ESTRO

Abstract

Title

Statistical discrepancies in GTV delineation for H&N cancer across expert centers

Authors

Amaury Leroy[1,5], Nikos Paragios[1], Eric Deutsch[2], Vincent Grégoire[3], Diana Mitrea[4], Adeline Pêtre[3], Roger Sun[5], Yun Gan Tao[4]

Authors Affiliations

[1]Therapanacea, Artificial Intelligence, Paris, France; [2]Gustave Roussy, Paris-Saclay University, Inserm 1030, Molecular Radiotherapy and Therapeutic Innovation, Villejuif, France; [3]Centre Léon Bérard, Radiation Oncology, Lyon, France; [4]Gustave Roussy, Radiation Oncology, Villejuif, France; [5]Gustave Roussy, Paris-Saclay University, Inserm 1030, Molecular Radiotherapy and Therapeutic Innovation, Villejuif, France

Purpose or Objective

Accurate delineation of the primary tumor GTV is a decisive early step for radiotherapy since it impacts dose prescription, overall treatment toxicity, patient outcome and lifelong sequels. The aim of our work is to assess variability in GTV definition for H&N cancer through a statistical study involving two independent centers with observers of different experiences each. We also focus on the benefit of a consensus in the clinical routine and the need to incorporate multimodal imaging to add biological and functional insight in target volume delineation.

Materials and Methods

We have settled a retrospective cohort made of 45 patients, for which was provided a contrast enhanced CT acquisition and the report from endoscopy with photographic images and clinical data. For each center, junior and senior radiotherapists independently delineated the GTV with standardized rules. Initial statistical comparisons were conducted, such as volume, Dice score and Hausdorff distance, to assess inter-observer variability both in terms of center and experience. Next, we asked the senior practitioners to review each patient towards possible consensus. Based on their discussion, we updated the statistics as they were able either to find a common target volume or to stick to their original assessment, thus confirming disagreement.

Results

Table 1 reports an initial Dice score of 0.68 and Hausdorff distance of 12.1mm between senior observers. This strong disagreement warns us about the lack of standardization in treatment. Within the same center, lower variability between junior and senior (Dice of 0.71 for A and 0.73 for B) highlights bias in routine practice characteristic to each institution. The main difference between juniors and seniors lays in the tumor volume, bigger for juniors (≈31cm^3 against ≈24cm^3 for seniors), who usually prefer to avoid false-negative signals. During consensus, discussions lead to three main remarks: for 33% of patients, one observer aligned with his colleague's decision. 44% of cases were still in disagreement, the main explanation being that one center often excluded peritumoral edema from GTV. Finally, 23% of patients had similar delineations, becoming equal when extending to CTV. We computed statistics on updated volumes, with a new Dice score of 0.78 and Hausdorff distance of 7.4mm. Figure 1 shows a typical example of disagreement.

Dice \ Hausdorff Distance (mm)	Junior A	Senior A	Junior B	Senior B
Junior A		10.8 ± 2.7	14.2 ± 3.9	13.5 ± 4.1
Senior A	0.71 ± 0.08		13.8 ± 3.1	**Before consensus** **12.1 ± 3.8** / **7.4 ± 1.4** **After consensus**
Junior B	0.63 ± 0.09	0.65 ± 0.09		9.5 ± 1.9
Senior B	0.67 ± 0.10	**Before consensus** **0.68 ± 0.06** / **0.78 ± 0.04** **After consensus**	0.73 ± 0.09	

Figure 1: Dice score (bottom-left) and Hausdorff distance (top-right) between observers

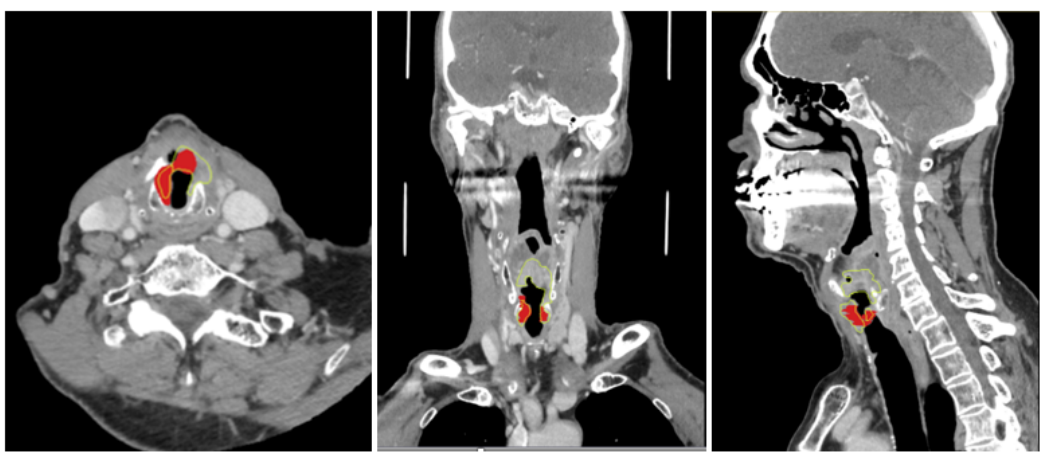

Figure 1: Extreme case of disagreement where senior A (filled red) chose not to include the upper area close to thyroid cartilage, while senior B did (contour yellow)

Conclusion

A significant deleterious inter-observer variability appears for GTV delineations, which can be explained by differences in interpretation of the endoscopy, level of experience, or working practice proper to each institution. An improved agreement was found after consensus as discussions acted as a sanity check and showed benefit for clinical routine. This study reinforces the need for multimodality when dealing with target volume definition, like multiparametric functional imaging or biopsies. Moreover, the development of artificial intelligence solutions for standardization and treatment automation could also be of great help.