Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark?

Gian Carlo Mattiucci, Luca Boldrini, Giuditta Chiloiro, Giuseppe Roberto D'Agostino, Silvia Chiesa, Fiorenza De Rose, Luigi Azario, Danilo Pasini, Maria Antonietta Gambacorta, Mario Balducci & Vincenzo Valentini

ORIGINAL ARTICLE

Automatic delineation for replanning in nasopharynx radiotherapy: What is the agreement among experts to be considered as benchmark?

GIAN CARLO MATTIUCCI[1], LUCA BOLDRINI[1], GIUDITTA CHILOIRO[1], GIUSEPPE ROBERTO D'AGOSTINO[1], SILVIA CHIESA[1], FIORENZA DE ROSE[1], LUIGI AZARIO[2], DANILO PASINI[1], MARIA ANTONIETTA GAMBACORTA[1], MARIO BALDUCCI[1] & VINCENZO VALENTINI[1]

[1]*Radiation Oncology Department, Università Cattolica del Sacro Cuore, Rome, Italy, and* [2]*Institute of Physics, Università Cattolica del Sacro Cuore, Rome, Italy*

Abstract

Background and purpose. Anatomic changes during head and neck radiotherapy require replanning. The primary aim of this study is the definition of the agreement among experts in the head and neck automatic delineation frame to use as benchmark. The secondary goal is to assess the reliability of automatic delineation for nasopharynx radiotherapy and time saving. *Material and methods.* A computed tomography (CT) scan was acquired in 10 nasopharynx patients along intensity-modulated radiotherapy (IMRT) treatment for replanning. Deformable registration with replanning autocontouring of the structures was performed using VelocityAI 2.3© software defining Structure Set A. The optimization of these contours was obtained through revision by a skilled operator, drawing Structure Set B. An ex novo Structure Set C was segmented on the replanning CT-scan by an expert delineation team. The mean Dice's Similarity Index (mDSI) was calculated between Structure Set A and B, A and C, and between B and C for each volume. All segmentation times for organs at risk (OARs) and clinical target volume (CTV) were recorded and compared. *Results.* We validated the replanning autocontoured Structure Sets for 10 patients. For volumetric analysis we observed mDSI values of 0.87 for the OARs, 0.70 for nodes, 0.90 for CTV in the Structure Set A-B comparison and respectively of 0.74, 0.63 and 0.78 for the Structure Set A-C one, and 0.78, 0.78 and 0.85 for Structure Set B-C, which represents the existing expert based benchmark. We calculated a mean saved time in Structure Set B of 30 minutes. *Conclusions.* Autocontouring procedures offer considerable segmentation time saving with acceptable reliability of the contours, even if an independent check procedure for their optimization is still required to increase their adherence to referential benchmark gold standard among experts, which stands at a 0.80 DSI value.

Radiotherapy (RT) represents one of the most important therapeutical approaches in the management of head and neck cancers. Literature data suggest that clinically significant internal geometric and volumetric changes occur throughout RT course, potentially causing underdosage of the target volumes and overdosage of normal tissues [1–3]. The dosimetric changes occurring during intensity-modulated RT (IMRT) treatment are generally more drastic than in conventional RT, due to the sharp dose gradients between the boundary of target volumes and critical normal tissues. A replanning CT scan, with a new clinical target volume (CTV) and organs at risk (OARs) delineation, is therefore recommended [4,5]. Delineation is a critical and time consuming step during treatment planning even for experienced head and neck radiation oncologists [6–8].

Deformable registration algorithms (DRAs) allow for the alignment of data sets that are mismatched in a non-uniform way consenting contours transfer between the original simulation CT scan and the replanning one, and decreasing the recontouring time, as well as reducing the intra- and inter-observer variability during plan reoptimization [9–12]. Furthermore, literature still lacks robust evaluation tools

G. C. Mattiucci et al.

and benchmark values in order to assess the reliability of these software.

The aim of this study is to present an actual benchmark for independent check procedures of the autocontoured structures and to quantify the replanning autocontouring reliability and the segmentation time saved with this approach.

Material and methods

Patients' characteristics

We enrolled 10 consecutive patients with nasopharynx cancer treated with IMRT. All patients had primary Stage III–IV tumors, and no patient underwent neck surgery before the RT treatment. The median age was 53.9 years (range 30–82); there were eight males and two females. The mean prescription dose to CTV1 (primary lesion and pathological nodes) was 70.20 Gy, to CTV-E (represented by the negative drainage nodes) was 50.40 Gy. The median number of days between the simulation CT scan and the replanning one was 42 (range 32–56) and the median delivered dose was 30.20 Gy, mean value 36 Gy (range 21.6–59.4 Gy). A conventional helical CT scanner (GE HiSpeed DX/i Spiral) was used for image acquisition (slice thickness 2.5 mm; no IV contrast agent).

Manual contouring, deformable coregistration and replanning autocontouring

Our regions of interest (ROIs) were manually contoured on each axial slice of the simulation CT scan using a commercial TPS (Eclipse®, Varian). ROIs included 25 volumes per patient:

1. Fourteen OARs: brain, brainstem, right and left eye, right and left parotid, oral cavity, jaw, larynx, spinal cord, right and left clavicle and right and left humeral head.
2. Ten nodal stations: Ia, Ib, I (Ia + Ib), IIa, IIb, II (IIa + IIb), III, IV, V, VI according to the "CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines" [13].
3. One CTV-E which included all the lymphatic drainage stations (from Ia to VI).

Primary tumor and pharyngeal nodes were not included because of their expected macroscopic volumetric changes during RT. A replanning CT scan was then acquired for each patient meanly during the fourth week of treatment, when a dose of 36 Gy was reached. Using a commercially available software (VelocityAI 2.3©, Velocity Medical Solutions Inc.), the structures were firstly aligned via a rigid body registration between the simulation CT scan and the replanning one in order to reduce mismatch and to initialize the deformable registration [14]. As a third step, the deformable registration [15] was fulfilled obtaining the automatic deformation of the original contours on the replanning CT. The new contours proposed by VelocityAI (Structure Set A) were then manually corrected on the TPS as new Structure Sets (Structure Set B). This procedure represents our QA workflow as in our institution every structure set is contoured by a resident and then corrected by a second, skilled, physician. In this study the autocontouring software plays the role of the resident. The replanning CT-scan was then recontoured ex novo for each OAR, nodal station and CTV-E with the common agreement of all the operators (Structure Set C) in order to limit inter- and intraobserver variability. The initial delineator of the Structure Set B has always been excluded from the consensus to limit his influence in Structure Set C delineation. Five skilled investigators performed the selection, delineation, deformation and correction steps of two patients each. To quantify time saving, the optimization time of each volume was recorded and compared to manual ex novo segmentation of the Replanning CT-scan data.

Dice's Similarity Index (DSI)

As the DSI has been widely used for this purpose [9,16,17], we setup an in house made software for the DSI calculation. Having A as the automatically contoured surface, B as the manually corrected surface and C as the ex novo contours, the DSI were defined as:

$$\mathrm{DSI} = \frac{2|A \cap B|}{|A| + |B|}, \frac{2|A \cap C|}{|A| + |C|}, \frac{2|B \cap C|}{|B| + |C|}$$

DSI is a scalar coefficient with a value between 0 and 1. A value of 0 indicates that the considered volumes are completely disjoint, whereas a value of 1 is reached when delineations are identical.

In order to quantify the volume overlap, the mean DSI (mDSI) was calculated for each volume between the automatically replanning autocontouring segmentation proposed by the software (A) and the ex novo Structure Set (C) (A vs. C) and between the segmentation manually corrected by the investigators (B) and the ex novo Structure Set C (B vs. C) (Table I). A further comparison between Structure Set A and B described the entity of the manual correction of the autocontoured structures. Summing up, the B-C comparison represents the agreement among the experts and the benchmark value, while the A-B one describes the practice of the single operator and the A-C the peer team approach.

Table I. Mean DSI values for Structure Set B-C comparison for OARs, nodal levels and CTV-E.

OARs	DSI	Nodal levels	DSI	CTV-E	DSI
Brain	0.97	I	0.81	Mean	0.78
Brainstem	0.85	Ia	0.73		
Right eye	0.89	Ib	0.81		
Left eye	0.88	II	0.82		
Right parotid	0.79	IIa	0.82		
Left parotid	0.78	IIb	0.79		
Oral cavity	0.80	III	0.80		
Jaw	0.89	IV	0.77		
Larynx	0.83	V	0.76		
Spinal cord	0.79	VI	0.74		
Right clavicle	0.89				
Left clavicle	0.90				
R. hum. head	0.88				
L. hum. head	0.87				

Statistics

The software SPSS v. 17 was used for data analysis. The Wilcoxon test was performed in order to evaluate statistically significant differences between the considered variables with the null hypothesis that the two samples belong to the same population ($p > 0.05$) and the alternative one that the two samples come from two different populations ($p < 0.05$). The comparison was performed between the time values of B and C samples and the DSI calculated between A and C, B and C and A and B.

Results

DSI analysis

The median DSI of each volume was calculated between Structure Set A and C, between Structure Set B and C and between Structure Set A and B, obtaining mean values of 0.71, 0.80 and 0.82, respectively, for the whole structure set.

DSI analysis for CTV and nodal stations

A statistically significant advantage for B-C versus A-C comparison was recognized in CTV-E analysis with mDSI value of 0.78 versus 0.74 ($p = 0.0039$) and in nodal stations as singles ($p = 0.002$ for stations IA, IV, V, VI − 0.039 nodal station II). Considering the CTV A-B comparison, a mDSI value of 0.90 has been described ($p = 0.002$) if considered versus the B-C one (0.78). A statistically significant advantage has also been seen for nodal station Ia ($p = 0.05$), III ($p = 0.02$) and V ($p = 0.02$).

DSI analysis for organs at risk

A statistically significant advantage for B-C versus A-C comparison is described for all the OARs ($p = 0.002$ for right clavicle, oral cavity, jaw and right parotid − 0.039 for right humeral head); exceptions were observed only for brain ($p = 0.8$) and brainstem ($p = 0.1$). However, a statistically significant advantage for A-B versus B-C has been recognized ($p = 0.002$ for brain and brainstem − 0.009 for left parotid, spinal cord and left clavicle). No statistically significant advantage was recorded for the other volumes. See supplementary Tables I–II (to be found online at http://informahealthcare.com/doi/abs/10.3109/0284186X.2013.813069) for volume specific mDSI values in Structure Set A–C and A–B comparisons.

Optimization time recording and time saving calculation

Structure Set A. The mean total coregistration and contours propagation time (Structure Set A delineation) was of 2.4 minutes (range 2–3.2 minutes).

Structure Set B. The mean optimization time for the replanning autocontouring Structure Set B was calculated both as OARs plus all nodal stations or OARs plus CTV-E. The OARs plus all nodal stations value was of 43.9 minutes (range 23.1–79 minutes), while for OARs plus CTV-E, it was 34.9 minutes (range 17.4–58.6 minutes). The mean optimization time for CTV-E was 12.5 minutes (range 6.3–25.7 minutes). The mean optimization time for a single nodal station was 2.1 minutes (range 0.3 minutes of IA–3.5 minutes of III). The mean optimization time for all nodal stations was 21.5 minutes (range 10.5–35.9 minutes). The mean optimization time for each OAR was 1.6 minutes (range 0.4 minutes for left eye–3.9 minutes for jaw). The mean optimization time for all OARs was 22.2 minutes (range 10.6–44 minutes). See supplementary Table III (to be found online at http://informahealthcare.com/doi/abs/10.3109/0284186X.2013.813069) for Structure Set B segmentation time values

Structure Set C. The mean optimization time for the whole replanning autocontouring Structure Set C was calculated as OARs plus all nodal stations or OARs plus CTV-E. For OARs plus all nodal stations, the mean optimization time value was 73.4 minutes (range 58.4–82.6 minutes), while for OAR plus CTV-E, it was 69.7 minutes (range 52.2–79.4 minutes). The mean delineation time for CTV-E was 39.9 minutes (range 31.5–48.2 minutes) (Table II). The mean optimization time for a single nodal station was 4.4 minutes (range 0.2 minutes for Ia–11.3 minutes for II). The mean delineation time for nodal stations was 43.7 minutes (range 37.3–53.4 minutes) (Table II). The mean optimization time for each OAR was 2.1 minutes (range 0.4 minutes for left eye–4.9 minutes for jaw). The mean delineation time for OARs was 29.7 minutes (range 20.5–35.2 minutes) (Table II).

Table II. Mean optimization time values for Structure Set C ex novo segmentation.

OARs	Time (m)	Nodal levels	Time (m)	CTV-E	Time (m)
Brain	2.6	I	3.4	Pt. 1	48.2
Brainstem	1.9	Ia	0.3	Pt. 2	44.7
Right eye	0.5	Ib	3.3	Pt. 3	31.5
Left eye	0.5	II	7.7	Pt. 4	37.2
Right parotid	2.3	IIa	5.1	Pt. 5	42.1
Left parotid	1.9	IIb	5.1	Pt. 6	37.2
Oral cavity	2.9	III	5.1	Pt. 7	46.3
Jaw	3.9	IV	4.6	Pt. 8	32.6
Larynx	2.3	V	4.9	Pt. 9	31.7
Spinal cord	2.1	VI	4.0	Pt. 10	48.1
Right clavicle	2.3				
Left clavicle	2.2				
R. hum. head	2.0				
L. hum. head	2.0				

A statistically significant advantage (p = 0.005) was observed also for the CTV-E delineation time for Structure Set B versus C. A statistically significant difference was seen for the nodal stations, with an advantage for B versus C for Lfn I (p = 0.006), Ib (p = 0.01), II (p = 0.004), IIa (p = 0.002), IIb (p = 0.01), IV (p = 0.002) and Lfn VI (p = 0.004). An advantageous trend has been recorded for Lfn III (p = 0.049) and V (p = 0.049). No statistically significant difference has been recorded for Lfn Ia (p = 0.08).

For what concerns the OARs there is a statistically significant advantage in terms of contouring time for Structure Set B versus C only for the oral cavity (p = 0.022), for the brain (p = 0.02), for the brainstem (p = 0.002) and for the larynx (p = 0.027). The mean ex novo manual delineation time for Structure Set C was approximately 74 minutes when inclusive of OARs and single nodal stations and 70 minutes when the CTV-E was considered as a single volume. Therefore, the mean segmentation saved time in Structure Set B was of 30 and 35 minutes, respectively.

Discussion

The need of replanning imaging in head and neck cancer is supported by several anatomical and dosimetric observations: e.g. Barker et al. showed that parotids are subject to significant shrinkage, while Wu et al. observed that if no replanning is done, the mean delivered dose to the parotids would be almost 10% higher than the initially planned one [5,18,19]. However, the huge amount of time spent recontouring target volumes and OARs cannot be omitted: that is why industry has recently started proposing image registration and autosegmentation software. These software still suffer from a lack of standard reliability indices and clinical validations.

In our study the comparison of Structure Set B versus C offered us the possibility to measure the agreement among senior doctors, to use as benchmark in the evaluation of the auto-segmentation software, exploiting the same DSI tool used for measuring the software performance. Quite surprisingly, we observed a mDSI of 0.80 for all Structure Set B and C comparisons. Considering that in Chao's [20] experience a value of 0.60–0.80 was recorded even between physician drawn contours, our results confirm that reaching a mDSI of 1 is not possible even between expert delineators. This observation is relevant in the validation process of auto-segmentation software: providing "0.8" with the meaning of a good performance value and not looking at the value of 1 as the needed benchmark of performance.

In our analysis we observed a better performance by mDSI analysis in the revision of the CTV proposal of the software when a single senior doctor reviewed it instead of the whole group of senior reviewers: the CTV passed from a mDSI of 0.9 when considering Structure Set A-B comparison, to 0.78 for A-C. This difference could be related to the consensus mechanism, which influenced of the reviewers final CTV-E delineation (Structure Set C), amplifying the differences among reviewers. Furthermore, we have to consider that in this study we are in a replanning setting, that means that the segmentation propagated in the replanning CT simulation was originated from an agreed segmentation. Similar considerations could be followed for the OARs: the mDSI was 0.87 for A-B comparison and 0.74 for A-C. The situation is different when looking at the nodal segmentation: the mean DSI difference between Structure Set B and C was 0.78 for CTV against 0.7 A-B and 0.63 A-C.

The software showed a lower performance in the segmentation of the individual nodal subsites, where mDSI values lower than 0.60 have been registered. This could be explained with the software difficulties to find some density-based reference to properly identify the outlines.

Overall, the manual correction of the Structure Set A allows a significant time saving with a mean value of almost 30 minutes (37% considering our ex novo head and neck mean segmentation time), agreeing the intra-patient automatic recontouring values of 26–47% expected by Chao et al. [20].

If we should consider the delineation of a single CTV-E, the advantage would be even greater: the mean ex novo segmentation (Structure Set C) time was 39.9 minutes (range 31.5–48.2 minutes) when compared to the manually corrected one (Structure Set B), which was of 12.5 minutes (range 6.3–25.8 minutes), with a statistically significant difference (p = 0.002).

Manual correction resulted to be better than the automatically proposed model also for the OARs with a lower advantage in terms of time spent which is statistically significant only for the oral cavity (p = 0.022), the encephalic trunk (p = 0.005) and the larynx (p = 0.028).

Literature is paying increasing attention to the need of establishing contouring independent check (IC) quality assurance procedures in order to guarantee more accurate treatments and dose assessment in RT planning and delivery [21,22,23]. In our department we have adopted for many years a contouring IC workflow where the initial manual segmentation by resident is always revised by a skilled physician. Using an autocontouring software can replace the role of the first delineator, and this could be even more advantageous in institutions lacking of in training physician personnel.

From this study we can define the manual correction of the Structure Set approach (option B) as the most convenient due to the important time saving [24], especially for nodal stations. When CTV-E for nodal subsite is considered at whole, the Structure Set A performance is anyway consistent with the DSI value of 0.8 among the senior doctors. A study to evaluate the dosimetric impact of planning the new dose distribution using Structure Set A vs B is ongoing.

In conclusion, we measured by mDSI an interobserver variability among senior doctors dedicated to delineation of a value of 0.80. In the frame of replanning it could represent a benchmark value for further investigations of autosegmentation software. The performance of the software was consistent with this benchmark, when whole CTV-E is considered. The time sparing supported the possibility to skip the first delineation by a young doctor in the frame of a quality assurance program for delineation, limiting the daily practice only at the independent check of the automated delineation. A dosimetric study is ongoing to evaluate the need of the peer review control.

Declaration of interest: The authors report no conflicts of interest. The authors alone are responsible for the content and writing of the paper.

References

[1] Cheng HCY, Wu VWC, Ngan RKC, Tang KW, Chan CCL, Wong KH. A prospective study on volumetric and dosimetric changes during intensity-modulated radiotherapy for nasopharyngeal carcinoma patients. Radiother Oncol 2012;104:317–23.

[2] Nishi T, Nishimura Y, Shibata T, Tamura M, Nishigato N, Okumura M. Volume and dosimetric changes and initial clinical experience of a two-step adaptive intensity modulated radiation therapy (IMRT) scheme for head and neck cancer. Radiother Oncol 2013;106:85–9.

[3] Ottosson S, Zackrisson B, Kjellen E, Nilsson P, Laurell G. Weight loss in patients with head and neck cancer during and after conventional and accelerated radiotherapy. Acta Oncol 2013;52:711–8.

[4] Zhao L, Wan Q, Zhou Y, Deng X, Xie C, Wu S. The role of replanning in fractionated intensity modulated radiotherapy for nasopharyngeal carcinoma. Radiother Oncol 2011;98:23–7.

[5] Schwartz DL, Garden AS, Shah SJ, Chronowski G, Sejpal S, Rosenthal DI, et al. Adaptive radiotherapy for head and neck cancer – Dosimetric results from a prospective clinical trial. Radiother Oncol 2013;106:80–4.

[6] Tsuji SY, Hwang A, Weinberg V, Yom SS, Quivey JM, Xia P. Dosimetric evaluation of automatic segmentation for adaptive IMRT for head-and-neck cancer. Int J Radiat Oncol Biol Phys 2010;77;3:707–14.

[7] Brouwer CL, Steenbakkers RJHM, van den Heuvel E, Duppen JC, Navran A, Bijl HP, et al. 3D variation in delineation of head and neck organs at risk. Radiat Oncol 2012;7:32.

[8] Moghaddasi L, Bezak E, Marcu LG. Current challenges in clinical target volume definition: Tumour margins and microscopic extensions. Acta Oncol 2012;51:984–95.

[9] Zhang T, Chi Y, Meldolesi E, Yan D. Automatic delineation of on-line head-and-neck computed tomography images: Toward on-line adaptive radiotherapy. Int J Radiat Oncol Biol Phys 2007;68:522–30.

[10] Castadot P, Lee JA, Parraga A, Geets X, Macq B, Grégoire V. Comparison of 12 deformable registration strategies in adaptive radiation therapy for the treatment of head and neck tumors. Radiother Oncol 2008;89:1–12.

[11] Geraghty JP, Grogan G, Ebert MA. Automatic segmentation of male pelvic anatomy on computed tomography images: A comparison with multiple observers in the context of a multicentre clinical trial. Radiat Oncol 2013;8:106.

[12] Stoiber EM, Schwarz M, Debus J, Bendl R, Giske K. An optimized IGRT correction vector determined from a displacement vector field: A proof of principle of a decision-making aid for re-planning. Acta Oncol Epub 2013 Apr 25.

[13] Grégoire V, Levendag P, Ang KK, Bernier J, Braaksma M, Budach V, et al. CT-based delineation of lymph node levels and related CTVs in the node-negative neck: DAHANCA, EORTC, GORTEC, NCIC, RTOG consensus guidelines. Radiother Oncol 2003;69:227–36.

[14] Castadot P, Geets X, Lee JA, Christian N, Grégoire V. Assessment by a deformable registration method of the volumetric and positional changes of target volumes and organs at risk in pharyngo-laryngeal tumors treated with concomitant chemo-radiation. Radiother Oncol 2010;95:209–17.

[15] van Kranen S, Mencarelli A, van Beek S, Rasch C, Sonke J, van Herk M. The accuracy of deformable registration for adaptive radiotherapy of head and neck cancer. Int J Radiation Oncol Biol Phys 2009;75:S573.

[16] Zou KH, Warfield SK, Bharatha A, Tempany CM, Kaus MR, Haker SJ, et al. Statistical validation of image segmentation quality based on a spatial overlap index. Acad Radiol 2004;11:178–89.

[17] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26:297–302.

[18] Barker JL, Garden AS, Ang KK, O'Daniel JC, Wang H, Court LE, et al. Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system. Int J Radiat Oncol Biol Phys 2004; 59: 960–70.

[19] Mohan R, Zhang X, Wang H, Kang Y, Wang X, Liu H, et al. Use of deformed intensity distribution for on-line modification of image-guided IMRT to account for interfractional anatomic changes. Int J Radiat Oncol Biol Phys 2005;61:1258–66.

[20] Chao KS, Bhide S, Chen H, Asper J, Bush S, Franklin G, et al. Reduce in variation and improve efficiency of target volume delineation by a computer-assisted system using a deformable image registration approach. Int J Radiat Oncol Biol Phys 2007;68:1512–21.

[21] Gambacorta MA, Valentini C, Dinapoli N, Boldrini L, Caria N, Barba MC, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using autosegmentation computed system. Acta Oncol Epub 2013 Jan 21.

[22] Breunig J, Hernandez S, Lin J, Alsager S, Dumstorf C, Price J, et al. A system for continual quality improvement of normal tissue delineation of radiation therapy treatment planning. Int J Radiat Oncol Biol Phys 2012; 83:e703–8.

[23] Rodrigues G, Louie A, Videtic G, Best L, Patil N, Hallock A, et al. Categorizing segmentation quality using a quantitative quality assurance algorithm. J Med Imaging Radiat Oncol 2012;56:668–78.

[24] Hardcastle N, Tomé WA, Cannon DM, Brouwer CL, Wittendrop PWH, Dogan N, et al. A multi-institution evaluation of deformable image registration algorithms for automatic organ delineation in adaptive head and neck radiotherapy. Radiat Oncol 2012;7:90.

Supplementary material available online

Supplementary Tables I to III.