Evaluation of a Deep Learning-Based Auto-Contouring Tool for Brain Metastases and Organs-at-Risk on MRI

Marvin Kinz[1,2], Vaisakh Nappady Joy[4], Jürgen Hesser[2], Ayal A. Aizer[1], Kamal Singhrao[1], Christian V. Guthier[3], Atchar Sudhyadhom[1]

1 Department of Radiation Oncology, Brigham and Women's Hospital, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA; 2 Mannheim Institute for Intelligent Systems in Medicine, Medical Faculty Mannheim, Department of Physics and Astronomy, Interdisciplinary Center for Scientific Computing (IWR), Central Institute for Computer Engineering (ZITI), CZS Heidelberg Initiative for Model-Based AI, Heidelberg University, 69117 Heidelberg, Germany; 3 Department of Radiation Oncology, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215, USA; 4 Siemens Healthineers, Forchheim 91301, Germany

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

ABSTRACT

Purpose: To evaluate the clinical accuracy and utility of a commercial deep learning-based auto-segmentation tool by Siemens for delineating brain metastases (GTV) and organs-at-risk (OAR) on contrast-enhanced MRI.
Methods: A retrospective analysis was conducted using a large clinical dataset of 786 structure sets. For the brain metastasis analysis, a filtered dataset of 436 patients (520 structure set comparisons) containing 1,933 clinical metastases was utilized. Ground truth was defined by clinical manual contours. Performance metrics included sensitivity, specificity, False Positive (FP) rate, Dice Similarity Coefficient (DSC), and Hausdorff Distance (HD95). GTV analysis was stratified by lesion volume (<0.1 cc vs. >0.1 cc) to assess detection limits. OAR segmentation performance was evaluated against manual contours using volume difference, DSC, and HD95.
Results: For GTV detection, the overall sensitivity was 89.6%. Performance was strongly volume-dependent: sensitivity was 93.3% for lesions >0.1 cc but decreased to 61.5% for lesions <0.1 cc. Specificity followed a similar trend (92.3% vs. 58.5%). The median false positive rate was 1 per case, with 59.4% of cases exhibiting FPs. The median DSC for detected GTVs was 0.79. For OARs, the algorithm demonstrated robust performance (high DSC, low HD95) for the brainstem, eyes, and hippocampus. However, discrepancies were noted in the optic apparatus, where the AI sometimes produced disconnected contours for the optic nerves and chiasm.
Conclusion: The AI auto-contouring tool demonstrates high sensitivity for detecting brain metastases larger than 0.1 cc and provides accurate segmentation for major intracranial OARs, supporting its use to improve efficiency in clinical workflows. Performance metrics are likely an underestimate, especially for lesions <0.1 cc, as the clinical ground truth was not curated for this retrospective analysis. Based on these validation results, the tool is currently being implemented and evaluated as a quality assurance measure in a prospective same-day, MRI-only stereotactic radiosurgery trial.

CONTACT

Marvin Kinz
Mass General Brigham Cancer Institute and Harvard Medical School
Department of Radiation Oncology
75 Francis St, Boston, MA 02115
mkinz@bwh.harvard.edu

INTRODUCTION

Context: Safe and effective stereotactic radiosurgery (SRS) for brain metastases requires precise delineation of the Gross Tumor Volume (GTV) and surrounding Organs-at-Risk (OARs) on MRI.
Problem: Manual contouring is time-consuming and subject to human error or missed lesions, posing a significant risk during the accelerated timelines of same-day, adaptive radiotherapy workflows.
Solution: Evaluate the clinical accuracy of a commercial deep learning auto-segmentation tool to determine its viability as an automated Quality Assurance (QA) "safety net" to prevent missed lesions and improve efficiency.

METHODS AND MATERIALS

Dataset: A large-scale retrospective analysis of 436 patients (520 structure set comparisons) comprising 1,933 clinical metastases delineated on contrast-enhanced MRI.
Ground Truth: Clinical manual contours served as the reference standard (note: weakly curated retrospective clinical data).
Evaluation Metrics: Sensitivity, specificity, False Positive (FP) rate, Dice Similarity Coefficient (DSC), and Hausdorff Distance (HD).
Stratification: GTV analysis was stratified by lesion volume (<0.1 cc vs. >0.1 cc) to assess the AI's detection limits for small targets. OARs were evaluated using volume differences, DSC, and HD.

RESULTS & DISCUSSION

Overall GTV Detection: The tool achieved an overall sensitivity of 89.6%, with a median DSC of 0.79 for detected lesions.
Volume-Dependent Accuracy: Performance was strongly tied to lesion size. For lesions >0.1 cc, sensitivity and specificity were excellent (93.3% and 92.3%, respectively). For micro-lesions <0.1 cc, sensitivity dropped to 61.5% and specificity to 58.5%.
False Positives (FP): The median FP rate was 1 per case overall. However, for clinically significant lesions (>0.1 cc), the median FP rate dropped to 0, demonstrating high reliability.
OAR Segmentation: The algorithm demonstrated robust, highly accurate performance (high DSC, low HD95) for the brainstem, eyes, and hippocampus.
Algorithm Limitations: Discrepancies were observed in the optic apparatus (disconnected contours for the optic nerves/chiasm). Additionally, spinal cord scores were artificially lowered due to the MRI field-of-view frequently cutting off compared to standard CT contours.

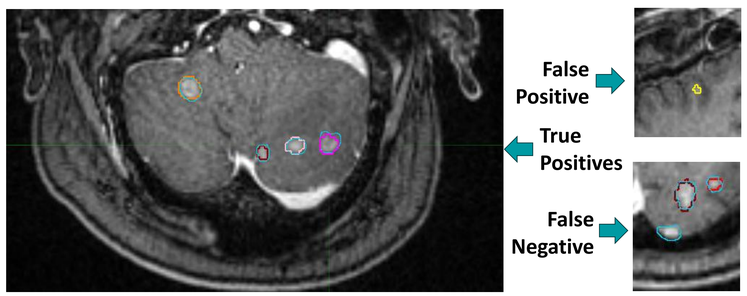

Figure 1. Representative axial MRI slices showing true positives (lesions correctly detected by the AI), a false negative (lesion missed by the AI), and a false positive (AI contour with no corresponding lesion); single illustrative cases. Clinical contours are thin in blue; AI contours are thick in other colors.

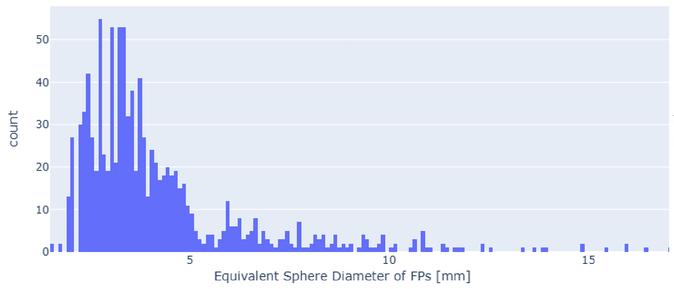

Figure 2. False positive AI lesions by diameter (Dice = 0, N = 1098, x-axis cut-off).

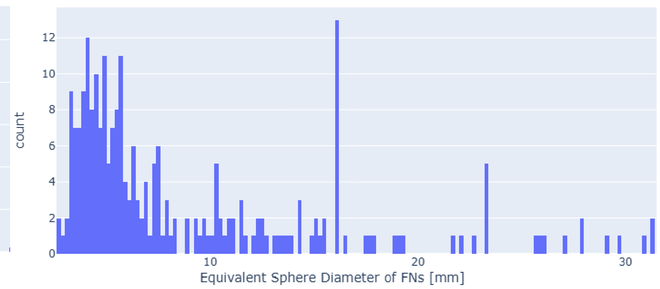

Figure 3. Missed clinical lesions by diameter (Dice = 0, N = 277, x-axis cut-off).

	Sensitivity =TP/(TP+FN)	Specificity =FP/(TP+FP)	False Positive Rate	False Positive Rate Median	% of cases with at least one FP	% of cases with at least one FN	% of cases with at least one FP or FN	False Negative Rate	False Negative Rate Median	Median Dice	Mean Dice	Median HD 95 Volume	Mean HD 95 Volume	Median HD Average Volume	Mean HD Average Volume
GTV <0.1 cc (6 mm)	61.5%	58.5%	1.3+-1.1	1	75.2%	67.4%	78.7%	1.0+-1.2	1	0.67	0.62	0.99	1.04	0.30	0.37
GTV >0.1 cc	93.3%	92.3%	0.2+-0.5	0	15.5%	12.0%	23.6%	0.2+-0.6	0	0.83	0.81	0.98	1.16	0.17	0.27
Overall	89.6%	76.8%	1.1+-1.2	1	59.4%	22.0%	65.0%	0.4+-1.2	0	0.79	0.75	0.98	1.12	0.21	0.31

Table 1. GTV detection performance stratified by lesion volume — sensitivity, specificity, false-positive and false-negative rates, and Dice and HD95 agreement against clinical contours. Dice and HD95 metrics are for non-IMRT, non-cavity GTVs <50 cc with Dice > 0.

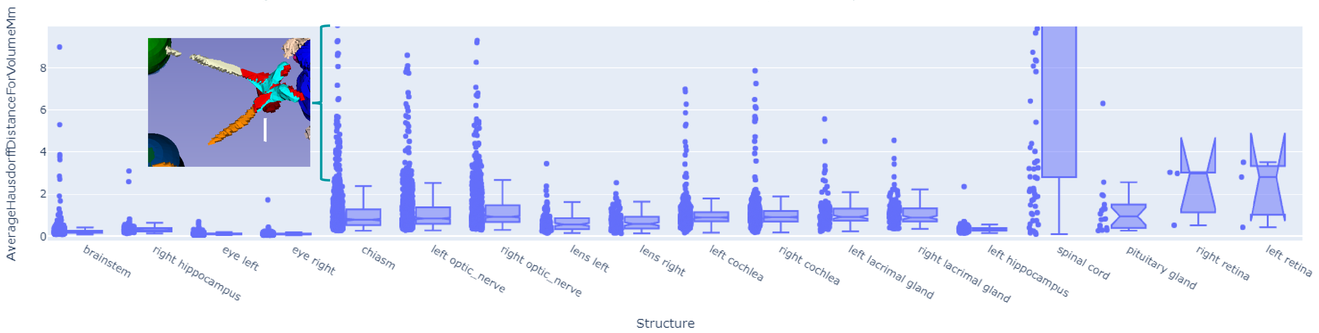

Figure 4. Average Hausdorff distance between AI and clinical contours, by structure. Spinal cord results are unreliable due to the MRI field-of-view frequently cutting off compared to standard CT. The image insert displays a representative case of disconnected optics, leading to the outliers observed in the box plot.

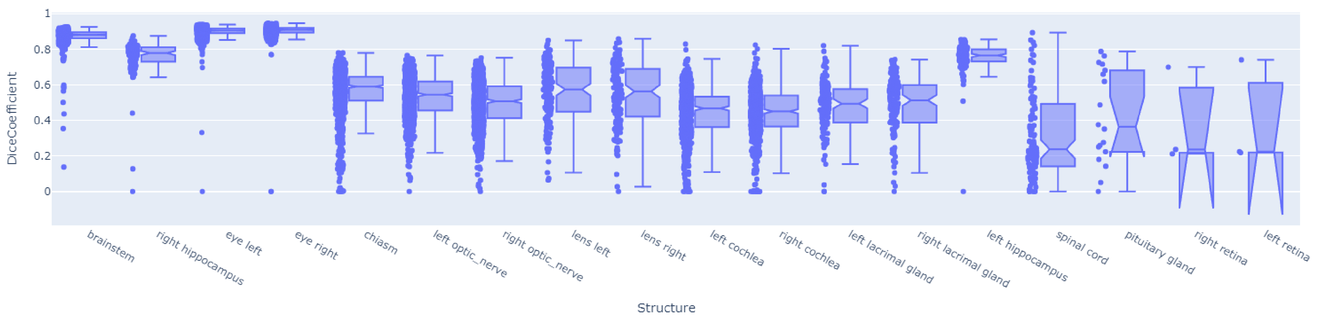

Figure 5. Dice overlap of AI vs. clinical OAR contours; high for brainstem, eyes, and hippocampus, lower for optic nerves and chiasm.

CONCLUSIONS

- The commercial AI auto-contouring tool demonstrates high sensitivity for detecting brain metastases larger than 0.1 cc and provides accurate, reliable segmentation for major intracranial OARs.
- Because the clinical ground truth was not retrospectively curated, the reported performance metrics (especially for lesions <0.1 cc) are likely an underestimate of the tool's true accuracy.
- Establishing the sensitivity and specificity profiles for small metastases proves this AI tool can effectively serve as a rapid QA safety net, reducing the risk of missed lesions during accelerated planning.

Next Steps: Based on these validation results, the tool is currently being implemented and evaluated as a prospective QA measure in a novel same-day, MRI-only stereotactic radiosurgery clinical trial (see poster SU-315-GPD-T-107 for details).