A blinded prospective evaluation of clinical applicability of deep learning-based auto contouring of OAR for Head & Neck radiotherapy

Introduction

Contouring Organs at Risk (OAR) is time-consuming and highly inhomogeneous among physicians; it affects the accuracy of high precision image-guided radiotherapy. Artificial intelligence (AI) can accelerate OAR delineation and homogenize volume definition. This study aims at blindly evaluating two versions of an AI-based automatic delineation software for OAR.

Material and Methods

The software tested is a CE-marked software for automatic contouring of more than 80 OAR harnessing a unique combination of anatomically preserving and deep learning annotation concept. This study involved 100 patients with head and neck tumors, retrospectively selected from two French Cancer Centers, for whom clinical expert's annotations that were used for treatment were retrieved. Two subsets of data were randomly created, the first mixing 50% of expert-delineated contours and 50% of software v1.0-generated contours, the second mixing (1/3 each) expert-contours and software v1.0 and v2.0 contours. v1.0 was trained using on average 6,000 cases per organ, while v2.0 used 21,000, in both cases after data augmentation. Contours of 16 OARs were generated and scored by 5 experts and then 4 OARs (mandible, M; brainstem, BS; optic nerve, ON; submandibular gland, SG) were scored again by two experts (PB & VG), as A/ acceptable, B/ acceptable after minor corrections, C/not acceptable. Dice similarity coefficient (DSC) and Hausdorff distance (HD) were also computed.

Results

For the first set of data, 96% of all manual contours were classified as clinically useable (75% and 21% in A and B categories, respectively), compared to 95% for auto-contours (56 % and 39 % in A and B, respectively).

Using software v2.0, contours classified as clinically useable (A + B) increased significantly, reaching 100% for M, 98% for BS, 98% for ON and 92% for SG, versus 100%, 97%, 63% and 50% for v1.0, respectively.

When the two datasets were compared, intra- and inter-observer rating (score A, B or C) reproducibility was rather poor, ranging from 26% to 78% for the 4 OARs. When only looking at score A+B vs C the reproducibility among observers increased, ranging between 50% and 98%. For ON and SG, mean DSC improved from 0.53 to 0.70 and 0.70 to 0.78 between v1.0 and v2.0 of the software, whereas mean HD decreased by 30% and 17%, respectively.

Conclusion

This study illustrates the potential of AI for automatic contouring of OAR in radiotherapy planning. Automatic contouring with this CE-marked software was very close to expert contouring and clinically usable in the vast majority of cases. Evaluation of automatic algorithms requires objective metrics as illustrated by the disagreement between experts. Evaluation of the impact of contour delineation heterogeneity on dose distribution remains is in progress.

Authors: Pierre Blanchard1, Vincent Grégoire2, Claire Petit1, Nicolas Milhade2, Albina Allajbej2, France Nguyen1, Sofia Bakkar1, Geoffroy Boulle1, Edouard Romano1, Wael Zrafi1, Aurélien Lombard3, Guillaume Beldjoudi3, Alexandre Munoz3, Eugénie Ullmann3, Nikos Paragios3,5, Eric Deutsch1,4,5, Charlotte Robert1,4,5.

Institutions:

1. Gustave Roussy Cancer Campus – Paris-Saclay University, Department of radiotherapy, Villejuif, France.
2. Léon Bérard Cancer Center, Department of radiotherapy, Lyon, France.
3. Therapanacea, Paris, France.
4. Molecular radiotherapy and innovative therapeutics, INSERM UMR1030, Gustave Roussy Cancer Campus, Université Paris Saclay, Villejuif, France
5. Gustave Roussy-CentraleSupélec-TheraPanacea, Center of Artificial Intelligence in Radiation Therapy and Oncology, Gustave Roussy Cancer Campus, Villejuif, France,