Research article

Beyond geometrical overlap: a Dosimetrical Evaluation of automated volumes Adaptation (DEA) in head and neck replanning

Gian Carlo Mattiucci [a], Luca Boldrini [a,*], Lorenzo Placidi [b], Luigi Azario [b], Nicola Dinapoli [a], Giuditta Chiloiro [a], Danilo Pasini [a], Danila Piccari [a], Maria Antonietta Gambacorta [a], Mario Balducci [a], Giovanna Mantini [a], Vincenzo Valentini [a]

[a] Gemelli Advanced Radiation Therapy Center, Fondazione Policlinico Universitario "A. Gemelli", Università Cattolica del Sacro Cuore, Rome, Italy
[b] Medical Physics Department, Fondazione Policlinico Universitario "A. Gemelli", Università Cattolica del Sacro Cuore, Rome, Italy

ARTICLE INFO

ABSTRACT

Introduction: Automated target volumes adaptation could be useful in H&N replanning, but its dosimetric impact has not been analyzed.
Primary aim of this investigation is dose coverage assessment in fully automated and edited PTV adaptation settings, compared to manual benchmark.
Materials and methods: Ten IMRT patients were selected and replanning CTs were acquired.
A deformable registration with PTV adaptation was performed defining PTVA.
PTV B was obtained through manual editing and a benchmark PTV C was manually segmented by a delineation team.
The Dice Similarity Index (DSI) and the mean Hausdorff Distance (mHD) were calculated between PTV A and PTV C, and between PTV B and PTV C.
One IMRT plan was realized for each PTV: the plans optimized on PTV A and PTV B were proposed on PTV C to evaluate their dosimetric reliability compared to the benchmark plan in terms of PTV V95% dose coverage.
Results: The comparisons between PTV A with PTV C and PTV B with PTV C showed that the better DSI (high) and mHD values (low) are, the smaller difference when compared to PTV C V95% is described.
Evaluating plan A and B, PTV C V95% reduced by 6.1 ± 3.0% and by 4.1 ± 2.3% respectively when compared to plan C PTV C V95%.PTV B reaches acceptable dose coverage values (PTV V95% >95%) when DSI is >0.91 and a mHD < 0.17 mm and it has better results when compared to PTV A in 70%.
Discussion: The results show a correlation between the DSI-mHD and the PTV V95% variation, in the comparisons PTV A and PTV B vs PTV C.
Furthermore, we observed that PTV V95% coverage is higher in PTV B than in PTV A: the use of automated propagation may not be definitive and requires manual correction.
© 2017 Published by Elsevier Ireland Ltd on behalf of European Society for Radiotherapy & Oncology. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Introduction

The current therapeutic approach for head and neck (H&N) malignancies often requires radiation therapy treatments. Modern radiotherapy offers intensity modulated techniques (IMRT) that allow a steep gradient dose distribution on the target volumes (primary gross target volumes and/or nodal clinical target volumes), optimizing its coverage and increasing the preservation of the surroundings organs at risk (OaR) [1].

A reliable segmentation of the therapy subvolumes is therefore mandatory as it clearly has clinical consequences both for target and OaR [2–4].

This procedure is extremely time consuming for daily clinical activity, as it could last up to three hours for each patient [5].

Furthermore, it is prone to random and systematic errors linked to the existing and, to this day, still unavoidable inter-intraobserver contouring variability.

In this frame, the recently released autosegmentation software represent a step forward, offering an improvement of segmentation quality, a tighter adherence to the chosen segmentation guidelines which are recognized as benchmark ontology and a significative reduction of the daily activity time burden [6–13].

* Corresponding author at: Gemelli Advanced Radiation Therapy Center, Fondazione Policlinico Universitario "A. Gemelli", Università Cattolica del Sacro Cuore, Largo A. Gemelli, 8, 00168 Rome, Italy.
E-mail address: lucaboldrini@hotmail.it (L. Boldrini).

Some papers have been published in the last years about the morphological adequacy of contours obtained through the use of autosegmentation software in various anatomical districts, usually comparing them to manually expert drawn versions [14–18].

Analyzing these studies from a geometrical point of view we can affirm that these software seem to offer anatomically reliable contours, even if authors agree that a manual editing of them is still strongly recommended before their clinical use [5–8,13,14].

On the other hand, only few investigations take into account the dosimetrical aspects of the autocontouring approach [5,13,14–18].

Aim of this paper is to offer a dosimetric perspective besides the geometrical-anatomical similarity described in our previous experience about head and neck automated propagation, evaluating the link between similarity indices values (Dice similarity index and Mean Hausdorff distance) and dose coverage (in terms of PTV V95% and D99%) for target volumes in edited and non-edited automatic adaptation of target volumes compared to a selected manual benchmark in H&N replanning setting [7].

Materials and methods

Patient selection and volume definition

Ten consecutive patients with nasopharynx cancer treated with IMRT technique and selected for a previous study published by Mattiucci et al. were enrolled for this dosimetric investigation [7].

All patients were staged as Stage III–IV, and no neck surgery was performed before radiation therapy.

The median age was 53.9 years (range 30–82); eight were males and two females.

A helical CT scanner (GE HiSpeed DX/i Spiral) was used for image acquisition (slice thickness was 2.5 mm; no IV contrast agent was administered, according to our internal simulation protocol).

Five investigators performed the selection, delineation, deformation and correction steps of two patients each.

The simulation CTs were manually contoured on each axial slice using a commercial TPS (Eclipse®, Varian): ten nodal stations were segmented (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" [10].

An additional single CTV volume included all the lymphatic drainage stations (from Ia to VI) (see Supplementary Table 2 for CTV volume values).

The mean prescription dose to CTV primary (primary tumor) was 70.2 Gy, to CTV (drainage nodal stations) was 50.4 Gy, conventionally fractionated at 1.8 Gy.

A replanning CT was acquired during IMRT treatment: the median delivered dose was 30.6 Gy, mean value 36 Gy (range 21.6–59.4 Gy), considering a total prescribed dose of 70.2 Gy for the high dose volumes) [7].

A deformable registration between the simulation CT and the replanning one was then performed using VelocityAI 2.3© (Velocity Medical Solutions Inc.) transferring the CTV structures and obtaining CTV A which represents therefore the automatic propagation of the original target volumes.

These contours were then manually revised and edited by a skilled operator, drawing CTV B.

Furthermore, an independent ex novo CTV C was segmented on the replanning CT-scan by an expert delineation team, in order to limit inter- and intra-observer variability.

The delineator of CTV B has not been involved in this task to limit its influence.

A 5 mm margin was added to the three CTVs defining the nodal elective PTV A, PTV B and PTV C volumes.

All the dosimetric observations of this investigation have been done on these volumes (total prescription dose 50.4 Gy with 1.8 Gy fractionation) as they represent standard therapy volumes in all head and neck cancer RT treatments (nodal CTV) and are usually not anatomically affected by the presence of primary tumors.

The Dice Similarity Index (DSI) [14,19,20] and the mean Hausdorff Distance (mHD) [13,19] were calculated between A and C, and between B and C for each volume in order to quantify the existing geometric similarity between the different PTVs.

The Dice Similarity Index (DSI) is an overlap similarity index widely used for pairwise volume comparisons, while the Hausdorff distance (HD) is a similarity index that measures how far apart two groups of points are in a metric space, offering the exact topographic identification and visualization of disagreement areas [13].

In this study we decided to use the mean values of these indices in order to describe two different aspects of the existing similarity between contours.

Plan optimization and dosimetrical comparisons

An IMRT plan was realized for each PTV, obtaining plan A, B and C.

The dosimetrical acceptability of each plan was met when at least 95% of the dose covered the 95% of the target volume and maximum the 5% of the target volume received 105% of the dose.

Plan C was recognized as benchmark plan for dosimetrical comparisons.

Plans A and B were then proposed on PTV C in order to test the dosimetrical reliability of plans optimized on fully automatically adapted contours (PTV A) or after their manual editing (PTV B) and if they can achieve values obtained on a manual benchmark (Plan C) (Fig. 1).

For each of the three plans, dose coverage was evaluated computing the PTV C volume receiving at least 95% of the prescribed
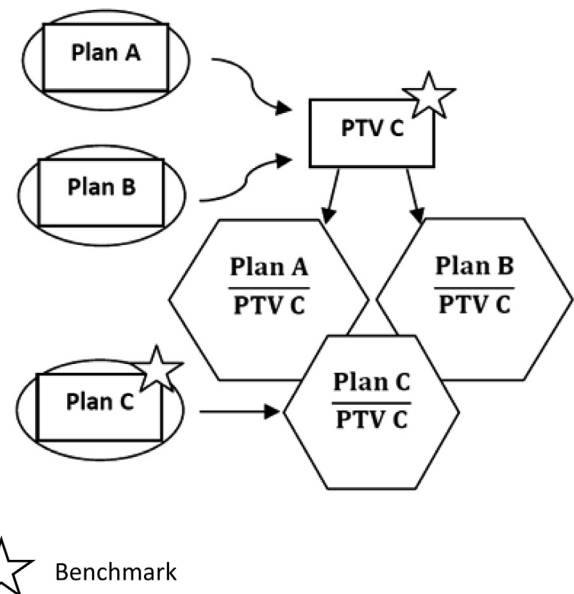

Fig. 1. IMRT plan realization and evaluation of plan A and B on benchmark PTV C.

dose (V95%) and the dose received by 99% of the PTV C itself (D99%) [5].

The PTVs dose coverage were calculated for each plan DVH using a in house software and the differences between plan A and B versus plan C was computed always using the benchmark C as the minuend of the subtraction.

Statistics

Statistical data analysis was executed using R software [21].

Wilcoxon test was used in order to evaluate statistically significant differences between DSI, mHD and PTV volume variation between PTV A and B vs PTV C.

Coefficient of determination or R^2 was applied to assess how data points fit a linear regression, in particular between DSI and mHD.

A non-parametric Spearman's correlation test was performed to estimate the correlation between observed difference in the PTV C V95% and D99% vs DSI and mHD.

Results

Geometrical evaluation

Statistically significant differences ($p < 0.05$ resulted by a Wilcoxon signed rank test) were observed in the evaluation of the geometric overlap (DSI and mHD) between PTV A and B in comparison with the benchmark PTV C respectively. The registered mean values and ranges are reported in Table 1.

A strong linear correlation has been identified between DSI and mHD ($R^2 = 0.91$) (Fig. 2), considering all data point from the overlap between PTV A and PTV B, compared to PTV C.

A statistical significant correlation was found ($p = 0.001$) in a Spearman's correlation test.

Fig. 3 shows the correlation between DSI and PTV volume reduction (cc) of PTV A and PTV B, compared to PTV C.

A statistical significant correlation was found ($p = 0.006$) in a Spearman's correlation test.

The mean PTV volume variation (cc) between PTV C and PTV A is 60.6 cc (range 25.7–130.1 cc) and between PTV C and PTV B is 37.8 cc (range 5.9–72.2 cc).

Dosimetrical evaluation

The observed mean V95% (±SD) for PTV C was 97.9% ± 1.2 and its dose coverage difference (both in terms of V95% and D99%) between Plan C vs Plan A and Plan C vs Plan B resulted to be correlated to the DSI and mHD values of their comparisons, as shown in Figs. 4 and 5.

In Table 2 the dose coverage differences, obtained through the plan A and plan B evaluation on PTV C, are reported in terms of PTV C ΔV95% and ΔD99%.

Evaluating plan A and B, PTV C V95% reduced by 6.1 ± 3.0% and by 4.1 ± 2.3% respectively when compared to plan C PTV C V95% (Fig. 6).
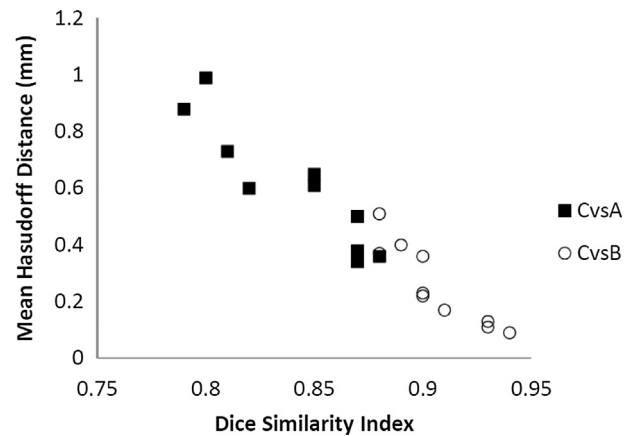
Fig. 2. Correlation between the DSI and mHD distance comparing the PTV C-PTV A (■) and PTV C-PTV B (○) overlap. Linear correlations between the DSI and mHD comparing PTV C-PTV A and PTV C-PTV B result, respectively, in a $R^2 = 0.84$ and $R^2 = 0.82$.
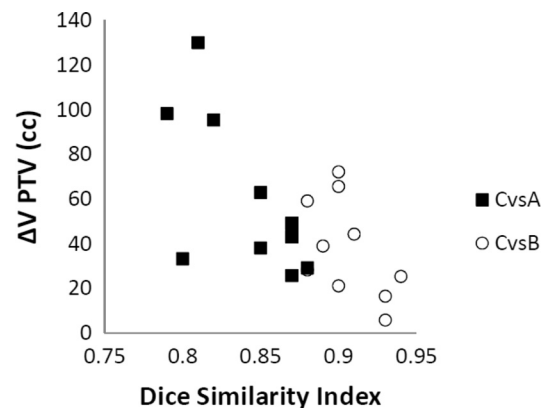
Fig. 3. Correlation between the DSI and PTV volume reduction (cc) comparing PTV C and PTV A (■) and between PTV C and PTV B (○).
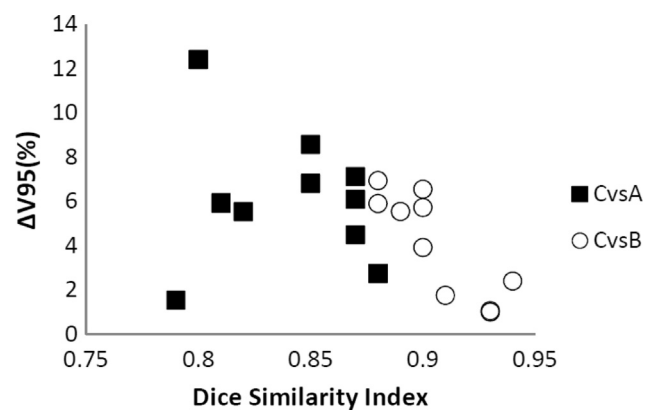
Fig. 4. Correlation between the DSI and PTV C V95% coverage difference for Plan A (■) and for Plan B (○).

On the other hand, no substantial variations (<1.0%) were observed on the mean reduction in PTV C D99% between plan C and A, and between plan C and B.

Therefore, the editing of the automatically propagated contours leads to a better optimization of the PTV V95% when compared to fully autopropagated structures: the PTV C V95% is higher in plan B than in plan A in 70% of the cases, while in the remaining 30% only

Table 1
Summary of the mean DSI (mDSI) and mHD (mm) between the autocontoured (A) and the manually edited (B) PTVs, compared to the manual ex novo ones (C) for the ten patients.

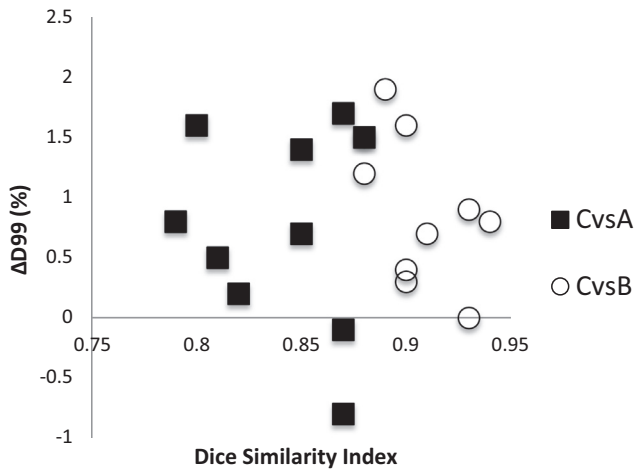|	PTV C vs A mean (range)	PTV C vs B mean (range)
mDSI	0.86 (0.88/0.79)	0.91 (0.94/0.88)
mHD (mm)	0.61 (0.99/0.34)	0.26 (0.51/0.09)

Fig. 5. Correlation between the DSI and PTV C D99% coverage difference for Plan A (■) and for Plan B (○).

Table 2
Summary of PTV C V95% and D99% coverage difference between Plan C and Plan A and between Plan C and Plan B; all evaluated on PTV C.

	Plan C vs Plan A on PTV C mean (range)	Plan C vs Plan B on PTV C mean (range)
ΔV95 (%)	6.1 (12.4/1.5)	4.1 (6.9/1.0)
ΔD99 (%)	0.8 (1.7/-0.8)	0.9 (1.9/0.0)

a small difference (1–2%) can be seen. However, PTV B does not always reach the PTV C V95% dose coverage and never reaches the manual benchmark values (Plan C on PTV C).

No trends can be instead observed for PTV C D99% and the mean variation between plan B and A D99% is 0.4% (range −0.1/1.1%).

In conclusion, the PTV B approach is clinically more reliable (especially when evaluated on the PTV V95% parameters), but never reaches Plan C. In terms of PTVC V95%, PTV B has better results when compared to A (3.3% average increase of dose coverage).

A statistically significant correlation was found in the PTV C V95% but not in the PTV C D99% versus the DSI ($p = 0.005$ and $p > 0.1$ respectively) when comparing Plan C vs Plan B.

Similar results were observed when comparing the PTV C V95% and PTV C D99% versus the mHD ($p = 0.013$ and $p > 0.1$ respectively).

On the other hand, no statistically significant correlation was found comparing Plan C and Plan A even in terms of PTV C V95% and D99% versus the DSI and the mHD (p values always higher than 0.1).

Mean DSI (mHD) resulted to be positive (negative) correlated to PTV V95% and D99% in 70% of the cases.

D2% values are reported in supplementary Table 1.

This correlation shows that the better DSI and mHD values are, the smaller benchmark PTV C V95% difference is described when comparing Plan A and Plan B to Plan C.

More specifically, PTV B reaches a PTV C V95% >95% when DSI between PTV B and PTV C is higher than 0.91 and the correlated mHD, smaller than 0.17 mm.

Table 3 summarizes the volumetric data of PTV C and the V95% data of its comparisons.

Conclusions

The segmentation of the therapy volumes for head and neck malignancies in IMRT setting is one of the most demanding and time consuming tasks in daily clinical practice.

This situation has brought the industry to propose the use of the autocontouring software as a reliable clinical tools.

As we demonstrated in our previous experience, autocontouring can lead to a significant time saving with a mean value of almost 30 min (37% considering our ex novo mean segmentation time), confirming the intra-patient automatic recontouring values of 26–47% expected by Chao et al. [7–22].

This reduction of time burden could also be linked to a segmentation quality improvement as we observed that the existing DSI for CTV between the manually edited autosegmented contours and the manual ex novo ones was 0.78, which represents a good score for this similarity index, against a value of 0.74 for the
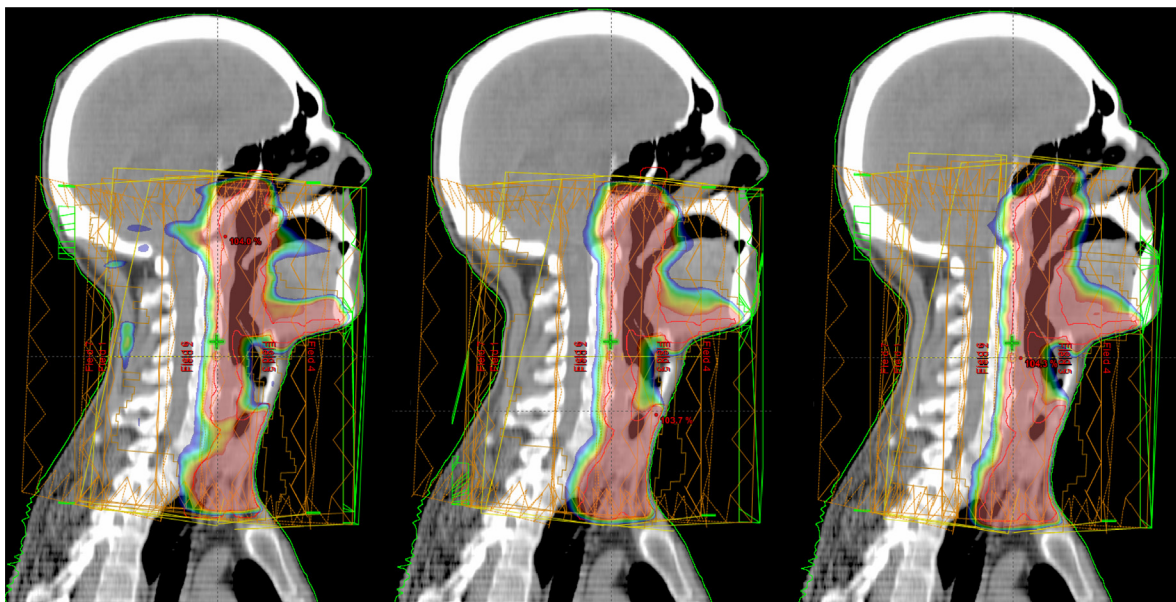

Fig. 6. PTV dose coverage (isodose levels: 75–95%) for different plans. Left: plan A, middle: plan B, right: plan C (benchmark) on benchmark structure set.

Table 3
PTVs volumetric data and V95% values.

	PTV C volume (cc)	PTV A volume (cc)	PTV B volume (cc)	PTV C V95%	PTV A vs C V95%	PTV B vs C V95%
Patient 1	871.7	776.3	812.6	98.4	92.8	92.4
Patient 2	781.2	732.1	759.9	96.5	89.4	90.8
Patient 3	777.7	734.3	712.1	99.1	93.0	92.5
Patient 4	583.5	550.1	555.1	95.1	82.7	88.1
Patient 5	816.7	686.6	744.5	97.4	91.4	93.5
Patient 6	603.1	641.3	619.8	98.6	91.8	97.6
Patient 7	756.9	731.1	750.9	98.1	93.6	97.0
Patient 8	888.4	825.5	849.3	99.3	90.8	93.8
Patient 9	972.3	874.1	927.8	97.9	96.3	96.1
Patient 10	852.2	822.9	826.8	99.5	96.8	97.1

comparison between autosegmentation and manual ex novo contouring [7].

Therefore, in our previous experience, the DSI showed us that there is a large amount of similarity between the proposed contours but we did not analyze its dosimetrical aspects.

Thus we decided to test them, in order to clarify the real clinical safety, the limits and the main strengths of the autosegmentation software.

The results we obtained show that some dosimetric differences exist between a treatment plan optimized on manual contours edited after their automated propagation and the corresponding benchmark manual ones, even if the available similarity indices show an encouraging overlap in an anatomic site where very high interobserver contouring variation has been described for common simulation CT imaging [23].

Regarding the PTV volumes difference (cc) correlated with DSI (or mHD) when comparing PTV C with A and B, PTV B showed to be better than PTV A and always closer to PTV C, even if it never reaches the benchmark values.

For what concerns PTV volumetric differences (cc), DSI (which appeared to be closely related to mHD) is able to discriminate PTV A and B in two clear regions, as showed in Fig. 3.

A similar statement could be made regarding PTV C ΔV95% and ΔD99%.

Mean DSI values (or, once more, mHD ones) are able to discriminate plan A and B in two different regions, if considering PTV C ΔV95% or ΔD99%, as shown in Figs. 4 and 5.

Nevertheless, different PTV C ΔV95% and ΔD99% values can be correlated to a single DSI value, strongly limiting its dosimetrical predictive power.

On the other hand PTV B shows a correlation between similarity indices values (more specifically DSI and mHD) and dosimetric parameters (such as PTV C V95%) and seems to be coherent with the assumption that for high values of DSI and low mHD ones, low values of ΔV95% or ΔPTV C cc are associated, at least in 70% cases.

This observation cannot be done for PTV A, where no statistically significant correlation was registered. Furthermore D99%, by its very definition, does not correlate with the used similarity indices values (DSI and mHD).

The manual editing of the automatically propagated contours is therefore mandatory in order to make the proposed segmentation more adherent to the recognized manual benchmark and to allow a more reliable and clinically safer dose distribution.

On the other hand, as stated also by Voet et al. [5], having a high geometrical overlap cannot adequately predict a reliable dose coverage, which in our experience appeared to be met only above a DSI of 0.91 and with mHD smaller than 0.17 mm, which represent extreme values very hard to be reached in daily clinical practice.

With the limits of a small patients sample, we did not even recognize a safe dose distribution for DSI of 0.85 which represents the expert based interobserver delineation variability benchmark value for target volumes in our institution [7].

Autosegmentation software can therefore provide a quick solution to the segmentation time burden in daily activity but, to date, they cannot offer target volumes suitable for direct plan calculation in H&N replanning setting and a careful manual editing remains mandatory in order to get as close as possible to manual benchmark dose distributions.

Furthermore the Dice similarity index and the mean Hausdorff distance, which represent the most used similarity indices at the moment, despite offering very different information about the existing overlap between two structures, do appear to be strongly correlated.

This correlation limits their dosimetrical predictive power and they should be integrated with other indices that take into account the spatial disposition of the existing non overlapping areas and tested on more numerous contour samples.

Conflict of interest

The authors of this article report no conflict of interest.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.tipsro.2017.06.002.

References

[1] Nelms BE, Tomé WA, Robinson G, Wheeler J. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. Int J Radiat Oncol Biol Phys 2012;82:368–78.
[2] Teguh DN, Levendag PC, Voet PW, et al. Clinical validation of atlas-based autosegmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys 2011;81:950–7.
[3] Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother Oncol 2012;102:68–73.
[4] Young AV, Wortham A, Wernick I, Evans A, Ennis RD. Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. Int J Radiat Oncol Biol Phys 2011;79:943–7.
[5] Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother Oncol 2011;98:373–7.
[6] Gambacorta MA, Valentini C, Dinapoli N, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. Acta Oncol 2013;52:1676–81.
[7] Mattiucci GC, Boldrini L, Chiloiro G, et al. Automatic delineation for replanning in nasopharynx radiotherapy: what is the agreement among experts to be considered as benchmark? Acta Oncol 2013;52:1417–22.
[8] La Macchia M, Fellin F, Amichetti M, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. Radiat Oncol 2012;7:160.
[9] Eriksen JG, Salembier C, Rivera S, et al. Four years with FALCON – an ESTRO educational project: achievements and perspectives. Radiother Oncol 2014;112:145–9.

[10] Grégoire V, Ang K, Budach W, et al. Delineation of the neck node levels for head and neck tumors: a 2013 update. DAHANCA, EORTC, HKNPCSG, NCIC CTG, NCRI, RTOG, TROG consensus guidelines. Radiother Oncol 2014;110():172–81.

[11] Nijkamp J, de Haas-Kock DF, Beukema JC, et al. Target volume delineation variation in radiotherapy for early stage rectal cancer in the Netherlands. Radiother Oncol 2012;102:14–21.

[12] Boersma LJ, Janssen T, Elkhuizen PH, et al. Reducing interobserver variation of boost-CTV delineation in breast conserving radiation therapy using a preoperative CT and delineation guidelines. Radiother Oncol 2012;103:178–82.

[13] Boldrini L, Damiani A, Valentini V. In: Principles and clinical applications of autocontouring software. Milano: FrancoAngeli; 2014.

[14] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiother Oncol 2014;112:317–20.

[15] van der Leij F, Elkhuizen PH, Janssen TM, et al. Target volume delineation in external beam partial breast irradiation: less inter-observer variation with preoperative-compared to postoperative delineation. Radiother Oncol 2014;110:467–70.

[16] Conson M, Cella L, Pacelli R, et al. Automated delineation of brain structures in patients undergoing radiotherapy for primary brain tumours: from atlas to dose-volume histograms. Radiother Oncol 2014;112:326–31.

[17] Walker GV, Awan M, Tao R, et al. Prospective randomized double-blind study of atlas-based organ-at-risk auto segmentation-assisted radiation treatment planning in head and neck cancer. Radiother Oncol 2014;112:321–5.

[18] 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;77:707–14.

[19] Fotina I, Lütgendorf-Caucig C, Stock M, Pötter R, Georg D. Critical discussion of evaluation parameters for inter-observer variability in target definition for radiation therapy. Strahlenther Onkol 2012;188:160–7.

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

[21] R (software), The R foundation for Statistical Computing, version 3.1.2.

[22] Chao KS, Bhide S, Chen H, 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.

[23] Rasch CR, Steenbakkers RJ, Fitton I, et al. Decreased 3D observer variation with matched CT-MRI, for target delineation in Nasopharynx cancer. Radiat Oncol 2010;15(5):21.