Beyond Global Metrics: Unmasking Clinically Meaningful Deviations Using a Novel Local Metric In Regression Testing of Deployed Auto-Segmentation Models

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

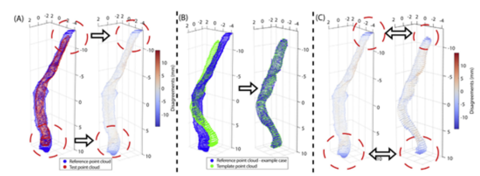
Jingwei Duan, PhD[1], Roya Barati[1], Yao Zhao, PhD[1], Song Gao, PhD[1], Mr. Laurence Edward Court, PhD[2], Jared D. Ohrt, MS[1], Peter Balter, PhD[1], Libing Zhu, PhD[3], Yi Rong, PhD[4], Quan Chen, PhD[3] and Jinzhong Yang, PhD[1]
(1)The University of Texas MD Anderson Cancer Center, Houston, TX, (2)Department of Radiation Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, (3)Mayo Clinic Arizona, Phoenix, AZ, (4)Mayo Clinic Arizona, Department of Radiation Oncology, Phoenix, AZ

INTRODUCTION

- Auto-segmentation is now routinely used in clinical practice. A core concern during implementation is how to continuously monitor segmentation performance over time. Domain shift between the training environment and real-world settings—driven by both aleatoric and epistemic uncertainty—can lead to suboptimal performance. Common clinical drivers include anatomical variability and differences in clinical practice, such as institutional protocols and guideline-dependent contouring conventions. To address domain shift, vendors and developers frequently release upgraded models intended to improve performance. However, because deep-learning (DL) behavior can change in non-intuitive ways, each upgrade still requires a dedicated clinical acceptability assessment and a clear understanding of how it differs from the prior version. Otherwise, users may incorrectly transfer their expectations and "mental model" from the previous system to the upgraded model, potentially overlooking new or shifted failure modes
- A well-known limitation of current evaluation metrics is that most provide global summaries, whereas clinically relevant contour discrepancies are often local and anisotropic, with region-dependent patterns across the surface. Motivated by this gap, we present a regression-testing framework to support commissioning of upgraded DL auto-segmentation models in commercial software. In this study, we first quantify performance changes between the previous and upgraded model versions using common metrics such as DSC, HD, and MSD. We then apply a method to quantify and visualize localized discrepancies between the two models

METHOD

A recent study introduced Deformable Point Cloud Registration–Based Bidirectional Local Distance (DPCR-BLD) [1], which is designed to reveal localized disagreements by shifting the analysis from global summary differences to region-specific discrepancies[1]. We applied DPCR-BLD to the same dataset.



(a) BLD calculation on a single-pair point clouds (test vs reference). (b) The reference point cloud being deformably registered to the template contour using CPD algorithm. (c) Mapping local deviations from the reference contour to the template contour.

RESULTS

- A total of 36 patients were included in the regression-testing cohort, including 18 head-and-neck (HN) and 18 male pelvis cases that had previously been used for initial model commissioning and were reused here to re-test the upgraded model. We first compared the initial and upgraded model outputs using standard geometric metrics, including DSC, HD95, and MSD. Paired t-tests were performed to assess differences, with statistical significance defined as $p < 0.05$. As shown in Figure 4, no statistically significant differences were observed for most organs at risk (OARs), except for the prostate (MSD), right cochlea (DSC), mandible (DSC, HD95, and MSD), and spinal cord (DSC).
- Unlike common geometric metrics that provide only global summaries, DPCR-BLD visualizes localized differences and explicitly incorporates spatial relationships. For example, the prostate achieved a DSC of 0.77 with the initial model and 0.80 with the upgraded model, but these aggregate values do not indicate where the discrepancies occur (Fig1). DPCR-BLD revealed that the initial model tends to over-contour the posterior prostate adjacent to the rectal centerline/wall, whereas the upgraded model tends to under-contour peripheral regions in contact with the rectum. Similarly, for the brainstem, DPCR-BLD indicated that the upgraded model slightly over-contoured the superior peri-superior region compared with the initial model (Fig2). These differences were not captured by the common metrics reported in Figure 4.
- Meanwhile, DPCR-BLD facilitates statistical outlier detection. For the rectum, common geometric metrics showed no statistically significant differences(Fig. 4); however, the baseline model failed to contour a case with a rectal balloon. DPCR-BLD clearly localized and highlighted these discrepancies (Fig3). Notably, although the colormap is analyzed using the entire dataset, the error case is sufficiently pronounced that it remains clearly visible in the aggregate evaluation.
- Figure 4 summarizes the DPCR-BLD analyses for most OARs evaluated in this study. These results indicate that relying solely on global geometric metrics can miss clinically meaningful, spatially localized differences. The capability to quantify and visualize anisotropic deviations is essential for the continuous monitoring of auto-segmentation tools in a clinical setting. Future work will (1) expand the evaluation to additional datasets and institutions and (2) investigate how DPCR-BLD outputs can be operationalized for routine clinical monitoring and QA.

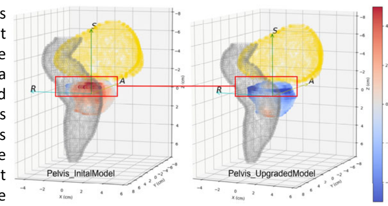


Figure 1. Local prostate contour disagreements between the two models. Red indicates over-contouring and blue indicates under-contouring, each relative to the reference contours.

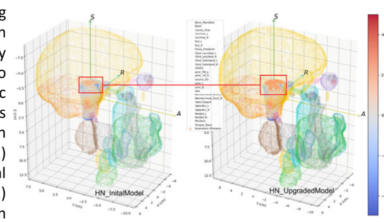


Figure 2. Local Brainstem contour disagreements between the two models.

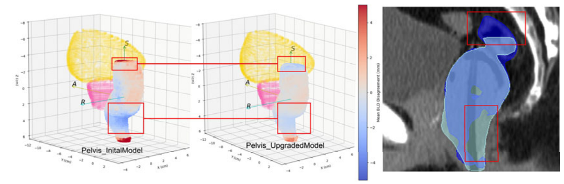


Figure 3. Local rectum contour disagreements between the two models. The sagittal view of the outlier case is shown. Green – initial model; Blue – upgraded model

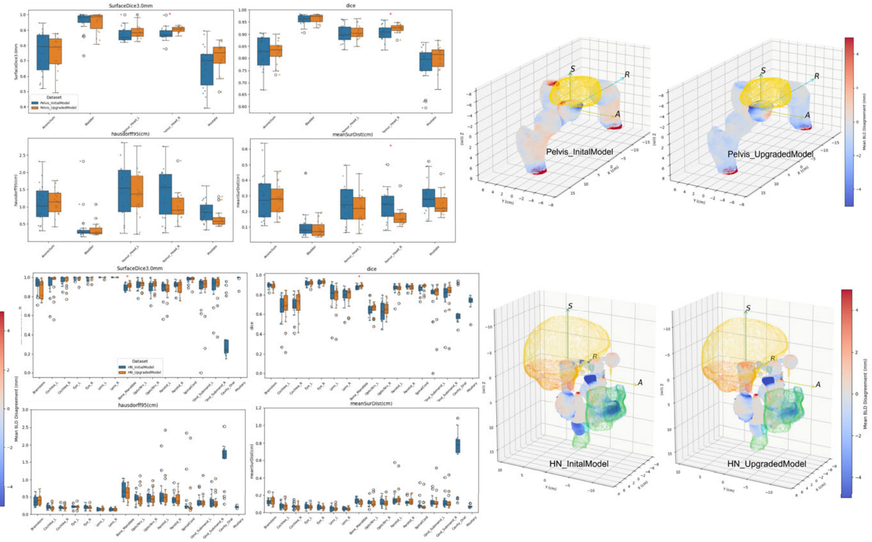


Figure 4. Most of OARs contour disagreements between the two models for both HN and Pelvis sites. Some OARs (e.g., the spinal cord, Brain, Bladder) are not shown for visualization clarity.

CONCLUSIONS

- Global metrics values did not indicate where discrepancies occurred. DPCR-BLD localized prostate differences: the initial model over-contoured posterior prostate adjacent to the rectal wall, whereas the upgraded model under-contoured peripheral regions in contact with the rectum.
- Combining conventional global metrics with DPCR-BLD provides actionable, spatially resolved regression testing for upgraded DL auto-segmentation models, helping prevent inappropriate transfer of expectations between versions and supporting continuous clinical monitoring and QA.

REFERENCES

1. Duan, Jingwei, et al. "Deformable Point Cloud Registration-Based Bidirectional Local Distance (DPCR-BLD): A Methodology for Systematic Evaluation and Visualization of Local Disagreements in Clinical Auto-Contouring." International Journal of Radiation Oncology* Biology* Physics (2026).