Eugene M. & Christine E. Lynn Cancer Institute
BAPTIST HEALTH SOUTH FLORIDA

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

Expert Agreement on Deep-Learning Auto-Segmentation for Head and Neck Organs at Risk

Nishan Shrestha[1], Shyam Pokharel[1], Noufal Manthala Padannayil[1], Rachel Jacques[1], Zachariah Khan[1], Nebi Demez[1], Ranjith Cholakkara Poyil[1], Suresh Rana[1,2]

[1]Radiation Oncology, Lynn Cancer Institute, Boca Raton Regional Hospital, Boca Raton, FL, United States of America
[2]Oncological Sciences, Florida International University, Miami, FL, United States of America

Objective

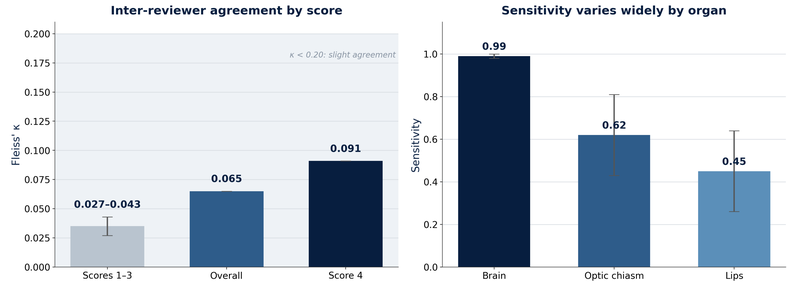
Deep-learning (DL) auto-segmentation may improve efficiency in head-and-neck (HN) planning, but clinical use depends on how consistently experts judge contour acceptability and how performance varies by organ. We evaluated a DL-based auto-segmentation system for HN organs-at-risk (OARs), emphasizing expert agreement and complementary quantitative metrics.

Methods

- ❑ The DL system (Limbus version 1.8) generated 34 OARs per case for fourteen patients.
- ❑ Seven HN-experienced reviewers scored each OAR on a 1–4 scale based on the estimated fraction of axial slices requiring correction: 4 (<3%), 3 (3–10%), 2 (10–40%), 1 (>40%).
- ❑ Inter-reviewer agreement was assessed using Fleiss' kappa (overall and score-specific). Quantitative agreement with clinical contours was characterized using Dice similarity coefficient (DSC) and sensitivity.
- ❑ The relationship between volume difference (Δvol) and DSC was examined to interpret cases where overlap metrics ma misleading.

Results

Across 3332 scores (14 patients × 34 OARs × 7 reviewers), inter-reviewer agreement was low but non-random (Fleiss' κ = 0.065, p < 0.0001) and highest for clearly acceptable contours (Score 4, κ = 0.091). Scoring tendencies varied widely between reviewers (Score-4 counts: 436 vs 234). Geometric accuracy was strongly organ-dependent (mean DSC 0.22–0.98; sensitivity 0.45–0.99), and volume difference (Δvol) did not track DSC. See the figures below for score-level agreement, organ sensitivity, and the Δvol–DSC comparison.



Organ-level performance: DSC and Δvol are not monotonic

Organ	DSC	Sensitivity	Δvol (cc)
Brain	0.98 ± 0.01	0.99 ± 0.01	−33.9 ± 13.7
Optic chiasm	0.51 ± 0.16	0.62 ± 0.19	−0.2 ± 0.4
Lips	—	0.45 ± 0.19	—

Brain: large volume difference yet high DSC. Optic chiasm: near-zero Δvol but only moderate DSC — DSC alone can mislead for small structures.

Key findings

- Inter-reviewer agreement was low but non-random (Fleiss' κ = 0.065, p < 0.0001) and best for clearly acceptable contours (Score 4, κ = 0.091).
- Reviewers differed sharply in scoring tendency (Score-4 counts of 436 vs 234), supporting the need for standardized scoring guidance.
- Geometric accuracy was strongly organ-dependent: mean DSC 0.22–0.98, and sensitivity from 0.45 (lips) to 0.99 (brain).
- Δvol and DSC were not monotonic — small structures can show tiny volume error yet only moderate overlap, so DSC alone can mislead.
- DL contours are useful starting points but must be expert-reviewed before treatment-plan optimization.

Conclusion

- ❑ DL auto-segmentation can provide useful starting contours for HN OARs, but expert agreement on acceptability is limited and organ-dependent.
- ❑ Reliability measures, sensitivity, and Δvol–DSC behavior add clinically relevant context beyond DSC alone.
- ❑ The observed reviewer-to-reviewer variability supports standardized scoring guidance, and DL-generated OARs should be reviewed before treatment plan optimization.

Corresponding Author

Suresh Rana, PhD; AVP and Chief of Medical Physics
- *Email: sureshr@baptisthealth.net or suresh.rana@gmailcom*
- *Phone: 1-405-795-6697*