Quantitative versus qualitative and dosimetric evaluation of automated segmentations

J. PURSLEY, G. MAQUILAN, and G. SHARP
Massachusetts General Hospital and Harvard Medical School, Boston, Massachusetts

JULY 12–16 2020 VIRTUAL
JOINT AAPM | COMP MEETING
EASTERN TIME [GMT-4]

INTRODUCTION

- ❑ With AI-based algorithms, automated segmentation of normal tissues moving from research to clinical use
- ❑ Clinicians must evaluate commercial products to select one for implementation
  - Most studies use quantitative metrics like DICE score[1,2]
- ❑ **Showing a correlation between quantitative and qualitative metrics would establish a scientific basis to use quantitative metrics for comparisons**
  - Then improvements in a quantitative metric could be said to indicate an improvement in clinical acceptability

AIM

- ❑ Devise qualitative scoring system to evaluate auto. contours
- ❑ Compare qualitative scores to quantitative metrics between auto. and physician-approved contours
- ❑ Compare dosimetric changes to other metrics

METHODS

- ❑ Evaluated three disease sites:
  - Auto. contours generated for 20 prostate, 10 abdomen, and 10 head and neck patients
- ❑ Auto. contours generated using MiM deformable atlases[2]
  - Three atlases of previously contoured patients
  - Combined 4 deformed subjects using STAPLE strategy (Simultaneous Truth and Performance Level Estimation)[3]
  - Performed post-processing steps such as smoothing; clipped spinal cord to extent of physician-approved contour
- ❑ Three raters (the authors) evaluated qualitative scores
  - 5 score system preferred over 3 for wider range

Score = 5	Score = 4	Score = 3	Score = 2	Score = 1
Clinically acceptable	Minor edits required to be acceptable	Moderate edits overall or major edits required to only part of the structure to be acceptable	Major edits required to be acceptable	Completely unacceptable

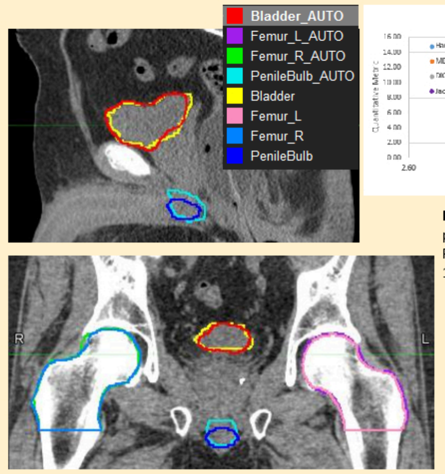
Table 1: *Qualitative rating system used by three reviewers to evaluate auto-segmented contours.*

- ❑ Scores averaged for comparison to quantitative metrics
  - Performed Spearman rank correlation
  - Correlations better than 0.70 highlighted in results
- ❑ Quantitative metrics evaluated in MiM
  - Two distance metrics: Hausdorff distance (HD) and Mean Distance to Agreement (MDA)
  - Two overlap metrics: DICE Similarity Coefficient (DSC) and Jaccard Index (JI)
- ❑ Recomputed dosimetric metrics for 10 prostate auto. contours using approved treatment plan

RESULTS

Prostate results

- ❑ Evidence of automation bias[4]
  - Prostate atlas used clinically since 2017
- ❑ 10 patients where auto. contours were provided to physicians:
  - Minimal or no changes made to femurs; Hausdorff distance best metric
- ❑ 10 patients where auto. contours not provided: results in Table 2
  - Femur quantitative metrics no longer correlate with qualitative scores
  - Only bladder metrics showed high correlation, with no preferred metric
- ❑ Calculated mean and max organ doses for auto. contours with approved plan
  - Did not observe any significant correlations between change in mean/max organ dose and either qualitative or quantitative metrics



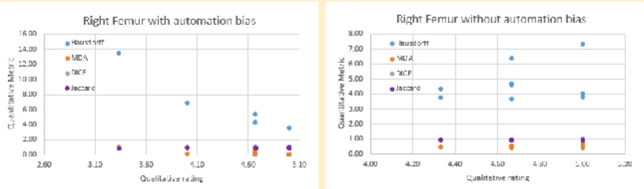


Figure 1. Left: Comparison of automated to physician segmentations for one prostate patient without automation bias. Above: Qualitative rating of the auto-segmented Right Femur contours for 10 prostate patients versus the four quantitative metrics for 10 subjects with (left) and without (right) automation bias.

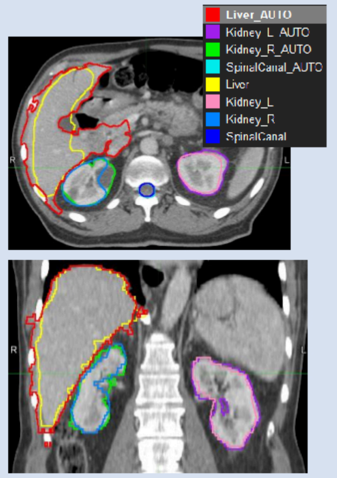
Prostate Spearman Correlations

Metric	Bladder	Left Femur	Right Femur	Penile Bulb
HD	-0.88	0	0.29	-0.19
MDA	-0.85	0.23	0.25	-0.30
DSC	0.81	-0.13	-0.01	0.33
JI	0.82	-0.13	0	0.33

Table 2: Spearman Rank Correlations of quantitative to qualitative metrics for 10 prostate patients without automation bias.

Abdomen results

- ❑ Higher correlation of qualitative and quantitative observed than for prostate
  - Exception: spinal canal. Auto. contour usually considered clinically acceptable
  - Quantitative metrics may not be good predictors of quality here
- ❑ Liver and kidney auto. contours ranged from moderate edits (shown in Figure 2) to completely unacceptable as judged by qualitative raters
  - Both distance and overlap metrics highly correlated with qualitative scores
  - Liver correlations lower than kidneys; may be due to large volume of liver
  - For large organs, quantitative metrics may not be as sensitive to clinically significant modifications of contour boundary



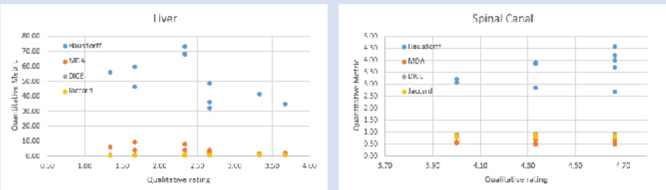


Figure 2. Left: Comparison of automated to physician segmentations for one abdomen patient. Above: Qualitative rating of the auto-segmented Liver and Spinal Canal contours versus the four quantitative metrics for 10 abdomen patients.

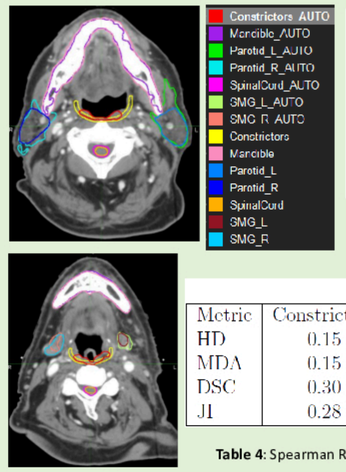
Abdomen Spearman Correlations

Metric	Liver	Left Kidney	Right Kidney	Spinal Canal
HD	-0.53	-0.63	-0.80	0.45
MDA	-0.70	-0.91	-0.93	-0.03
DSC	0.76	0.87	0.91	0.26
JI	0.78	0.87	0.92	0.27

Table 3: Spearman Rank Correlations of quantitative to qualitative metrics for 10 abdomen patients.

Head and Neck results

- ❑ Mandible showed similar results as prostate femurs
  - Bony anatomy largely clinically acceptable despite quantitative variations
- ❑ Similar results for H&N spinal cord as abdomen spinal canal
  - Variations not considered clinically significant by qualitative raters
- ❑ Other organ results are mixed; need more patient data to confirm
  - Right parotid gland showed high correlations while left parotid did not; most likely explanation is fluctuation of results in a few patient cases
  - Submandibular glands (SMG) did show high correlation between qualitative and overlap metrics
  - Constrictors, small soft tissue organ, did not show correlation with any metrics



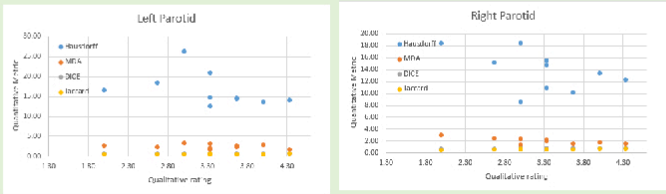


Figure 3. Left: Comparison of automated to physician segmentations for one head and neck (H&N) patient. Above: Qualitative rating of the auto-segmented Left and Right Parotid contours versus the four quantitative metrics for 10 H&N patients.

Head and Neck Spearman Correlations

Metric	Constrictors	Mandible	Lt Parotid	Rt Parotid	Spinal Cord	Lt SMG	Rt SMG
HD	0.15	-0.32	-0.40	-0.50	0.18	-0.16	-0.09
MDA	0.15	-0.26	-0.28	-0.75	-0.05	-0.70	-0.65
DSC	0.30	0.12	0.17	0.72	0.14	0.78	0.76
JI	0.28	0.12	0.20	0.70	0.16	0.76	0.73

Table 4: Spearman Rank Correlations of quantitative to qualitative metrics for 10 H&N patients.

CONCLUSIONS

- ❑ Qualitative scoring system was fast, easy to use, and consistent among three reviewers from the same institution
  - Standard deviation on average scores ranged from 0 to 0.9
  - But evaluating hundreds of patients this way would be time-consuming
- ❑ Automation bias observed when auto. contours provided to physician
  - Higher correlation between qualitative and quantitative metrics when physician directly modifies auto. contours
- ❑ Quantitative metrics must be used with caution to compare quality of automated segmentations
  - MDA, DSC, and JI showed high correlation with qualitative scores for larger soft tissue organs: liver, kidneys, bladder, salivary glands
  - But there was low correlation with qualitative scores for bony anatomy and small organs: femurs, mandible, spinal cord or canal, constrictors
  - Increase in quantitative metric may not correspond to more acceptable contour
- ❑ **Study limitations**:
  - Limited number of patients and disease sites analyzed
  - Auto. contours generated from deformable atlases not AI-based algorithm
  - Variation in physician-approved contours may contribute to noise
  - Preliminary dosimetric analysis showed no correlation with qualitative scores, calling into question whether qualitative scores represent clinical significance
- ❑ **Recommendations**:
  - Use quantitative metrics to compare algorithms for large number of patients
  - Analyze a small number (10-20) qualitatively to cross-check quantitative results

REFERENCES

1. J Yang, H Veeraraghavan, SG Armato 3rd, *et al*. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. *Med Phys*. 2018;45(10):4568-4581.
2. M La Macchia, F Fellin, M Amichetti, *et al*. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol* **7**, 160 (2012).
3. W Wong, L Leung, and D Kwong. Evaluation and optimization of the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy. *Br J Radiol*. January 2016;89(1057).
4. A Aselmaa, M van Herk, Y Song, and RHM Goossens. The influence of automation on tumor contouring. *Cogn Tech Work* 2017;19:795-808.

CONTACT INFORMATION

Jennifer Pursley, jpursley@mgh.harvard.edu