while leaving the typical case untouched [25]. Infant segmentation studies have mostly emphasised algorithm development or headline accuracy. A recent BOBs-based study quantified how segmentation error from adult-trained pipelines applied off the shelf propagates into volume and fractal-dimension estimates and called for evaluation of infant-trained models [3]; the comparability of the segmentation benchmark itself, with pipelines trained on infant data, has not been examined.

We therefore asked, under one subject-disjoint protocol: how large are the observed differences among pipelines, and are they larger than retraining variation (RQ1, RQ1.1) and stable across cross-validation partitions (RQ1.2); whether the interpretation is consistent across mean and rank-based summaries (RQ2.1) and how strongly rare failures influence comparisons and boundary-error estimates (RQ2.2); how structure size and surface-to-volume ratio affect overlap and surface metrics (RQ3.1); and whether a simple cross-fitted consensus improves robustness over its constituents (RQ4.1). Two exploratory questions addressed age and image-contrast effects (EQ1) and longitudinal behaviour in the repeated-visit subset (EQ2). The rationale and literature behind each question are given in Online Resource 1 (Section S1). We distil the results into a reporting template (Table 6) that lets a reader judge how far an AI segmentation tool can be trusted.

2 Materials and methods

2.1 Study design, cohort and cross-validation

This was a retrospective computational evaluation of publicly available, deidentified infant MRI and expert-corrected reference segmentations (Table 1; Fig. 1). Three design choices govern everything that follows. First, the primary estimand was fixed before any metric was computed: each metric was averaged over all case–structure observations belonging to an infant, and the 51 infant-level values were weighted equally, so that the 17 infants with repeat visits cannot dominate any summary; all headline values, paired tests and fold summaries are subject weighted. Second, all visits from an infant were assigned to the same fold using one shared split definition, so every prediction entering the main comparison was out of fold; the evaluation case list was verified against the split. Third, folds were balanced on nominal session labels because chronological age metadata were retrieved later; nominal label and true age were highly rank-concordant (Spearman $\rho = 0.984$), sensitivity analyses gave unchanged headline results under either definition, and the original split was retained (details in Online Resource 1, Section S2).

Chronological age was taken from the participant and session metadata. Age at scan ranged from 1.0 to 9.4 months (mean 5.34, SD 2.31 across 71 visits; mean 5.33, SD 2.02 across 51 infants). Thirty-one infants were female and 20 male. Gestational age at birth was 39.85 weeks (SD 1.11, range 37.1–42.0), with no birth below 37 weeks. All 71 metadata records matched the evaluated case list.

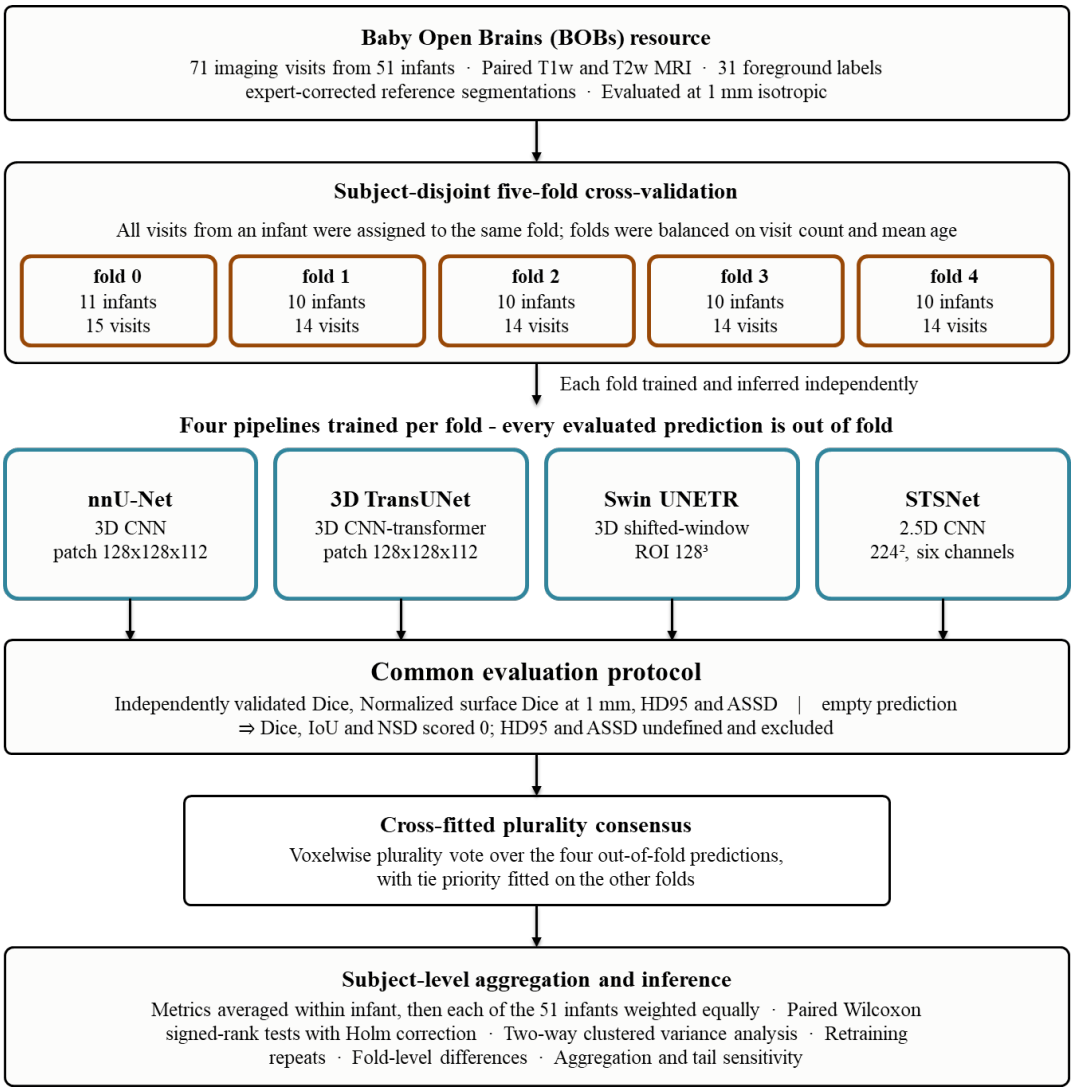

Figure 1: Study and evaluation flow. Selection of the Baby Open Brains resource, the 51-infant / 71-visit hierarchy, subject-disjoint assignment into five folds with exact infant and visit counts, the four pipelines trained and inferred independently per fold, construction of the cross-fitted plurality consensus, the common evaluation protocol including the empty-prediction policy, and subject-level aggregation.

Table 1: Cohort, imaging and evaluation characteristics.

Characteristic	Value
Source resource	Baby Open Brains (BOBs), OpenNeuro ds005450 [9]
Site	University of Minnesota site, Baby Connectome Project [10]
Infants / imaging visits	51 / 71
Infants with repeat visits	17 (37 visits)
Age at scan, per visit	5.34 months (SD 2.31), range 1.0–9.4
Sex, per infant	31 female / 20 male (60.8% female)
Gestational age at birth	39.85 weeks (SD 1.11), range 37.1–42.0; no births <37 weeks
Scanner / coil	Siemens 3 T Prisma / 32-channel
T1-weighted sequence	MPRAGE; TR 2400 ms, TE 2.24 ms, TI 1600 ms, FA 8°, 0.8 mm isotropic
T2-weighted sequence	Turbo spin-echo; TR 3200 ms, TE 564 ms, turbo factor 314, 0.8 mm isotropic
Evaluation space	1 mm isotropic
Foreground labels evaluated	31
Case–structure units per pipeline	2,201 (71 visits × 31 structures)
Cross-validation	Subject-disjoint, five folds
Validation fold sizes (infants)	11, 10, 10, 10, 10
Validation fold sizes (visits)	15, 14, 14, 14, 14

2.2 Reference standard

BOBs segmentations were initialised automatically, manually corrected by trained annotators consulting both contrasts, and reviewed and revised by expert reviewers [9]. The labels were not re-annotated for this study and no inter-rater reliability estimate was available, so no annotation ceiling can be quoted against which pipeline agreement could be judged. Our local label verification was a sanity check rather than an anatomical re-review: volume and left–right symmetry checks on 20 sampled cases and 40 paired hemispheric comparisons across 4 cases and 5 labels per hemisphere.

2.3 Segmentation pipelines, training and inference

All pipelines received paired T1- and T2-weighted input. nnU-Net and 3D TransUNet used two input channels with a $128 \times 128 \times 112$ patch; Swin UNETR used a $128 \times 128 \times 128$ region of interest with RAS orientation, 1 mm spacing, foreground cropping and channel-wise nonzero normalisation; STSNet used a 2.5D six-channel representation formed from the T1- and T2-weighted slices at $z - 1$, z and $z + 1$, padded to 224×224. Training budgets were matched to approximately 62,500–63,700 optimiser updates per fold. Configuration, augmentation and inference settings are given in Table 2; preprocessing scripts were not identical across pipelines and are reported separately there. The 3D pipelines used sliding-window inference with pipeline-specific overlap; STSNet used slice-wise inference.

Table 2: Pipeline configuration, augmentation and inference settings.

	nnU-Net	3D TransUNet	Swin UNETR	STSNet
Representation	3D CNN	3D CNN–transformer	3D shifted-window transformer	2.5D CNN
Input channels	2	2	2	6 (adjacent slices)
Output classes	32	32	32	32
Patch / ROI	$128 \times 128 \times 112$	$128 \times 128 \times 112$	128^3	224×224
Batch size	2	1 (as executed)	1	16
Loss	Dice + CE	Dice + CE	Dice + CE	Cross-entropy
Optimiser	SGD ($m = 0.99$, Nesterov)	AdamW	AdamW	AdamW
Learning rate	0.01	3×10^{-4}	1×10^{-4}	1×10^{-4}
Weight decay	3×10^{-5}	3×10^{-5}	1×10^{-5}	1×10^{-5}
Schedule	Polynomial	Warm-up + cosine	Cosine	Cosine
Epochs × iters	250×250	250×250	1096×57	100×637
Total updates	62,500	62,500	62,472	63,700
Augmentation	Rotation, scaling, additive noise, blur, brightness, contrast, simulated low resolution, gamma, mirroring	Pipeline default (not independently verified)	Random spatial crops, axis flips, intensity scaling and shifting	Horizontal and vertical flips, shift–scale–rotation
Seed	none set	12345 (hard-coded)	2026, deterministic	none set
Inference	Sliding window, overlap 0.5	Sliding window, overlap 0.3	Sliding window, overlap 0.25	Slice-wise

CNN, convolutional neural network; CE, cross-entropy; ROI, region of interest. Preprocessing differed by pipeline: nnU-Net and 3D TransUNet used two-channel $128 \times 128 \times 112$ patches; Swin UNETR used RAS orientation, 1 mm spacing, foreground cropping and channel-wise nonzero normalisation; STSNet used a six-channel 2.5D representation from adjacent slices padded to 224×224. Training ran under shared GPU conditions, so wall-clock times are not reported as an efficiency benchmark.

2.4 Cross-fitted consensus

To assess whether complementary errors could be reduced by aggregation, a consensus segmentation was formed from the four out-of-fold predictions by voxelwise plurality voting, with ties broken using only the remaining folds to avoid information leakage. Because it combines existing predictions without additional training, it was evaluated as a robustness baseline rather than as a segmentation method.

2.5 Evaluation metrics

Performance was measured with the Dice similarity coefficient (Dice), normalised surface Dice at a 1 mm tolerance (NSD), the 95th-percentile Hausdorff distance (HD95) and the average symmetric surface distance (ASSD), computed on the physical voxel grid resampled to 1 mm isotropic spacing. Empty predictions were handled by a predefined policy rather than exclusion: when the reference structure was non-empty but the prediction contained no voxels, Dice, intersection-over-union and NSD were assigned zero, whereas HD95 and ASSD, which require a predicted surface, were treated as undefined and excluded from their summaries. Valid denominators therefore differ across metrics and pipelines; empty-prediction counts and valid sample sizes are reported alongside the results (Table 3). The metric implementation was validated against MONAI and an independent SciPy cKDTree implementation on 24 non-empty case–structure combinations (maximum absolute deviation below 1×10^{-6}); this does not cover empty-mask or degenerate-geometry cases.

Table 3: Main subject-weighted performance, mean (95% confidence interval) over $n = 51$ infants.

Pipeline	Dice ↑	NSD@1 mm ↑	HD95 (mm) ↓	ASSD (mm) ↓	Empty	Valid
Cross-fitted consensus	0.8867 (0.8802–0.8932)	0.9467 (0.9402–0.9531)	1.8521 (1.5069–2.1973)	0.4374 (0.3831–0.4917)	0	2201
3D TransUNet	0.8805 (0.8674–0.8936)	0.9416 (0.9283–0.9549)	2.0310 (1.3820–2.6801)	0.4916 (0.3544–0.6289)	1	2200
Swin UNETR	0.8772 (0.8707–0.8836)	0.9388 (0.9319–0.9457)	1.9046 (1.6346–2.1746)	0.4690 (0.4232–0.5149)	0	2201
nnU-Net	0.8722 (0.8551–0.8893)	0.9321 (0.9142–0.9500)	2.6134 (1.5116–3.7153)	0.6044 (0.3671–0.8418)	5	2196
STSNet (2.5D)	0.7868 (0.7795–0.7942)	0.8414 (0.8330–0.8497)	2.7946 (2.5590–3.0301)	0.7939 (0.7437–0.8442)	2	2199

Each metric was first averaged within infant and then across the 51 infants with equal weighting. "Empty" denotes case–structure predictions containing no voxels. For empty predictions, Dice, IoU, and NSD were assigned zero, whereas HD95 and ASSD were undefined and excluded from their respective summaries. "Valid" denotes the number of non-empty case–structure observations contributing to the HD95 and ASSD calculations, out of 2,201 total case–structure observations.

2.6 Statistical analysis

Metrics were averaged within infant and then across the 51 infants with equal weighting; 95% confidence intervals (CI) used the t distribution with 50 degrees of freedom. Pairwise pipeline comparisons used within-infant paired differences ($n = 51$): mean difference with CI, median, infant-level win count, Wilcoxon signed-rank test with Holm correction across pipeline pairs within each metric, and Cohen's d. Pipeline contribution was assessed with the linear model metric $\sim \log_{10}(\text{volume}) + \text{surface/volume} + \text{age} + C(\text{pipeline})$ fitted over case–structure observations, reporting the type-II sum-of-squares share of the pipeline term under two-way cluster-robust inference (clusters: infant and structure). Retraining variation was the SD of the subject-weighted fold-level mean across repeated runs of an identical configuration on an identical fold; cross-validation partition variation was the set of fold-specific paired differences. Influential scans were flagged by ranking case-level paired differences and cross-referenced against the empty-prediction list, with a sensitivity re-analysis excluding them. Secondary analyses comprised boundary-error fractions at 1 and 2 mm, structure-level correlations, age and contrast regressions with subject clustering, and a descriptive longitudinal analysis. All tests were two-sided with $\alpha = 0.05$. The procedures are formalised as Algorithms S1 and S2 in Online Resource 1. Evaluation code, the split definition, configuration files and the scripts that regenerate every table are available from the corresponding author on reasonable request; all reported numbers were regenerated from a single declared set of hashed input files.

3 Results

3.1 Cohort and evaluation completeness

Fifty-one infants contributing 71 visits entered the five-fold subject-disjoint evaluation; 17 contributed repeat visits. Every prediction in the main comparison was out of fold, and 31 foreground structures were evaluated per visit, giving 2,201 case–structure units per pipeline. Across the four pipelines, 8 empty case–structure predictions were recorded; the consensus produced none.

3.2 Pipeline performance and effect (RQ1)

The consensus achieved the highest mean Dice (0.8867, 95% CI 0.8802–0.8932), followed by 3D TransUNet (0.8805, 0.8674–0.8936), Swin UNETR (0.8772, 0.8707–0.8836) and nnU-Net (0.8722, 0.8551–0.8893); STSNet was clearly lower (0.7868, 0.7795–0.7942) (Table 3; subject-level distributions in Online Resource 1, Fig. S2). The intervals of the three 3D pipelines overlap substantially; that of STSNet overlaps none of them. Boundary metrics were not ordered like Dice: Swin UNETR had a lower mean HD95 (1.9046 mm) than 3D TransUNet (2.0310 mm) despite lower Dice, and nnU-Net had a markedly higher HD95 (2.6134 mm) and ASSD (0.6044 mm) with wide intervals. Empty predictions numbered 5 for nnU-Net, 2 for STSNet, 1 for 3D TransUNet and none for Swin UNETR or the consensus, so the HD95 and ASSD summaries of the affected pipelines rest on reduced denominators.

Among nnU-Net, 3D TransUNet and Swin UNETR, the pipeline term accounted for 0.134% of the modelled Dice sum of squares and was not statistically detected under two-way clustered inference ($p = 0.177$); the NSD share was 0.171% ($p = 0.176$). Adding STSNet raised the share to 14.645% for

Dice ($p = 3.62 \times 10^{-7}$) and 17.293% for NSD ($p = 1.15 \times 10^{-6}$), showing that the design could detect a difference of roughly 0.09 Dice, without identifying which of the many differences between STSNet and the 3D pipelines produced it. Non-detection among the 3D pipelines is not an equivalence result: no smallest meaningful difference or equivalence margin was prespecified and no equivalence test was performed, so the data show that a mean pipeline effect was not detected under this design, not that the pipelines perform equivalently.

3.3 Retraining variation and cross-validation partition variation (RQ1.1, RQ1.2)

Retraining variation was small where it could be measured (Table 4; Fig. 2A): the SD of the subject-weighted fold-level mean Dice was 0.00053 across five nnU-Net fold-0 runs (95% CI for the SD 0.00032–0.00153), 0.00156 across two nnU-Net fold-1 runs and 0.00099 across two Swin UNETR fold-0 runs; the two-run estimates cannot establish pipeline- or fold-invariant retraining variance. Fold-to-fold variation was larger (Fig. 2B): the subject-weighted nnU-Net minus Swin UNETR difference across the five held-out folds was +0.0093, +0.0022, −0.0490, +0.0169 and −0.0058 (SD 0.0258, three sign changes), so the direction of that comparison depended on the fold, whereas within fold 0 all pairwise comparisons between repeated runs remained positive (+0.0072 to +0.0100).

Table 4: Retraining variation and cross-validation partition variation.

(a) Repeated training of an identical configuration on an identical fold

Configuration	Runs	SD of fold mean Dice	Range	95% CI for SD
nnU-Net, fold 0	5	0.00053	0.00146	0.00032–0.00153
nnU-Net, fold 1	2	0.00156	0.00220	not estimable ($n = 2$)
Swin UNETR, fold 0	2	0.00099	0.00141	not estimable ($n = 2$)

(b) Subject-weighted nnU-Net − Swin UNETR Dice difference by held-out fold

Fold 0	Fold 1	Fold 2	Fold 3	Fold 4	SD across folds	Sign changes
+0.0093	+0.0022	−0.0490	+0.0169	−0.0058	0.0258	3

3D TransUNet was not independently repeated because its trainer hard-codes the seed. The nnU-Net configuration used no manually fixed seed; repeated runs therefore varied only through stochastic training while the fold, configuration, planner, and training budget were held fixed. The two panels are not estimates of directly comparable variance components: panel (A) reflects repeated training, whereas panel (B) also reflects changes in held-out and training-set composition and independently fitted fold models.

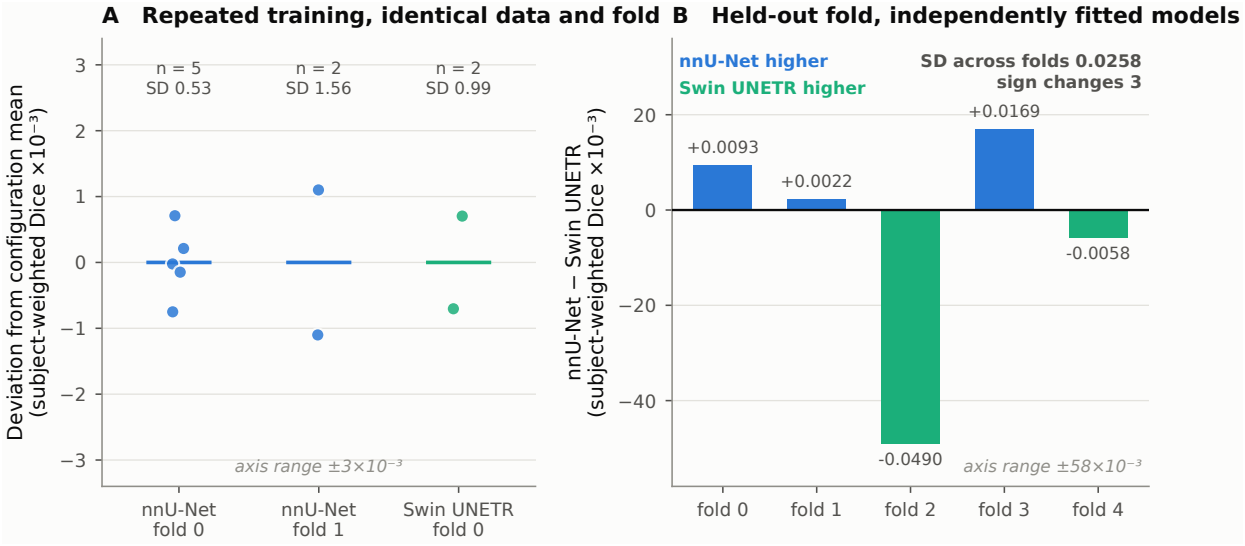

Figure 2: Retraining variation versus cross-validation partition variation. (**A**) Subject-weighted fold-level mean Dice for repeated training of an identical configuration on an identical fold, as deviation from each configuration's own mean, with run count and standard deviation above each. (**B**) The nnU-Net minus Swin UNETR difference for each held-out fold, where each fold uses independently fitted models. **The panels do not share an axis range** ($\pm 3 \times 10^{-3}$ versus $\pm 58 \times 10^{-3}$, as annotated) and the two quantities are not estimated from designs of comparable completeness: panel A rests on five, two and two runs with no 3D TransUNet repeat, while panel B confounds held-out composition, training composition and the fitted models. No ratio between panels is reported.

3.4 Paired ordering and aggregation sensitivity (RQ2.1)

For nnU-Net versus Swin UNETR the mean Dice difference was −0.0050 (95% CI −0.0206 to +0.0106) while the median was +0.0072, and nnU-Net was higher in 42 of 51 infants (Holm-adjusted $p = 0.0045$) (Table 5; Fig. 3). For 3D TransUNet versus Swin UNETR the mean was +0.0033 (−0.0074 to +0.0141), the median +0.0092, and TransUNet was higher in 46 of 51 infants ($p < 0.0001$). For nnU-Net versus 3D TransUNet the mean was −0.0083 (−0.0213 to +0.0047; $p = 0.2769$). The same mean–median sign disagreement appeared for NSD, HD95 and ASSD in the nnU-Net versus Swin UNETR comparison. Every paired mean-difference CI among the three 3D pipelines includes zero, including the TransUNet–Swin UNETR comparison whose rank test is significant at $p < 0.0001$; the CI concerns the mean paired difference and the Wilcoxon test the signed-rank distribution, so these are different estimands rather than a contradiction, and together they indicate a consistent directional ordering of small and imprecisely estimated magnitude. Every paired standardised effect among the 3D pipelines was small ($|d| \leq 0.22$) for all four metrics.

Table 5: Subject-level paired comparisons among the three 3D pipelines (n = 51 infants).

Metric	Comparison (A − B)	Mean diff. (95% CI)	Median	Favouring A	p_{Holm}	d
Dice	nnU-Net − 3D TransUNet	-0.0083 (-0.0213 to +0.0047)	-0.0001	22/51	0.2769	-0.180
	nnU-Net − Swin UNETR‡	-0.0050 (-0.0206 to +0.0106)	+0.0072	42/51	0.0045*	-0.090
	3D TransUNet − Swin UNETR	+0.0033 (-0.0074 to +0.0141)	+0.0092	46/51	<0.0001*	+0.087
NSD@1 mm	nnU-Net − 3D TransUNet	-0.0095 (-0.0233 to +0.0043)	-0.0011	17/51	0.0174*	-0.193
	nnU-Net − Swin UNETR‡	-0.0067 (-0.0231 to +0.0097)	+0.0072	40/51	0.0174*	-0.115
	3D TransUNet − Swin UNETR	+0.0028 (-0.0081 to +0.0137)	+0.0086	46/51	<0.0001*	+0.072
HD95 (mm)	nnU-Net − 3D TransUNet	+0.5824 (-0.1793 to +1.3441)	+0.0123	19/51	0.0191*	+0.215
	nnU-Net − Swin UNETR‡	+0.7088 (-0.3880 to +1.8057)	-0.0713	40/51	0.0135*	+0.182
	3D TransUNet − Swin UNETR‡	+0.1264 (-0.4868 to +0.7397)	-0.0861	45/51	<0.0001*	+0.058
ASSD (mm)	nnU-Net − 3D TransUNet	+0.1128 (-0.0496 to +0.2752)	+0.0025	21/51	0.2302	+0.195
	nnU-Net − Swin UNETR‡	+0.1354 (-0.0974 to +0.3682)	-0.0329	43/51	0.0006*	+0.164
	3D TransUNet − Swin UNETR‡	+0.0226 (-0.1049 to +0.1501)	-0.0371	47/51	<0.0001*	+0.050

Paired Wilcoxon signed-rank tests on within-infant differences, with Holm correction across pipeline pairs within each metric. ∗ Holm-adjusted $p < 0.05$; ‡ mean and median differences have opposite signs. For HD95 and ASSD, lower is better, so negative differences favour pipeline A; win counts follow each metric's direction of improvement.

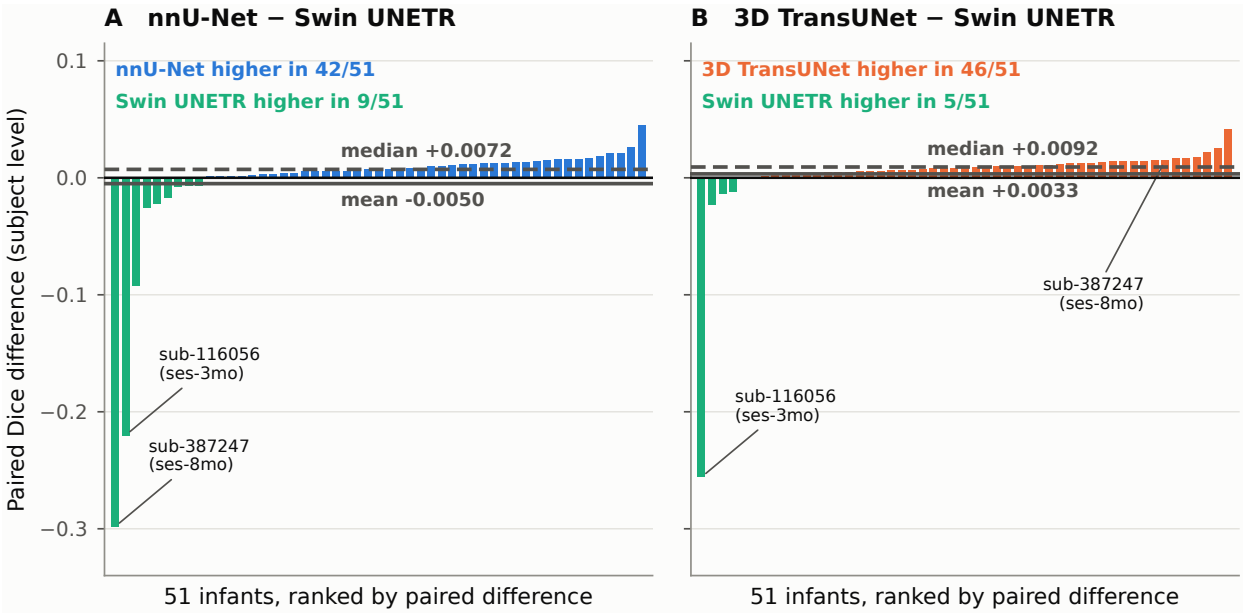

Figure 3: Paired subject-level differences and the mean–median reversal. Within-infant paired Dice differences for (**A**) nnU-Net minus Swin UNETR and (**B**) 3D TransUNet minus Swin UNETR, with the 51 infants ranked by difference and bars coloured by which pipeline was higher. The solid line marks the mean, the dashed line the median. In panel A the two summaries fall on opposite sides of zero: nnU-Net was higher in 42 of 51 infants with a median of +0.0072, yet two large negative cases pull the mean to −0.0050. The two influential scans are labelled.

3.5 Rare failures, influential scans and boundary-error localisation (RQ2.2)

Two scans exerted disproportionate influence (Fig. 4). In `sub-387247_ses-8mo` the nnU-Net minus Swin UNETR case-level Dice difference was −0.2982, and all five nnU-Net empty predictions occurred in this scan (Accumbens-A, Amygdala-A, Caudate-A, Pallidum-A, Putamen-A). In `sub-116056_ses-3mo` the difference was −0.2201, and the only 3D TransUNet empty prediction occurred there. Five empty structures within one scan is more consistent with a case-level failure mode than with independent structure-specific weaknesses. The same two scans contribute to the mean–median reversal, the empty-prediction counts and the reduced pooled boundary fraction below; these are overlapping manifestations of one tail behaviour rather than three independent findings.

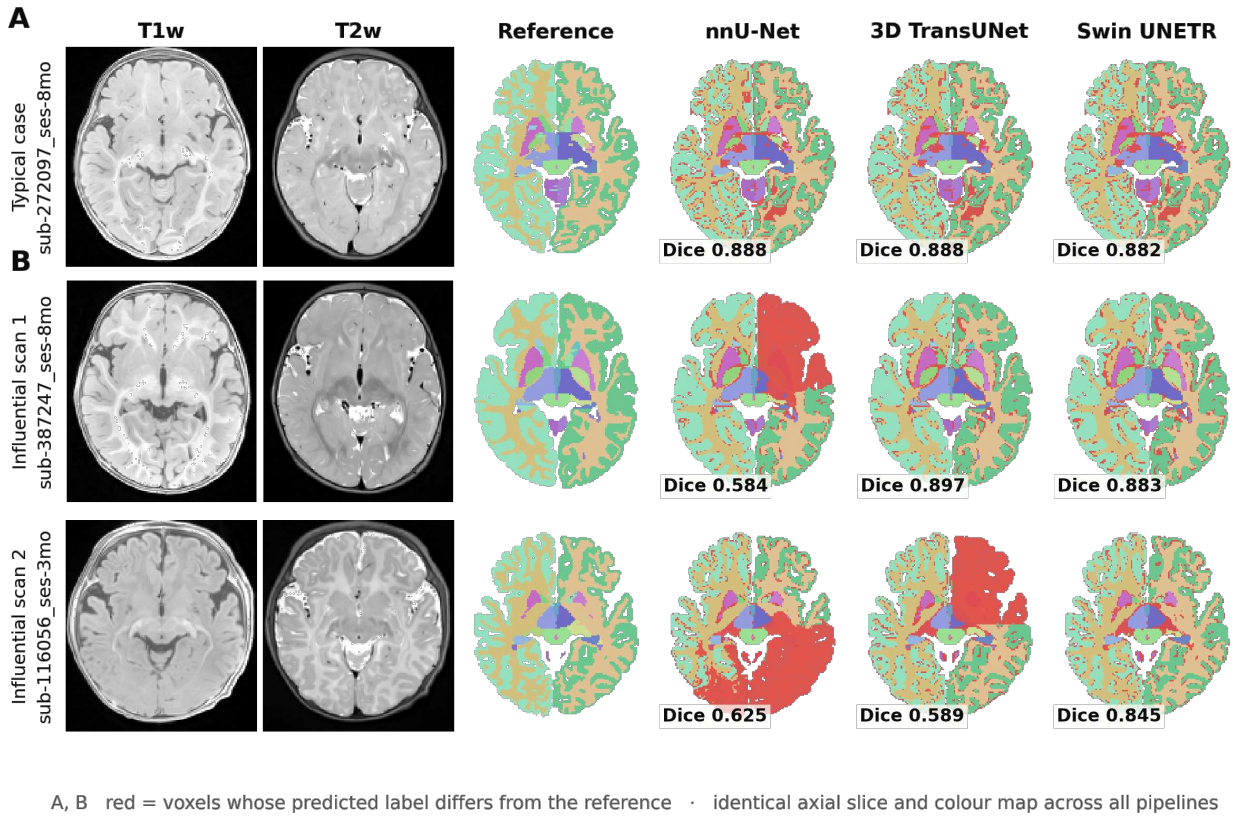

Figure 4: A typical case and two influential scans. Each row shows the same axial slice and label colour map across T1-weighted and T2-weighted images, the expert-corrected reference, and out-of-fold predictions from the three 3D pipelines. Red denotes voxels whose predicted label differs from the reference; anatomical label colours are illustrative and exclude the red hue band. Per-case mean Dice is shown below each prediction. The typical case (top) shows close agreement among pipelines. In `sub-387247_ses-8mo` (middle), nnU-Net shows a large hemispheric failure, while the other two pipelines remain largely unaffected. In `sub-116056_ses-3mo` (bottom), nnU-Net and 3D TransUNet show substantial failures, whereas Swin UNETR remains comparatively intact.

Boundary-error fractions were examined under two aggregation schemes (Online Resource 1, Table S2). With equal weighting of structures, the fraction of error within 1 mm of the reference surface was 84.1% for 3D TransUNet, 82.9% for Swin UNETR, 82.4% for nnU-Net and 72.2% for STSNet; with pooled surface counts it was 88.7%, 90.3%, 74.9% and 63.9%. The schemes disagreed on ordering at 1 mm: pooled aggregation placed nnU-Net well below the other 3D pipelines, equal weighting did not. The qualitative conclusion that 3D-pipeline error was predominantly boundary localisation held under both schemes. In a post-hoc sensitivity analysis, excluding the two influential scans raised the nnU-Net pooled 1 mm fraction from 74.9% to 82.8%.

3.6 Structure complexity and metric behaviour (RQ3.1)

Across the 31 structures, Dice was strongly negatively associated with surface-to-volume ratio (Spearman $\rho = -0.836$, $p = 4.83 \times 10^{-9}$) and positively with reference volume ($\rho = +0.438$, $p = 0.0138$); NSD at 1 mm was more weakly associated with surface-to-volume ratio ($\rho = -0.325$, $p = 0.0741$) and not with volume ($\rho = +0.046$, $p = 0.806$) (Online Resource 1, Fig. S1). The metrics nevertheless ordered structures similarly overall (rank agreement $\rho = +0.632$, $p = 1.36 \times 10^{-4}$), with individual structures moving substantially: the cerebellar cortex labels gained 18 and 20 rank positions under NSD relative to Dice, while the fourth ventricle and Amygdala-A lost 14 and 12. Dice was therefore more strongly associated with structure geometry than NSD in this dataset; the data do not demonstrate size independence of NSD.

3.7 Age and image-contrast analyses (EQ1)

We detected no association between Dice and chronological age (subject-clustered slope +0.00246 per month, $p = 0.302$) or between Dice and the prespecified joint contrast-to-noise index (slope −0.01275, $p = 0.262$). The nominal session label gave the same conclusion (+0.00293 per month, $p = 0.187$), and substituting true age in the variance model left the headline result unchanged (Dice pipeline share 0.134%, $p = 0.177$). Of four contrast terms, only T2 contrast was nominally associated with Dice ($p = 0.038$) and did not survive Holm correction ($p = 0.153$). T1 and T2 contrast were themselves strongly age-dependent (Spearman $\rho = +0.888$ and -0.780), indicating substantial age-related contrast change without a detected association with performance.

3.8 Consensus (RQ4.1)

The consensus improved subject-weighted mean Dice by +0.0062 over 3D TransUNet (95% CI −0.0031 to +0.0155) and produced no empty predictions. The paired improvement over TransUNet was detected for Dice ($p < 0.0001$) and ASSD ($p < 0.0001$) but not for NSD ($p = 0.066$) or HD95 ($p = 0.295$); against nnU-Net all four metrics improved with detection (Dice $p = 0.0001$; NSD $p = 0.0013$; HD95 $p = 0.0114$; ASSD $p = 0.0015$). Vote composition and the cross-fitted tie priority, which confirmed that cross-fitting was active, are given in Online Resource 1 (Section S3). The +0.0062 gain is small, its CI includes zero, it is smaller than the +0.0083 nnU-Net-to-TransUNet gap, and the consensus multiplies inference cost fourfold.

3.9 Exploratory longitudinal analysis (EQ2)

Among the 17 infants contributing 20 within-subject intervals, the predicted direction of structure-wise volume change agreed with the reference in 87.4% of cases for nnU-Net, 87.1% for STSNet, 86.9% for 3D TransUNet and 86.0% for Swin UNETR, with median absolute errors in the rate of volume change of 62–75 mm^3/month. These 620 structure–interval observations are statistically dependent and come from a small subset; the result is descriptive and does not establish developmental sensitivity.

4 Discussion

Under a single subject-disjoint protocol on 71 visits from 51 infants, the observed differences among three competently configured 3D pipelines were small: the pipeline term explained 0.134% of the modelled Dice sum of squares and was not statistically detected, every paired standardised effect was small and every paired mean-difference CI included zero, whereas the 2.5D pipeline separated clearly (RQ1). Retraining variation was small where measurable, while fold-to-fold differences were larger and changed sign (RQ1.1, RQ1.2). Mean and rank-based summaries disagreed in sign for one pipeline pair (RQ2.1), and two scans drove that reversal, the empty-prediction failures and the depressed pooled boundary fraction (RQ2.2). Dice was far more strongly associated with structure geometry than NSD (RQ3.1). The consensus gave a small mean gain but eliminated empty predictions (RQ4.1). No age or contrast association was detected (EQ1), and the longitudinal subset was too small for inference (EQ2).

The two parts of the RQ1 result belong together. Average separation among the 3D pipelines was small, yet 3D TransUNet exceeded Swin UNETR in 46 of 51 infants: a consistent direction across subjects can

coexist with a small mean difference when individual differences are small and variable. The mean–median disagreement makes the same point from the other side. nnU-Net exceeded Swin UNETR in 42 of 51 infants, yet two large negative cases reversed the sign of the mean; a benchmark reporting only the cohort mean would conclude that Swin UNETR was better, one reporting only the median or win count the opposite. Both are arithmetically correct and capture different aspects of the paired distribution, which supports reporting distributions, paired differences, win counts and failure counts rather than one central-tendency statistic, consistent with the literature on ranking instability [17, 21, 20].

The contrast between retraining SDs of 0.0005–0.0016 and a fold-level SD of 0.0258 requires qualification. Repeats were sparse and covered selected folds and pipelines; 3D TransUNet was not repeated; a training seed is not the only source of nondeterminism [23, 24]; and changing the fold changes both the held-out subjects and the training set, producing a separately fitted model. The two quantities are not directly comparable variance components, and no ratio between them is reported. Nevertheless, repeated-seed variability alone would not have captured the variability visible in this cohort, so fold-level results are useful complementary information for small datasets. Likewise, because the pipelines differed in dimensionality, patch geometry, optimiser, schedule, augmentation, batch size, preprocessing and implementation, the results support claims about end-to-end pipelines under a common dataset and split, not causal claims about architectural components; the approximately 0.09 Dice separation of STSNet demonstrates detectable power but does not isolate 2.5D representation as its cause.

The strong negative association of Dice with surface-to-volume ratio, against the weaker associations of 1 mm NSD, is consistent with the geometric sensitivity of overlap metrics: for a given boundary displacement the share of misassigned voxels grows with surface-to-volume ratio, so small or complex structures score lower even when boundary discrepancies are small [19, 20]. We do not conclude that one metric is superior; the implication is that reports should specify metric, physical tolerance, aggregation scheme, empty-mask policy and evidence of implementation validation, particularly because surface metrics can be implemented incorrectly [21, 22].

A pipeline can achieve high cohort-level Dice while failing substantially on individual scans. That all five nnU-Net empty structures fell within one scan, and that two scans drove several summary statistics, argues for case-level quality control, explicit failure flags and visual review of tail cases whenever segmentations feed downstream morphometry [25]. This was a public research dataset rather than a deployment study, and we make no claim about clinical safety. The cross-fitted consensus removed the empty predictions, but its Dice gain over 3D TransUNet was small with a CI including zero, it was not detected for NSD or HD95, it requires four inferences and it was not externally validated; it is a robustness baseline, not a new method.

For paediatric radiologists the overall picture is encouraging. Three widely available pipelines, trained under five-fold cross-validation in a cohort of 51 infants, segmented 31 structures with mean Dice above 0.87 in infants aged 1.0 to 9.4 months, and no association with age or image contrast was detected despite the large contrast changes across the isointense period; a simple consensus of existing predictions removed every empty-structure failure without additional training. These observations support cautious optimism that AI-derived infant segmentations can be reliable inputs to developmental research, with two provisos that apply to any AI tool offered for children's imaging. First, the absence of a detected age effect in 51 infants is not evidence that none exists, and external validation on other scanners and populations is outstanding. Second, trust should rest on how a tool was evaluated rather than on its headline score: whether the evaluation was subject-disjoint, how much repeated training and different partitions changed the result, how many scans failed and whether those failures were inspected. Table 6 turns these into a reporting template; it is inexpensive relative to the cost of training the models it evaluates, it allowed the aggregation reversal and failure concentration in this cohort to be detected rather than averaged away, and it is consistent with existing reporting guidance for medical imaging AI [26, 25].

Table 6: Reporting template for benchmarks of segmentation pipelines on small paediatric imaging cohorts with repeated measures, with the item's location in this study.

	Item to report	Why it matters, and where it appears here
1	Subject-disjoint split locked before training and released as a file	Repeat visits of one infant in train and test inflate every metric; Table 1, Fig. 1
2	Every reported prediction out of fold; any non-out-of-fold reference clearly separated	Leaderboard scores from in-fold predictions are not comparable; Methods
3	Metric code validated against an independent implementation, with the agreement reported	Surface metrics are easy to implement incorrectly; Methods
4	Explicit empty-prediction policy and valid denominators per metric	Silent exclusion changes denominators between pipelines; Table 3
5	Repeated visits aggregated at the subject level before inference	Infants with repeat visits otherwise dominate cohort summaries; Methods
6	Retraining variation and fold-level variation, where available, not pooled means alone	Fold means changed sign here while retraining SD was 0.0005–0.0016; Table 4, Fig. 2
7	Paired differences with confidence intervals, effect sizes, medians and win counts, not significance alone	Mean and median disagreed in sign for one pair; Table 5, Fig. 3
8	Failures counted, inspected, displayed, and tested for concentration within cases	Two scans drove three summary statistics; Fig. 4
9	Split, code, configurations and regeneration scripts archived	Every number here regenerates from one declared set of hashed inputs; Methods

A study published while this analysis was being finalised used the same 71 BOBs visits to quantify how segmentation error from adult-trained pipelines applied off the shelf propagates into volume and fractal-dimension estimates, reported that segmentation-related limits of agreement frequently exceeded published between-group developmental effects, and identified evaluation of infant-trained frameworks as future work [3]. The two studies are complementary rather than comparable: the estimands differ (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 whereas all pipelines here used the same paired T1- and T2-weighted input (Online Resource 1, Section S4). We therefore do not compare absolute values between the studies; together they argue that segmentation quality is a substantive determinant of infant morphometry rather than a technical preprocessing step.

This study has limitations. It is a single cohort from a single site without external validation, and the sample of 51 infants with a limited age range means that a non-detected age or contrast association is not evidence that no developmental effect exists. The compared systems are end-to-end pipelines differing beyond architecture, so no causal architectural conclusion is available. Retraining coverage is sparse, with no 3D TransUNet repeat. No equivalence margin was prespecified and no equivalence test was performed, so non-detection of a pipeline effect must not be read as equivalence. HD95 and ASSD are undefined for empty predictions, producing metric-specific denominators. Two influential scans contribute to several findings that should not be treated as independent evidence. The reference labels were expert-corrected rather than independently rated, no inter-rater estimate was available, and the local label checks were sanity checks. Training ran under shared GPU conditions, so wall-clock figures are not a controlled efficiency benchmark. The longitudinal analysis is small, dependent and exploratory. Finally, folds were balanced on nominal session labels rather than true age, although the two were highly rank-concordant and substituting true age left the headline results unchanged.

5 Conclusion

Under a common subject-disjoint protocol, three well-configured three-dimensional deep learning pipelines segmented the infant brain with high accuracy and small mutual differences that were dwarfed by fold-to-fold

variation and by two individual scans, while a simple consensus removed every failure. Benchmarks of infant brain segmentation should report retraining and fold variation, paired effect sizes with confidence intervals, failure counts and aggregation sensitivity alongside average overlap, and external validation is required before these findings are generalised beyond this cohort.

Data availability

The source MRI and reference segmentations are publicly available through the Baby Open Brains resource on OpenNeuro (accession ds005450), subject to its stated licence. Data generated during this analysis, the subject-disjoint split definitions, evaluation code, configuration files and the scripts used to regenerate the reported tables are available from the corresponding author on reasonable request. Generative AI was used solely for language editing of human-written text; no AI tool was used for data analysis, figure generation or content creation.

References

[1] Li G, Wang L, Yap PT, Wang F, Wu Z, Meng Y et al (2019) Computational neuroanatomy of baby brains: a review. NeuroImage 185:906–925. https://doi.org/10.1016/j.neuroimage.2018.03.042

[2] Makropoulos A, Robinson EC, Schuh A, Wright R, Fitzgibbon S, Bozek J et al (2018) The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. NeuroImage 173:88–112. https://doi.org/10.1016/j.neuroimage.2018.01.054

[3] 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

[4] Mhlanga ST, Viriri S (2023) Deep learning techniques for isointense infant brain tissue segmentation: a systematic literature review. Front Med 10:1240360. https://doi.org/10.3389/fmed.2023.1240360

[5] Bastiani M, Andersson JLR, Cordero-Grande L, Murgasova M, Hutter J, Price AN et al (2019) Automated processing pipeline for neonatal diffusion MRI in the developing Human Connectome Project. NeuroImage 185:750–763. https://doi.org/10.1016/j.neuroimage.2018.05.064

[6] Wang L, Nie D, Li G, Puybareau 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

[7] Zöllei L, Iglesias JE, Ou Y, Grant PE, Fischl B (2020) Infant FreeSurfer: an automated segmentation and surface extraction pipeline for T1-weighted neuroimaging data of infants 0–2 years. NeuroImage 218:116946. https://doi.org/10.1016/j.neuroimage.2020.116946

[8] Hendrickson TJ, Reiners P, Moore LA, Lundquist JT, Fayzullobekova B, Perrone AJ et al (2026) BIBSNet: a deep learning baby image brain segmentation network for MRI scans. Dev Cogn Neurosci 79:101706. https://doi.org/10.1016/j.dcn.2026.101706

[9] Feczko E, Stoyell SM, Moore LA, Alexopoulos D, Bagonis M, Barrett K et al (2025) Baby Open Brains: an open-source dataset of infant brain segmentations. Sci Data 12:1423. https://doi.org/10.1038/s41597-025-05404-y

[10] Howell BR, Styner MA, Gao W, Yap PT, Wang L, Baluyot K et al (2019) The UNC/UMN Baby Connectome Project (BCP): an overview of the study design and protocol development. NeuroImage 185:891–905. https://doi.org/10.1016/j.neuroimage.2018.03.049

[11] 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

[12] Isensee F, Wald T, Ulrich C, Baumgartner M, Roy S, Maier-Hein K, Jäger PF (2024) nnU-Net revisited: a call for rigorous validation in 3D medical image segmentation. In: Medical Image Computing and Computer Assisted Intervention, MICCAI 2024. Lecture Notes in Computer Science. Springer, Cham, pp 488–498. https://doi.org/10.1007/978-3-031-72114-4_47

[13] Chen J, Mei J, Li X, Lu Y, Yu Q, Wei Q et al (2024) TransUNet: rethinking the U-Net architecture design for medical image segmentation through the lens of transformers. Med Image Anal 97:103280. https://doi.org/10.1016/j.media.2024.103280

[14] Hatamizadeh A, Nath V, Tang Y, Yang D, Roth HR, Xu D (2022) Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. In: Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries, BrainLes 2021. Lecture Notes in Computer Science. Springer, Cham, pp 272–284. https://doi.org/10.1007/978-3-031-08999-2_22

[15] Hatamizadeh A, Tang Y, Nath V, Yang D, Myronenko A, Landman B, Roth HR, Xu D (2022) UNETR: transformers for 3D medical image segmentation. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp 1748–1758. https://doi.org/10.1109/WACV51458.2022.00181

[16] Zhao L, Wang T, Chen Y, Zhang X, Tang H, Lin F et al (2025) A novel framework for segmentation of small targets in medical images. Sci Rep 15:9924. https://doi.org/10.1038/s41598-025-94437-9

[17] 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

[18] Bilic P, Christ P, Li HB, Vorontsov E, Ben-Cohen A, Kaissis G et al (2023) The Liver Tumor Segmentation Benchmark (LiTS). Med Image Anal 84:102680. https://doi.org/10.1016/j.media.2022.102680

[19] Taha AA, Hanbury A (2015) Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. BMC Med Imaging 15:29. https://doi.org/10.1186/s12880-015-0068-x

[20] 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

[21] 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

[22] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J et al (2021) Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res 23:e26151. https://doi.org/10.2196/26151

[23] Åkesson J, Töger J, Heiberg E (2024) Random effects during training: implications for deep learning-based medical image segmentation. Comput Biol Med 180:108944. https://doi.org/10.1016/j.compbiomed.2024.108944

[24] Bosma JS, Peeters D, Alves N, Saha A, Saghir Z, Jacobs C, Huisman H (2024) Reproducibility of training deep learning models for medical image analysis. In: Medical Imaging with Deep Learning, Proceedings of Machine Learning Research, vol 227, pp 1269–1287. https://proceedings.mlr.press/v227/bosma24a.html

[25] Maleki F, Moy L, Forghani R, Ghosh T, Ovens K, Langer S et al (2025) RIDGE: reproducibility, integrity, dependability, generalizability, and efficiency assessment of medical image segmentation models. J Imaging Inform Med 38:2524–2536. https://doi.org/10.1007/s10278-024-01282-9

[26] Tejani AS, Klontzas ME, Gatti AA, Mongan JT, Moy L, Park SH et al (2024) Checklist for Artificial Intelligence in Medical Imaging (CLAIM): 2024 update. Radiol Artif Intell 6:e240300. https://doi.org/10.1148/ryai.240300

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- ESM1.pdf