Title

Quality Assurance and Clinical Acceptability for AI-driven Automatic Contouring of Organs at Risk

Authors

Nina Dissler[1], Sotiri Stathakis[2], Aurélien Lombard[3], Nikos Paragios[4], Guillaume Klausner[5], Lucien Lahmi[5], William Jones III[6], Elizabeth Maani[6]

Authors Affiliations

[1]TheraPanacea, Functional & Clinical Specifications, Paris, France; [2]MD Anderson Cancer Center, Mays Cancer Center, Radiation Oncology and Radiology, San Antonio, TX, USA; [3]TheraPanacea, Machine Learning Research & Development, Paris, France; [4]TheraPanacea, President & Chief Executive Officer, Paris, France; [5]Université de médecine Pierre et Marie Curie, Paris Sorbonne Université, Radiation Oncology and Radiology, Paris, France; [6]South Texas Veterans Health Affairs, Radiation Oncology, San Antonio, TX, USA

Purpose or Objective

The aims of this study are three-fold: (i) to evaluate the generalization properties of automatic contouring solution across different annotation guidelines (ESTRO for training, ASTRO for testing), (ii) to compare the precision between experts' annotations (inter-variability) and the ones obtained by AI solution (iii) to evaluate the relevance of commonly adopted quality assurance metrics with respect to the clinical reality.

Materials and Methods

An automatic contouring CE/FDA-cleared solution - based on an ensemble deep learning models trained for contouring 100+ organs at risk with ESTRO guidelines on +25,000 patients after anatomically preserving data augmentation – is blindly evaluated on ~140 ASTRO guidelines patients coming from the Mays Cancer Center (60%) and the cancer imaging archive (40%) for pelvis, abdomen head & neck and brain anatomies. In addition, a distinct set of 60+ patients from the pelvis and head & neck anatomies is retrospectively blindly annotated from a second qualified radiation oncologist.

Results

The absence of ASTRO-contoured patients in training doesn't seem to impact the generalization of the trained models to the ASTRO treated population cohort in terms of Dice coefficient (DSC) and Hausdorff distance (HD) (with marginal consistent improvement on the ASTRO patient cohort). Generalization is demonstrated by an average DSC among 34 organs of 0.74, higher value of 0.97 for the lungs and lower value of 0.31 for the right brachial plexus. In terms of acceptability, DSC score which is the most commonly adopted and recommended metric fails short to depict the complexity of the objective. On top of the fact that the measure is biased towards organs with significant volume, it was observed a clear relationship between the observed DSC and the volume of the organs on all anatomies (except for bone structures). Finally, inter-expert variability (average DSC among 9 organs of 0.54, higher value of 0.82 for the rectum and lower value of 0.13 for the penile bulb) seems to follow in terms of DSC discrepancies the same trend of behavior as the one observed for the automatic contouring solution while in average showing constantly lower similarities than the ones observed between the automatic solution and the expert's annotation (see Figure 1).

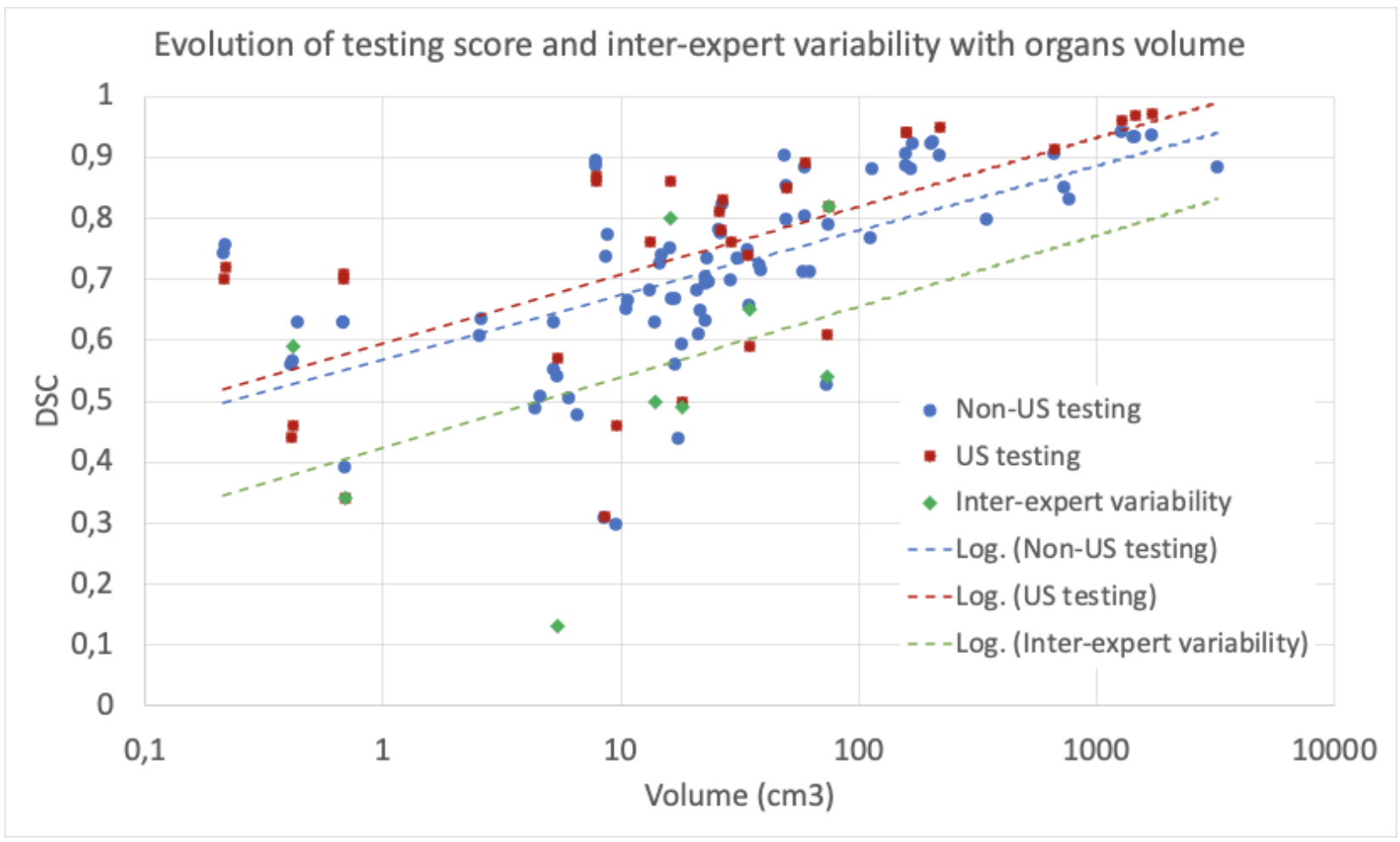

Figure 1: *Automatic contouring DSC in function of the volume of the organ with respect to (i) ESTRO-guidelines patients, (ii) ASTRO-guidelines patients and iii) against inter-expert variability.*

Conclusion

This blinded quantitative evaluation of experts' and AI-based delineation shows that AI-driven contours when obtained after training on massive multi-centric cohorts are clinically usable. It also shows that existing quality assessment tools such as DSC fail to reflect the reality of the clinical setting and negatively bias the adoption of these methods in clinical practices. Last, but not least it also shows that inter expert variability could be more significant than the one observed between experts and the automatic solution.