Physics and Imaging in Radiation Oncology 32 (2024) 100654

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/physics-and-imaging-in-radiation-oncology

Original Research Article

Head and neck automatic multi-organ segmentation on Dual-Energy Computed Tomography

Anh Thu Lê [a], Killian Sambourg [a], Roger Sun [a,b], Nicolas Deny [b], Vjona Cifliku [a], Rahimeh Rouhi [a], Eric Deutsch [a,b], Nathalie Fournier-Bidoz [b,1], Charlotte Robert [a,b,*,1]

[a] *Université Paris-Saclay, Gustave Roussy, Inserm, Molecular Radiotherapy and Therapeutic Innovation, U1030, 94800 Villejuif, France*
[b] *Department of Radiation Oncology, Gustave Roussy Cancer Campus, Villejuif, France*

ARTICLE INFO

Keywords:
Dual-Energy Computed Tomography
Automatic segmentation
Multi-organ segmentation
Head-and-neck cancer
Virtual monoenergetic images

ABSTRACT

Background and purpose: Deep-learning-based automatic segmentation is widely used in radiation oncology to delineate organs-at-risk. Dual-energy CT (DECT) allows the reconstruction of enhanced contrast images that could help with manual and auto-delineation. This paper presents a performance evaluation of a commercial auto-segmentation software on image series generated by a DECT.
Material and methods: Different types of DECT images from seventy four head-and-neck (HN) patients were retrieved, including polyenergetic images at different voltages [80 kV reconstructed with a kernel corresponding to the commercial algorithm DirectDensity™ (PEI80-DD), 80 kV (PEI80), 120 kV-mixed (PEI120)] and a virtual-monoenergetic image at 40 keV (VMI40). Delineations used for treatment planning were considered as ground truth (GT) and were compared with the auto-segmentations performed on the 4 DECT images. A blinded qualitative evaluation of 3 structures (thyroid, left parotid, left nodes level II) was carried out. Performance metrics were calculated for thirteen HN structures to evaluate the auto-contours including dice similarity coefficient (DSC), 95th percentile Hausdorff distance (95HD) and mean surface distance (MSD).
Results: We observed a high rate of low scores for PEI80-DD and VMI40 auto-segmentations on the thyroid and for GT and VMI40 contours on the nodes level II. All images received excellent scores for the parotid glands. The metrics comparison between GT and auto-segmented contours revealed that PEI80-DD had the highest DSC scores, significantly outperforming other reconstructed images for all organs ($p < 0.05$).
Conclusions: The results indicate that the auto-contouring system cannot generalize to images derived from DECT acquisition. It is therefore crucial to identify which organs benefit from these acquisitions to adapt the training datasets accordingly.

1. Introduction

Most head-and-neck cancer (HNC) patients are treated with radiotherapy (RT). Following the treatment, the patient can be subject to locoregional recurrences [1] but also several side effects that can affect their quality of life [2]. To prevent those complications, ongoing efforts are being made to define organs at risk (OAR) and target volumes more precisely [3]. However, this stage is a time-consuming task, subjected to high intra and inter-observer variability [4]. Variation in delineation can have many origins, in particular difficulties in interpretation due to the limited contrast of conventional CT acquisitions [5], generating disagreement about the disease extensions [6]. To address this variability, different guidelines for HNC delineation exist [7] but differences in clinical practices still remain. Intra and inter-observer variability can be evaluated with metrics like the dice similarity coefficient (DSC). According to Van der Veen et al. (2021) [8], results in DSC can vary considerably between physicians ($n = 22$) even for structures with good mean agreement such as submandibular glands [median DSC = 0.8 (0.5–0.9)]. For target volumes, several recent studies have attempted to measure this variability [9,10,11] with a comparison of the clinical target volumes delineations and the same conclusion can be drawn, highlighting the need to use complementary imaging modalities (Magnetic Resonance Imaging, Positron Emission Tomography) and to refine the guidelines. However, the use of a complementary imaging modality

[*] Corresponding author at: Department of Radiotherapy, Gustave Roussy Cancer Campus, 114 rue Edouard Vaillant, 94800 Villejuif, France.
E-mail address: CH.ROBERT@gustaveroussy.fr (C. Robert).
[1] Nathalie Fournier-Bidoz and Charlotte Robert have contributed equally to this work and share last authorship.

https://doi.org/10.1016/j.phro.2024.100654
Received 24 June 2024; Received in revised form 17 September 2024; Accepted 26 September 2024
Available online 30 September 2024

introduces additional uncertainties due to multi-modal image registration.

Dual-energy CT (DECT) has already been used in diagnostic radiology since the mid-2000 s and is now subject to a growing interest in RT [12–14]. By acquiring images at 2 complementary voltages, this technology makes it possible to reconstruct different types of images such as polyenergetic images, mixed images, or virtual-monoenergetic images (VMI) [15]. A mixed image corresponds to a linear combination of the two acquired image stacks, making it possible to recover 120 kVp standard CT. VMI of low energies have an increased contrast for iodine, soft tissue and nodes while VMI of higher energies reduce artifacts around bones or high density materials at the expense of the contrast [16]. Usually, VMI at 40–65 keV are presented as the images with the most optimal contrast for the visualization of structures such as lymph nodes (LN), vessels and tumors [17,18].

The clinical implementation of this technology could facilitate and harmonize manual segmentation [19], but also opens up new prospects for auto-segmentation [20]. Nowadays, deep-learning (DL) based auto-segmentation is broadly implemented in clinics for OARs [21]. However, such tools were mainly trained on single energy CT (SECT) images [22], with limited soft tissue contrast [23]. As an example, in the study by Heilemann et al. (2023) [24], three auto-segmentation tools were tested in all OARs and compared to physicians' segmentation. The mean DSC for all organs were above 0.74 for each tool but small or thin structures like chiasm, cochlea and optic nerves had a DSC below the acceptance level (0.7). This shows that auto-segmentation still has room for improvement especially with the prospect of target volumes delineation [25].

This study was conducted to test, in a single-center setting, the performance of a commercial auto-segmentation software on images derived from a DECT scanner. The aim of this work is to evaluate the ability of auto-segmentation tools trained on conventional polyenergetic images to adapt to DECT images.

2. Material and methods

2.1. Patient cohort and DECT image series description

A retrospective cohort of seventy four HNC patients with squamous cell lesions treated in the RT department of Gustave Roussy (Villejuif, France) between October and December 2023 was gathered (Table 1). All patients benefited from a DECT acquisition with the Siemens SOMATOM go.Sim® CT scanner (Siemens Healthineers, Forcheim, Germany). The go.Sim® is a single source CT on which the acquisition is performed in 2 consecutive passages: 80 kVp and 140 kVp with a tin filter for spectral separation. All acquisitions were performed in treatment position. A 90 mL biphasic contrast medium injection of 320 mg Iodine/mL (Visipaque, GE Healthcare, Velizy, France) was used for 87.8 % of the patients. First, 40 mL at 1.5 mL/s was injected followed by a 3 min break, then the remaining 50 mL of injection was administered at 1.5 mL/s. The DECT acquisition started 5 s before the end of the second injection. Acquisition pitch and rotation time were 0.8 and 0.5 s respectively. For all reconstructions, slice thickness and increment were set to 2 mm, and iMAR® option was applied for metal dental implant artifact correction.

Table 1
Description of the patient cohort.

Characteristics	Cohort (n = 74)
Tumor Location	
Oral cavity	10 (13.5 %)
Oropharynx	34 (45.9 %)
Larynx	8 (10.8 %)
Hypopharynx	4 (5.4 %)
Nasopharynx	6 (8.1 %)
Salivary glands	3 (4.1 %)
Other	9 (12.2 %)
TNM Stage	
Stage I	2 (2.7 %)
Stage II	1 (1.4 %)
Stage III	16 (21.6 %)
Stage IVA	37 (50.0 %)
Stage IVB	10 (13.5 %)
NA	8 (10.8 %)
Surgery	
Yes	32 (43.2 %)
No	42 (56.8 %)
CT with injected contrast	
Yes	65 (87.8 %)
No	9 (12.2 %)

Overall, 4 series per patient were retrieved for the study (Supplementary Fig. S1):

- a polyenergetic image at 80 kV with Sm36 kernel corresponding to the Direct-Density (DD) algorithm® (PEI80-DD). This algorithm transforms Hounsfield units (HU) to get a unique HU-to-mass density lookup table that is independent of the acquisition kV. DD kernel reduces contrast of high density materials so the image resembles a non-contrast enhanced image. The field-of-view (FOV) was set to 600 mm to include external body contours and immobilization devices.
- a polyenergetic series at 80 kV with Br36 kernel (PEI80): the Br36 (Body regular) filter restores the iodine contrasting effects. In clinics, this series is used by physicians to see vessels and should enable the doctors to check and correct auto-segmentation. The FOV was set to 500 mm.
- a VMI at 40 keV with Qr40 kernel (VMI40): this spectral image series is reconstructed with a Quantitative regular kernel; it was selected because of its high contrast potential for soft tissues, especially to visualize the gross tumor volumes (GTV) [17].
- a mixed image equivalent to a SECT image at 120 kV with Qr40 (PEI120); we chose to include this series in our study as 120 kV is broadly used in RT for delineation and treatment planning.

The utilization of the retrospective cohort was performed under the General Data Protection Regulation (GDPR) and approved by the Institutional Review Board (n° IRB2023-278 Gustave Roussy cancer campus).

2.2. Organs at risk delineation

OARs were delineated with the help of ART-Plan® Annotate™ (TheraPanacea, France), a CE-marked solution for automatic annotation of OARs in RT. In this study, ART-Plan v1.11.5 was used. According to the information provided by the manufacturer, the training set (>300 patients per structure) included high-voltage scanners (100–140 kVp) performed with and without injection of contrast medium for organs located in the HN region.

In the clinical workflow, OARs auto-segmentation was performed on PEI80-DD series then imported into our treatment planning software, RayStation V2023B (RaySearch, Stockholm, Sweden). A resident verified and corrected the contours when necessary using PEI80 and VMI40 according to atlases such as Gregoire et al. (2014) [26]. Then, a senior physician double checked the contours. The final contours obtained, attached to PEI80-DD, served for dose calculation and are referred to as GT (Ground Truth) in this paper. A total of 5 residents and 5 seniors were involved in the contouring process.

2.3. Structures of interest

A list of thirteen structures of interest for the study was derived from discussions with the physicians: 10 left and right LN areas (I, II, III, IV and V), the parotid glands and the thyroid. The parotid glands were considered as reference volumes, given the good performance of auto-contouring software reported in the literature [27]. LN areas are often poorly segmented by automatic segmentation tools, and are therefore almost always corrected by doctors [28]. The same goes for the thyroid which can be wrongly auto-segmented due to its proximity with blood vessels.

2.4. Qualitative assessment

First, we established a blinded evaluation to qualitatively assess the contours. We evaluated clinical contours and those resulting from auto-segmentation on PEI80-DD, PEI80, PEI120, and VMI40 (i.e., 5 contours per organ) for 3 OARs (thyroid, left parotid, and left LN II). Thirty patients from the 74 included in the cohort were selected for this analysis, resulting in the evaluation of one hundred fifty contours. Two doctors (NS referred as doctor 1: 4 years of expertise; RS referred as doctor 2: 3 years of expertise) were asked to grade the contours (A = no correction needed, B = minor corrections, C = major corrections needed). They had access to the 4 DECT series without knowing on which series the contours had been auto-segmented and were allowed to choose the CT series on which they would evaluate the contours.

2.5. Metrics for quantitative performance evaluation

To assess the segmentation results, 3 metrics were analyzed: the DSC [29], the 95th percentile Hausdorff distance (HD95) [30] and the mean surface distance (MSD) [31]. All the metrics were computed using the deepmind Python library [32].

2.6. Statistical analysis

For each OAR, the distribution of DSC, 95HD and MSD of the 4 image series were compared using a Wilcoxon signed-rank test [33] with a level of significance set to 0.05.

3. Results

3.1. Qualitative assessment of the automatic delineation of OARs by doctors

Doctor 1 chose to work on both PEI80 and VMI40 to evaluate the 5 contour sets while doctor 2 preferred PEI80. Scores resulting from the blinded evaluation of all the contours are reported in Fig. 1. There was a difference in marks between doctor 1 (Fig. 1a) and 2 (Fig. 1b), with more B marks for doctor 1 in general. C scores attributed by the 2 doctors were close (±2) except for GT and VMI40 node contours. For the thyroid, the auto-contour based on PEI80-DD presented the most C scores (doctor 1 and 2: 12), followed by VMI40 (doctor 1: 10, doctor 2: 8); and PEI80 the most A-score (doctor 1: 9, doctor 2: 12) beside the GT (doctor 1 and 2: 13). GT contours were categorized as C in 7 and 5 cases by doctor 1 and doctor 2 respectively. The left parotid had 1 to 3 C-scores for each contour. Concerning the left LN area II, GT got the most C (doctor 1: 10, doctor 2: 6), closely followed by the auto-contours of VMI40 (doctor 1: 9, doctor 2: 6).

3.2. Performance metrics and statistical analysis

Four percent of auto-contoured structures had a DSC ≥ 0.99 and 95HD and an MSD ≤ 0.01 mm on the PEI80-DD series, suggesting that they were not corrected by the doctors.

The auto-segmentation results summarized in Fig. 2 show box plots of DSC (a), 95HD (b) and MSD (c) from the seventy four patients per structure. This comparison reveals that GT vs PEI80-DD had the highest scores in terms of median DSC ranging from 0.68 for IV R to 0.91 for parotid R (Supplementary Table S2). The statistical analysis (Fig. 3) confirms that GT vs PEI80-DD DSC results were significantly different from all the other DSC distributions for all organs ($p < 0.05$). The second series which had the closest contours to the GT in terms of DSC was PEI80 with median scores varying from 0.64 for structure IV R to 0.87 for parotid R. However, GT vs PEI80 DSC distribution cannot be considered significantly different from PEI120 and VMI40 for LN areas ($p > 0.05$).

Fig. 2(b) and (c) show 95HD and MSD results without the outliers (95HD > 45 mm and MSD > 20 mm). GT vs PEI-80DD holds the shortest median 95HD for majority of the structures ranging from 2.3 to 9.1 mm respectively for parotid R and IV R LN areas. The results could not be

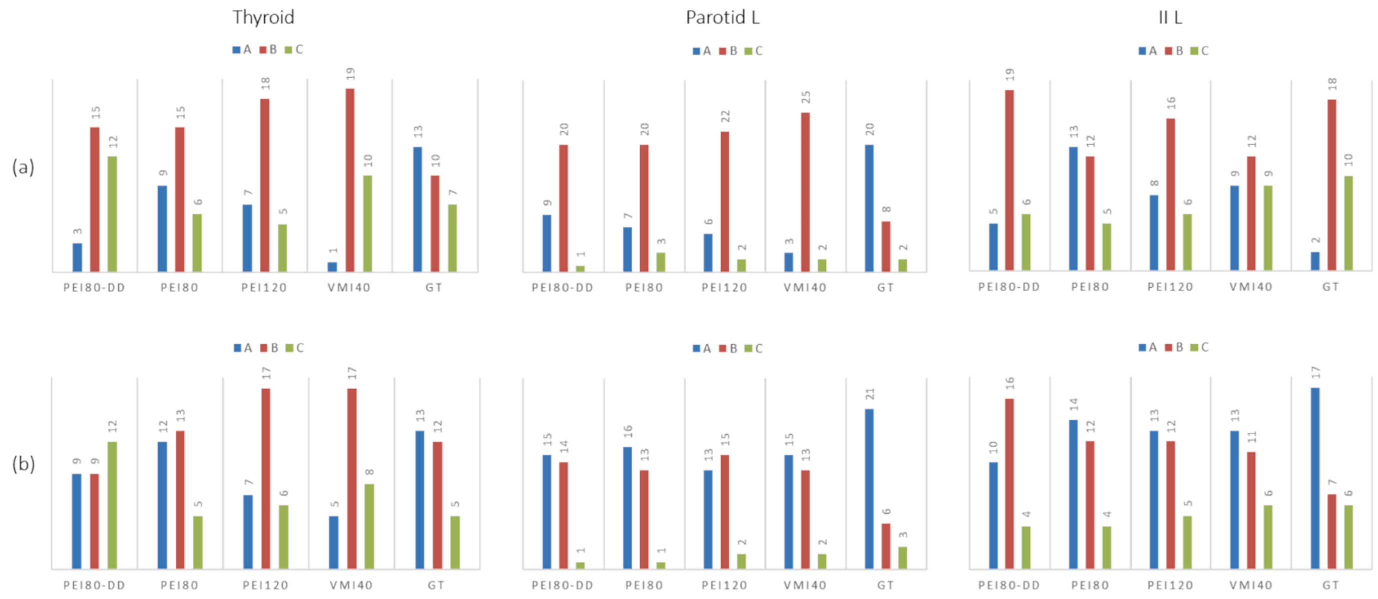

Fig. 1. Results of blinded qualitative assessment of automatic contours as obtained by ART-PLAN® and GT contours on the thyroid, parotid L and II L (L: Left) lymph node area by doctor 1 (a) and doctor 2 (b) – data from thirty patients were analyzed in this study.

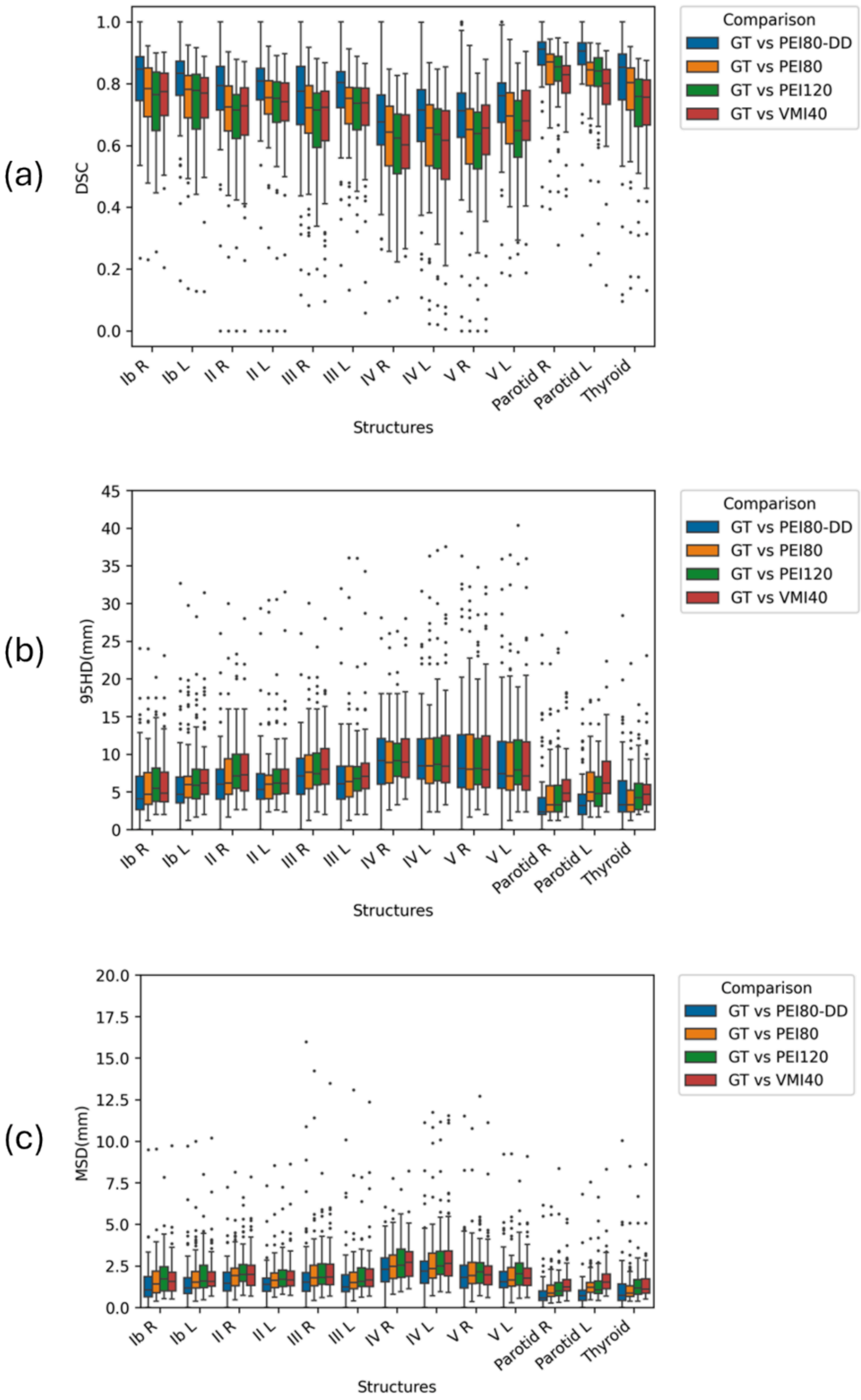

Fig. 2. Performance results of automatic segmentations on the series set from 13 different structures in terms of DSC, 95HD and MSD metrics, respectively shown in sub-figures (a), (b) and (c) (L: Left, R: Right).

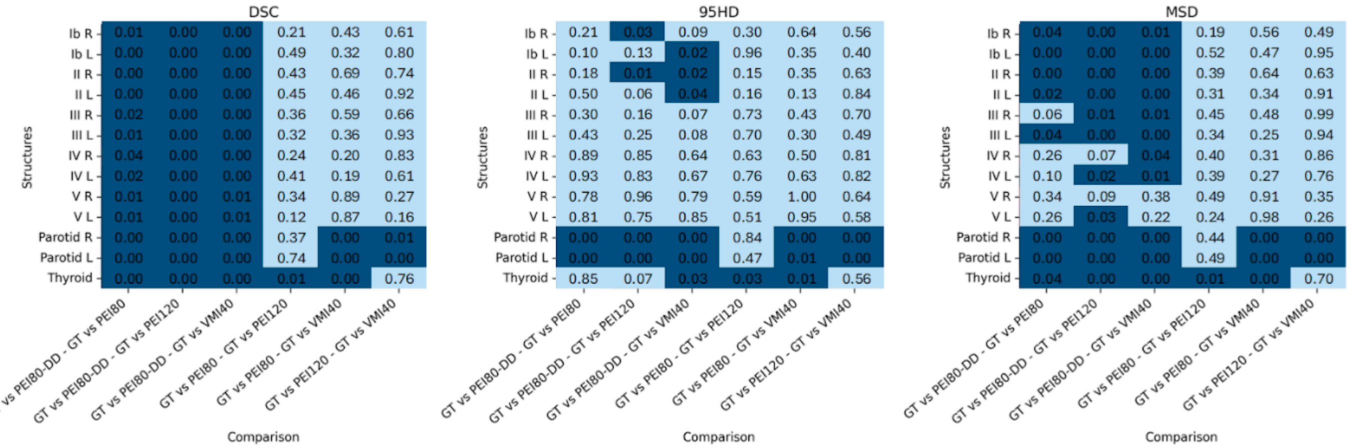

Fig. 3. P-values (Dark blue: $p < 0.05$, light blue $p > 0.05$) resulting from the Wilcoxon test on all distribution metrics (DSC, 95HD and MSD) for all 13 structures.

considered significantly different ($p > 0.05$) for most LN areas and thyroid between GT vs PEI-80DD and GT vs the other images (Fig. 3). In terms of MSD, the difference in performance between GT vs PEI80-DD and the other series set seemed similar to DSC results for all the structures except LN III, IV and V for which PEI80-DD was not significantly different from the others ($p > 0.05$): its median ranged between 1.1–1.5 mm for the nodes level I, II and III and was <0.8 mm for the parotids and thyroid.

Fig. 4 shows illustrative cases with the greatest positive and negative difference of DSC between GT vs PEI80-DD and GT vs the other series, and an unmodified case (DSC = 0.99) for the thyroid. In Fig. 4(a), corresponding to the best DSC case for PEI80-DD (DSC = 0.89), the auto-segmentation on the other images shows underestimated volumes. Fig. 4(b) illustrates one of the cases where auto-segmentation performed poorly on PEI80-DD (DSC = 0.66) on the thyroid compared to the other series (DSC = 0.81–0.87). PEI80 auto-segmentation included the carotid artery. In Fig. 4(c), being the unmodified PEI80-DD contour case (DSC = 0.99), the thyroid is not enhanced in any of the series. The algorithm had also included the carotid artery in the segmentation for PEI80-DD, PEI80 and PEI120 that was not corrected by the physician. In the present case,

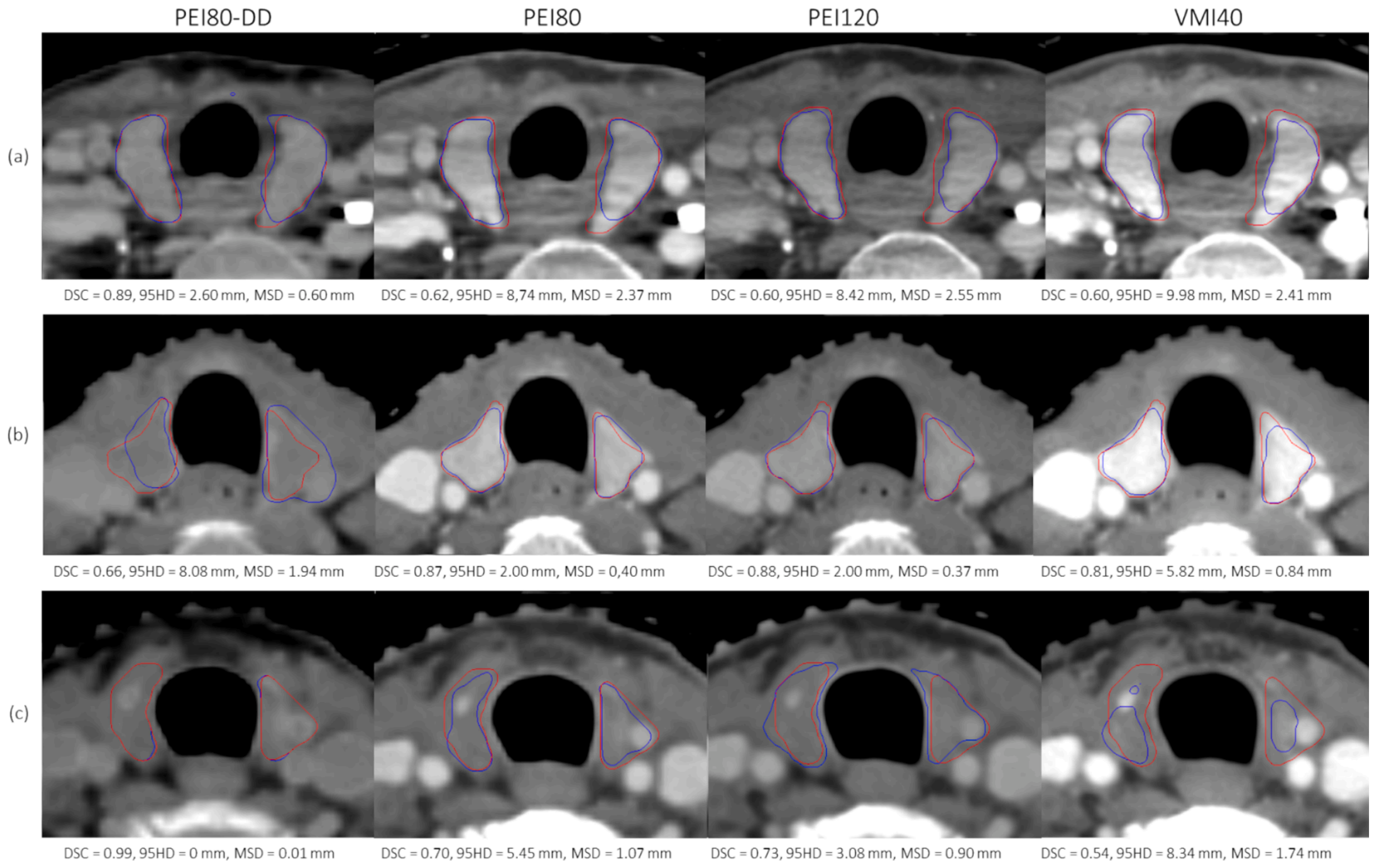
PEI80-DD
PEI80
PEI120
VMI40
(a)
DSC = 0.89, 95HD = 2.60 mm, MSD = 0.60 mm
DSC = 0.62, 95HD = 8,74 mm, MSD = 2.37 mm
DSC = 0.60, 95HD = 8.42 mm, MSD = 2.55 mm
DSC = 0.60, 95HD = 9.98 mm, MSD = 2.41 mm
(b)
DSC = 0.66, 95HD = 8.08 mm, MSD = 1.94 mm
DSC = 0.87, 95HD = 2.00 mm, MSD = 0,40 mm
DSC = 0.88, 95HD = 2.00 mm, MSD = 0.37 mm
DSC = 0.81, 95HD = 5.82 mm, MSD = 0.84 mm
(c)
DSC = 0.99, 95HD = 0 mm, MSD = 0.01 mm
DSC = 0.70, 95HD = 5.45 mm, MSD = 1.07 mm
DSC = 0.73, 95HD = 3.08 mm, MSD = 0.90 mm
DSC = 0.54, 95HD = 8.34 mm, MSD = 1.74 mm

Fig. 4. Samples of ground truth (= clinical) segmentation (red) and segmentation output from ART-PLAN® Annotate™ (blue) on PEI80-DD, PEI80, PEI120, VMI40 and their metrics scores depicting respectively different patients for whom ART-PLAN Annotate auto-segmentation obtained the best (a), the worst (b) and close to 1 (c) DSC scores for PEI80-DD compared to the other series for the thyroid (Contrast window width/level: 400/100 PEI80-DD, 600/100 PEI80, 500/100 PEI120 and 800/100 VMI40).

automatic contouring based on VMI40 series performed poorly, returning a fragmented structure (DSC = 0.54).

4. Discussion

This study used a retrospective cohort of seventy four patients to assess the ability of a DL-based commercial solution to generalize to non-conventional CT images. The performance of ART-PLAN® Annotate™ was tested by analyzing thirteen structures for head-and-neck RT on 4 different DECT images.

In the qualitative study, the analysis revealed that there are more B-scores from doctor 1 than doctor 2 demonstrating the inter-observer variability between doctors. Additionally, the blind examination of the GT contours for retrospective cases raised concerns regarding the accuracy of LN delineation: doctor 1 assigned a C-grade to 10 patients, and doctor 2 to 6 patients, which highlights the inherent difficulty of the delineation task. Concerning the thyroid, the performance of auto-contouring on PEI80-DD was found poor by both doctors. As illustrated on Fig. 4(b), the auto-segmentation of the thyroid on PEI80-DD wrongly included the carotid artery and needed correction while the auto-segmentation on PEI80, PEI120 and VMI40 were equivalent. DD may be the cause for the lack of contrast for those structures. Both statements suggest that PEI-80DD might not be the best series for segmenting structures close to blood vessels like the thyroid.

In a rather contradictory way, quantitative results (Fig. 2) showed the closest results to the GT on PEI80-DD auto-segmentation followed by PEI80 while VMI40 seems to have the most different results to the GT. At the time, we integrated the VMI40 to our clinical protocol based on literature about contrast enhancement for the target volume, which leans towards VMI of lower energies (40–65 keV) [17,18,34]. Wang et al. (2019) [35] investigated the optimal energy leading to the best contrast-noise-ratio (CNR) in head-and-neck, for 4 OARs including the parotid. They used VMI ranging from 40 to 190 keV with 5 keV increment and obtained the maximum CNR for the VMI of 80 keV. Hence, it is possible that VMI40 is not optimal for OAR segmentation of HNC and that a higher energy could be used as a supplement with the purpose of improving OAR delineation [36]. Also, and most importantly, ART-PLAN® Annotate™ segmentation algorithm is not trained with any type of contrast-enhanced images derived from a DECT. Indeed, ART-PLAN® Annotate™ training dataset is solely composed of conventional CT from 100 to 140 kVp which could explain the suboptimal response to the VMI. It is possible that VMI image quality might be out-of-distribution data for the DL-algorithm. Monoenergetic images could still reveal great potential for segmentation if the gain in contrast could be effectively used in a DL algorithm.

Our study has several limitations. As our physicians focus their corrections around the PTV and are less demanding with structures afar, the use of the clinical contours as GT can constitute a bias. Moreover, variability in scores for the GT in the qualitative results (Fig. 1) already suggests that the physicians are not completely confident with the clinical contour of certain structures. This raises doubts about the contours acceptability which needs to be evaluated according to the treatment planning objectives. For each patient, 2 physicians inspected the results of the auto-segmentation and corrected them, with a total of 10 physicians for the whole cohort. The use of a Simultaneous Truth and Performance Level Estimation [37] as the ground-truth for each organ could have made the study more robust. Moreover, doctors may use a set of images for segmentation but they delineate structures on PEI80-DD in the clinical workflow, which also constitute a bias. In this context, DSC comparing PEI80-DD to GT reflects the amount of modification made by the physicians on each OAR. Since the implementation of the DECT protocol in our RT department, PEI80 has been praised by physicians for its image quality for enhancing soft tissue contrast. It is possible to imagine a segmentation performed on PEI80 instead of PEI80-DD. However, a risk analysis needs to be conducted as for all RT protocols we use a unique DD LUT for dose calculation. If a structure needs correction, the physician would most likely use PEI80 and rarely move on to VMI40 by habit as the previous protocol only used PEI80 as a help for segmentation. Also, the contrast in VMI40 can be really strong especially near bone structures which can be disruptive for the physician. At the moment, doctors have not changed their practices and only use the VMI40 series as a last resort to delimit the target volume when complementary imaging modalities are not satisfactory. This type of image needs specific contrast window adjustments based on HU of the structure of interest and its surroundings. This raises a question about the need of training physicians to be more familiar with the use of DECT images in the clinical workflow. Finally, our study did not allow to determine with confidence which image would best help the doctor in his delineation task. However, we highlighted the performance limitations of a commercial software auto-segmentation for non-conventional CT images. The next step would be to identify which DECT images for which organ could improve the performance of the industrial tools.

In conclusion, our results have highlighted that not only auto-segmentation is influenced by image quality, but radiation oncologists are influenced by the image series that is used for treatment planning on which they validate their contours. DECT scanners allow the reconstruction of a large choice of images at different energies. It is therefore important to identify the structures that would benefit from these non-routine images and to discuss with manufacturers so that they re-train their DL algorithms on the image qualities best suited for the organ concerned. Finally, there is an urgent need for editors of treatment planning systems to facilitate the use of several series of CT images different from the series used to calculate the dose.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Gustave Roussy Cancer Campus received funding from Therapanacea for ART-PLAN® Annotate™ evaluation (clinical usage) and has several research contracts ongoing with Therapanacea. Lê A.T.'s thesis is being financially supported by Siemens Healthineers.

Acknowledgement

We would like to acknowledge the Siemens Heathineers company for its financial support.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2024.100654.

References

[1] Chang JH, Wu CC, Yuan KS, Wu ATH, Wu SY. Locoregionally recurrent head and neck squamous cell carcinoma: incidence, survival, prognostic factors, and treatment outcomes. Oncotarget 2017;8(33):55600–12. https://doi.org/10.18632/oncotarget.16340.

[2] Brook I. Late side effects of radiation treatment for head and neck cancer. Radiat Oncol J 2020;38(2):84–92. https://doi.org/10.3857/roj.2020.00213.

[3] Lo Faso EA, Gambino O, Pirrone R. Head-neck cancer delineation. Appl Sci 2021;11(6):2721. https://doi.org/10.3390/app11062721.

[4] Li F, Li Y, Wang X, Zhang Y, Liu X, Liu S, et al. Inter-observer and intra-observer variability in gross tumor volume delineation of primary esophageal carcinomas based on different combinations of diagnostic multimodal images. Front Oncol 2022;12:817413. https://doi.org/10.3389/fonc.2022.817413.

[5] McCollough CH, Leng S, Yu L, Fletcher JG. Dual- and multi-energy CT: principles, technical approaches, and clinical applications. Radiology 2015;276(3):637–53. https://doi.org/10.1148/radiol.2015142631.

[6] Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. J Med Phys 2008;33(4):136–40. https://doi.org/10.4103/0971-6203.44472.

[7] Brouwer CL, Steenbakkers RJ, Gort E, Kamphuis ME, Van der Laan HP, Van't Veld AA, et al. Differences in delineation guidelines for head and neck cancer result in inconsistent reported dose and corresponding NTCP. Radiother Oncol 2014;111(1):148–52. https://doi.org/10.1016/j.radonc.2014.01.019.

[8] Van der Veen J, Gulyban A, Willems S, Maes F, Nuyts S. Interobserver variability in organ at risk delineation in head and neck cancer. Radiat Oncol 2021;16(1):120. https://doi.org/10.1186/s13014-020-01677-2.

[9] Van der Veen J, Gulyban A, Nuyts S. Interobserver variability in delineation of target volumes in head and neck cancer. Radiother Oncol 2019;137:9–15. https://doi.org/10.1016/j.radonc.2019.04.006.

[10] Bollen H, Gulyban A, Nuyts S. Impact of consensus guidelines on delineation of primary tumor clinical target volume (CTVp) for head and neck cancer: Results of a national review project. Radiother Oncol 2023;189:109915. https://doi.org/10.1016/j.radonc.2023.109915.

[11] Trignani M, Argenone A, Di Biase S, Musio D, Merlotti A, Ursino S, et al. Inter-observer variability of clinical target volume delineation in definitive radiotherapy of neck lymph node metastases from unknown primary. A cooperative study of the Italian Association of Radiotherapy and Clinical Oncology (AIRO) Head and Neck Group. Radiol Med 2019;124(7):682–92. https://doi.org/10.1007/s11547-019-01006-y.

[12] Wang Y, Peng Y, Wang T, Li H, Zhao Z, Gong L, et al. The evolution and current situation in the application of dual-energy computed tomography: a bibliometric study. Quant Imaging Med Surg 2023;13(10):6801–13. https://doi.org/10.21037/qims-23-467.

[13] Kruis MF. Improving radiation physics, tumor visualisation, and treatment quantification in radiotherapy with spectral or dual-energy CT. J Appl Clin Med Phys 2022;23(1):e13468.

[14] Van Elmpt W, Landry G, Das M, Verhaegen F. Dual energy CT in radiotherapy: current applications and future outlook. Radiother Oncol 2016;119(1):137–44. https://doi.org/10.1016/j.radonc.2016.02.026.

[15] Greffier J, Villani N, Defez D, Dabli D, Si-Mohamed S. Spectral CT imaging: Technical principles of dual-energy CT and multi-energy photon-counting CT. Diagn Interv Imaging 2023;104(4):167–77. https://doi.org/10.1016/j.diii.2022.11.003.

[16] Goo HW, Goo JM. Dual-energy CT: new horizon in medical imaging. Korean J Radiol 2017;18(4):555–69. https://doi.org/10.3348/kjr.2017.18.4.555.

[17] Albrecht MH, Scholtz JE, Kraft J, Bauer RW, Kaup M, Dewes P, et al. Assessment of an advanced monoenergetic reconstruction technique in dual-energy computed tomography of head and neck cancer. Eur Radiol 2015;25(8):2493–501. https://doi.org/10.1007/s00330-015-3627-1.

[18] Roele ED, Timmer VCML, Vaassen LAA, Van Kroonenburgh AMJL, Postma AA. Dual-energy CT in head and neck imaging. Curr Radiol Rep 2017;5(5):19. https://doi.org/10.1007/s40134-017-0213-0.

[19] Noid G, Zhu J, Tai A, Mistry N, Schott D, Prah D, et al. Improving structure delineation for radiation therapy planning using dual-energy CT. Front Oncol 2020;10:1694. https://doi.org/10.3389/fonc.2020.01694.

[20] Wang T, Lei Y, Roper J, Ghavidel B, Beitler JJ, McDonald M, et al. Head and neck multi-organ segmentation on dual-energy CT using dual pyramid convolutional neural networks. Phys Med Biol 2021;66(11). https://doi.org/10.1088/1361-6560/abfce2.

[21] Liu P, Sun Y, Zhao X, Yan Y. Deep learning algorithm performance in contouring head and neck organs at risk: a systematic review and single-arm meta-analysis. Biomed Eng Online 2023;22(1):104. https://doi.org/10.1186/s12938-023-01159-y.

[22] Tang H, Chen X, Liu Y, Lu Z, You J, Yang M, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. Nat Mach Intell 2019;1:480–91. https://doi.org/10.1038/s42256-019-0099-z.

[23] Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. Med Phys 2014;41(5):050902. https://doi.org/10.1118/1.4871620.

[24] Heilemann G, Buschmann M, Lechner W, Dick V, Eckert F, Heilmann M, et al. Clinical implementation and evaluation of auto-segmentation tools for multi-site contouring in radiotherapy. Phys Imaging Radiat Oncol 2023;28:100515. https://doi.org/10.1016/j.phro.2023.100515.

[25] Shi J, Wang Z, Ruan S, Zhao M, Zhu Z, Kan H, et al. Rethinking automatic segmentation of gross target volume from a decoupling perspective. Comput Med Imaging Graph 2024;112:102323. https://doi.org/10.1016/j.compmedimag.2023.102323.

[26] Grégoire V, Ang K, Budach W, Grau C, Hamoir M, Langendijk JA, 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(1):172–81. https://doi.org/10.1016/j.radonc.2013.10.010.

[27] Van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. Int J Radiat Oncol Biol Phys 2019;104(3):677–84. https://doi.org/10.1016/j.ijrobp.2019.02.040.

[28] Chen A, Deeley MA, Niermann KJ, Moretti L, Dawant BM. Combining registration and active shape models for the automatic segmentation of the lymph node regions in head and neck CT images. Med Phys 2010;37(12):6338–46. https://doi.org/10.1118/1.3515459.

[29] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302. https://doi.org/10.2307/1932409.

[30] Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. IEEE Trans Pattern Anal Mach Intell 1993;15(9):850–63. https://doi.org/10.1109/34.232073.

[31] Yeghiazaryan V, Voiculescu I. Family of boundary overlap metrics for the evaluation of medical image segmentation. J Med Imaging (Bellingham) 2018;5(1):015006. https://doi.org/10.1117/1.JMI.5.1.015006.

[32] DeepMind (2022), Surface distance metrics. https://github.com/deepmind/surface-distance.

[33] Siegel S. Nonparametric statistics for the behavioral sciences. J Nerv Ment Dis 1957;125(3):497. https://doi.org/10.1097/00005053-195707000-00032.

[34] Wichmann JL, Nöske EM, Kraft J, Burck I, Wagenblast J, Eckardt A, et al. Virtual monoenergetic dual-energy computed tomography: optimization of kiloelectron volt settings in head and neck cancer. Invest Radiol 2014;49(11):735–41. https://doi.org/10.1097/RLI.0000000000000077.

[35] Wang T, Ghavidel BB, Beitler JJ, Tang X, Lei Y, Curran WJ, et al. Optimal virtual monoenergetic image in "TwinBeam" dual-energy CT for organs-at-risk delineation based on contrast-noise-ratio in head-and-neck radiotherapy. J Appl Clin Med Phys 2019;20(2):121–8. https://doi.org/10.1002/acm2.12539.

[36] Lam S, Gupta R, Levental M, Yu E, Curtin HD, Forghani R. Optimal virtual monochromatic images for evaluation of normal tissues and head and neck cancer using dual-energy CT. AJNR Am J Neuroradiol 2015;36(8):1518–24. https://doi.org/10.3174/ajnr.A4314.

[37] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004;23(7):903–21. https://doi.org/10.1109/TMI.2004.828354.