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

P Blanchard[1], V Grégoire[2], C Petit[1], N Milhade[2], A Allajbej[2], F Nguyen[1], S Bakkar[1], G Boulle[1], E Romano[1], W Zrafi[1], A Lombard[3], G Beldjoudi[2], A Munoz[2], E Ullmann[3], N Paragios[3], E Deutsch[1,4], C Robert[1,4].

1. Gustave Roussy Cancer Campus, Villejuif, France; 2. Léon Bérard Cancer Center, Lyon, France; 3. Therapanacea, Paris, France; 4. Molecular radiotherapy and innovative therapeutics, INSERM UMR1030, Gustave Roussy Cancer Campus, Université Paris Saclay, Villejuif, France

PURPOSE/OBJECTIVE(S)

- 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 & 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.** v1.0 was trained using on average 6,000 cases per organ, while v2.0 used 21,000, in both cases after data augmentation.
- **One hundred 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 mixed **50% of expert-delineated contours and 50% of software v1.0-generated contours**,
  - the second mixed (1/3 each) **expert-contours and software v1.0 and v2.0 contours**.
- 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) (Table 1).
- 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 (Table 2).
- 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.

Figure: Examples of auto-contoured HN volumes

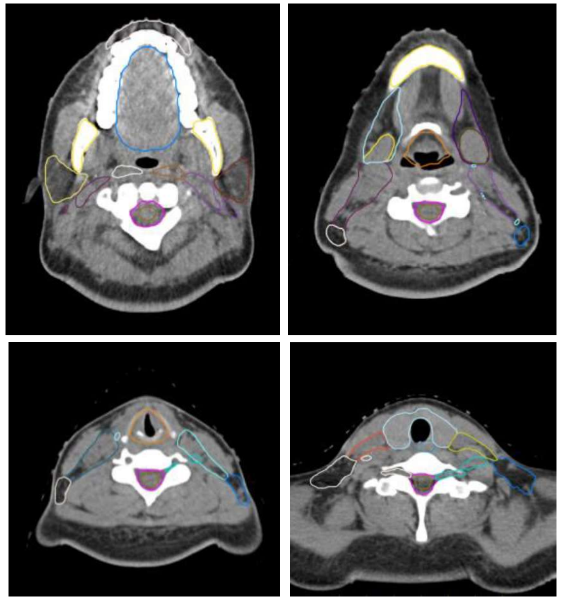

Table 1: Percentage of clinically useable contours between manual and autocontouring (v1.0 of the software – 1st evaluation)

% of A+B	Right Parotid	Left Parotid	Mandibule	Spinal Cord	Brainstem	Right Eyeball	Left Eyeball
Manual contouring	97%	96%	97%	89%	89%	100%	97%
Autocontour v1.0	96%	96%	99%	94%	98%	96%	96%

% of A+B	Right optic nerve	Left optic nerve	Oral Cavity	Larynx	Thyroid	R sub mandib Gl	L sub mandib Gl
Manual contouring	99%	99%	94%	94%	99%	99%	100%
Autocontour v1.0	89%	92%	99%	100%	97%	87%	86%

Table 2: Percentage of clinically useable contours between manual and 2 versions of autocontouring (v1.0 & v2.0) for 4 selected OAR (2nd evaluation)

% of A+B	Manual contouring	Software v1.0	Software v2.0
Mandibule	87%	100%	100%
Brainstem	69%	97%	98%
Optic nerve	93%	63%	98%
Sub-mandibular gland	97%	50%	92%

SUMMARY/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.

#ASTRO20

@PBlanchardMD