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

Large Scale Analysis of Practice Patterns Emerging from an Automated Real-Time Evaluation of a Commercial Artificial Intelligence Auto-Contouring Solution

Riya Prashad[1], Piotr Dubrowski[1], Sara A Richter[1], Matthew Kim[2], Yong Yang[1] and Dimitre Hristov[1]

[1]Department of Radiation Oncology, Stanford University, Stanford, CA, [2]University of California, San Francisco, San Francisco, CA

Stanford MEDICINE
Health Care
Radiation Oncology

INTRODUCTION

AI segmentation has improved planning efficiency but requires evaluation in real time. We implemented a commercial AI auto-contouring tool across a multi-site academic network and developed an unsupervised evaluation framework to assess segmentation for every patient.

METHODS & MATERIALS

Evaluation Tool:
Autonomously compares clinically approved and initial AI structures geometrically and dosimetrically by evaluating the dose distribution on both structure sets (**Figure 1**).

Dataset:
Data collection between July 2024 and September 2025 provided 2000+ patient plans with 40,000+ structures.

- After QC: 36,332 structures analyzed among 2018 patients

Quality Check (QC):
Removed test records (~600), duplicated records (~1500), records with mislabeled structures or missing metrics (~60), and plans with partial scans (~300). Contours with absolute Relative Volume Difference (RVD) greater than 50% or large % differences in Dice/volume were also excluded (~1000).

Evaluation Metrics:

- Modified
  - Dice Score (DSC) or absolute Relative Volume Difference (absRVD) were outside of 3 standard deviations of their respective reference variations.
- Optimized
  - An optimization objective existed for the structure within the treatment plan.
- 15 other geometrical and dosimetric evaluation metrics

Software Upgrade:
Auto-contouring software was updated in March 2025, and our dataset was divided at that timepoint to conduct a comparative analysis.

- 51% of records were pre-upgrade, and 49% post-upgrade.

Breast Subset:

- n=7000+ records; 472 patients
- Included regional lymph nodes (LNs) and planning target contours (PTCs).
- Relative Mean Clinical Dose (rMCD) > 80% used to infer planning intent for PTCs.

Physician Groups:

- Total of 29 providers included (2 joined post-upgrade)
- 2 breast specialists contributed 67% of breast records

CONTACT

Piotr Dubrowski
piotr.dubrowski@stanford.edu
Department of Radiation Oncology
Stanford University

RESULTS

Physician Adoption and Trends

- Nearly all physicians adopted use of the auto-segmentation software after clinical implementation (**Figure 2)**
  - 25/29 physicians used it for >65% cases, 7 physicians for >90% cases
- Modification rates and Dice scores were **independent** of patient volume
- Modification rates were **higher** in physicians with **lower overall adoption** (18% vs 12%, P < 0.0001) suggesting increased **scrutiny** by reluctant users

Overall Editing

- **Modifications** were performed in **15%** of structures processed
- **Median Dice Score (DSC)** ranged from a high of **0.989** for the Bag_Bowel structure to a low of **0.7-0.65** for Cochlea and Lens
- Auto segmentation is template based:
  - **42%** of structures exhibited **significantly different** editing across site **templates**
  - **For example:** Left and right Brachial Plexus structures were edited ~40% in the Head and Neck template, but only ~19% in the Thorax template and barely edited in the Breast template (p<0.0001)

Dosimetric Impact and Optimization Summaries

- **Figure 3** illustrates the relationship of mean clinical dose to a structure with Dice scores and modification rates. As **dose** to a structure **increases, DSC decreases.** As dose to a structure increases, **modification rates increase**. (p < 0.0001)
- Additionally, modification rates also increased when they were included in the plan's **optimization objectives** (**Figure 4)**

Software Upgrade

- After an upgrade, median Dice Score **improved** from 0.881 to 0.902 (p<0.0001) while modification rates increased slightly
- **Figure 5** highlights structures with the **largest positive and negative differences** in modification rates after the upgrade

Breast Subset

- PTCs (structures used to guide forward planning) were **edited more often** than non-PTCs (35% vs 7%)
- Most PTC edits **reduced contour volume** on average
- Site specialists edited PTCs **more often** (42% vs 24%) and **smaller structures** on average (~200c vs ~600 cc)

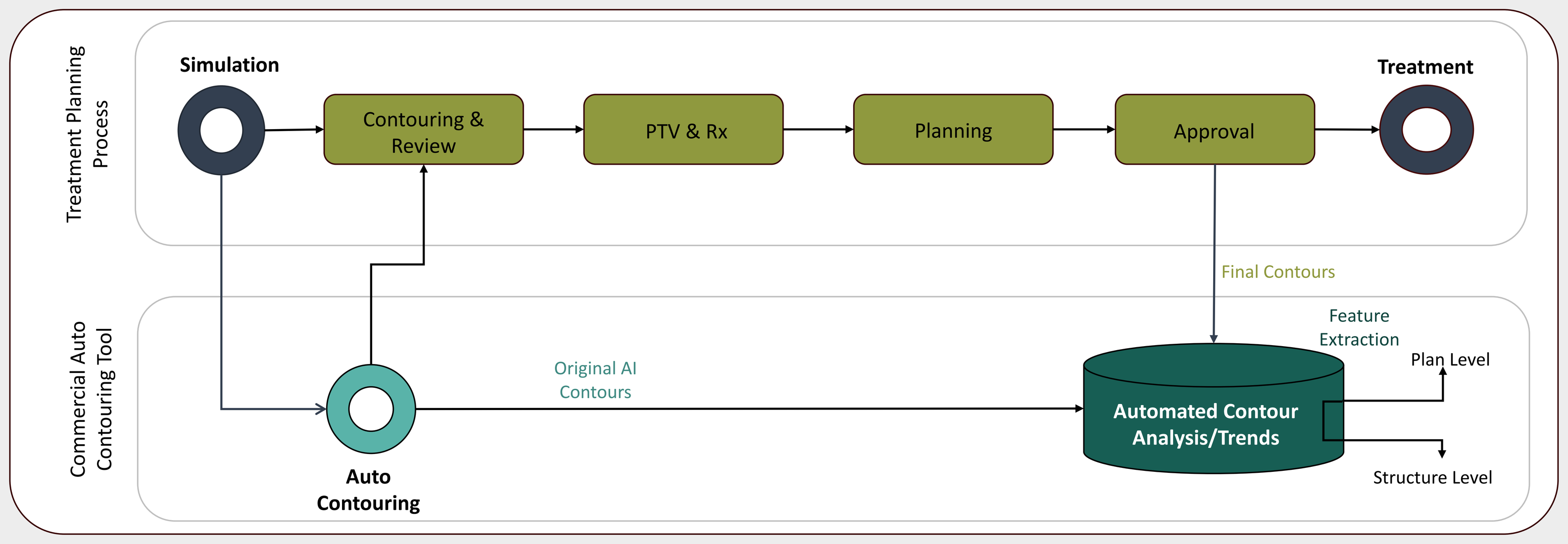

Figure 1. Overview of the treatment planning with AI auto contour workflow and incorporation of our evaluation tool. A copy of the initial AI-generated auto contours is compared to the final physician-approved contours to produce comparisons at the plan and structure levels. Features are extracted and analyzed utilizing geometric and dosimetric evaluations.

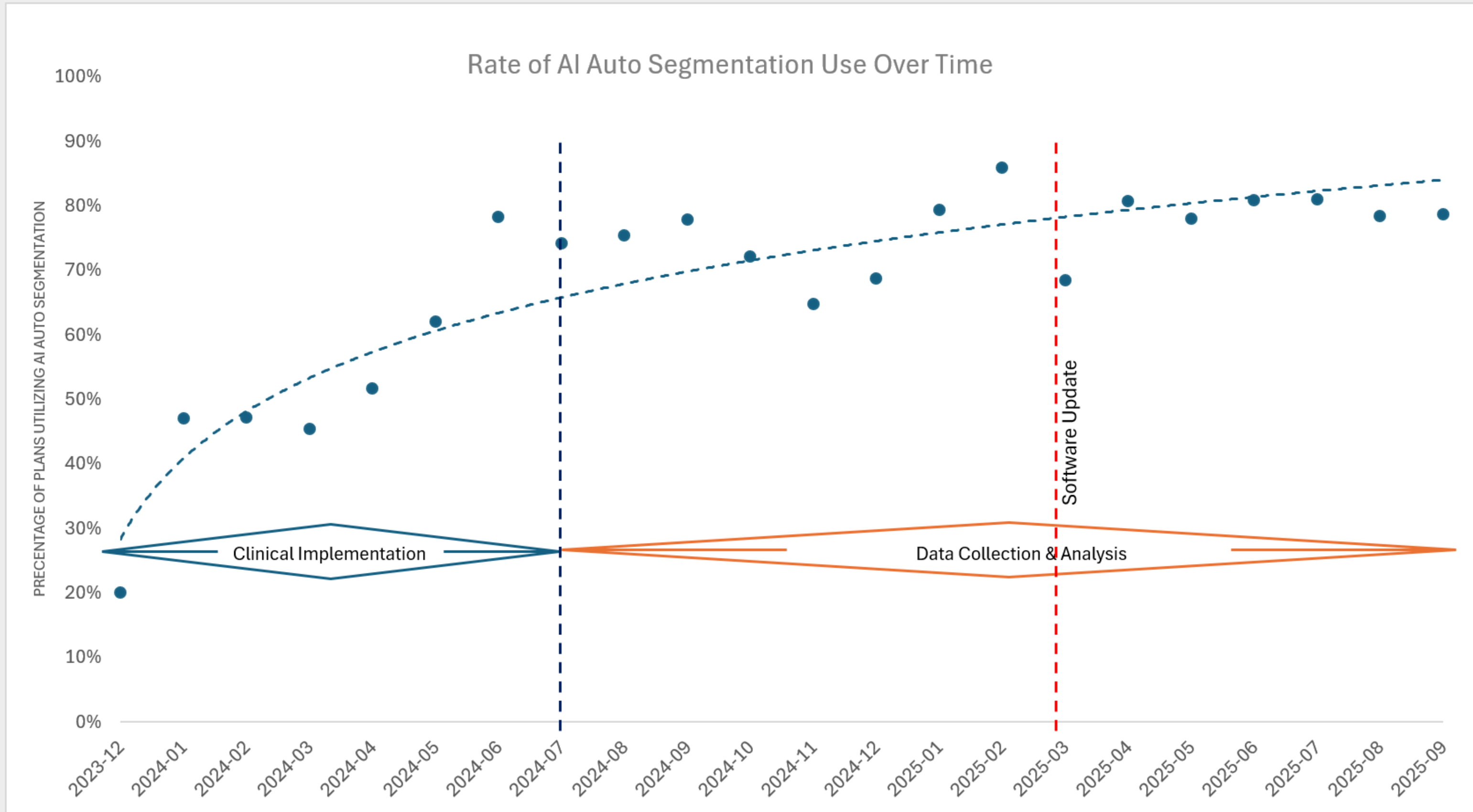

Figure 2. Physician adoption rates. Proportion of physicians' usage of commercial segmentation software in overall case volume.

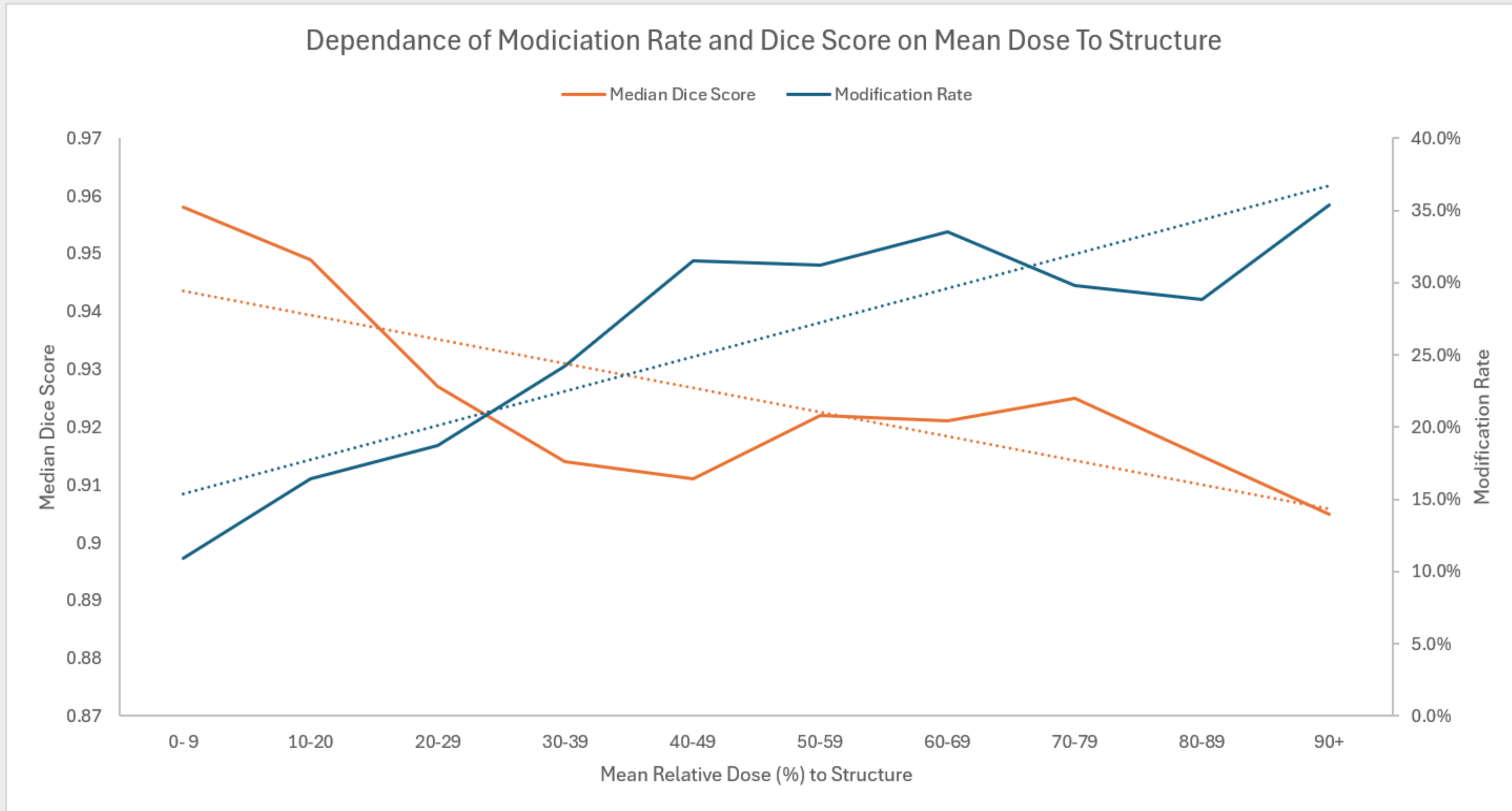

Figure 3. (left) DSC and modification rate by mean relative dose bins. As mean dose increases, median DSC decrease and modification rates increase significantly (p<0.0001).

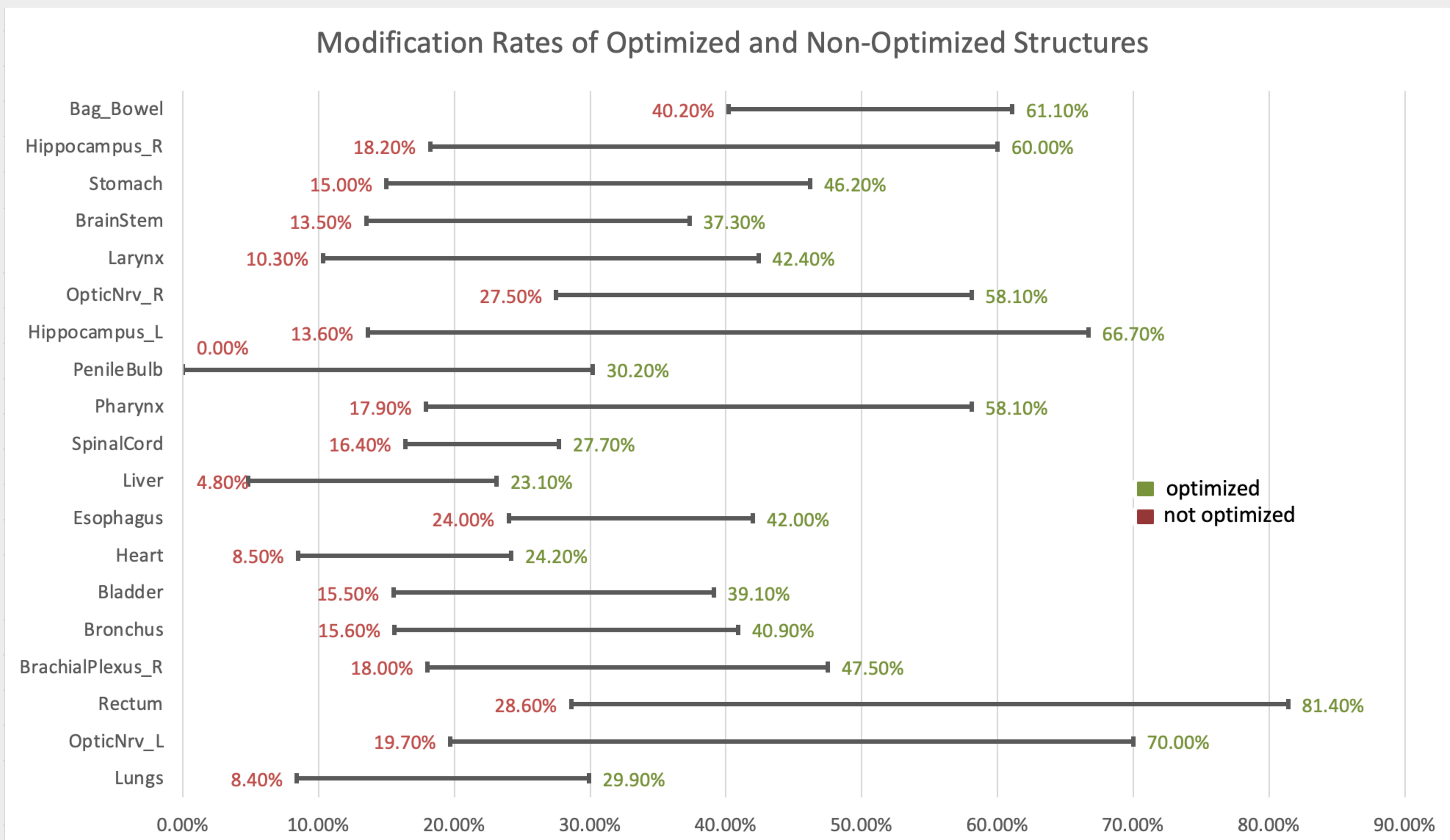

Figure 4.(left) Modification rates of optimized vs non-optimized structures. Comparison of significantly different modification rates when structures were optimized versus not optimized.

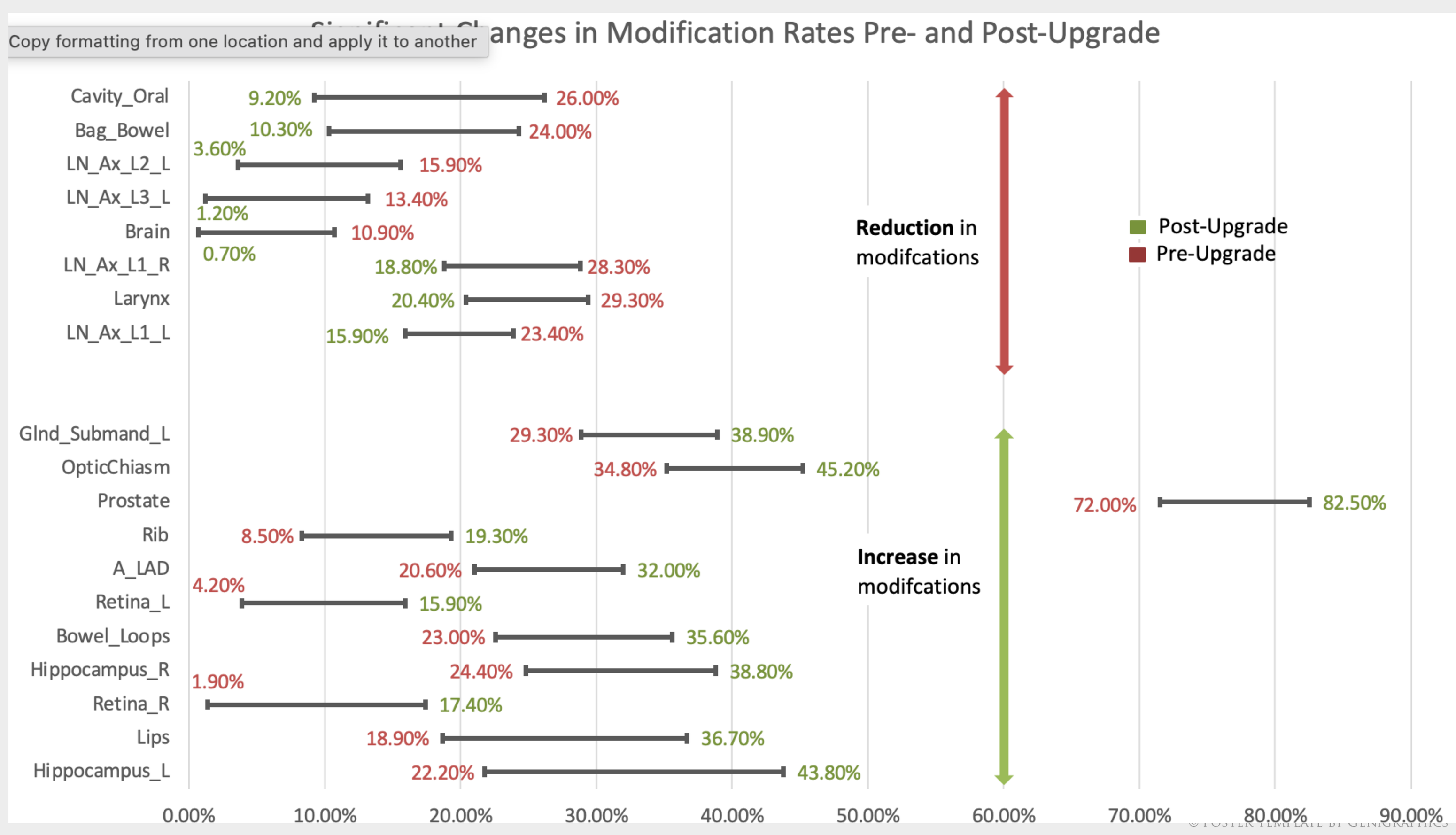

Figure 5. (left) Significant changes in modification rates after upgrade. Largest positive and negative changes in modifications per structure are represented.

DISCUSSION

- AI auto segmentation deployed in a multi-site clinic was **widely accepted** and utilized by the majority of providers
- Physician modifications **increased with lower use** suggesting increased scrutiny.
- **DCS and modification rates** were highly correlated with mean dose to a structure and its use in inverse **optimization** suggesting increased attention rather than purely anatomical basis
- Higher DSC post-upgrade indicates better **overall agreement** in the newer models despite small overall increase in editing rate.
- Site **specialists** seem **more comfortable with complex**, **smaller** anatomy than generalists, indicating an area for practice standardization and education