Clinical Experience with AI-Assisted Gross Tumor Volume Contouring In the Setting of Stereotactic Radiosurgery for Brain Metastases

Dylan Hsu PhD[1], Luke R.G. Pike, MD DPhil[2], Brandon Imber, MD[2], James Janopaul-Naylor, MD[2], Melissa Zinovoy, MD[2], Jean M. Moran, PhD[1], Ase M. Ballangrud, PhD[1] and Michalis Aristophanous, PhD[1]

[1]Department of Medical Physics, Memorial Sloan Kettering Cancer Center, New York, NY

[2] Department of Radiation Oncology, Memorial Sloan Kettering Cancer Center, New York, NY

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

INTRODUCTION

Stereotactic radiosurgery (SRS) is increasingly used to treat larger numbers of brain metastases (BMs) in a single course, improving quality of life and cognitive outcomes relative to whole-brain radiotherapy while raising the burden of lesion identification and contouring.

Although many AI algorithms can detect and segment BMs, clinical adoption has been limited by concerns about safety and explainability. Safe introduction of novel clinical software benefits from structured commissioning and risk analysis, including failure mode and effects analysis (FMEA) as recommended by AAPM Task Group 100.

Here we describe prospective clinical implementation of AI-assisted gross tumor volumes (AIGTVs) for BM SRS, the risk-mitigation workflow used to support physician review, and results from the initial clinical evaluation period.

METHODS AND MATERIALS

Patients. We included 129 patients simulated for BM SRS between October 2024 and January 2026 (141 treatment courses; 11 patients retreated). Median age was 60 years; primary sites were predominantly lung, breast, and melanoma. Approximately half had prior brain radiotherapy. A median of 5 lesions were treated per course (range 1–50).

AI model. An existing 3D V-Net deep learning model was applied to skull-stripped, co-registered CT and T1 post-contrast MR images to generate AIGTV structure sets, delivered through MIM Maestro for contour review before Eclipse/ARIA VMAT SRS planning.

Safety and workflow. Commissioning included an FMEA that identified highest-risk failure modes: unreviewed or inadequately edited AIGTVs and missed lesions due to overreliance on AI. Mitigation included a mandatory MIM "checker" workflow requiring classification of each AIGTV (false positive; true positive–treat; true positive–do not treat), reminders to add false negatives, automated QA reports, and physician training/credentialing. Final treatment volumes were always edited and approved by an attending radiation oncologist.

Analysis. Each AIGTV was compared with the final GTV approved by the MD to quantify true/false positives and false negatives. Contour quality for treated true positives was assessed with Dice coefficient, Hausdorff distance, and edited volume as a percent of GTV volume.

Patient, treatment and history of disease for the patients included in this study.

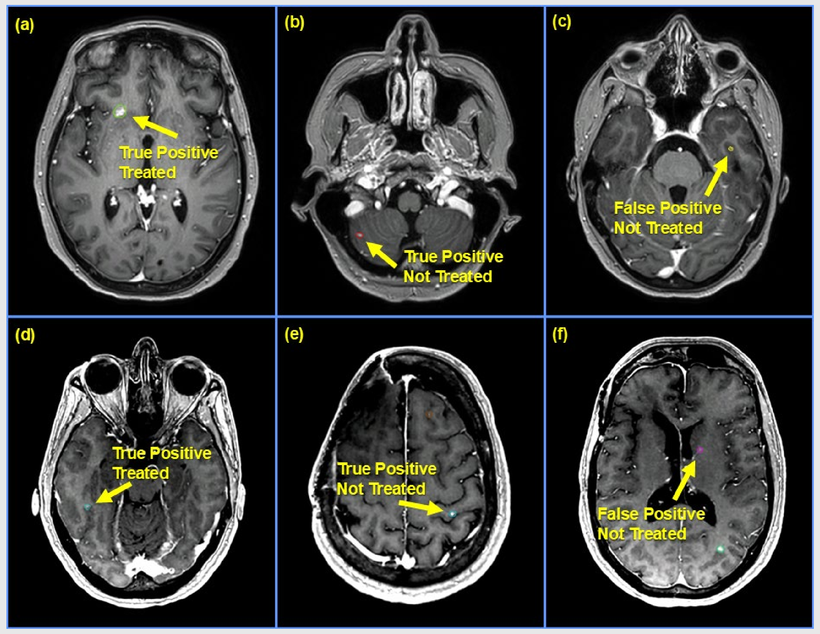
Demographics		Primary Site		Treatment History	
Age	60 (21–90)	Lung	40 (31%)	Prior RT (any)	100 (71%)
Sex		Breast	34 (26%)	Prior brain RT	69 (49%)
Female	76 (59%)	Melanoma	16 (12%)		
Male	53 (41%)	RCC	12 (9%)	**Lesions Treated**	
		GI	8 (6%)	Median (range)	5 (1–50)
		GU	5 (4%)	1–2	40 (28%)
		GYN	4 (3%)	3–5	33 (23%)
		Other	10 (8%)	6–10	26 (18%)
				11–20	29 (21%)
				21+	13 (9%)

RESULTS

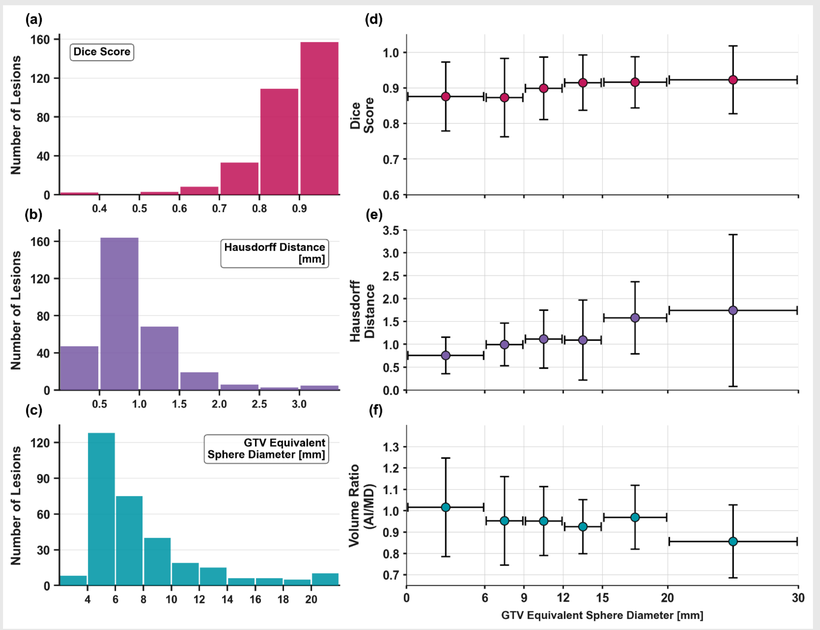
During clinical testing, **no adverse events related to AIGTVs** were observed, and no healthy structures (e.g., vessels) were approved as treatment targets.

At the default 70% confidence threshold, **1739 AIGTVs** were reviewed: **1418 true positives**, **321 false positives** (2.3/case), and **137 false negatives** (1.0/case), corresponding to sensitivity **91.2%** (95% CI 89.7–92.6) and PPV **81.5%** (79.6–83.3). At a 90% confidence threshold, sensitivity was 78.7%, PPV 93.5%, and false positives fell to 0.60/case.

Patients were treated to a median of **5 BMs** (IQR 2–11). Of 1206 treated BMs, 1064 (**88.2%**) had been detected as AIGTVs. Treated true-positive AIGTVs had median Dice **0.92** (IQR 0.85–0.97), median Hausdorff distance **1.2 mm** (IQR 0.8–1.6), and median edit volume **17%** of the final GTV (IQR 6–30). The median AI-to-physician volume ratio was **0.99**.



Examples of true positive AI GTVs that were and were not treated, as well as false positive AI GTVs. In the top row, (a)-(c) are from one patient case; in the bottom row, (d)-(f) are from another patient case. (e) is an example of a previously radiated lesion and (f) is an example of a T1-bright blood vessel.



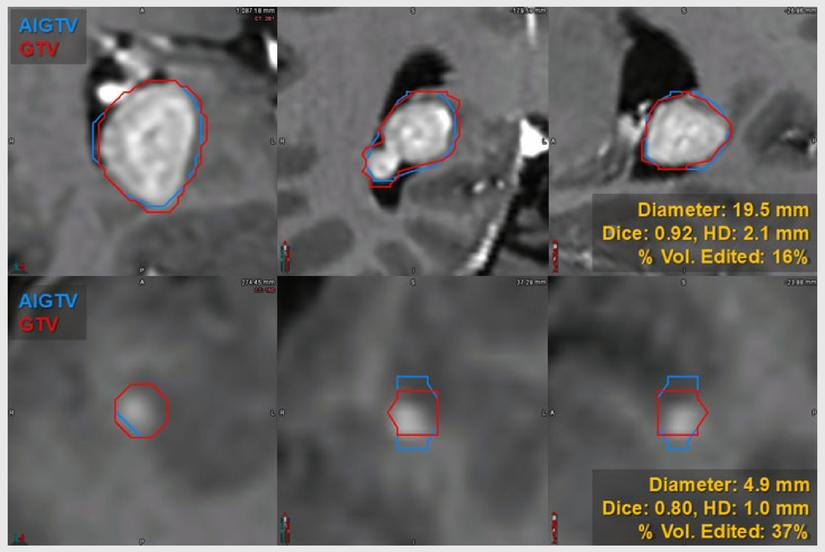
Segmentation quality metrics and size for AIGTVs that were modified and treated. (a)-(c) show distributions of Dice score, Hausdorff distance, and approved GTV size. (d)-(f) show the Dice score, Hausdorff distance, and the AI/MD volume ratio as a function of GTV size; vertical error bars are standard deviations.

DISCUSSION

This work represents one of the first prospective clinical applications of AI target-volume proposals for BM radiosurgery, with formal risk analysis and continuous process monitoring.

False positives can still add review effort (e.g., corroboration with radiology or secondary imaging). Some systematic under-contouring relative to final GTVs reflects institutional practice of including dura for dural-based or dura-adjacent lesions.

Remaining limitations include false positives from vascular and dural structures and incomplete sensitivity. Importantly, workflow safeguards prevented model weaknesses from reaching the patient.



AIGTVs before physician edits, overlaid with the final approved GTV volumes after physician edits, along with information on MD and AI contour agreement.

CONCLUSIONS

- AI-based BM detection and segmentation **was safely introduced** into prospective SRS planning to assist radiation oncologists.
- Performance was clinically useful with **high sensitivity in lesion detection**, **high dice** coefficient and **low Hausdorff distance** between AI and MD contours, with modest edits.
- An FMEA-informed checker workflow mitigated key risks; **no AIGTV-related adverse events** occurred.
- Future work will improve ease of use and BM model performance (fewer false positives/negatives).

REFERENCES

1. 1. Huq MS, Fraass BA, Dunscombe PB, et al. The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management. **Med Phys.** 2016;43(7):4209-4262.
2. 2. Moran JM, Paradis KC, Hadley SW, et al. A Safe and Practical Cycle for Team-Based Development and Implementation of In-House Clinical Software. **Adv Radiat Oncol.** 2021;7(1):100768.
3. 3. Hsu DG, Ballangrud Å, Shamseddine A, et al. Automatic segmentation of brain metastases using T1 magnetic resonance and computed tomography images. **Phys Med Biol.** 2021;66(17):175014.

CONTACT

Dylan Hsu, PhD
Memorial Sloan Kettering Cancer Center
Email: hsud3@mskcc.org