Built for Children: A Pediatric-Trained Hybrid Autosegmentation Pipeline for Craniospinal Irradiation

Zachary T. Wooten, PhD[1], Catherine Prabish[2], Andrew Boria, PhD[2], Chia-ho Hua, PhD[2], Thomas E. Merchant, DO, PhD[2] and Ozgur Ates, PhD[2]

[1]Department of Biostatistics, [2]Department of Radiation Oncology, St. Jude Children's Research Hospital, Memphis, TN

BACKGROUND

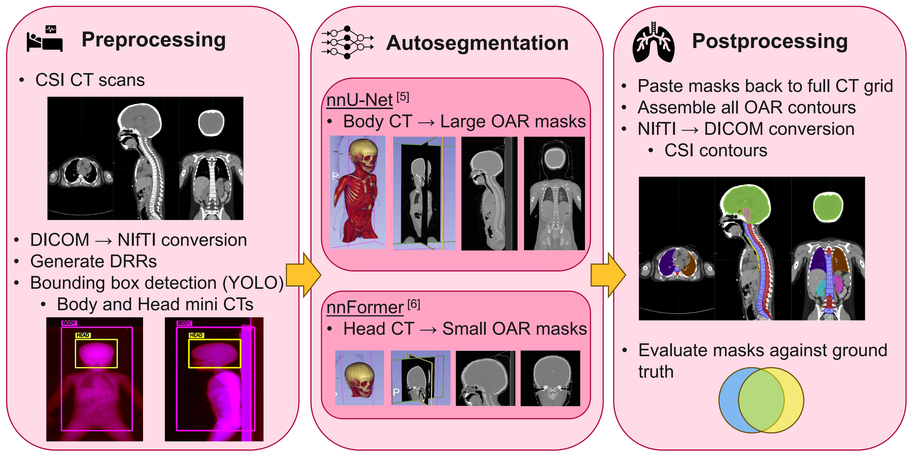
- Autosegmentation accelerates radiation planning by reducing manual contouring time and inter-observer variability, improving consistency and workflow efficiency across patients. [1,2]
- Commercial tools are adult-trained and can omit key craniospinal irradiation (CSI) structures limiting pediatric applicability. [3,4]
- We hypothesized that a pediatric-trained AI pipeline would produce more anatomically consistent contours and reduced boundary error compared with adult-trained and atlas-based tools, bridging the gap between research-stage AI and clinically deployable pediatric segmentation.

OBJECTIVES

1. Develop a modular and scalable in-house AI pipeline to capture both large and small structures.
2. Compare performance against an atlas-based method and a commercial AI using 13 overlap and boundary metrics, analyzed with linear mixed-effects (LME) models to account for patient and organ variability.
3. Evaluate robustness across age (2–25 years) to determine whether pediatric-specific training maintains performance despite growth-related anatomical variation.

METHODS

- 211 CSI CTs total (138 train, 73 test; median age 9y, range 2–25y).
- 18 Organs at Risk (OAR): brain, brainstem, optic chiasm, esophagus, spinal canal, vertebrae the left and right structures for cochleae, eyes, lenses, lungs, kidneys, and optic nerves.
- Performed LME analysis to capture performance of methods, age, and organ variability



$$y_{ijk} = \beta_0 + \beta_{Method}(Method)_j + \beta_{Age}(Age)_i + \beta_{Method \times Age}(Method \times Age)_{ij} + u_i + v_k + \epsilon_{ijk}$$

- y_{ijk} : Metric value for patient *i*, method *j*, organ *k*
- β : Method, Age, and Method × Age interaction effects
- u_i : Random intercept for patient
- v_k : Random intercept for organ
- ϵ_{ijk} : Residual error

RESULTS

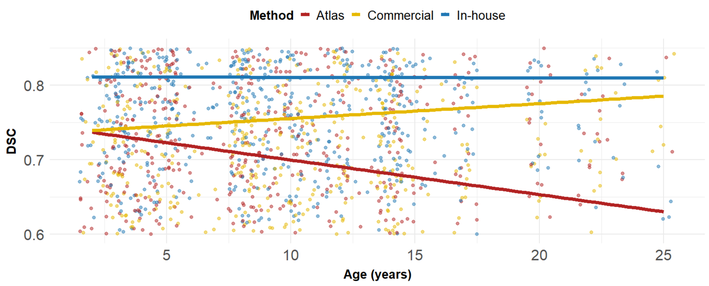


Figure 1: Fixed effects from segmentation methods. Dice similarity coefficients (DSCs) were plotted with age interaction from the linear mixed-effects model. Results were stratified by the segmentation method: atlas (red), commercial AI (gold), and in-house AI model (blue)

- The LME showed a method's performance is affected by patient age, $\beta_{Method \times Age}$ ($p < 0.01$).
- Atlas performance decreased with age, while commercial AI improved with older patients.
- In-house AI maintained consistent performance across all ages.

RESULTS

Performance Metrics: Dice Similarity Coefficient (DSC); Jaccard Index (JAC); True Positive Rate (TPR); True Negative Rate (TNR); Positive Predictive Value (PPV); Rand Index (RI); Mean Distance to Agreement (MDA); Hausdorf Distance (HD).

- Our proposed model shows superior performance across all 13 metrics.

Comparing β_{Method} differences:

- In-house AI vs Atlas: Dice +0.11 (95% CI [0.10–0.12], $d = 0.73$, $p < 0.01$).
- In-house AI vs Commercial AI: Dice +0.06 (95% CI [0.04–0.07], $d = 0.37$, $p < 0.01$)

Metric	Atlas	Commercial	In-house
DSC	0.70 ± 0.27	0.74 ± 0.25	**0.81 ± 0.18**
JAC	0.60 ± 0.28	0.64 ± 0.28	**0.71 ± 0.21**
TPR	0.71 ± 0.28	0.74 ± 0.28	**0.81 ± 0.20**
TNR	0.92 ± 0.09	0.94 ± 0.07	**0.96 ± 0.05**
RI	0.89 ± 0.10	0.88 ± 0.11	**0.93 ± 0.06**
PPV	0.73 ± 0.28	0.80 ± 0.23	**0.83 ± 0.18**
Std HD	2.26 ± 4.16	1.55 ± 2.30	**1.29 ± 2.67**
Min HD	2.10 ± 13.02	0.57 ± 7.82	**0.24 ± 3.55**
Median HD	3.82 ± 16.39	1.61 ± 10.90	**0.62 ± 3.86**
Mean HD	4.36 ± 16.37	2.00 ± 10.79	**1.11 ± 4.10**
HD95	8.71 ± 20.95	5.17 ± 13.67	**3.50 ± 8.88**
Max HD	14.69 ± 23.44	**9.38 ± 15.75**	**9.38 ± 13.97**
MDA	4.05 ± 14.89	1.89 ± 10.14	**1.11 ± 4.10**

Table 1: Thirteen quantitative metrics (mean ± SD) averaged over 73 test patients comparing atlas-based, commercial AI, and in-house AI methods.

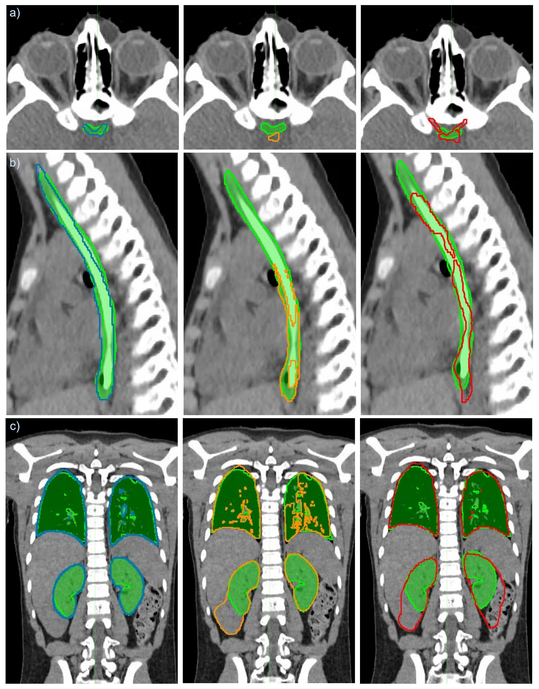


Figure 2: DSC were determined by comparing an in-house AI model (blue), commercial AI (gold), and atlas-based method (red) to ground truth (green) in images of the following structures: (a) optic chiasm (b) esophagus; and (c) lungs and kidney

- In-house AI performed exceptionally well in contouring the optic chiasm, compared to commercial AI, with an average improvement of DSC +0.31 ($d = 1.45$).
- When compared to Atlas, the in-house AI made significant improvements in segmenting kidneys, DSC +0.18 ($d = 1.29$), and produced tighter boundary and cleaner edges for the esophagus, Median HD −0.75 mm ($d = 1.61$).
- We aggregated two-sample t-tests comparing in-house AI vs Atlas, and in-house vs. commercial AI. Across 234 tests (13 metrics × 18 organs), the in-house AI was superior in 42% of cases, while commercial AI led in only 4%; others showed no difference.

- Including organ variability is critical for accurate method comparisons.
- Organs naturally cluster in performance.
  - **Low (<0.40):** Optic chiasm
  - **Mid (0.40–0.75):** Optic nerves, lenses, cochleae
  - **High (>0.75):** Brain, brainstem, lungs, kidneys, vertebrae
- Our model shows improved performance for difficult organs and reduces null segmentations (DSC = 0).

Figure 3: Comparison of predicted versus observed DSCs from the linear mixed-effects model. Results were stratified by organ group (top panel) and then by segmentation method (atlas, commercial AI, in-house AI) (bottom panel). The dashed line indicates perfect agreement between the predicted and observed DSC values.

CONCLUSION

- The in-house, pediatric-trained, AI consistently outperformed both atlas-based and commercial AI methods across 18 organs, achieving higher overlap and lower boundary error.
- Performance remained stable across ages (2–25 years), demonstrating robustness to potential age-related anatomy changes, which supports pediatric deployment.

Next Steps:

- Integrate automated QA into the pipeline, using CT-intensity and shape-based validation
- Assess workflow gains, including edit time reduction and planning efficiency in clinical use.
- Assess dosimetric impact of segmentation accuracy on clinical outcomes.

[1] L. Court, The Radiation Planning Assistant: addressing the global gap in radiotherapy services, Lancet Oncology 24, e363 (2023).
[2] F. Vaassen et al., Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy, Physics and Imaging in Radiation Oncology 13, 1–6 (2020).
[3] O. Ates et al., Comparative Analysis of Atlas and Neural Network Autosegmentation Methods for Pediatric Craniospinal Irradiation With the Development of a Knowledge-Based Quality Assurance Tool, Advances in Radiation Oncology 10, 101847 (2025).
[4] H. Wan, Automated contouring using neural networks, MIM Software Inc. White Paper, 2020.
[5] F. Isensee et al., nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation, Nature Methods 18, 203–211 (2021).
[6] H.-Y. Zhou et al., nnFormer: Volumetric Medical Image Segmentation via a 3D Transformer, 2021, Preprint.

The Wooten group provides statistical support and machine-learning based methodological innovations for diagnosis and treatment planning.