Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images

Madalina Costea, Alexandra Zlate, Anne-Agathe Serre, Séverine Racadot, Thomas Baudier, Sylvie Chabaud, Vincent Grégoire, David Sarrut, Marie-Claude Biston

Evaluation of different algorithms for automatic segmentation of head-and-neck lymph nodes on CT images

Madalina Costea[a, b], Alexandra Zlate[c], , Anne-Agathe Serre[a], Séverine Racadot[a], Thomas Baudier[a,b], Sylvie Chabaud[d] , Vincent Grégoire[a] , David Sarrut[a,b], Marie-Claude Biston[a, b,*]

[a]Centre Léon Bérard, 28 rue Laennec 69373 LYON Cedex 08 - France

[b]CREATIS, CNRS UMR5220, Inserm U1044, INSA-Lyon, Université Lyon 1, Villeurbanne - France

[c] MedEuropa, Strada Turnului 8, Brașov 500152 - Romania

[d]Unité de Biostatistique et d'Evaluation des Thérapeutiques, Centre Léon Bérard, Lyon 69373, France

*Corresponding author. Department of Radiation Oncology, Centre Léon Bérard, 28 rue Laennec, 69373 Lyon Cedex 08, France

E-mail address: marie-claude.biston@lyon.unicancer.fr

Short title: Atlas-based and deep learning algorithms for HN lymph nodes automatic segmentation

Abstract

Purpose: To investigate the performance of 4 atlas-based (multi-ABAS) and 2 deep learning (DL) solutions for head-and-neck (HN) elective nodes (CTVn) automatic segmentation (AS) on CT images.

Material and Methods: Bilateral CTVn levels of 69 HN cancer patients were delineated on contrast-enhanced planning CT. Ten and 49 patients were used for atlas library and for training a mono-centric DL model, respectively. The remaining 20 patients were used for testing. Additionally, three commercial multi-ABAS methods and one commercial multi-centric DL solution were investigated. Quantitative evaluation was assessed using volumetric Dice Similarity Coefficient (DSC) and 95-percentile Hausdorff distance ($HD_{95\%}$). Blind evaluation was performed for 3 solutions by 4 physicians. One recorded the time needed for manual corrections. A dosimetric study was finally conducted using automated planning.

Results: Overall DL solutions had better DSC and $HD_{95\%}$ results than multi-ABAS methods. No statistically significant difference was found between the 2 DL solutions. However, the contours provided by multi-centric DL solution were preferred by all physicians and were also faster to correct (1.1min vs 4.17min, on average). Manual corrections for multi-ABAS contours took on average 6.52min Overall, decreased contour accuracy was observed from CTVn2 to CTVn3 and to CTVn4. Using the AS contours in treatment planning resulted in underdosage of the elective target volume.

Conclusion: Among all methods, the multi-centric DL method showed the highest delineation accuracy and was better rated by experts. Manual corrections remain necessary to avoid elective target underdosage. Finally, AS contours help reducing the workload of manual delineation task.

Introduction

According to the global cancer statistics, in 2020, more than 1.5 million Head and Neck (HN) cancer cases were diagnosed, which represents 7.9% of all cancer diagnoses, and over 510 000 deaths worldwide [1]. Overall, the main treatment options are surgery, radiation therapy (RT), chemotherapy, and targeted therapy. Approximately 74% of the HN cancer patients benefit from external beam RT either prescribed alone or in combination with other treatment strategies [2]. In external RT, highly conformal dose distributions with steep dose gradients are achieved today, through the use of intensity-modulated radiation therapy (IMRT) techniques. In particular, by allowing continuous gantry rotation of the linear accelerator around the patient during treatment delivery, volumetric modulated arc therapy (VMAT) technique results in plans of similar or improved quality compared with fixed-field IMRT while reducing the treatment time per fraction [3].

Therefore, accurate delineation of both organs-at-risk (OARs) and targets is a crucial step, particularly for HN cancer, where numerous organs with strict dose objectives are involved. Manual contouring is time-consuming and although international guidelines exist [4–6], large inter (IOV) and intra-observer variation are observed [7–9] that can negatively impact patient doses [10,11]. To assist organ differentiation and increase the image contrast, the patients should be injected with an iodine contrast agent before the simulation computed tomography (CT) scans [12]. To reduce the delineation time, improve consistency and accuracy of volume definition, automatic segmentation (AS) solutions have received great interest [13–15]. In the recent years, the performances of AS methods for HN cancer were mainly focus on OARs contouring, but few studies were focused on the clinical target volumes (CTV) [16–19]. Whereas important anatomical variations make gross tumor volumes difficult candidates for AS [20], the healthy HN lymph nodes levels (LN) have well-established anatomical borders [5,6] and are often irradiated prophylactically (i.e. with a preventive intent) as secondary nodal target (CTVn).

While the performance evaluation of an AS method is not yet standardized, the most recent recommendations suggest the use of several complementary metrics [21]. Among the most widely used indices comparing the geometric accuracy of AS versus expert contours is the volumetric Dice similarity coefficient (DSC) [22]. However, DSC alone does not represent a direct estimate of the clinical impact on radiation doses, nor the clinical workflow/labor reduction [23]. Therefore, a dosimetric study is highly recommended as well as the assessment of the time needed to perform manual corrections on these contours.

Among the AS solutions, atlas-based AS (ABAS) methods are attractive as they require only one or few (multi-ABAS) patients as prior information (in form of an atlas library), but they are limited

to the range of patient anatomical representation. Few studies have demonstrated the superiority of multi-ABAS vs single-ABAS strategies for CTVn segmentation [24–28]. It was shown that using 11vs1 atlas enabled to decrease the manual delineation time from 42.3min to 21.4min vs 30.1min, respectively [25]. In another study, the range of DSC results between ABAS and expert contours, was 0.29-0.78 depending on the CTVn level [27]. One multi-ABAS study (N=10 atlases) evaluated dosimetric plan quality when using AS contours obtained with a commercially available solution (ABAS, Elekta AB), and demonstrated that despite mean DSC>0.80, non-edited CTVn contours can cause large underdosage in target volumes [11]. Hybrid approaches combining multi-ABAS and machine learning features have also been explored [29–31]. *Qazi et al.* evaluated a model-based algorithm (N=15 atlases) and achieved mean DSC of 0.74 (LN level 1-4) [31]. Their results were superior to *Chen et al*. [30] who created an active shape model (N=14 atlases) that reached mean DSC=0.69 (one volume covering LN 2-4) and superior to that of *Gorthi et al.* [29] with an active contour-based model (N=9 atlases) that reached maximum DSC of 0.58 (individual LN levels 1-6). In these studies, the combination of the individual LN volumes probably had an important contribution in the differences in the DSC results.

Alternatively, deep learning (DL) solutions could increase accuracy and efficiency in AS at the cost of more efforts involved in gathering and curating manual contours databases for training. Promising results were obtained particularly for OARs in HN patients and several solutions are commercially available [15,17,32–35]. From the 3 studies evaluating their accuracy in segmenting HN CTVn levels, *Wong et al*. investigated a commercial DL-based segmentation software (Limbus Contour) trained with publicly available annotated data (on average 328 CT scans/organ) [17]. One single CTVn volume (including 6 LN) was auto-segmented. The mean DSC against the experts' contours was 0.72 which was inferior to the IOV (DSC=0.79). Another study investigated a 3D-convolutional neural network (CNN) trained on 69 patients (mono-centric data), for AS of 10 separated CTVn levels (one volume for LN 2-4) [18]. Compared with 2 experts manual contours the mean DSC ranged between 0.46-0.82, in function of the considered CTVn level. The manual delineation time was reduced from 52 to 35min when editing AS contours. Moreover, it was shown that using the DL solution enabled to significantly improve the IOV (DSC=0.92vs0.79). Lastly, *Strijbis et al*. [19] trained 3 different Unet networks on 70 patients for AS of individual LN 1-5. They showed that an ensemble of networks provided the best contours with mean DSC>0.85 for the LN 1, 2 and 3, but mean DSC<0.72 for LN 4 and 5.

In this context, the objective of the present study was to evaluate the performance of 4 multi-ABAS and 2 DL solutions for the individual segmentation of left (L) and right (R) LN levels 2 (CTVn2), 3

(CTVn3) and 4 (CTVn4), following the formerly performed work on HN OARs segmentation [33]. Five out of the six solutions were investigated for the first time on HN CTVn segmentation. Notably, non-commercial solutions (one hybrid-ABAS and one mono-centric DL solution) were compared with 4 commercial solutions (three multi-ABAS and one multi-centric DL solution). All 6 solutions were evaluated based on geometrical accuracy. A clinical scoring of the contours was performed by 4 physicians on the 3 most accurate AS solutions. One physician recorded the time spend on manual corrections. Lastly, an auto-planning solution based on a priori multi-criterial optimization (MCO) algorithm was used to generate treatment plans using manual and AS CTVn contours with and without manual corrections.

Material and methods

Patient data

Sixty-nine HN cancer patients treated with radiation therapy between 2018-2022 were included in the study, which was approved by the hospital ethics committee. Each patient was immobilized using personalized head cushion (Moldcare®, Qfix, Avondale, USA) and a 5-points thermoplastic mask (MacroMedics, Moordrecht, The Nederlands). CT-scan acquisitions were performed on a SOMATOM go.Sim scanner (CT) (Siemens, Munich, Germany), after 2-phase injection of iodine solution, following recommendations [12]. Bilateral CTVn2, CTVn3 and CTVn4 were manually delineated according to international delineation guidelines [5] by a senior expert physician, on 512x512 and 2mm-thick CT slices with maximum physical in-plane resolution of 1.17mm. Forty-nine non-operated patients with standard anatomy, were used to train a mono-centric DL model. Ten of these patients were subsequently used to form an atlas library for the multi-ABAS solutions. The patients' selection was arbitrary and based on their body mass index (BMI) which was intended to cover a large variety of patient anatomies (atlas library BMI range: 19.9-26). Identical atlas libraries were created within MIM-Maestro (MIM-Software; Cleveland, USA) and the research version of ADMIRE software (ADMIREv3.41, Elekta AB; Stockholm, Sweden). Conversely, the second DL solution was trained by the vendor (Therapanacea, France) with large amount of data (>1000) from multiple centers. The remaining 20 patients with different tumors and anatomies (BMI range: 17.9–33.7), were used for testing of the 6 AS solutions. In addition to reference contours for CTVn, the test cohort (Table 1) included expert delineations for OARs and primary tumor volumes.

Automatic segmentation solutions

Three multi-ABAS solutions integrated in the research version of Monaco treatment planning system (TPS) (Monaco 5.59.11 with ADMIREv3.41) and another one available in MIM-Maestro (MIM Software Inc., Cleveland, OH) were investigated (Table 2). Two DL solutions were considered: one mono-centric (ADMIRE-DL, data from this study) and one commercial multi-centric solution (ART-Plan, Table 2).

All AS solutions have been fully described in a previous work [33]. Briefly, ABAS.1 uses a traditional method for atlas fusion based on expectation-maximization algorithm [36]. ABAS.2 uses voxel intensity information to obtain a weighted average of the atlases' contours [37]. ABAS.3 algorithm trains a voxel classifier on the fly using the registered atlases as training data [38]. ABAS.4 performs the voxel annotation based on labels predicted by majority of the atlases [39]. For the 3 ABAS solutions used in ADMIRE, for each test patient, a reference atlas was selected from the library, upon the closest BMI. In MIM-Maestro, to create the atlas library, one atlas was chosen as template patient (based on BMI) and was registered to the 9 remaining atlases [39].

Among the DL solutions, DL.1 is a CNN where the high-resolution image features captured in the encoding part are preserved with the help of short-range connectors in the decoding part for a label map corresponding to the input image size [34,40]. The DL.2 solution uses a set of organ-specific networks with an original combination of data-driven and decisional artificial intelligence that enforces anatomical consistency [35,41].

Geometric evaluation

The quantitatively evaluation of the 6 AS solution was performed per LN level and per their union, based on volumetric DSC and 95-percentile Hausdorff Distance ($HD_{95\%}$) [22], similar to [33]. Results are presented as mean ± 1 standard deviation (SD).

Clinical acceptability assessment and time required for manual editing

The union of bilateral CTVn contours (CTVn_union) was further examined for the three most accurate solutions (DL.1, DL.2, ABAS.2) on 11 patients who had all 3 LN levels (CTVn2+CTVn3+CTVn4) involved in the RT treatment. To get a trend from different HN cancer experts a blinded evaluation was made by 4 physicians choosing for each CTVn_union one of the following options:

a) clinically acceptable without corrections

b) clinically acceptable with minor corrections

c) clinically acceptable with major corrections

d) not acceptable for clinical use

Then, the AS contours were manually adjusted on Elekta ProKnow® (Elekta AB, Stockholm) platform and the time spent on corrections was recorded for each of the 3 solutions. Note that, out of the four physicians, one single physician performed the reference contours whereas another one performed manual corrections of the AS contours.

Dosimetric end-points using auto-planning solution

For the 11 patients, 7 VMAT treatment plans (one reference plan + 6 experimental plans) were calculated automatically using mCycle auto-planning solution (Monaco 5.59.11, Elekta AB; Stockholm, Sweden) [42]. All plans were designed using 2 arcs and a simultaneous integrated boost technique to deliver 70Gy to the primary planned target volume (PTV_70Gy) and 54.25Gy to the prophylactic nodal target (PTV_54.25Gy), in 35 fractions. The reference plan was created using exclusively manually delineated contours of OARs and CTVs. Three experimental plans were created by replacing the manual CTVn contours with CTVn contours obtained by ABAS.2, DL.1 and DL.2 solutions, the other three being plans obtained with corrected AS contours (ABAS.2+corr, DL.1+corr and DL.2+corr). The PTV_54.25Gy was created for each plan from CTVn_union and the prophylactic target plus 4mm margin. The resulting 7 dose distributions were all analyzed on the reference manual contours by evaluating the dose differences between the reference (Dref) and the experimental plan (Dexp) ($\Delta D = Dref - Dexp$). From the dose-volume histograms (DVHs) clinically relevant dosimetric endpoints were extracted according to the French Society of Radiation Oncology recommendations [12].

Statistics

For all 6 solutions and for each CTVn, Kruskal-Wallis test was performed to assess if the distribution of geometric indices (DSC and $HD_{95\%}$) and the volume of the contours corresponding to the different AS methods tested were statistically significant. Furthermore, post-hoc Dunn's with Bonferroni correction for multiple testing was performed to detect between which pairs of algorithms the differences were statistically significant. For the dosimetric study, paired Wilcoxon signed-rank test was performed to assess significant dose differences between the treatment plans (reference vs experimental plans). The statistical tests were performed with level of significance set <0.05.

Results

Computational time for one patient was on average 6min, 9min and 10min for ABAS.1, ABAS.2 and ABAS.3, respectively. For ABAS.4, the computational time was on average 1min, whereas creating the atlas library (registering of the atlases) took around 7min. DL.1 and DL.2 provided segmentations in <1 and <2min, respectively.

DSC and $HD_{95\%}$ results obtained for each CTVn level and for CTVn_union are presented in Fig.1. Overall DL solutions (mean DSC:0.62-0.87) were more accurate than multi-ABAS methods (mean DSC:0.50-0.79), with no statistically significant difference between DL.1 and DL.2 ($p>0.14$). However, DL.1 performed better on CTVn4 than DL.2 (mean DSC:0.72vs0.64), whereas DL.2 had the lowest mean$HD_{95\%}$ distance (6.4±5.3mm) among all the methods.

Considering the multi-ABAS solutions, no statistically significant difference was observed between ABAS1, ABAS.2 and ABAS.3 ($p=1$). ABAS.4 provided overall the worst results but differences in both DSC and $HD_{95\%}$ compared with other multi-ABAS methods were statistically significant only on CTVn2_L ($p<0.01$).

Differences were statistically significant between DL.1 algorithm and ABAS.1, ABAS.2 and ABAS.3 solutions only in DSC for the CTVn3_L/R ($p<0.02$). Compared with ABAS.4, DL.1 had significantly better DSC results for all CTVn levels ($p<0.001$) but CTVn4_L ($p=0.07$) and statistically better $HD_{95\%}$ results for CTVn2_L ($p<0.001$) and CTVn3_L/R ($p<0.04$).

Similarly, DL.2 had significantly better DSC results compared with ABAS.1, ABAS.2 and ABAS.3 for CTVn2_L/R ($p<0.03$) and CTVn3_L/R ($p<0.001$) and significantly better $HD_{95\%}$ for CTVn3_L/R ($p<0.004$) and CTVn4_R ($p<0.02$). Moreover, compared with ABAS.4, DL.2 had significantly better DSC results for all levels ($p<0.04$) but CTVn4_L ($p=1$) and significantly better $HD_{95\%}$ for all CTVn ($p<0.01$).

An additional geometric analysis of the CTVn_union resulted in increased conformity to the manual reference, particularly for the multi-ABAS solutions, for which the contour unification enabled DSC results to reach values up to 0.81 (ABAS.2). Finally, union of DL.2 contours obtained the best conformity to the reference (mean DSC:0.86±0.03; mean$HD_{95\%}$:4.1±1.3mm) (Fig.1 Panel C and D). Moreover, the blinded study results (Fig.2) showed that all physicians clinically approved DL.2 contours without or with only minor corrections. Contrarily, majority of DL.1 and ABAS.2 contours were clinically accepted with minor or major corrections, and only one physician accepted few contours from DL.1 without corrections. Moreover, some ABAS.2 and DL.1 contours were rejected by two physicians.

One patient case is illustrated in Fig. 3, for a visual representation of ABAS.2, DL.1 and DL.2 CTVn contours versus reference contours. Two more cases are presented in Fig.1 and Fig.2 of supplementary data.

Furthermore, manual correction time was in average 6.52min, 4.17min and 1.1min for ABAS.2, DL.1 and DL.2 respectively. Contours' accuracy improved significantly only for ABAS.2 solution ($p=0.01$) (Fig.3 of supplementary data). In general, the volume of AS contour was smaller than the reference and for DL.1 the difference was significant (volume difference of 8.7%, $p=0.001$). After performing manual corrections, the differences was smaller for all solutions, but were still significant for DL.1 ($p=0.02$).

The dosimetric results obtained per PTV and per CTVn level are presented in Table 3. No statistically significant difference was observed on the dosimetric results obtained for PTV_70Gy ($p>0.1$). Conversely, compared with reference plans, all experimental plans created with AS CTVn contours experienced a significant loss in PTV_54.25Gy coverage (mean$V_{95\%}$ reduction up to 5.9%, $p<0.001$). In addition, using ABAS.2 contours to perform the plans lead to significant dose differences regarding all the analyzed DVH parameters ($p<0.05$). Using ABAS.2+corr contours, dose differences were still significant for all CTVn4 DVH parameters ($p<0.02$), $V_{95\%}$ for CTVn3 ($p=0.003$) and $D_{98\%}$ and $V_{95\%}$ for CTVn2 ($p<0.01$).

For DL.1, with the exception of the $D_{50\%}$ to the CTVs ($p>0.26$), the differences in $D_{98\%}$ and $V_{95\%}$ between the reference plans and the plans obtained with AS contours were significant ($p<0.03$). Using DL.1+corr contours did not improve the dose agreement to CTVn4 contours ($p<0.0003$), but for the other CTVn levels, meanΔD was smaller, with difference in $D_{98\%}$ becoming not significant ($p=0.16$).

The differences in DVH parameters between the reference and plans performed with DL.2 AS contours were significant for all CTVn ($p<0.05$), with the exception of $D_{98\%}$ to the CTVn3 ($p=0.21$). After correction of DL.2 contours, no significant difference between all DVH parameters was observed for the CTVn3 level ($p>0.37$). In addition, using DL.2+corr contours improved the dose agreement on CTVn2 regarding $D_{98\%}$ ($p=0.15$) and $D_{50\%}$ ($p=0.54$). At the same time, while the meanΔD was smaller on CTVn4, the differences in the DVH parameters remained significant ($p<0.001$).

For the OARs, in general, better agreement was observed between reference doses and doses obtained with ABAS.2 and DL.2 versus DL.1 AS contours, with no significant difference compared with the reference doses for 10 and 10 vs 8 out of 14 DVH parameters (Table 1 of the

supplementary data). After correction of the CTVn, for all solutions, 10 out of 14 DVH parameters were not significantly different from the reference. Interestingly, the correction of the CTVn tended to introduce significant differences in DHV parameters for certain OARs. Hence, on parotids, using the corrected CTVn contours generally resulted in an increase in the meanΔD and statistically significant differences compared with reference doses, whatever the algorithm. Thus for such OARs located at the very close vicinity of the PTVs (submandibular glands, parotids), small differences in the delineation of the contours might lead to significant dose differences. The same observation applied to D_{mean} to the thyroid where significant differences were still observed (or appeared) after correction of the contours, for the 3 algorithms ($p<0.02$). On the opposite, for the 3 solutions, there was no statistically significant difference in D_{mean} to the oral cavity, esophagus and brainstem, and $D_{5\%}$ to the larynx. Finally, the CTVn contours corrections generally lead to smaller meanΔD values for theses OARs, which might reflect a smaller IOV and/or larger distance of the OARs to the PTVs.

An example of dose distribution obtained for one patient case to illustrate where important loss in PTV_54.25Gy coverage can be observed at each CTVn level is provided in Fig.4.

Discussion

In this study, we evaluated several atlas-based and deep learning segmentation solutions for automatic delineation of nodal clinical target volumes for HN RT on planning CT images. We observed that overall DL solutions had better accuracy compared with multi-ABAS methods. With regard to the geometric indices, the two DL solutions were not statistically different. In general DSC results were better for DL.2 on CTVn2 and CTVn3, and better for DL.1 on CTVn4. When evaluating CTVn_union, DL.2 provided better conformity to manual reference. All physicians accepted the contours without or with minor corrections, and they were quickly corrected (about 1min/patient) by one of the physicians involved. Thus, considering that between 15-20 min are required to manually delineate CTVn2-4 by a skilled physician, up to 90% time reduction for contouring can be reached with DL.2 solution. Conversely, most of DL.1 contours were accepted with minor/major corrections which resulted in more important manual correction times (4min/patient). Overall, manual corrections were necessary on the lateral borders of the contours following the sternocleidomastoid muscle (SCM) as well as on the space between the SCM and the parotid glands. Corrections were also needed for the caudal limit of CTVn4 (2cm from sternal manubrium), particularly for DL.2 contours.

DL.1 model was trained with a relatively small number of patients (N=49) delineated exclusively by one expert, which ensured data uniformity. Similar with the previous work on OARs, we showed in this study that accurate results can be obtained with a limited but uniform training database. A similar mono-centric model (N=69) was trained by *van der Veen et al.* for the segmentation of 10 CTVn levels [18]. Regarding union of LN 2-4, they found mean DSC of 0.76 and 0.82 against 2 observers, whereas a mean DSC of 0.83 was obtained with DL.1 solution in our study. DL.2 solution, trained with much more patients from multiple centers, obtained a mean DSC of 0.86 in this study, which indicated a good generalizability of the model.

Regarding multi-ABAS methods, this study showed that accurate results can be obtained when using a library of only 10 patients. The two ABAS methods evaluated had an atlas selection strategy based on BMI. Using the same atlas library, we showed that the ADMIRE algorithm provided better results compared with MIM Maestro. Although the workflow to process the data was less efficient, to choose a reference atlas having the closest BMI to the test patient before each AS provided more accurate results than choosing one patient representing average anatomy over the ATLAS patients to perform all AS of the test cohort.

Moreover, performing the CTVn_union, enabled an DSC≥0.80 (ABAS.1, ABAS.2 and ABAS.3), which suggested that most of the AS contour discrepancies happened at the junction of the level. Manual corrections took on average 6.52min/patient for ABAS.2 method. Contrary to the previous study on OARs segmentation, the superiority of the new ABAS.3 solution over the commercial ABAS.1 and ABAS.2 solutions was not demonstrated for the segmentation of CTVn volumes [33]. Nonetheless, DL methods outperformed multi-ABAS algorithms for both OARs and CTVn segmentation.

The heterogeneity of literature studies makes comparisons difficult (Table 2 supplementary data). Some studies considered a total volume of the CTVn whereas others considered independent contours per LN. Moreover, guidelines suggest that a dosimetric study should be included when evaluating new AS methods [21]. Only one study performed a dosimetric evaluation and attested that editing ABAS contours of HN CTVn was required to avoid large reduction in target coverage (mean reduction in $V_{95\%}$ of 7.2%) [11].

In this study, we reached the same conclusions. We showed that, using the AS contours for treatment planning, the mean$V_{95\%}$ to PTV_54.25Gy was between 93.1%-94.7% compared with 98.7% for the reference plan. After manual correction of the AS contours, PTV_54.25Gy mean$V_{95\%}$ was increased to 94.8%-95.4%. Note that reference contours and manual corrections were done

by two different physicians. Although the DSC results showed a strong agreement between the 2 physicians (DSC>0.85 after manual corrections), significant differences in mean$V_{95\%}$ to PTV_54.25Gy were observed between reference plans and plans performed with AS+corr contours. There are two main reasons for this observation. The first one is IOV, which was noticeable in the delineation of the CTVn4. This is also where AS solutions have been less accurate (DSC≤0.72). Despite significant differences still being present between the reference doses and the doses obtained with the AS+corr contours, the meanΔDs were generally lower, particularly for DL.2. This is mainly due to the fact that, especially for DL.2, the AS of CTVn4 contours were missing in the bottom slices. The other reason of the differences observed in $V_{95\%}$ to PTV_54.25Gy was the use of the auto-planning tool which was demonstrated to provide highly conformed plans with steeper dose gradients than manual ones [42]. Hence, even a small IOV in contouring of CTVn could lead to important dose distributions discrepancies on PTVs.

The consequences on the dose distributions to the OARs were different according their proximity to the PTVs. We showed that small differences in the delineation of the CTVn contours might lead to significant meanΔD on OARs such as submandibular glands, parotids or thyroid. For other OARs such as the esophagus (D_{mean}) or the larynx ($D_{5\%}$), which was also close to the PTVs, there was no significant difference in the meanΔD either when comparing the reference dose to the treatment plan performed with AS contours, or the one performed with AS+corr contours. This both reflected better accuracy of the algorithm in the delineation of CTVn3 and better agreement between the contours of the two physicians.

To our knowledge, the present study investigated for the first time 5 AS methods (ABAS.2, ABAS.3, ABAS.4, DL.1 and DL.2) for segmenting 3 distinct CTVn levels. Additionally, auto-planning was used to assess dosimetric consequences of using the AS contours from one multi-ABAS and 2 DL solutions. This allowed decreased labor, to eliminate inter or intra operator variability in planning, and to focus on the dosimetric effect coming from the CTVn contours only. Overall the results were similar among the AS methods and showed no significant impact on the primary PTV and OARs. However, despite the use of a CTV-to-PTV margin of 4mm, significant reduction in coverage of the elective PTV (up to 5.9%, $p<0.006$, see Table 3) was observed for all the AS solutions, which was consistent with the literature [11]. The effect was more pronounced on the CTVn4 which could be related with the worse results previously identified in the geometrical indices (mean DSC<0.8 and $HD_{95\%}$>5mm). Moreover, the blinded study showed that, overall, manual corrections still need to be performed. However, DL.2 contours were preferred by the physicians. When considering both computational and manual correction time, DL solutions

proved the most promising in decreasing the manual delineation time. A mean DSC of 0.85 between the two experts who did reference and corrected contours was observed when manual corrections were performed on the AS contours (Fig.3 supplementary data). IOV between manual delineations among the experts was not assessed in this study. Other study showed that performing manual adjustments on AS contours enabled to improve IOV [18].

While differences in geometric indices were not statistically significant between DL solutions, DL.2 contours were better rated by all the 4 physicians, and time for correcting the contours was shorter. This strongly suggests that DSC/$HD_{95\%}$ alone are not sufficient to characterize the performances of a solution. Therefore, the interplay between the training cohort size and a DL model architecture could be further investigated by training DL.1 on a larger cohort (N>50 patients). In our previous work, 63 patients were used for training the same model on OARs, which provided consistent result over the majority of structures. While on CTVn delineation, DSC≥0.82 were obtained for CTVn2, more training data could potentially improve the accuracy on CTVn3 and CTVn4. At the same time, DL.2 was trained on large patient database and the mean DSC for CTVn4 was inferior to DL.1 model. Overall, both multi-ABAS and DL results showed decreased accuracy from CTVn2 to CTVn3 and CTVn4 which is consistent with the literature [19,27].

Conclusion

DL solutions provided contours with better geometric accuracy and were better rated by experts than multi-ABAS methods for CT-based AS of CTVn levels. Lower mean DSC were observed for CTVn4 compared with CTVn2 and CTVn3. However, DL contours were faster to compute and to manually correct. With few patient data, multi-ABAS methods provided good conformity to reference contours, but with decreased workflow efficiency.

Acknowledgements

Elekta AB is acknowledged for having involved the CLB team in this research project. We are grateful to Nicolette O'Connell, employed by Elekta, who had a key role in the development of our ADMIRE deep learning model. This work was performed within the framework of the SIRIC LYriCAN Grant INCa-INSERM-DGOS-12563, and the LABEX PRIMES(ANR-11-LABX-0063) of Université de Lyon, within the program Investissements d'Avenir (ANR-11-IDEX-0007) operated by the ANR.

References

[1] Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A, et al. Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. CA Cancer J Clin 2021;71:209–49. https://doi.org/10.3322/CAAC.21660.

[2] Delaney GP, Barton MB. Evidence-based Estimates of the Demand for Radiotherapy. Clin Oncol 2015;27:70–6. https://doi.org/10.1016/j.clon.2014.10.005.

[3] Otto K. Volumetric modulated arc therapy: IMRT in a single gantry arc. Med Phys 2008;35:310–7. https://doi.org/10.1118/1.2818738.

[4] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, et al. Head and neck guidelines CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines 2015. https://doi.org/10.1016/j.radonc.2015.07.041.

[5] 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;110:172–81. https://doi.org/10.1016/j.radonc.2013.10.010.

[6] Grégoire V, Eisbruch A, Hamoir M, Levendag P. Proposal for the delineation of the nodal CTV in the node-positive and the post-operative neck. Radiother Oncol 2006;79:15–20. https://doi.org/10.1016/J.RADONC.2006.03.009.

[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;7:1–10. https://doi.org/10.1186/1748-717X-7-32/FIGURES/4.

[8] Awan M, Kalpathy-Cramer J, Gunn GB, Beadle BM, Garden AS, Phan J, et al. Prospective assessment of an atlas-based intervention combined with real-time software feedback in contouring lymph node levels and organs-at-risk in the head and neck: Quantitative assessment of conformance to expert delineation. Pract Radiat Oncol 2013;3:186–93. https://doi.org/10.1016/J.PRRO.2012.11.002.

[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] Tao CJ, Yi JL, Chen NY, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: A multi-institution clinical study. Radiother Oncol 2015;115:407–11. https://doi.org/10.1016/j.radonc.2015.05.012.

[11] Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. 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. https://doi.org/10.1016/j.radonc.2010.11.017.

[12] Grégoire V, Boisbouvier S, Giraud P, Maingon P, Pointreau Y, Vieillevigne L. Management and work-up procedures of patients with head and neck malignancies treated by radiation. Cancer/Radiotherapie 2022;26:147–55. https://doi.org/10.1016/j.canrad.2021.10.005.

[13] Lim JY, Leech M. Use of auto-segmentation in the delineation of target volumes and organs at risk in head and neck. vol. 55. Taylor and Francis Ltd; 2016. https://doi.org/10.3109/0284186X.2016.1173723.

[14] 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. https://doi.org/10.1118/1.4871620.

[15] Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. vol. 47. John Wiley and Sons Ltd; 2020. https://doi.org/10.1002/mp.14320.

[16] Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. Front Oncol 2017;7. https://doi.org/10.3389/FONC.2017.00315.

[17] Wong J, Fong A, McVicar N, Smith S, Giambattista JJ, Wells D, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol 2020;144:152–8. https://doi.org/10.1016/j.radonc.2019.10.019.

[18] J van der V, S W, H B, F M, S N. Deep learning for elective neck delineation: More consistent and time efficient. Radiother Oncol 2020;153:180–8. https://doi.org/10.1016/J.RADONC.2020.10.007.

[19] Strijbis VIJ, Dahele M, Gurney-Champion OJ, Blom GJ, Vergeer MR, Slotman BJ, et al. Deep Learning for Automated Elective Lymph Node Level Segmentation for Head and Neck Cancer Radiotherapy. Cancers (Basel) 2022;14:5501. https://doi.org/10.3390/CANCERS14225501/S1.

[20] Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep Deconvolutional Neural Network for Target Segmentation of Nasopharyngeal Cancer in Planning Computed Tomography Images. Front Oncol 2017;7:315. https://doi.org/10.3389/fonc.2017.00315.

[21] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[22] Taha AA, Hanbury A. Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. BMC Med Imaging 2015. https://doi.org/10.1186/s12880-015-0068-x.

[23] Sherer M V., Lin D, Elguindi S, Duke S, Tan LT, Cacicedo J, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. Radiother Oncol 2021;160:185–91. https://doi.org/10.1016/J.RADONC.2021.05.003.

[24] X H, MS H, PC L, LS H, DN T, P V, et al. Atlas-based auto-segmentation of head and neck CT images. Med Image Comput Comput Assist Interv 2008;11:434–41. https://doi.org/10.1007/978-3-540-85990-1_52.

[25] Sjöberg C, Lundmark M, Granberg C, Johansson S, Ahnesjö A, Montelius A. Clinical evaluation of multi-atlas based segmentation of lymph node regions in head and neck and prostate cancer patients. Radiat Oncol 2013;8. https://doi.org/10.1186/1748-717X-8-229.

[26] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation 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. https://doi.org/10.1016/j.ijrobp.2010.07.009.

[27] Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: A clinical validation. Radiat Oncol 2013;8. https://doi.org/10.1186/1748-717X-8-154.

[28] Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of Automatic Atlas-Based Lymph Node Segmentation for Head-and-Neck Cancer. Int J Radiat Oncol Biol Phys 2010;77:959–66. https://doi.org/10.1016/j.ijrobp.2009.09.023.

[29] Gorthi S, Duay V, Houhou N, Bach Cuadra M, Schick U, Becker M, et al. Segmentation of head and neck lymph node regions for radiotherapy planning using active contour-based atlas registration. IEEE J Sel Top Signal Process 2009;3:135–47. https://doi.org/10.1109/JSTSP.2008.2011104.

[30] 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;37:6338–46. https://doi.org/10.1118/1.3515459.

[31] Qazi AA, Pekar V, Kim J, Xie J, Breen SL, Jaffray DA. Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach. Med Phys 2011. https://doi.org/10.1118/1.3654160.

[32] Urago Y, Okamoto H, Kaneda T, Murakami N, Kashihara T, Takemori M, et al. Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models. Radiat Oncol 2021;16:175. https://doi.org/10.1186/s13014-021-01896-1.

[33] Costea M, Zlate A, Durand M, Baudier T, Grégoire V, Sarrut D, et al. Comparison of atlas-based and deep learning methods for organs at risk delineation on head-and-neck CT images using an automated treatment planning system. Radiother Oncol 2022;177:61–70. https://doi.org/10.1016/J.RADONC.2022.10.029.

[34] Yang J, Veeraraghavan H, Armato SG, Farahani K, Kirby JS, Kalpathy-Kramer J, et al. Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017. Med Phys 2018;45:4568–81. https://doi.org/10.1002/mp.13141.

[35] Ung M, Rouyar-Nicolas A, Limkin E, Petit C, Sarrade T, Carre A, et al. Improving Radiotherapy Workflow Through Implementation of Delineation Guidelines & AI-Based Annotation. Int J Radiat Oncol 2020;108:e315. https://doi.org/10.1016/J.IJROBP.2020.07.753.

[36] 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;23:903–21. https://doi.org/10.1109/TMI.2004.828354.

[37] Yang X, Jani AB, Rossi PJ, Mao H, Curran WJ, Liu T. Patch-Based Label Fusion for Automatic Multi-Atlas-Based Prostate Segmentation in MR Images n.d. https://doi.org/10.1117/12.2216424.

[38] Han X. Learning-Boosted Label Fusion for Multi-atlas Auto-Segmentation. Lect Notes Comput Sci (Including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 2013;8184 LNCS:17–24. https://doi.org/10.1007/978-3-319-02267-3_3.

[39] Lee H, Lee E, Kim N, Kim J ho, Park K, Lee H, et al. Clinical evaluation of commercial

atlas-based auto-segmentation in the head and neck region. Front Oncol 2019;9:1–9. https://doi.org/10.3389/fonc.2019.00239.

[40] Amjad A, Xu J, Thill D, Lawton C, Hall W, Awan MJ, et al. General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis. Med Phys 2022;49:1686–700. https://doi.org/10.1002/mp.15507.

[41] Robert C, Munoz A, Moreau D, Mazurier J, Sidorski G, Gasnier A, et al. Clinical implementation of deep-learning based auto-contouring tools-Experience of three French radiotherapy centers. Cancer/Radiotherapie 2021;25:607–16. https://doi.org/10.1016/j.canrad.2021.06.023.

[42] Biston M-C, Costea M, Gassa F, Serre A-A, Voet P, Larson R, et al. Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer. Phys Medica 2021;87:31–8. https://doi.org/10.1016/j.ejmp.2021.05.037.

Table 1

Characteristics of the testing cohort used for evaluation of the AS solutions

	Tumor localization	TNM	Delineated CTVn	BMI
Patient 1	Nasopharynx	T3 N1 M0	2 – 4 L/R	19
Patient 2	Nasopharynx	T2 N2 M0	3-4 L/R	21.4
Patient 3	Nasopharynx	T1 N1 M0	2-4 L, 3-4 R	33.7
Patient 4	Hypopharynx	T4 N3b M0	2-4 L	18.6
Patient 5	Hypopharynx	T1 N1 M0	2 – 4 L/R	23.4
Patient 6	Hypopharynx	T4b N0 M0	2 – 4 L/R	19
Patient 7	Larynx	T2 N0 M0	2 – 4 L/R	24
Patient 8	Larynx	T2 N0 M0	2-4 R	30.4
Patient 9	Larynx	T3 N0 M0	2 – 4 L/R	21.8
Patient 10	Larynx	T4a N0 M0	2 – 4 L/R	21.5
Patient 11	Oropharynx	T2 N1 M0	2 – 4 L/R	25.5
Patient 12	Oropharynx	T1 N1 M0	2-4 L, 3-4 R	17.9
Patient 13	Oropharynx	T2 N1 M0	2-4 R	23.8
Patient 14	Oropharynx	T2 N0 M0	2 – 4 L	32.4
Patient 15	Oropharynx	T3 N0 M0	2 – 4 L/R	31.5
Patient 16	Oropharynx	T2 N1 M0	2 – 4 L/R	32.6
Patient 17	Oral cavity	T4a N2c M0	2-4 R, 4 L	30.2
Patient 18	Oral cavity	T3 N1 M0	2-4 L, 3-4 R	24.2
Patient 19	Oral cavity	T2 N0 M0	2 – 4 L/R	22.1
Patient 20	Oral cavity	T3 N2a M0	2 – 4 R	21.0

Abbreviations: TNM= tumor-node-metastasis, BMI= body mass index

Table 2
Characteristics of the 6 automatic segmentation solutions investigated in the study

	Solution name	Software Vendor	Nr. of Atlases/ Nr. of training patients	Commercially available
1. ABAS.1	STAPLE* [36]	ADMIREv3.41 (Elekta AB, Stockholm, Sweden)	N=10	Yes
2. ABAS.2	Patch Fusion* [37]		N=10	Yes
3. ABAS.3	Random Forest* [38]		N=10	No
4. ABAS.4	Majority Voting [39]	MIM Maestro 7.0 (MIM Software Inc., Cleveland, OH)	N=10	Yes
5. DL.1	ADMIRE-DL [34,40]	ADMIREv3.41 (Elekta AB, Stockholm, Sweden)	N=49 mono-centric patient data	No
6. DL.2	ART-plan Annotate [35,41]	ART-plan (Therapanacea, France)	N>1000 multi-centric patient data	Yes

Abbreviations: STAPLE = Simultaneous Truth and Performance Level Estimation

*Same vendor

Table 3. Dosimetric study results. Per PTV and CTVn level average doses for the reference plans and dose differences (ΔD) observed between reference and experimental plans generated either with AS contours or AS+manually corrected contours. With * are highlighted the significant differences ($p<0.05$).

		mean+1SD [min, max]	meanΔD+1SD [ΔDmin, ΔDmax]			meanΔD+1SD [ΔDmin, ΔDmax]		
	DVH parameters	Reference doses	ABAS.2	DL.1	DL.2	ABAS.2 + corr	DL.1 + corr	DL.2 + corr
PTV_70Gy	$V_{95\%}$ (%)	97.4 ± 1.3 [95, 99.8]	-0.3 ± 0.9 [-2, 0.6]	-0.2 ± 0.7 [-2.1, 0.8]	-0.1 ± 0.6 [-1.3, 0.8]	-0.2 ± 0.6 [-1.7, 0.6]	-0.1 ± 0.9 [-2.5, 0.7]	-0.3 ± 1 [-3, 1.1]
	$D_{2\%}$(Gy)	72.7 ± 0.2 [72.4, 73.1]	0.2 ± 0.4 [-0.3, 1.3]	0 ± 0.3 [-0.4, 0.4]	0.1 ± 0.3 [-0.4, 0.7]	0 ± 0.3 [-0.6, 0.4]	-0.1 ± 0.3 [-0.6, 0.3]	-0.1 ± 0.3 [-0.6, 0.3]
	$D_{50\%}$(Gy)	70.4 ± 0.3 [69.9, 70.9]	0 ± 0.1 [-0.3, 0.2]	0 ± 0.2 [-0.5, 0.3]	0 ± 0.2 [-0.2, 0.4]	-0.1 ± 0.2 [-0.4, 0.3]	-0.1 ± 0.2 [-0.4, 0.3]	-0.1 ± 0.2 [-0.4, 0.3]
PTV_54.25Gy	$V_{95\%}$ (%)	98.7 ± 0.6 [97.7, 100]	5.7 ± 3 [0.7, 11.1] *	5.9 ± 4.2 [0.3, 13.2] *	4.1 ± 2.1 [1.1, 8.8] *	3.5 ± 1.8 [0.1, 5.6] *	4 ± 2 [0.2, 6.4] *	3.3 ± 1.8 [0.1, 5.7] *
	$D_{98\%}$(Gy)	52.2 ± 0.7 [50.7, 53.6]	8.3 ± 5 [0.2, 17.3] *	8.2 ± 6.3 [0.2, 21.3] *	6.1 ± 3.7 [0.6, 11.9] *	6.3 ± 6.9 [0.1, 25.7] *	5 ± 2.5 [0, 7.4] *	4.9 ± 3.5 [0.1, 12.1] *
	$D_{50\%}$(Gy)	56.5 ± 1.5 [55.7, 60.9]	0.2 ± 0.2 [0, 0.6] *	0.2 ± 0.2 [-0.1, 0.4] *	0.2 ± 0.2 [0, 0.5] *	0.1 ± 0.2 [-0.2, 0.3] *	0.2 ± 0.1 [0, 0.3] *	0.1 ± 0.1 [-0.2, 0.3] *
CTVn2	$V_{95\%}$ (%)	100 ± 0 [99.9, 100]	1.3 ± 1.2 [0.3, 4.8] *	1.7 ± 2.3 [0, 7.3] *	0.3 ± 0.3 [0, 0.8] *	0.4 ± 0.5 [0, 1.7] *	0.6 ± 0.6 [0, 1.7] *	0.3 ± 0.4 [0, 1.2] *
	$D_{98\%}$(Gy)	53.5 ± 0.3 [52.7, 54]	1.1 ± 1.5 [-0.2, 5.5] *	1.8 ± 3.1 [-0.2, 9.8] *	0.3 ± 0.3 [-0.1, 0.7] *	0.4 ± 0.8 [-0.1, 2.9] *	0.5 ± 0.6 [-0.2, 1.6] *	0.1 ± 0.3 [-0.3, 0.8]
	$D_{50\%}$(Gy)	55.9 ± 0.4 [55.4, 56.6]	0.1 ± 0.2 [-0.1, 0.5] *	0 ± 0.1 [-0.2, 0.2]	0.1 ± 0.2 [-0.2, 0.7] *	0.1 ± 0.2 [-0.2, 0.4]	0 ± 0.2 [-0.2, 0.4]	0 ± 0.2 [-0.3, 0.4]
CTVn3	$V_{95\%}$ (%)	100 ± 0 [99.9, 100]	1.6 ± 2.3 [0, 8] *	1.5 ± 2.7 [0, 8.9] *	0.1 ± 0.3 0.2 [0, 0.9] *	0.1 ± 0.1 0.2 [0, 0.5] *	0.1 ± 0.1 0.2 [0, 0.4] *	0 ± 0 [-0.1, 0.1]
	$D_{98\%}$(Gy)	53.6 ± 0.6 [52.4, 54.7]	2.2 ± 4.4 [-0.1, 14.9] *	2.4 ± 5.3 [-0.2, 17.2] *	0.1 ± 0.4 [-0.4, 1]	0.1 ± 0.1 [-0.1, 0.3]	0.1 ± 0.2 [-0.3, 0.6]	-0.1 ± 0.3 [-0.6, 0.4]
	$D_{50\%}$(Gy)	56.5 ± 3 [55, 65.5]	0.1 ± 0.1 [-0.1, 0.3] *	0.1 ± 0.2 [-0.2, 0.4]	0.1 ± 0.2 [-0.2, 0.5] *	0.1 ± 0.1 [-0.1, 0.4]	0.1 ± 0.2 [-0.2, 0.3]	-0.1 ± 0.2 [-0.4, 0.2]
CTVn4	$V_{95\%}$ (%)	100 ± 0 [99.9, 100]	2.8 ± 3.1 [0, 8.7] *	1.9 ± 2.4 [0.1, 8.3] *	8.4 ± 11.9 [0.1, 39.7] *	1.7 ± 2.7 [0, 7.5] *	2.6 ± 3.7 [0, 11.4] *	2.8 ± 6 [0, 20.1] *
	$D_{98\%}$(Gy)	53.5 ± 0.5 [52.8, 54]	4.3 ± 5.5 [0, 15.4] *	2.3 ± 3.7 [-0.1, 12.3] *	5.4 ± 6.8 [-0.1, 20] *	1.3 ± 2 [-0.7, 5.7] *	1.7 ± 2.3 [-0.3, 7.1] *	1.6 ± 2.8 [-0.4, 8.4] *
	$D_{50\%}$(Gy)	56 ± 0.4 [55.5, 56.7]	0.3 ± 0.2 [-0.1, 0.5] *	0.1 ± 0.2 [-0.3, 0.5]	0.6 ± 0.9 [-0.3, 3] *	0.2 ± 0.3 [-0.2, 0.6] *	0.2 ± 0.3 [-0.3, 0.8] *	0.3 ± 0.5 [-0.2, 1.6] *

Figure 1 panel A

Click here to access/download;Figure;Figure1.Pannel- A_revision.jpg

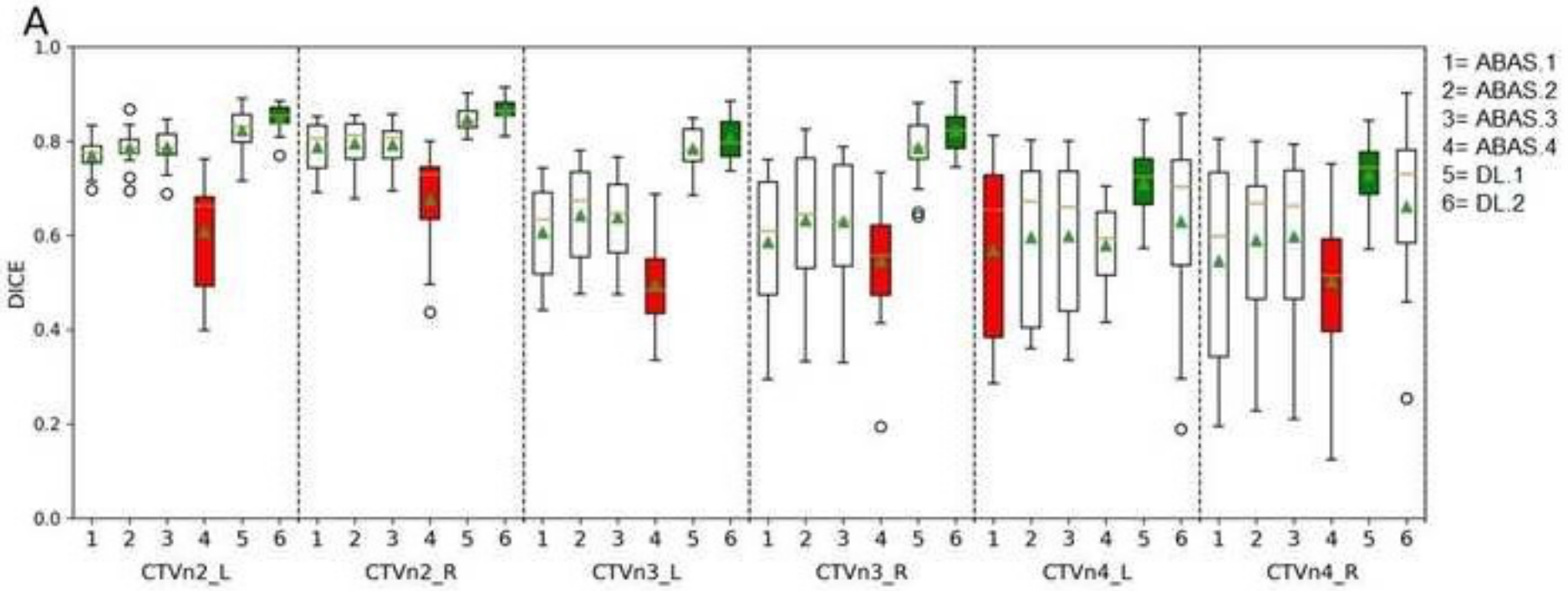


Figure 1 panel B

Click here to access/download;Figure;Figure1.Pannel- B_revision.jpg

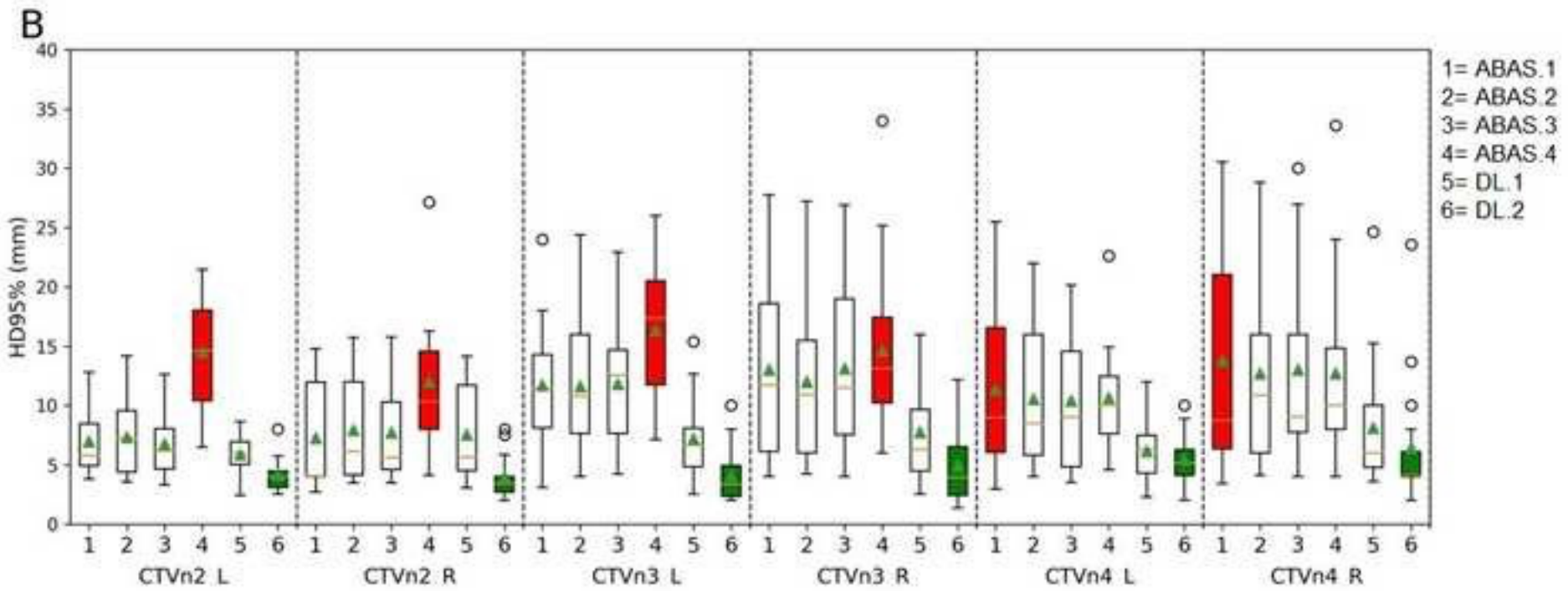


Figure 1 panel C

Click here to access/download;Figure;Figure1. Pannel- C_revision.jpg

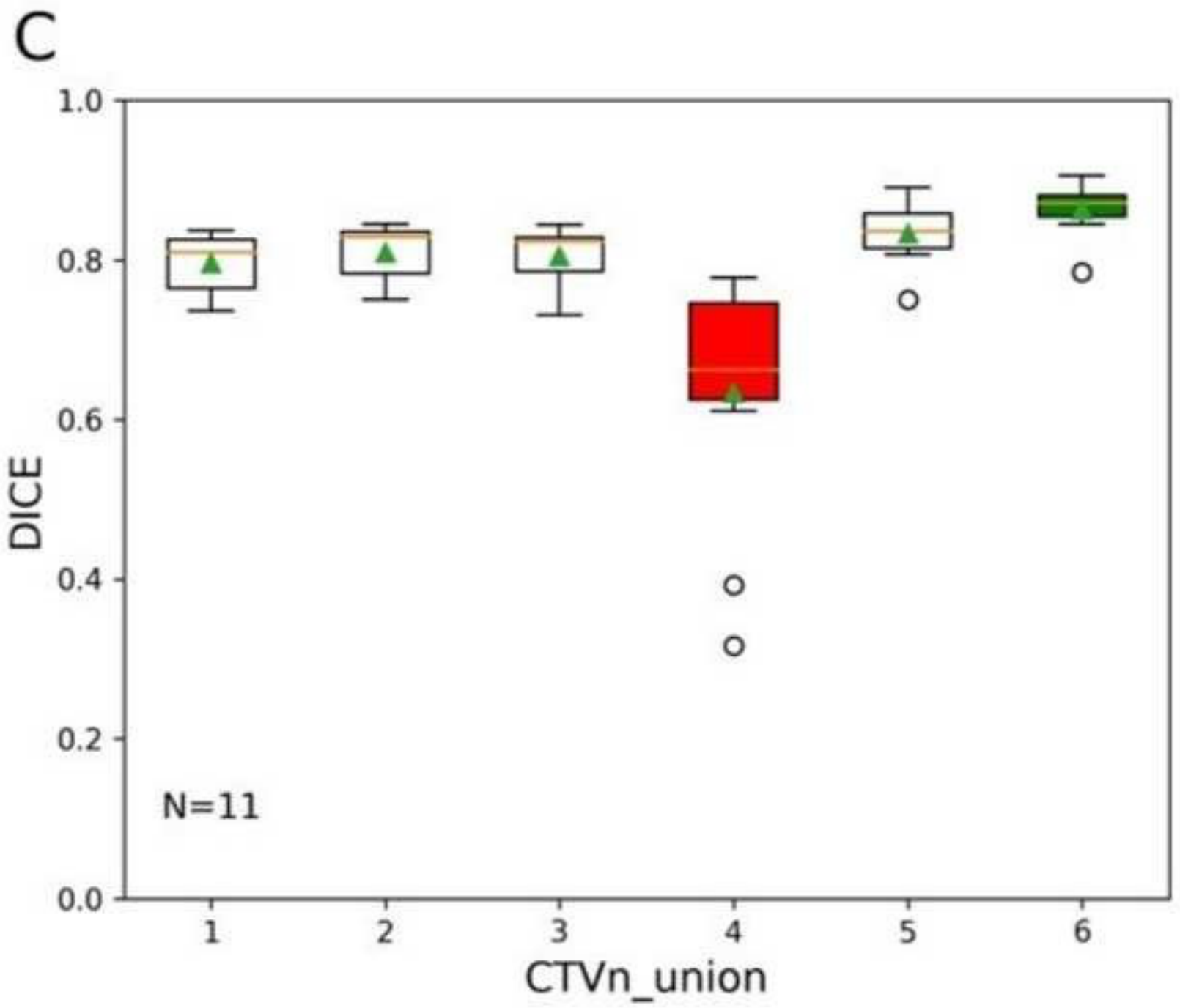


Figure 1 panel D

Click here to access/download;Figure;Figure1.Pannel- D_revision.jpg

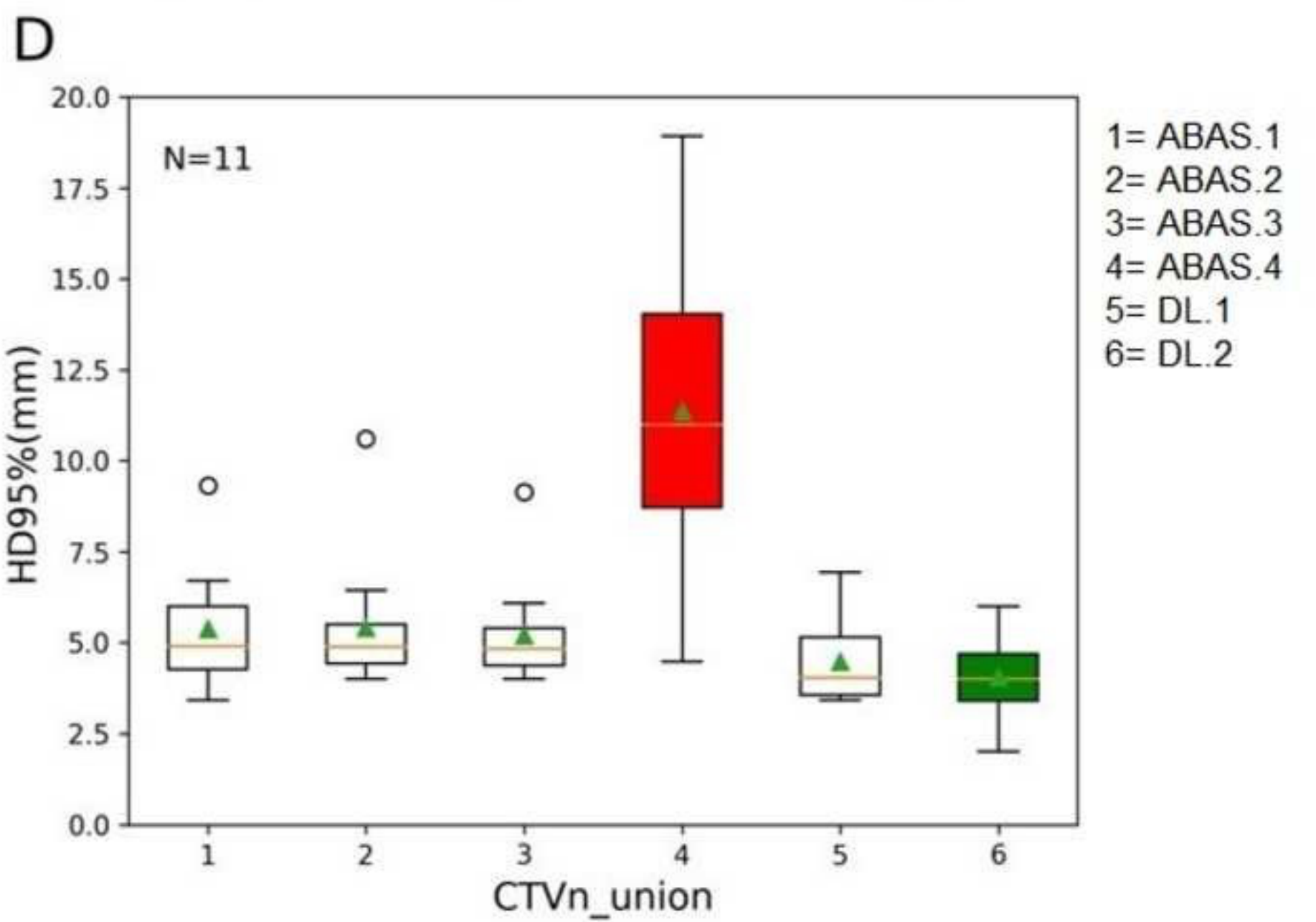


Figure 2

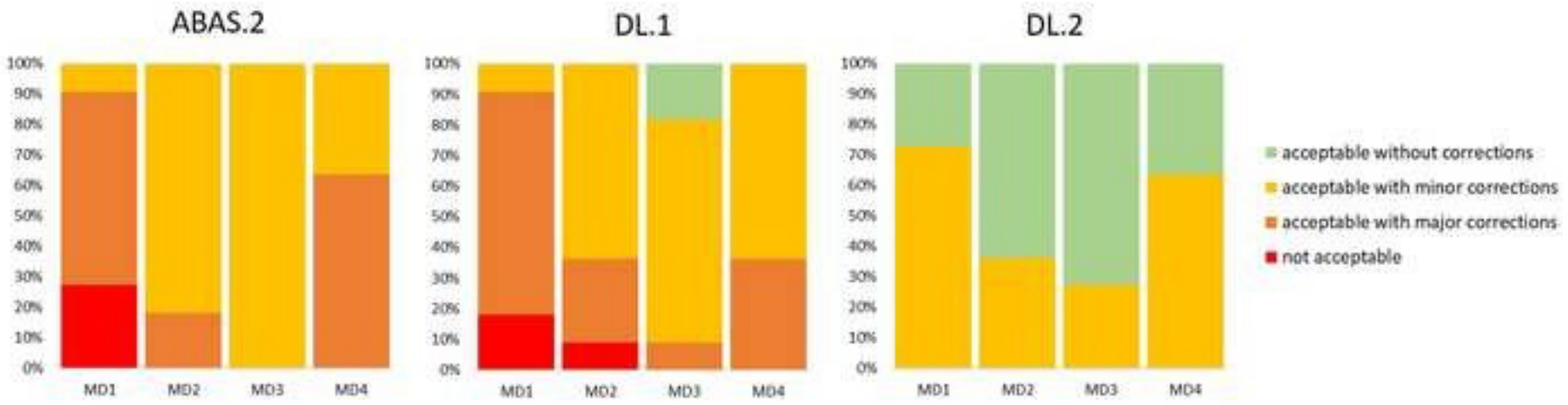


Figure 3

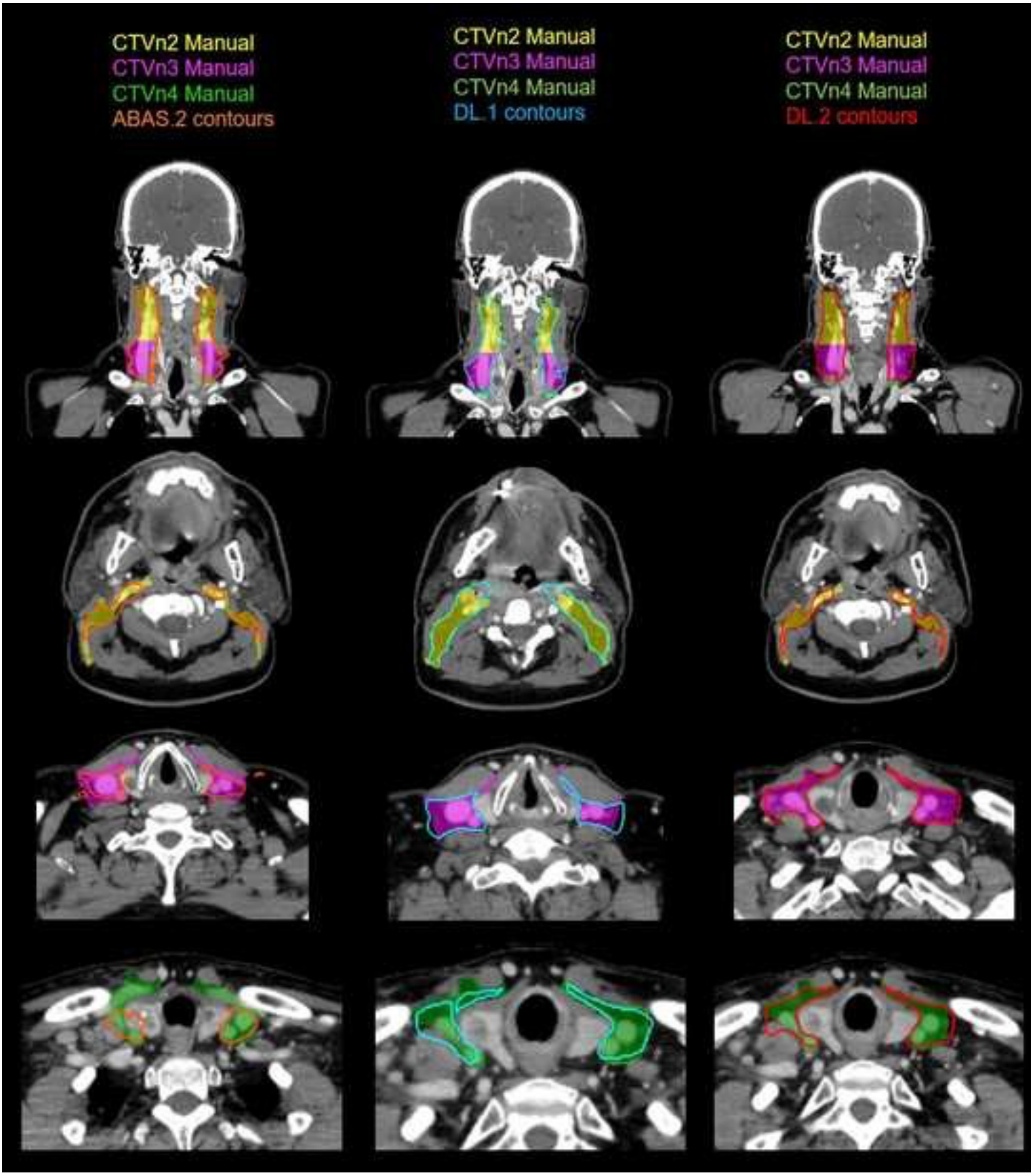
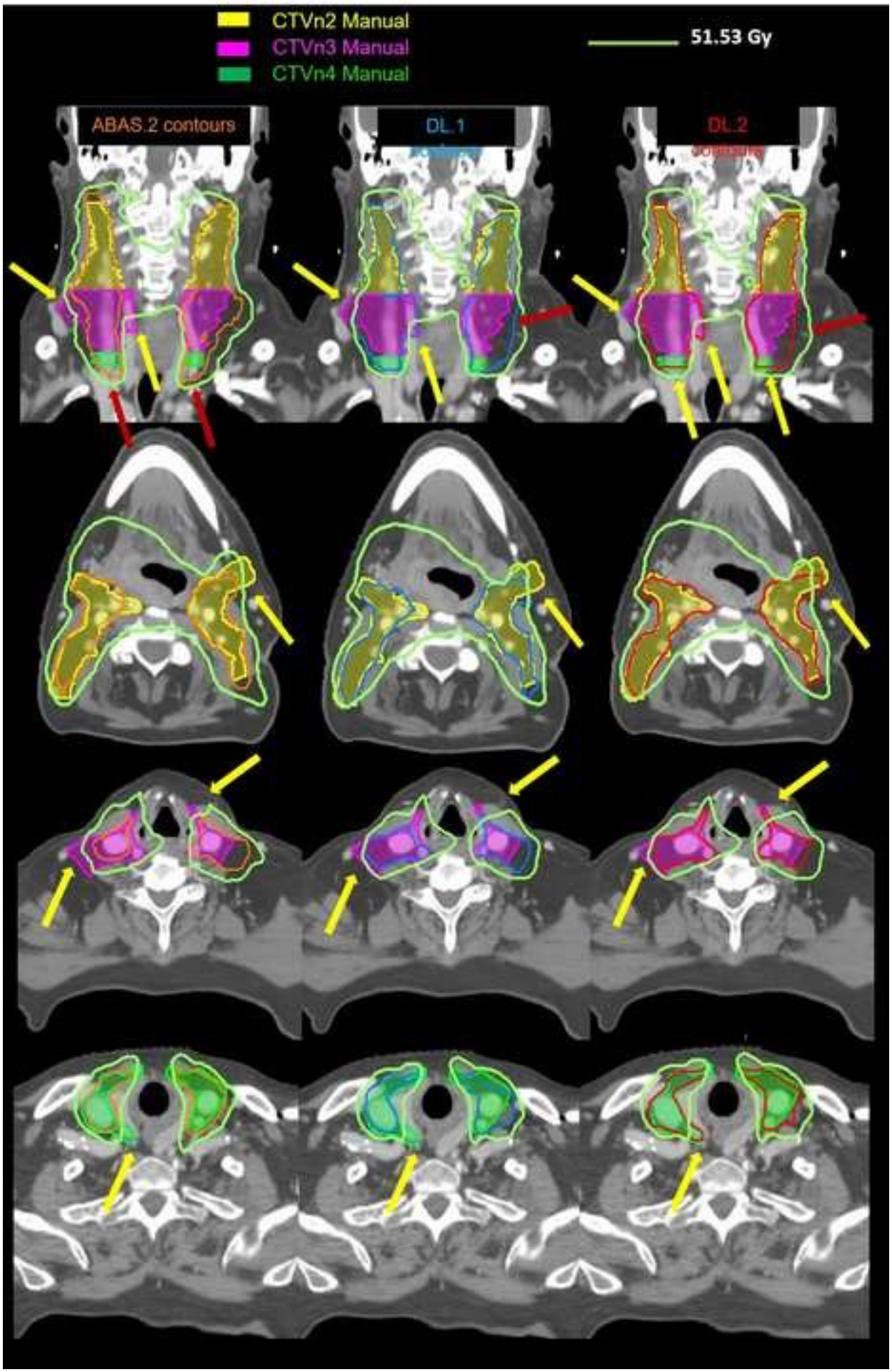


Figure 4

Click here to access/download;Figure;FIgure 4.jpg



Conflicts of interest statement

This work was performed in the framework of a research cooperation agreement with Elekta AB.