JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]
Improving Health QUALITY, Increasing Global IMPACT

Automatic CT segmentation for radiotherapy treatment planning: how good is good enough?

W. Scott Ingram and Lei Dong
University of Pennsylvania, Philadelphia, PA

Contact:
williamscott.ingram@pennmedicine.upenn.edu

Penn Medicine
RADIATION ONCOLOGY
PENN MEDICINE

INTRODUCTION

Manual segmentation of anatomical structures on CT scans for radiotherapy treatment planning is a time-consuming process, and inter-observer variability introduces uncertainty into plan optimization and evaluation. Autosegmentation has the potential to mitigate these issues, provided the segmentation is accurate enough.

Autosegmented structures will always differ from contours drawn by an experienced clinician. Small differences will have minimal impact on plan optimization and evaluation, but at what magnitude does this impact become significant? Despite a large number of publications and the availability of commercial products, this question does not have a clear answer.

Evaluation of segmentation accuracy is currently limited to two approaches: qualitative assessment of acceptability, or quantitative metrics based on structure overlap or surface distance. However, it is unknown whether these metrics provide any information about the dosimetric impact of segmentation errors.

The purpose of this work is to investigate the relationship between two commonly-used segmentation quality metrics and the dosimetric impact of the underlying structure variations, with the goal of determining if there are threshold values beyond which the dosimetric impact is negligible. This is accomplished by simulating a large number of autosegmented structures to determine if there is any correlation.

SIMULATION OF AUTOSEGMENTATION

Five head and neck radiotherapy plans were selected. All had the same dose prescriptions for the gross disease and nodal volumes, and all used 6 MV VMAT. Three structures and their dosimetric endpoints were used for evaluation:

- Mean dose to the right parotid gland
- D0.03cc to the brainstem
- D95% to the nodal PTV

These structures were chosen because each has a different type of dosimetric endpoint and each is close to steep dose gradients in all plans used. The nodal PTV is an expansion of elective nodal volumes, which are suitable for autosegmentation.

Autosegmentation was simulated by applying random 3D deformations to the manually-segmented reference structures:

1. The contours of the reference structure were converted to a polygon mesh.
2. A sparse grid of randomly-oriented deformation vectors was created covering the extent of the mesh.
3. Trilinear interpolation was used to apply a local deformation at each vertex of the mesh.

Steps 2 and 3 were repeated 400 times per structure, generating a total of 2000 deformed versions of each structure across the five plans. In order to produce a wide variety of deformations, the spacing of the sparse grid was varied between 10, 20 and 50 mm, and the magnitude of the vectors was varied between 2, 5, and 10 mm.

EVALUATION OF SEGMENTATION METRICS

For each deformed structure D and the manually-segmented reference structure R, two segmentation quality metrics were calculated:

- Dice similarity coefficient (DSC) $= \dfrac{2|D \cap R|}{|D|+|R|}$
- Average surface distance (ASD) $= \dfrac{\sum_{d \in D} \min\limits_{r \in R} \|d-r\| + \sum_{r \in R} \min\limits_{d \in D} \|r-d\|}{|D|+|R|}$

These values were compared to the dosimetric impact of the underlying structure deformation, which was quantified by the difference in the dosimetric endpoint when the dose-volume histogram was computed using the reference or deformed structure.

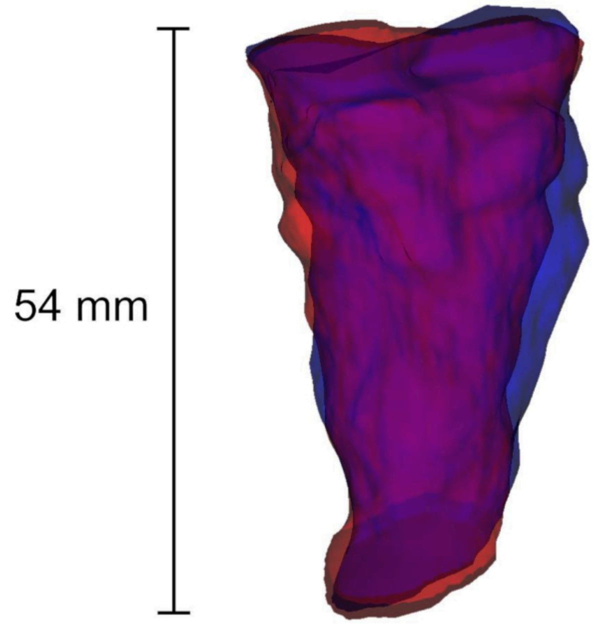

Figure 1: An example of reference and deformed polygon meshes for the brainstem. This 3D rendering shows the reference mesh in red and the deformed mesh in blue. The DSC is 0.86, and the ASD is 1.09 mm. The D0.03cc for the deformed mesh is 49.4 Gy, a 3.3% increase over the reference value of 47.8 Gy.

This deformed mesh was created with sparse grid spacing = 20 mm and deformation vector magnitude = 5 mm. The effect of varying the magnitude is intuitive. The grid spacing controls qualitative aspects of the deformations. When the spacing is large, the deformations vary slowly in space, creating systematic changes like translation, contraction, and expansion. When the spacing is small, the deformations vary quickly over short distances, resulting in more of a rippling effect on the surface.

RESULTS

The ranges of DSC and ASD were similar for all three structures, with DSC ranging from 0.60 to 0.98 and ASD from 0.12 to 3.74 mm. The distributions were heavily skewed, with roughly 50% having DSC > 0.9 and ASD < 0.9 mm. This is a desirable characteristic for the dataset, because the range of structures that could potentially be used for treatment planning without modification is sampled more densely.

The dosimetric endpoint differences were similarly skewed, with median and maximum differences of 2.6% and 47.0% for the right parotid mean dose, 1.6% and 35.0% for the brainstem D0.03cc, and 1.0% and 30.0% for the nodal PTV D95%. These changes were moderately correlated with DSC and ASD for the right parotid (Spearman's ρ = -0.59 and 0.58, respectively) and brainstem (ρ = -0.55 and 0.54). The correlations were stronger for the nodal PTV (ρ = -0.88 and 0.83).

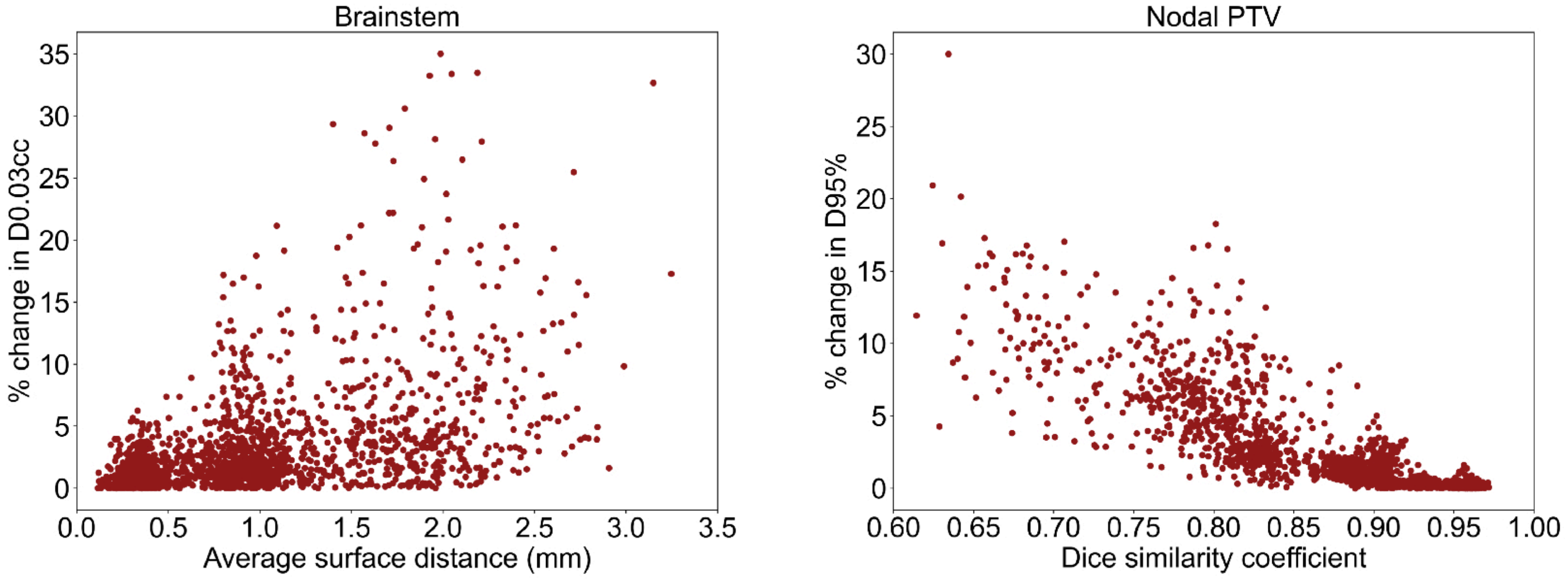

Figure 2: Scatter plots of dosimetric endpoint differences between reference and deformed structures vs. segmentation accuracy metrics for all 2000 deformed structures. The weakest correlation was Spearman's ρ = 0.54 between ASD and D0.03cc for the brainstem (left). The strongest correlation was ρ = -0.88 between DSC and D95% for the nodal PTV (right).

DISCUSSION

The segmentation quality metrics evaluated in this study do not generally correlate well with the dosimetric impact of the underlying structure deformations. The explanation for this is simple: the metrics contain no information about the spatial location of the deformations, while intensity-modulated radiotherapy treatment plans contain sharp dose gradients that are highly localized. The exception to the weak correlations is the nodal PTV, which might be expected given that the dose gradients generally conform to the shape of the PTV.

It is clear from the brainstem plot in figure 2 that there is no useful threshold of ASD below which one can be confident that structure variations can be ignored when evaluating the D0.03cc. This is likely an inherent limitation in ASD and all other metrics that are commonly used to evaluate autosegmentation algorithms. Future efforts should investigate novel metrics that characterize structure proximity to target volumes and dose gradients. Autosegmentation is just beginning to make its way into clinical use, and as more options become available, there will be a greater need to objectively characterize their performance in a meaningful way.