Clinical Implementation of a Knowledge-Based Quality Assurance Tool Using CT and Shape Radiomics for Autosegmentation in Pediatric Craniospinal Irradiation

Zachary T. Wooten, PhD; Isaac DeBord; Chia-Ho Hua PhD; Thomas E. Merchant DO, PhD; Ozgur Ates, PhD
St. Jude Children's Research Hospital, Memphis, TN, USA

2026 JULY 19–22 VANCOUVER, BC
JOINT AAPM | COMP MEETING

St. Jude Children's Research Hospital
Finding cures. Saving children.
ALSAC • DANNY THOMAS, FOUNDER

ABSTRACT

Purpose: Automated segmentation is increasingly integrated into radiotherapy planning workflows; however, ensuring reliable quality assurance (QA) remains challenging, particularly in pediatric patients due to substantial anatomical variability. In this study, we leveraged a large cohort of historical cases to develop a novel knowledge-based QA framework that integrates CT intensity- and shape-based radiomic features within a kernel density estimation (KDE) model. This framework is designed to flag erroneous autosegmentations in pediatric craniospinal irradiation (CSI).

Methods: KDE distributions of CT number intensity and five geometric shape metrics (eccentricity, sphericity, solidity, perimeter, and area) were derived for 16 organs-at-risk, accounting for bilateral structures (brain, brainstem, chiasm, esophagus, eyes, cochleae, kidneys, lenses, lungs, and optic nerves) using clinician-approved contours from 100 pediatric CSI patients (range: 2-25 years; median: 8 years) and served as baseline KDEs. For each auto-generated contour, agreements with organ-specific KDE baseline distributions were quantified, and acceptance thresholds were defined as ±2 standard deviations. Autosegmentations generated by three methods (atlas, commercial AI, and an in-house pediatric-trained AI) were evaluated across 47 patients with 2,256 contours. Segmentation performance between flagged and unflagged contours was compared using Wilcoxon rank-sum test.

Results: Across all organs, contours flagged by the proposed QA framework revealed significantly lower Dice similarity coefficients, sensitivity, accuracy, and precision ($p < 0.001$), while specificity did not differ significantly. Shape-based QA identified 132 contours with geometric inconsistencies, 99 of which were not detected by CT-based QA alone. Incorporating shape QA increased the total flagging rate from 13.6% (CT-based QA only) to 17.8%, improving detection of suboptimal segmentations.

Conclusion: Distinct radiomic patterns from historical pediatric CSI cases enabled the development of a mathematical KDE model integrating CT intensity- and shape-based analyses. Together, these methods form a quantitative, interpretable, and vendor-neutral QA framework for pediatric CSI that facilitates reliable detection of gross autosegmentation errors.

CONTACT

Ozgur Ates, PhD, DABR
St. Jude Children's Research Hospital
262 Danny Thomas Pl. Memphis TN 38105
ozgur.ates@stjude.org
901-595-2404

INTRODUCTION

The absence of pediatric-specific commercial auto-segmentation solutions highlights the need for robust quality assurance (QA) strategies [1,2,3]. Pediatric cases pose unique challenges due to limited data availability and wide variability in organ sizes, which hinder the reliable training and validation of AI algorithms.

It was found that kernel density estimation (KDE) provides a reliable mathematical framework for modeling radiomic patterns of critical organs, encompassing both CT number intensity and shape, across patients and independent of organ size variability. By establishing QA frameworks that leverage radiomic features such as CT intensity- and geometric shape-based patterns from large historical cases, clinicians can be guided toward suboptimal contours that deviate from expected segmentations.

METHODS AND MATERIALS

KDE-based QA models were developed using clinician-approved organ-at-risk contours from 100 pediatric craniospinal irradiation patients. Patient ages ranged from 2 to 25 years, with a median age of 8 years. 16 organ-at-risk structures were included. These included the brain, brainstem, chiasm, esophagus, left and right eyes, left and right cochleae, left and right kidneys, left and right lenses, left and right lungs, and left and right optic nerves.

For each clinician-approved contour, CT number intensity information was extracted from the corresponding CT image. KDE distributions of CT number intensity were generated to characterize the expected image-intensity profile of each organ. In addition, five geometric shape metrics were calculated for each contour: eccentricity, sphericity, solidity, perimeter, and area. These metrics were selected to capture complementary aspects of contour morphology, including elongation, compactness, convexity, boundary extent, and overall size. Organ-specific KDE baseline distributions were then created from the clinician-approved contours and used as reference models representing the expected intensity and geometric characteristics of each organ.

Auto-generated contours were evaluated by comparing their CT intensity and shape-feature distributions against the corresponding organ-specific baseline KDEs. For each autosegmented structure, agreement with the baseline distribution was quantified across the intensity and geometric feature domains. Acceptance thresholds were defined using the baseline population statistics, with contours falling outside ±2 standard deviations from the organ-specific reference distribution flagged as potentially unacceptable. Contours within this range were considered unflagged.

Autosegmentations generated by three methods were assessed: atlas-based segmentation, a commercial AI segmentation tool, and an in-house pediatric-trained AI model. The evaluation cohort consisted of 47 pediatric patients and was partitioned into a calibration subset of 14 patients and an independent test subset of 33 patients. Segmentation performance was compared between contours flagged by the KDE-based quality assurance approach and those not flagged. Differences in segmentation performance metrics between flagged and unflagged groups were assessed using the Wilcoxon rank-sum test. This nonparametric test was selected to compare group distributions without assuming normality.

RESULTS

Substantial organ-level variability was observed in segmentation performance and KDE agreement across the three autosegmentation models, with 99 contours evaluated per organ (**Table 1**). Large structures such as the brain and lungs achieved consistently strong performance, with mean Dice similarity coefficient (DSC) values exceeding 0.90 and accuracy above 0.97. These organs also demonstrated high KDE agreement for both CT intensity and shape features. In contrast, smaller and more anatomically challenging structures, including the optic nerves, cochleae, and optic chiasm, showed lower DSC values, often below 0.60, while maintaining relatively high accuracy, typically above 0.80. Across organs, CT-based KDE agreement remained high, exceeding 90% for most structures, with lower agreement primarily observed in small organs such as the optic chiasm and lenses. When aggregating results across organs, flagged segmentations performed worse than unflagged segmentations across multiple metrics (**Figure 1**).

To account for organ-specific anatomical variability, adaptive KDE agreement thresholds were used rather than a single global cutoff. For each organ and feature type, target agreement levels of 0.90 for CT-number agreement and 0.80 for shape agreement were specified. Across all organs, contours flagged by the proposed QA framework revealed significantly lower Dice similarity coefficients, sensitivity, accuracy, and precision ($p < 0.001$), while specificity did not differ significantly.

Table 1. Organ-level KDE percent agreement metrics across three autosegmentation models in the independent test cohort. Agreement percent values are reported as mean ± standard deviation and represent percent KDE agreement relative to clinician-approved ground truth. Metrics were computed from 99 autosegmentations per organ (33 patients × 3 models).

Organ	CT	Area	Perimeter	Solidity	Sphericity	Eccentricity
Brain	95.59 ± 12.83	92.86 ± 14.05	94.76 ± 6.17	86.88 ± 13.74	93.15 ± 6.91	94.33 ± 9.37
Brainstem	97.14 ± 7.11	92.02 ± 12.94	91.44 ± 13.20	88.57 ± 9.71	89.32 ± 15.00	92.86 ± 10.81
Chiasm	84.10 ± 25.03	86.67 ± 20.02	80.17 ± 23.84	77.89 ± 24.66	78.36 ± 27.52	91.29 ± 14.21
$Cochlea_L$	87.19 ± 19.66	83.61 ± 18.49	81.82 ± 18.94	65.15 ± 29.62	86.50 ± 18.71	85.18 ± 16.36
$Cochlea_R$	87.59 ± 16.60	84.69 ± 18.87	82.38 ± 19.23	70.88 ± 29.75	89.36 ± 16.30	88.16 ± 16.31
Esophagus	92.33 ± 12.76	91.71 ± 10.76	88.66 ± 11.54	78.60 ± 21.91	91.49 ± 12.57	92.21 ± 10.36
Eye_L	91.67 ± 16.02	92.58 ± 13.61	89.79 ± 12.99	64.06 ± 21.68	76.06 ± 15.47	83.68 ± 20.95
Eye_R	91.38 ± 14.89	87.82 ± 17.88	85.63 ± 16.90	64.80 ± 18.15	74.05 ± 15.24	80.93 ± 21.37
$Kidney_L$	95.11 ± 10.78	87.51 ± 16.95	87.65 ± 15.21	60.48 ± 31.54	83.18 ± 16.44	89.61 ± 10.42
$Kidney_R$	91.21 ± 16.88	83.31 ± 19.12	83.54 ± 18.17	68.04 ± 20.31	86.23 ± 13.50	80.00 ± 17.41
$Lens_L$	84.22 ± 22.05	89.14 ± 16.02	86.28 ± 15.98	70.74 ± 22.96	86.71 ± 14.38	82.52 ± 16.71
$Lens_R$	84.08 ± 22.56	90.41 ± 14.30	89.27 ± 14.38	65.98 ± 25.56	84.70 ± 19.99	81.39 ± 19.74
$Lung_L$	94.14 ± 12.99	93.69 ± 15.64	92.74 ± 8.98	94.69 ± 7.66	93.82 ± 11.94	92.17 ± 7.63
$Lung_R$	95.65 ± 10.21	92.33 ± 20.11	93.51 ± 9.54	92.00 ± 7.31	94.68 ± 9.22	93.48 ± 9.65
Optic $nerve_L$	83.49 ± 16.42	94.40 ± 12.77	93.02 ± 14.30	87.68 ± 16.47	86.00 ± 20.24	83.52 ± 12.72
Optic $nerve_R$	84.75 ± 17.43	95.06 ± 12.47	90.00 ± 18.47	91.28 ± 10.70	93.42 ± 14.82	86.42 ± 14.32

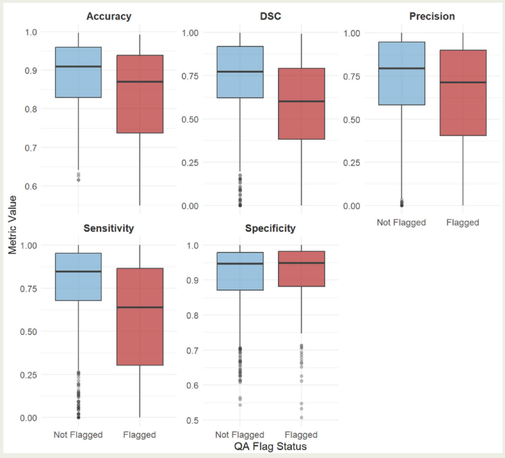

Figure 1. Boxplots of segmentation performance metrics across all organs, comparing contours flagged by the QA framework (red) and those not flagged (blue). Metrics shown include DSC, sensitivity, accuracy, precision, and specificity. Across all organs, flagged contours demonstrated significantly worse performance than unflagged contours, with median differences of 0.152 for DSC, 0.200 for sensitivity, 0.037 for accuracy, and 0.084 for precision (all $p < 0.001$); specificity showed no significant difference (median difference = 0.001).

Figure 2 illustrates the visual interpretability of combined CT-number and shape-based QA after autosegmentation. An example of a suboptimal right kidney autosegmentation from the atlas model is shown in red. The HU KDE deviates from the baseline distribution near −100 HU. The perimeter KDE also exceeds the expected baseline range, extending beyond approximately 400 mm. These deviations are consistent with oversegmentation and demonstrate how shape-based KDEs can indicate the nature of the failure rather than only flagging its presence.

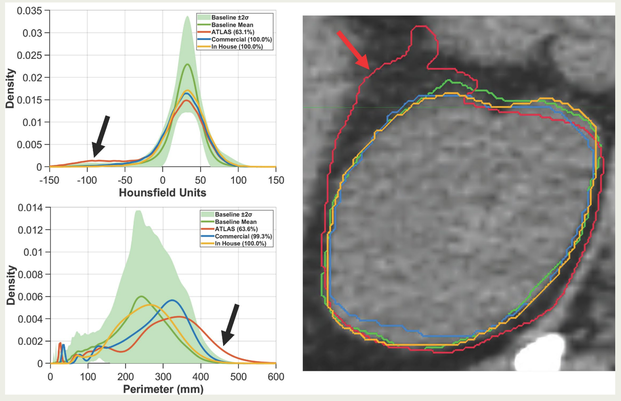

Figure 2. Combined CT-number and shape-based QA for a representative right kidney autosegmentation. Left panels: CT-number and perimeter metric KDEs with the baseline mean and ±2SD band from ground-truth cases (green) and autosegmentation KDEs for the atlas (red), commercial AI (blue), and in-house AI (yellow) segmentations, with agreement shown in the plot legends. Black arrows indicate regions out of agreement and red arrow shows corresponding segmentation error. Right panel: the corresponding CT slice with ground truth (green) and atlas (red), commercial (blue), and in-house (yellow) contours.

DISCUSSION

The proposed QA framework differs from prior approaches [2,3] in several important respects. Methodologically, it models CT-number intensity and geometric shape variability as separate organ-specific signals rather than collapsing them into a single latent representation or composite score. Separating these signals preserves interpretability and exposes failure modes that intensity-only checks may overlook. Shape-based QA identified 132 contours with geometric inconsistencies, 99 of which were not detected by CT-based QA alone. Incorporating shape QA increased the total flagging rate from 13.6% (CT-based QA only) to 17.8%, improving detection of suboptimal segmentations.

Clinical focus further distinguishes the framework. Pediatric CSI presents challenges related to age-dependent anatomy, small-volume structures, and wide inter-patient variability that limit the applicability of adult-derived or global QA thresholds. Learning organ-specific intensity and shape distributions directly from pediatric clinical contours enables a fully unsupervised QA strategy that can be applied across different autosegmentation models without retraining while still producing outputs that clinicians can interpret and act upon.

CONCLUSIONS

The proposed combined QA approach based on CT and shape features allows autosegmentation process to function not only as an efficiency-enhancing workflow but also as a safety net, enabling clinicians to prioritize careful review of critical structures for error. Ultimately, QA-driven autosegmentation can improve confidence, reduce variability, and streamline the contouring process in complex treatments such as craniospinal irradiation (CSI) and can be readily scaled to other anatomical regions.

REFERENCES

1. Thibodeau-Antonacci A, Popovic M, Ates O, Hua CH, Schneider J, Skamene S, Freeman C, Enger SA, Tsui JMG. Trade-off of different deep learning-based auto-segmentation approaches for treatment planning of pediatric craniospinal irradiation autocontouring of OARs for pediatric CSI. Med Phys. 2025 Jun;52(6):3541-3556. doi: 10.1002/mp.17782. Epub 2025 Apr 1.
2. Hanna EM, Sargent E, Hua CH, Merchant TE, Ates O. Performance analysis and knowledge-based quality assurance of critical organ auto-segmentation for pediatric craniospinal irradiation. Sci Rep. 2024 Feb 21;14(1):4251. doi: 10.1038/s41598-024-55015-7.
3. Ates O, Tsui JMG, Wooten Z, Hutcheson S, Zhang R, Becksfort J, Merchant TE, Hua CH. Comparative Analysis of Atlas and Neural Network Autosegmentation Methods for Pediatric Craniospinal Irradiation With the Development of a Knowledge-Based Quality Assurance Tool. Adv Radiat Oncol. 2025 Jul 28;10(9):101847. doi: 10.1016/j.adro.2025.101847.