

## ABSTRACT

### Purpose:

Two commercial AI auto-contouring tools, Siemens DirectORGANS (DO) and Radformation AutoContour (RAD) are compared for their structure-specific accuracy and clinical efficiency.

### Methods:

Twenty-two adult head-and-neck cancer patients who underwent CT simulation in 2025 were retrospectively analyzed. Sixteen structures were contoured by both AI tools and a board-certified radiation oncologist. Contouring accuracy was evaluated using both quantitative and qualitative metrics. Statistical significance was assessed using the Wilcoxon signed-rank test ( $p < 0.05$ ).

### Results:

AI contouring performance was strongly structure-dependent. Both tools demonstrated high performance for well-defined structures, including the brain, spinal cord, eyes, lenses, and salivary glands, while lower accuracy was observed in anatomically complex regions such as the lips and oral cavity. Overall, qualitative Likert scores were consistent with quantitative metrics, with comparable agreement across DSC, HD, and MDA in assessing relative performance between the two systems.

### Conclusion:

AI auto-contouring performance in head-and-neck radiotherapy is highly structure-dependent. While both systems perform well for select well-defined structures, challenging anatomy continues to require careful clinician review. Understanding these performance patterns may help optimize clinical workflow by prioritizing manual review for complex structures while leveraging AI for more reliable regions.

## CONTACT

Yuan Shao, PhD  
 Rush University Medical Center  
 500 South Paulina St., Chicago, IL 60612  
[Yuan\\_Shao@rush.edu](mailto:Yuan_Shao@rush.edu)

# Clinical Evaluation of Two AI Auto-Contouring Solutions for Head and Neck Radiotherapy

Yuan Shao, PhD; Yixiang Liao, PhD; Nikhil Joshi, MD; Julius Turian, PhD  
 Department of Radiation Oncology, Rush University Medical Center

## INTRODUCTION

Artificial intelligence (AI) has been rapidly integrated into radiation oncology to improve clinical efficiency by reducing the time required for labor-intensive tasks such as organ-at-risk (OAR) contouring. However, AI-generated contours may require manual review and editing before they can be used clinically.

This clinical study aimed to evaluate the performance of two AI-powered contouring tools using both qualitative and quantitative metrics to identify structures that could be reliably contoured by AI and those that remained challenging.

## METHODS AND MATERIALS

The two AI contouring tools evaluated were DirectORGANS (DO) (VB40, Siemens Healthineers, Erlangen, Germany) and AutoContour (RAD) (v2.5.24, Radformation Inc., New York, NY, USA).

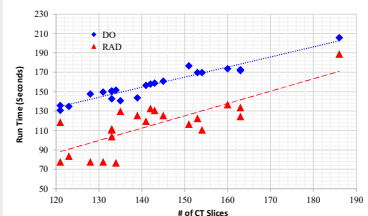
Twenty-two adult head-and-neck cancer patients who underwent CT simulation (Siemens SOMATOM go.Sim) in 2025 were retrospectively analyzed. As summarized in **Table 1**, sixteen structures were contoured by both AI tools and a board-certified radiation oncologist. The sample size varied by structure, ranging from 6 to 22.

Contour accuracy was evaluated qualitatively by a board-certified radiation oncologist using a five-point Likert scale. Quantitative metrics included Dice similarity coefficient (DSC), Hausdorff distance (HD), and mean distance to agreement (MDA), comparing AI-generated contours to manual reference ("ground truth") contours. Statistical significance was assessed using the Wilcoxon signed-rank test ( $p < 0.05$ ).

**Table 1.** Structures Supported by Each Tool and Corresponding Sample Sizes

|    | TG263 H&N Structures | DO | RAD | Sample Size |
|----|----------------------|----|-----|-------------|
| 1  | A_Carotid_L          | X  | X   |             |
| 2  | A_Carotid_R          | X  | X   |             |
| 3  | Bone_Mandible        | X  | X   | 21          |
| 4  | BrachialPlex_L       | X  | X   |             |
| 5  | BrachialPlex_R       | X  | X   |             |
| 6  | Brain                | X  | X   | 6           |
| 7  | Brainstem            | X  | X   | 22          |
| 8  | Cavity_Oral          | X  | X   | 21          |
| 9  | Cochlea_L            | X  | X   |             |
| 10 | Cochlea_R            | X  | X   |             |
| 11 | Esophagus            | X  | X   | 17          |
| 12 | Eye_L                | X  | X   | 15          |
| 13 | Eye_R                | X  | X   | 16          |
| 14 | Gland_Lacrimal_L     | X  | X   |             |
| 15 | Gland_Lacrimal_R     | X  | X   |             |
| 16 | Gland_Submand_L      | X  | X   | 17          |
| 17 | Gland_Submand_R      | X  | X   | 13          |
| 18 | Gland_Thyroid        | X  | X   |             |
| 19 | Hypopharynx_L        | X  | X   |             |
| 20 | Hypopharynx_R        | X  | X   |             |
| 21 | Joint_TM_L           | X  | X   |             |
| 22 | Joint_TM_R           | X  | X   | 19          |
| 23 | Larynx               | X  | X   |             |
| 24 | Larynx_SG            | X  | X   |             |
| 25 | Lens_L               | X  | X   | 10          |
| 26 | Lens_R               | X  | X   | 12          |
| 27 | Lips                 | X  | X   | 17          |
| 28 | Lobe_Temporal_L      | X  | X   |             |
| 29 | Lobe_Temporal_R      | X  | X   |             |
| 30 | Mus_Constrict        | X  | X   |             |
| 31 | OpticChiasm          | X  | X   |             |
| 32 | OpticNrv_L           | X  | X   |             |
| 33 | OpticNrv_R           | X  | X   |             |
| 34 | Parotid_L            | X  | X   | 22          |
| 35 | Parotid_R            | X  | X   | 20          |
| 36 | Pituitary            | X  | X   |             |
| 37 | Retina_L             | X  | X   |             |
| 38 | SpinalCord           | X  | X   | 22          |

As shown in **Figure 1**, total processing time for both tools scaled linearly with CT scan volume. RAD demonstrated shorter raw processing times (77-189 s) compared with DO (131-206 s). However, additional data transfer time was required for RAD, whereas DO generated contours directly during CT simulation.



**Figure 1.** Comparison of AI Contour Generation Time Versus CT Scan Volume

As shown in **Table 2**, both AI tools achieved a mean Likert score  $\geq 4$  for the esophagus, right eye, bilateral submandibular glands, and bilateral lenses.

Compared with DO, RAD achieved a mean Likert score  $\geq 4$  for five additional structures: the mandible, brain, brainstem, larynx, and spinal cord. The only structures for which DO outperformed RAD were the bilateral parotid glands.

Both tools struggled to accurately contour the lips, which received the lowest mean Likert scores (RAD: 2.65; DO: 2.00). The oral cavity also received relatively low mean Likert scores from both tools (RAD: 3.00; DO: 3.42).

**Table 2.** Structures Supported by Each Tool and Corresponding Sample Sizes

| OAR             | Grp | N  | HD (mm) <sup>2</sup> | MDA (mm) <sup>2</sup> | DSC (%) <sup>2</sup> | 5-point Likert    |
|-----------------|-----|----|----------------------|-----------------------|----------------------|-------------------|
|                 |     |    | Mean (Range)         | Mean (Range)          | Mean (Range)         | Mean (Range)      |
| Bone_Mandible   | RAD | 21 | 23.1 (6.3–237.2)     | 1.2 (0.5–7.4)         | 85% (51%–91%)        | <b>4.05 (3–5)</b> |
|                 | DO  | 21 | 30.2 (3.6–237.4)     | 1.8 (0.4–8.3)         | 80% (42%–92%)        | 2.43 (1–5)        |
| Brain           | RAD | 6  | 11.0 (8.3–13.5)      | 0.7 (0.6–0.8)         | 98% (97%–98%)        | <b>4.83 (4–5)</b> |
|                 | DO  | 6  | 23.0 (15.0–28.3)     | 0.8 (0.6–1.2)         | 97% (96%–98%)        | 3.67 (3–5)        |
| Brainstem       | RAD | 22 | 12.4 (8.0–30.7)      | 2.2 (1.3–9.2)         | 76% (24%–84%)        | <b>4.45 (1–5)</b> |
|                 | DO  | 22 | 12.5 (7.7–18.0)      | 2.4 (1.6–5.5)         | 73% (36%–81%)        | 3.64 (1–5)        |
| Cavity_Oral     | RAD | 21 | 24.7 (13.9–34.5)     | 5.3 (3.1–7.5)         | 70% (62%–80%)        | 3.00 (2–5)        |
|                 | DO  | 21 | 20.1 (13.1–33.0)     | 2.8 (1.9–4.9)         | 83% (69%–88%)        | 3.43 (2–5)        |
| Esophagus       | RAD | 17 | 102.6 (58.7–160.4)   | 25.4 (9.8–54.4)       | 44% (30%–62%)        | <b>4.29 (1–5)</b> |
|                 | DO  | 17 | 102.5 (59.2–152.3)   | 23.8 (10.2–46.8)      | 45% (31%–60%)        | <b>4.65 (1–5)</b> |
| Eye_L           | RAD | 15 | 7.1 (3.4–31.4)       | 2.4 (1.1–37.5)        | 75% (0%–83%)         | 3.86 (1–5)        |
|                 | DO  | 15 | 5.5 (3.6–8.6)        | 1.3 (1.0–2.1)         | 81% (67%–85%)        | 3.87 (2–5)        |
| Eye_R           | RAD | 16 | 5.4 (3.9–6.8)        | 1.3 (0.9–1.8)         | 81% (72%–86%)        | <b>4.00 (3–5)</b> |
|                 | DO  | 16 | 5.3 (3.1–8.1)        | 1.0 (0.7–1.4)         | 85% (79%–90%)        | <b>4.13 (3–5)</b> |
| Gland_Submand_L | RAD | 17 | 7.2 (3.0–24.2)       | 1.3 (0.6–5.0)         | 80% (39%–89%)        | <b>4.65 (2–5)</b> |
|                 | DO  | 17 | 6.1 (3.3–17.9)       | 1.1 (0.5–4.9)         | 82% (39%–92%)        | <b>4.71 (2–5)</b> |
| Gland_Submand_R | RAD | 13 | 7.2 (3.1–15.3)       | 1.2 (0.6–3.3)         | 80% (43%–90%)        | <b>4.62 (2–5)</b> |
|                 | DO  | 13 | 8.3 (3.5–24.7)       | 1.5 (0.6–4.4)         | 77% (31%–89%)        | <b>4.15 (1–5)</b> |
| Larynx          | RAD | 19 | 15.6 (6.4–24.3)      | 3.8 (1.8–21.2)        | 65% (44%–71%)        | <b>4.00 (2–5)</b> |
|                 | DO  | 19 | 34.9 (13.5–51.0)     | 9.3 (3.6–14.0)        | 51% (10%–39%)        | 2.68 (2–4)        |
| Lens_L          | RAD | 10 | 2.4 (1.4–3.2)        | 0.5 (0.4–0.7)         | 71% (62%–80%)        | <b>4.50 (3–5)</b> |
|                 | DO  | 10 | 2.9 (1.5–3.8)        | 0.7 (0.3–1.1)         | 64% (40%–82%)        | <b>4.00 (2–5)</b> |
| Lens_R          | RAD | 12 | 2.6 (1.3–3.5)        | 0.5 (0.4–0.8)         | 72% (60%–81%)        | <b>4.92 (4–5)</b> |
|                 | DO  | 12 | 2.9 (1.4–3.8)        | 0.6 (0.3–1.0)         | 66% (47%–81%)        | <b>4.25 (3–5)</b> |
| Lips            | RAD | 17 | 21.0 (10.5–54.9)     | 4.7 (2.0–20.2)        | 36% (10%–55%)        | 2.65 (1–5)        |
|                 | DO  | 17 | 21.9 (15.2–35.8)     | 4.7 (2.9–8.0)         | 36% (16%–47%)        | 2.00 (1–3)        |
| Parotid_L       | RAD | 22 | 15.2 (6.0–82.2)      | 3.1 (1.4–48.7)        | 77% (0%–97%)         | 3.85 (1–5)        |
|                 | DO  | 22 | 17.2 (6.1–44.0)      | 2.4 (0.8–12.8)        | 77% (13%–91%)        | <b>4.27 (1–5)</b> |
| Parotid_R       | RAD | 20 | 13.5 (6.6–25.1)      | 1.7 (1.2–2.2)         | 80% (72%–86%)        | 3.99 (2–5)        |
|                 | DO  | 20 | 16.5 (4.6–39.6)      | 1.9 (0.7–5.3)         | 81% (67%–91%)        | <b>4.60 (2–5)</b> |
| SpinalCord      | RAD | 22 | 8.9 (3.3–72.9)       | 0.9 (0.5–5.5)         | 86% (79%–93%)        | <b>5.00 (5–5)</b> |
|                 | DO  | 22 | 8.4 (4.8–63.4)       | 2.4 (1.5–7.2)         | 52% (43%–67%)        | 2.32 (2–4)        |

NOTE:

- Hausdorff Distance (HD): Maximum distance between the edges of two contours; smaller values indicate closer alignment.
- Mean Distance to Agreement (MDA): Average distance between the edges of two contours; smaller values indicate better overall agreement.
- Dice Similarity Coefficient (DSC): Ranges from 0 (no overlap) to 1 (perfect overlap) between two contours.
- 5-point Likert Scale: 5: Clinically acceptable. Could be used for treatment without change; 4: Minor edits that are not necessary. Not clinically important; 3: Minor edits that are necessary. Expected to have minimal effect on treatment outcome; 2: Major edits. Edits required to ensure appropriate treatment; 1: Unusable.

## RESULTS

Overall, conclusions derived from subjective Likert ratings were largely consistent with those based on the three quantitative metrics (DSC, HD, and MDA). In general, structures with higher Likert scores tended to demonstrate higher DSC values and lower HD and MDA values, indicating better geometric agreement with the reference contours.

As shown in **Table 3**, to evaluate agreement between quantitative metrics and Likert scores in assessing relative AI performance, the AI tool with the higher Likert score was designated as the "Likert winner" for each structure. The AI tool identified as better-performing by DSC and MDA agreed with the Likert winner in 14 of 16 structures (87.5%), whereas HD agreed in 13 of 16 structures (81.3%).

These findings suggest that DSC, HD, and MDA all show comparable agreement with subjective Likert ratings.

**Table 3.** Agreement of HD, MDA, and DSC with Likert-Based Ratings

| Structure       | Likert Winner (Reference) | HD Winner  | MDA Winner | DSC Winner |
|-----------------|---------------------------|------------|------------|------------|
| Bone_Mandible   | RAD                       | RAD        | RAD        | RAD        |
| Brain           | RAD                       | RAD        | RAD        | RAD        |
| Brainstem       | RAD                       | RAD        | RAD        | RAD        |
| Cavity_Oral     | DO                        | DO         | DO         | DO         |
| Esophagus       | DO                        | <b>RAD</b> | DO         | DO         |
| Eye_L           | DO                        | DO         | DO         | DO         |
| Eye_R           | DO                        | DO         | DO         | DO         |
| Gland_Submand_L | DO                        | DO         | DO         | DO         |
| Gland_Submand_R | RAD                       | RAD        | RAD        | RAD        |
| Larynx          | RAD                       | RAD        | RAD        | RAD        |
| Lens_L          | RAD                       | RAD        | RAD        | RAD        |
| Lens_R          | RAD                       | RAD        | Tie        | RAD        |
| Lips            | RAD                       | RAD        | Tie        | Tie        |
| Parotid_L       | DO                        | <b>RAD</b> | DO         | Tie        |
| Parotid_R       | DO                        | <b>RAD</b> | DO         | DO         |
| SpinalCord      | RAD                       | RAD        | RAD        | RAD        |

## CONCLUSIONS

In this retrospective study of 22 head-and-neck cancer patients, both AI contouring tools demonstrated generally high performance for well-defined structures, including the brain, brainstem, spinal cord, eyes, lenses, and salivary glands (parotids and submandibular glands), with lower accuracy observed in anatomically complex or low-contrast regions such as the lips and oral cavity.

Understanding the strengths and limitations of AI contouring tools across different anatomical structures can help guide clinical workflow prioritization.

## REFERENCES

- Goddard, Lee, et al. "Evaluation of multiple-vendor AI autocontouring solutions." *Radiation Oncology* 19.1 (2024): 69.
- Cant, Isabel, et al. "Comparison of Auto-Contouring Tools for Delineation of Normal Organs at Risk in Paediatric Patients Undergoing Radiotherapy." *Journal of Medical Radiation Sciences* 73 (2026): 562-572.
- Goddard, Lee, et al. "Evaluation of multiple-vendor AI autocontouring solutions." *Radiation Oncology* 19.1 (2024): 69.
- Ding, Serap Catti, et al. "The dosimetric impacts of ct-based deep learning autocontouring algorithm for prostate cancer radiotherapy planning dosimetric accuracy of DirectORGANS." *BMC urology* 25.1 (2025): 190.