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

Madalina Costea, Alexandra Zlate, Morgane Durand, Thomas Baudier, Vincent Grégoire, David Sarrut, Marie-Claude Biston

Title page

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

Madalina Costea[a, b] , Alexandra Zlate[c], Morgane Durand[a], Thomas Baudier[a,b], 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

*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

Abstract

Background and purpose: To investigate the performance of head-and-neck (HN) organs-at-risk (OAR) automatic segmentation (AS) using four atlas-based (ABAS) and two deep learning (DL) solutions.

Material and Methods: All patients underwent iodine contrast-enhanced planning CT. Fourteen OAR were manually delineated. DL.1 and DL.2 solutions were trained with 63 mono-centric patients and >1000 multi-centric patients, respectively. Ten and 15 patients with varied anatomies were selected for the atlas library and for testing, respectively. The evaluation was based on geometric indices (DICE coefficient and 95th percentile-Hausdorff Distance ($HD_{95\%}$)), time needed for manual corrections and clinical dosimetric endpoints obtained using automated treatment planning.

Results: Both DICE and $HD_{95\%}$ results indicated that DL algorithms generally performed better compared with ABAS algorithms for automatic segmentation of HN OAR. However, the hybrid-ABAS (ABAS.3) algorithm sometimes provided the highest agreement to the reference contours compared with the 2 DL. Compared with DL.2 and ABAS.3, DL.1 contours were the fastest to correct. For the 3 solutions, the differences in dose distributions obtained using AS contours and AS+manually corrected contours were not statistically significant. High dose differences could be observed when OAR contours were at short distances to the targets. However, this was not always interrelated.

Conclusion: DL methods generally showed higher delineation accuracy compared with ABAS methods for AS segmentation of HN OAR. Most ABAS contours had high conformity to the reference but were more time consuming than DL algorithms, especially when considering the computing time and the time spent on manual corrections.

Introduction

Manual contouring of organs-at-risk (OAR) is a time-consuming task that suffers from large intra- and inter-observer variations, especially for head-and-neck (HN) cancer patients, because of the complex anatomy and the number of OAR [1–4]. Contour variations may also result in important dosimetric differences [5]. Therefore, automatic segmentation (AS) methods are strongly sought after to increase contouring accuracy, improve the inter-observer variability, reduce delineation time, and facilitate treatment plan adaptation [6, 7].

Among the different methods, atlas-based segmentation (ABAS) uses one or more representative patients with carefully delineated OAR as reference atlas library for contouring new patients [8]. Those methods are widely spread because they require minimum of resources, but they do have several drawbacks: atlas selection strategy (single vs multi-atlas) [8]; performance plateau reached after 10-20 atlases [9]; poor performance for small and low contrast soft tissue structures [10]; increased computational time with each added atlas [11].

Data from multiple atlases (multi-ABAS) can be combined with the help of a fusion algorithm in order to reduce the risk of anatomical variability between the atlas and the new patient [12]. Additionally, hybrid approaches are developed to combine multi-ABAS with machine learning features [13–17]. Despite a higher computational time, multi-ABAS studies have consistently demonstrated improved conformity to the reference contours over the single atlas methods, with consequent reduction of the post-editing time [18, 19]. By adding image intensity information, other studies have shown improved accuracy for model-based methods particularly on large organs such as brainstem and spinal cord but lacking precision for tiny structures like cochlea [7], [13–15].

Another method issued from artificial intelligence (AI) research and challenging ABAS is the use of deep learning (DL) techniques [6], [7], [20–26]. DL contouring typically implies the training of a convolutional neural network (CNN) directly from a set of annotated reference data. Although the training phase requires extensive GPU computing power and work in data gathering and curation, once trained, the segmentation is very fast. Different network architectures are continuously investigated to reach the best predictions for multiple organ segmentation. While some models are accurate on most volumes, they may have difficulties in segmenting small volumes such as optical nerves or cochlea, or organs with low image contrast such as constrictor muscles. Comparison between different DL models is rather difficult due to differences in the data sets used. From the few studies analyzing the performance of different DL models

trained and tested on the same data sets, *Chen et al.* examined one multi-ABAS and three similar DL models following U-Net-like network architectures with distinctive differences in the configuration and loss functions [25]. While nnU-net [27] is a self-configuring network based on the training dataset, AnatomyNet [28] follows a defined scheme with squeeze-and-excitation residual blocks for better feature representation and a combination of two loss functions (DICE and Focal Loss). By using Ua-Net [29] for the HN model, that first performs an OAR detection module and then considers image features only within the detected regions, WBNet was superior to the other methods for most organs. Apart from the in-house developed models, several commercially available solutions have reported good agreement with physicians' manual contours and considerable time savings on the delineation task [23, 24], [30–32].

Most studies showed that DL methods outperformed ABAS methods [23–25]. However, there is still room for improvement in the AS of computed tomography (CT) images for small organs or with limited image contrast such as optic nerves, optic chiasm or cochlea [10], [20], [25], [32].

Generally, AS methods comparisons are based on geometric indices calculations only (DICE; Hausdorff distance (HD)) to compare the volume overlap between the reference and the automatically generated contour [33]. However, it is highly recommended to perform additionally a dosimetric evaluation by generating treatment plans with the AS contours [7], [34–36]. Nevertheless, this involves extensive time in generating treatment plans, and may also introduce inter or intra-planner variability [37, 38].

In this context, the objective of the present study was to evaluate the performances of 4 atlas-based algorithms and 2 DL solutions for the AS of 14 HN OAR. Three multi-ABAS algorithms and one DL solution are commercially available while one hybrid-ABAS algorithm and one center-specific DL solution were investigated for the first time on HN CT images. All six solutions were evaluated based on geometrical accuracy and computational time. The time spent for correcting the contours was measured for the most accurate three AS methods and an auto-planning solution based on a priori multi-criterial optimization (MCO) algorithm was used for the first time to derive doses from AS contours with and without manual correction.

Materials and methods

Patient data

Seventy-eight non-operated HN cancer patients treated with radiation therapy between 2018 and 2021 and who underwent iodine contrast-enhanced planning CT, were selected for this study, which was approved by the hospital ethics committee. Fourteen OAR (i.e. parotids, submandibular glands, oral cavity,

constrictor muscle, larynx, esophagus, trachea, thyroid, eyes, optical nerves, cochlea, brainstem, spinal cord, mandible) were manually delineated by a single expert physician (>30 years of experience), on 512x512 and 2mm-tick CT slices following HN delineation guidelines [3]. An overview of the study design is provided in Fig. 1. For the multi-ABAS approach, 10 patients from this database were selected based on their body mass index (BMI) (from 18.9 to 30.7) to form a heterogeneous library of atlases with various representative patient anatomies. The same 10 atlases were used to create a library in MIM-Maestro (MIM Software; Cleveland, USA) and in research version of the ADMIRE software (ADMIREv3.41, Elekta AB; Stockholm, Sweden). A mono-centric DL.1 model was trained using 63 patients with the same set of OAR excluding optical nerves and cochlea. Conversely, DL.2 model was trained on a large database of patients (>1000) collected from multiple centers including ours (Fig. 1). Fifteen patients having a BMI ranging from 12.1 to 34.7 were reserved for the testing phase. Characteristics of the test cohort are detailed in Table 1.

Multi-ABAS and DL methodologies

Three multi-ABAS solutions integrated in the research version of Monaco treatment planning system (TPS) [39] (Monaco 5.59.11 with ADMIREv.3.41) and another one available in MIM-Maestro (MIM Software Inc., Cleveland, OH) were investigated:

- ABAS.1: Simultaneous Truth And Performance Level Estimation (STAPLE) consists in estimating the optimal combination of the atlases segmentations by weighting each segmentation upon the estimated performance level based on expectation-maximization algorithm [12].
- ABAS.2: Patch Fusion (PF) algorithm computes the final probability of a voxel to belong to a structure as a weighted average of the atlases' contours based on voxel intensity information [40].
- ABAS.3: Random Forest (RF) is a supervised learning algorithm which constructs a voxel classifier for each structure using the registered atlases as training data [16].
- ABAS.4: Majority voting (MIM) [41].

For the ADMIRE software, out of the 10 atlases used, a reference patient was selected for each test patient based on the closest BMI of the atlas and the underlying patient. No individual atlas selection was required for MIM, but a general *template scan* (patient having an anatomy close to the mean BMI of the atlas cohort) was registered with all the atlases in the library.

Two DL models were investigated:

- DL.1: ADMIRE-DL (ADMIREv.3.41, Elekta AB, Stockholm) trained with N=63 patients from one center. It is a fully connected deep convolutional neural network (DCNN) with 3D U-net architecture and short-range residual connections developed from the ResUnet3D network [42]. While the encoding part is responsible for learning multi-scale multi-dimensional image features

in multiple levels, the combination of long and short-range connections allows the decoding part to preserve the high-resolution image features and produce a label map corresponding to the input image size [42, 43].

- DL.2: ART-plan Annotate (Therapanacea, France) trained on a large database with N>1000 patients obtained from several clinical sites. The model uses anatomy preserving DL ensemble networks that first detects organs through DL-based registration to a collection of whole-body annotated volumes. Then, the delineation of each anatomical structure is performed through an original combination of data-driven and decisional artificial intelligence that enforces anatomical consistency [30, 31].

Geometric evaluation of auto-segmentation solutions

To quantitatively evaluate the segmentation results, we used two geometric indices: volumetric DICE and 95thpercentile-Hausdorff distance ($HD_{95\%}$) [33]. DICE is a measure of the volumetric overlap between the ground truth contour (A) and the predicted segmentation (B), leading to a value between 0 (no overlap) and 1 (perfect overlap):

$$DICE = \frac{2x|A \cap B|}{|A|+|B|}$$

However, DICE is limited to the pixels overlap without considering the shape differences. Therefore, a second metric was used to indicate the magnitude of mislocalization of the prediction. The HD is a boundary-based metric that measures the surface distances between the predicted contour and the ground truth segmentation. To eliminate the possible outliers, we used $HD_{95\%}$:

$$HD_{95\%} = max_{k95\%}[d(A,B), d(B,A)]$$

$$d(A,B) = \max_{a \in A} \min_{b \in B} \|a - b\|$$

Where d(A,B) is the directed HD and A and B are the set of non-zero pixels in the images. HD metric has its own limitation that does not focus on the object itself therefore does not punish a prediction with a large hole inside or with a spotted pattern within the contour [44]. For the elongated organs (i.e., esophagus, trachea, constrictor muscle and spinal cord) the results were calculated only on the slices where both contours were present to avoid situations where the reference ground truth was missing.

Time needed for manual corrections

Three of the automatic solutions (ABAS.3, DL.1 and DL.2) were clinically reviewed and corrected by a dosimetrist and validated by a skilled physician on Monaco contouring workstation following the regular clinical routine. The time spent on correcting and validating each structure was recorded independently.

Dosimetric evaluation – automatic treatment plans

For each patient, and for ABAS.3, DL.1 and DL.2 solutions, 2 different plans were generated: one using the AS contours and another one using AS+manually corrected contours. The differences in dose distributions were then evaluated on the corrected contours. In total, 90 VMAT treatment plans were calculated with mCycle auto-planning solution (Monaco 5.59.11, Elekta AB, Stockholm). The software uses a lexicographic MCO which has been extensively described before [45]. All plans were performed using two 360° arcs. A simultaneous integrated boost technique was used for delivering 70Gy to the planned target volume (PTV) associated to the primary tumor and 54.25Gy to the PTV associated to prophylactic nodal target, in 35 fractions of 2Gy. Clinically relevant dosimetric endpoints for target volumes ($V_{95\%}$) and OAR (D_{mean}, $D_{2\%}$, $D_{5\%}$) were considered upon the clinical protocol and according to the recommendations of the French Society of Radiation Oncology [46].

Statistical analysis

Per organ and per algorithm, statistical differences between methods were assessed using the non-parametric Kruskal-Wallis test. Subsequently, to detect between which pairs of algorithms the differences were significant, the post-hoc Dunn's test with Bonferroni correction was applied. Similarly, the differences between radiotherapy doses derived from AS contours with or without corrections were tested for statistical significance. P-values <0.05 were considered significant. The statistical analysis was performed using the libraries (scipy 1.6 and scikit-posthocs 0.7) in Python 3.8.

Results

Computational time per patient was in average 10.3±1.6min, 10.5±0.6min and 12.1±0.6min for ABAS.1, ABAS.2 and ABAS.3, respectively. For ABAS.4, it was under 1min while the atlas registration took approximately 6min for a library of 10 atlases. DL.1 and DL.2 provided a solution in less than 1min and 2min, respectively. Per algorithm and per OAR, DICE scores and $HD_{95\%}$ distance results of all solutions are summarized in Fig. 2 and Fig. 3, respectively.

Overall, both DICE and $HD_{95\%}$ results indicated that DL algorithms were more accurate than ABAS algorithms for AS of HN OAR. The Kruskal-Wallis statistical test identified significant differences between the 6 AS methods. However, the post-hoc paired test showed no statistical difference in terms of DICE and $HD_{95\%}$ between the DL.1 and DL.2 and between ABAS.3 and the 2 DL solutions. With 11 common OAR, DL.1

had overall better contour overlap compared with DL.2 with a DICE average of 0.85±0.32 vs 0.82±0.06 and 11 vs 9 OAR having DICEs ≥0.8. Per organ differences however did not reach a statistically significant level.

Regarding ABAS solutions, ADMIRE ABAS algorithms had overall better DICEs and $HD_{95\%}$ than ABAS.4, which had the lowest DICE results for 7 out of 14 OAR. While all the ADMIRE solutions had DICE results ≥0.8 for 7 OAR, ABAS.3 contours were closer to the reference contours. Per OAR statistics revealed however no significant differences in DICE and $HD_{95\%}$ between ABAS.2 and ABAS.3 and, compared with ABAS.1, both ABAS.2 and ABAS.3 performed significantly better for the eyes ($p<0.02$). Moreover, compared with ABAS.4, ABAS.3 performed significantly better for parotids ($p<0.003$), mandible ($p<0.01$) and submandibular glands ($p<0.02$). Note that ABAS.3 did not segment optic nerves and cochlea owing the algorithm's limitation for such small structures.

Compared with DL.1, ABAS.3 had significantly better DICE for the mandible ($p=0.02$). Compared with DL.2, ABAS.3 had significantly better DICE for the eyes ($p=0.01$) and for the mandible ($p=0.01$). On the opposite, DL.2 had significantly better DICE for the esophagus ($p=0.04$) and significantly better $HD_{95\%}$ for the thyroid ($p=0.03$). Finally, the superiority of DL. 1 over ABAS.3 was not statistically demonstrated.

An example of AS contours from ABAS.3, DL.1 and DL.2 in contrast with the physicians' manual delineations is provided in Fig. 4.

Ten of the OAR obtained with ABAS.3, DL.1 and DL.2 (best solutions graded based on the geometric accuracy results) were thereafter carefully corrected by a dosimetrist and checked by a physician. Manual corrections were done organ by organ on all the CT slices. The targets were never displayed, to not influence the observers. The manual correction time per patient was in average 36min34sec, 17min54sec and 26min57sec for ABAS.3, DL.1, and DL.2, respectively. The contours generated by DL.1, were the fastest to correct. In general, manual corrections of eyes, spinal cord and brainstem were <2min for the 3 solutions while for oral cavity and esophagus correction times were >3min depending on the AS algorithm used (Fig. 5).

After manual corrections, the DICE scores of all OAR were improved, except for the oral cavity on all 3 solutions, and for the spinal cord on DL.2 solution (Table 3), thus highlighting inter-observer variability in contouring the oral cavity between the expert physician providing the reference contours, and the other physician performing manual corrections. At the same time, the $HD_{95\%}$ did not decrease consistently for all the structures after the manual corrections, confirming, once more, the variability in manual delineation between observers. While performing correction on the DL.1 contours did not significantly improve DICE

and $HD_{95\%}$ results, for DL.2 contours, results were significantly improved for the trachea ($p<0.001$). For ABAS.3, the improvements were statistically significant for esophagus ($p<0.001$) and thyroid ($p<0.001$).

The differences in doses on corrected OAR, between treatment plans generated using the AS contours, with or without manual corrections are presented in Table 4. No statistically significant difference was found between doses for the 3 solutions. For each patient, a minimum distance between each OAR and the targets was calculated. Among OAR having a maximum dose constraint, the mandible had the largest dose difference when it overlapped with the PTV. For the brainstem and spinal cord, the largest dose differences occurred when the OAR was located at a larger distance to the PTV (>30mm). For the parotids and for the submandibular glands, maximum differences occurred when the OAR overlapped with the PTV. For the oral cavity, for the eyes and for the esophagus, the maximum differences were generally observed at distances<20 mm from the PTV. However, for the esophagus, there were some outliers at larger distances from the target (>60mm) for DL.2. For the trachea, only in one patient case, and for DL.2, a large difference was observed but at a high distance from the target (80mm). Some illustrations of dose distributions with regard to corrected/non-corrected contours and PTV position are available in Fig. 1 of the Supplementary Material.

Discussion

We showed in this study that, overall, both DICE and $HD_{95\%}$ results indicated that DL algorithms performed better compared with the ABAS algorithms for automatic segmentation of HN OAR. Concerning the 2 DL solutions, out of 12 contours, DL.1 outperformed DL.2 solution in terms of DICE for 7 OAR, with, however, no statistically significant differences. Contrarily to DL.2, DL.1 was not tailored to automatically contour optic nerves and cochlea. Nevertheless, the correction of the AS contour of small organs generally takes more time than starting from scratch [47, 48] .Conversely, DL.2 was not trained to contour the constrictor muscle. However, the DL.1 results were highly inaccurate, showing the difficulty to get satisfying results for such organs with high anatomical variations and low image contrast. Therefore, consistent with the literature, OAR with good CT contrast were better segmented by ABAS and DL solutions compared with small and/or thin OAR such as optic nerves or cochlea, and OAR which do not have well-defined boundaries like constrictor muscles [23–25], [28], [49, 50].

Before this study, DL.1 and DL.2 algorithms had not been explored on HN site. The DL.1 algorithm was trained exclusively with manual delineations coming from one expert physician, providing uniformity of the training data. Ideally, there should be a consortium for the contour delineation between physicians

working in a radiotherapy department, which should rely on internationally published guidelines [3]. In this study, with a limited training dataset (N=63), we showed that a model can achieve consistent results for most of the structures in HN. Hence, with a minimum of work, centers can adapt a model to their standard delineation's practices. Similarly, high accuracy segmentation results were obtained with the DeepVoxNet and another CNN with networks trained on N=70 and N=50 samples, respectively [47], [51]. Other studies demonstrated that organs' pattern depends on the training sample size [52] and yet similar results can be obtained when training on a small set of carefully curated data compared with a larger set of more easily available routine-level clinical annotations [53]. On the opposite, DL.2 solution was trained with more than 1000 samples per organ collected from multiple centers and can segment 50 OAR and target volumes in HN. Despite this, highly accurate contours were obtained in this study. Proving that a multi-center study approach includes combination of manual contours from different physicians (easier to obtain), DL.2 results presented good conformity to new datasets and comparable performance to a model train with data from a single center.

We also showed that, using a carefully selected atlas of patients, ADMIRE multi-ABAS methods achieved good agreement with manual contours (DICE≥0.8) [8], [18] and, for some organs, similar or better agreement with the reference contours compared with DL models (i.e. oral cavity, mandible, eyes). Conversely, ABAS.4 had overall inferior performance. Among multi-ABAS algorithms, ABAS.3, which had not been explored before, produced the best results and had significantly better DICEs than DL.1 and DL.2 solutions for mandible and eyes, respectively. Therefore, with only 10 carefully selected atlases composed of non-operated patients with a wide range of BMI, ABAS.3 algorithm may serve as an AS solution easy to implement clinically. Note that using an enlarged library of 20 patients (data not shown) did not considerably improved the performances of ABAS.3 but drastically increased the computation time, demonstrating that the performance plateau phenomena still exists with this new ABAS method.

Many studies have reported the performances of different algorithms for HN OAR segmentation on CT images (Table 1 of supplementary data). All studies underlined limited performance on small organs, and the importance of both manual contours' quality, and training data size to obtain accurate segmentations and clinically acceptable treatment plans. It was also mentioned that, for noncritical OAR (i.e. far from PTV), manual corrections could be omitted [34]. Moreover, AS has shown to reduce inter-observer variability when observers performed manual editing on the automatically generated contours, which improves the consistency of manual delineation [5].

According to the recently published guidelines, together with geometric accuracy, studies should ideally report benefit in time saving and clinical acceptability in terms of patient dose evaluation, for assessing the benefit of an automatic segmentation method [33]. Both tasks involve exhaustive labor and are not systematically conducted first because of the time requested to be completed, and secondly because of the intra-observer factor, which could introduce a bias in the observations. In this study, both tasks were completed for the three best algorithms, and an auto-planning solution was used to perform treatment plans based on AS contours with or without corrections. This was a strength of this study, and an efficient way to isolate the consequences of contour variations on the radiotherapy doses and reveal more precisely which contours require greater attention [34, 35]. Among other methods, some authors proposed to superpose the original clinical plan onto the automatically delineated contours [5], [23], to use automated planning strategies such as knowledge-based planning (KBP) [34, 35] or to conserve the original beam configuration parameters [36]. To our knowledge, this is the first time that an *a priori* MCO auto-planning solution is used for contour evaluation.

We observed in our study that, for most structures, the correction time for DL.1 and DL.2 solutions was <1min (e.g., eyes, brainstem, submandibular glands) and <2min (e.g., mandible, parotid glands) demonstrating significant time saving versus starting from scratch, particularly for the dosimetrist, whose work represented, depending on the AS solution, from 60% to 70% of the total manual editing time. Correcting DL.1 contours was 18min and 9min faster compared with ABAS.3 and DL.2 contours, respectively. Generally, the oral cavity and esophagus took more time to be corrected. For the oral cavity, this may be correlated with the inter-observer variability since the DICE results were consistently smaller for all 3 solutions after the manual corrections. We finally observed that all dose-volume constraints and target objectives were respected in all plans and that manual corrections of the AS contours had no statistically significant impact on the dose distributions. The ΔD_{mean} for the investigated structures were <0.9Gy. Generally, the range of the $\Delta D_{2\%}$ were the highest for the spinal cord and for the brainstem for all the solutions, which may be an important factor in physician's decision when validating the treatment plan. Similar to other studies, for most organs, the difference in the delivered dose was not significant [34], [36]. The dose constraints and objectives were respected for all the plans automatically generated and thus, manual correction could be omitted.

Considering the organ position relative to the PTV, high dose differences could be observed when the OAR contour overlapped with the target volumes or was located in their short vicinity. However, this was not always interrelated. This was true for the parotid glands, but for the spinal cord and brainstem, the highest $\Delta D_{2\%}$ were located at a larger distance between the OAR and the PTV (>35mm and >15mm relative to PTV

70Gy and PTV 54.25Gy, respectively). At the same time, at short distances from the PTV (<5mm), the $\Delta D_{2\%}$ in brainstem and spinal cord was <2Gy. One possible reason is that, closer to targets, the AS contours were highly accurate. Although spinal cord and brainstem presented generally good agreement with the manual reference contour, the manual corrections which were nevertheless fast, proved clinically meaningful in certain patients.

Note that this study was deliberately focused on a center-specific approach. The goal was to investigate which of the 6 AS solutions available in our department were more accurate and required less resources in terms of patient data and manpower. In particular, the objective was to evaluate whether, with a relatively small database of homogeneous contoured patients, a center could easily implement an AS solution conformed to its own contouring practices, which, nevertheless, should respect international contouring guidelines. At the same time, we evaluated a solution that was trained on a multi-centric database of contours. Note that the reference contours used in this study belonged to only one expert physician, and also, the manual corrections were done by only one dosimetrist and one physician, both trained by the reference expert. Although the study could benefit from multiple observers involved in manual corrections of the contours, this was, nevertheless, reproducing the clinical workflow of our department. Moreover, the relatively small cohort of the test patients was composed of heterogeneous patients' anatomies and tumor locations, in order to challenge the different AS solutions. Including more patients will definitively strengthen the study, in particular, the statistical analysis. Finally, these findings suggest that, acknowledging their strengths and limitations, the investigated hybrid ABAS and DL methods improved our clinical workflow.

Conclusions

DL methods generally showed higher delineation accuracy compared with ABAS methods for AS segmentation of HN OAR. We showed that a DL model can provide accurate contours with a limited training dataset, provided that data comes from a single hospital, and if possible, only one expert physician is involved. Most ABAS contours had high conformity to the reference but were more time consuming than DL algorithms, especially when considering the computing time and the time spent on manual corrections. Finally, even if manual checks and modifications must not be ignored, all AS solutions allow reducing inter-observer variability when physicians perform manual editing of the AS contours.

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. We are also grateful to Sophie King for her involvement in this work and Sylvie Chabaud for her advice in statistical analysis.

References

[1] C. L. Brouwer *et al.*, "3D Variation in delineation of head and neck organs at risk," *Radiation Oncology*, vol. 7, no. 1, pp. 1–10, Mar. 2012, doi: 10.1186/1748-717X-7-32/FIGURES/4.

[2] B. E. Nelms, W. A. Tomé, G. Robinson, and J. Wheeler, "Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer," *Int J Radiat Oncol Biol Phys*, 2012, doi: 10.1016/j.ijrobp.2010.10.019.

[3] C. L. Brouwer *et al.*, "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," *Radiotherapy and Oncology*, vol. 117, no. 1, pp. 83–90, Oct. 2015, doi: 10.1016/J.RADONC.2015.07.041.

[4] M. Awan *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*, vol. 3, no. 3, pp. 186–193, Jul. 2013, doi: 10.1016/J.PRRO.2012.11.002.

[5] C. J. Tao *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," *Radiotherapy and Oncology*, vol. 115, no. 3, pp. 407–411, 2015, doi: 10.1016/j.radonc.2015.05.012.

[6] G. Sharp *et al.*, "Vision 20/20: Perspectives on automated image segmentation for radiotherapy," *Med Phys*, vol. 41, no. 5, 2014, doi: 10.1118/1.4871620.

[7] T. Vrtovec, D. Močnik, P. Strojan, F. Pernuš, and B. Ibragimov, "Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods," *Med Phys*, vol. 47, no. 9, pp. e929–e950, Sep. 2020.

[8] H. X *et al.*, "Atlas-based auto-segmentation of head and neck CT images," *Med Image Comput Comput Assist Interv*, vol. 11, no. Pt 2, pp. 434–441, 2008, doi: 10.1007/978-3-540-85990-1_52.

[9] A. Larrue, D. Gujral, C. Nutting, and M. Gooding, "The impact of the number of atlases on the performance of automatic multi-atlas contouring," *Physica Medica*, 2015, doi: 10.1016/j.ejmp.2015.10.020.

[10] M. la Macchia *et al.*, "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer," *Radiation Oncology*, 2012, doi: 10.1186/1748-717X-7-160.

[11] S. Gresswell, P. Renz, D. Werts, and Y. Arshoun, "(P059) Impact of Increasing Atlas Size on Accuracy of an Atlas-Based Auto-Segmentation Program (ABAS) for Organs-at-Risk (OARS) in Head and Neck (H&N) Cancer Patients," *International Journal of Radiation Oncology*Biology*Physics*, 2017, doi: 10.1016/j.ijrobp.2017.02.155.

[12] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Trans Med Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[13] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray, "Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach," *Med Phys*, 2011, doi: 10.1118/1.3654160.

[14] V. Fortunati *et al.*, "Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling," *Med Phys*, 2013, doi: 10.1118/1.4810971.

[15] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp, "Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours," *Med Phys*, vol. 41, no. 5, 2014, doi: 10.1118/1.4871623.

[16] X. Han, "Learning-boosted label fusion for multi-atlas auto-segmentation," 2013. doi: 10.1007/978-3-319-02267-3_3.

[17] G. v. Walker *et al.*, "Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer," *Radiotherapy and Oncology*, 2014, doi: 10.1016/j.radonc.2014.08.028.

[18] D. N. Teguh *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*, vol. 81, no. 4, pp. 950–957, Nov. 2011, doi: 10.1016/j.ijrobp.2010.07.009.

[19] P. C. Levendag *et al.*, "Atlas Based Auto-segmentation of CT Images: Clinical Evaluation of using Auto-contouring in High-dose, High-precision Radiotherapy of Cancer in the Head and Neck," *International Journal of Radiation Oncology*Biology*Physics*, vol. 72, no. 1, p. S401, Sep. 2008, doi: 10.1016/j.ijrobp.2008.06.1285.

[20] M. Kosmin *et al.*, "Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer," *Radiotherapy and Oncology*, vol. 135. Elsevier Ireland Ltd, pp. 130–140, Jun. 01, 2019. doi: 10.1016/j.radonc.2019.03.004.

[21] M. Field, N. Hardcastle, M. Jameson, N. Aherne, and L. Holloway, "Machine learning applications in radiation oncology.," *Phys Imaging Radiat Oncol*, vol. 19, pp. 13–24, Jul. 2021.

[22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," Jan. 2020.

[23] L. v. van Dijk *et al.*, “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring,” *Radiotherapy and Oncology*, vol. 142, pp. 115–123, Jan. 2020, doi: 10.1016/j.radonc.2019.09.022.

[24] Y. Urago *et al.*, “Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models,” *Radiation Oncology*, vol. 16, no. 1, p. 175, 2021, doi: 10.1186/s13014-021-01896-1.

[25] X. Chen *et al.*, “A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy,” *Radiotherapy and Oncology*, vol. 160, pp. 175–184, 2021, doi: 10.1016/j.radonc.2021.04.019.

[26] Y. Fu, Y. Lei, T. Wang, W. J. Curran, T. Liu, and X. Yang, “A review of deep learning based methods for medical image multi-organ segmentation,” *Physica Medica*, vol. 85, no. November 2020, pp. 107–122, 2021, doi: 10.1016/j.ejmp.2021.05.003.

[27] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, “nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation,” *Nature Methods 2020 18:2*, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.

[28] W. Zhu *et al.*, “AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy,” *Med Phys*, vol. 46, no. 2, pp. 576–589, Feb. 2019, doi: 10.1002/MP.13300.

[29] H. Tang *et al.*, “Clinically applicable deep learning framework for organs at risk delineation in CT images,” *Nature Machine Intelligence 2019 1:10*, vol. 1, no. 10, pp. 480–491, Sep. 2019, doi: 10.1038/s42256-019-0099-z.

[30] C. Robert *et al.*, “Clinical implementation of deep-learning based auto-contouring tools–Experience of three French radiotherapy centers,” *Cancer/Radiotherapie*, vol. 25, no. 6–7, pp. 607–616, 2021, doi: 10.1016/j.canrad.2021.06.023.

[31] M. Ung *et al.*, “Improving Radiotherapy Workflow Through Implementation of Delineation Guidelines & AI-Based Annotation,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 108, no. 3, p. e315, Nov. 2020, doi: 10.1016/J.IJROBP.2020.07.753.

[32] J. Wong *et al.*, “Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning,” *Radiotherapy and Oncology*, vol. 144, pp. 152–158, 2020, doi: 10.1016/j.radonc.2019.10.019.

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

[34] W. van Rooij, M. Dahele, H. Ribeiro Brandao, A. R. Delaney, B. J. Slotman, and W. F. Verbakel, “Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation,” *Int J Radiat Oncol Biol Phys*, vol. 104, no. 3, pp. 677–684, Jul. 2019, doi: 10.1016/J.IJROBP.2019.02.040.

[35] T. Y. Lim, E. Gillespie, J. Murphy, and K. L. Moore, “Clinically Oriented Contour Evaluation Using Dosimetric Indices Generated From Automated Knowledge-Based Planning,” *International Journal*

*of Radiation Oncology*Biology*Physics*, vol. 103, no. 5, pp. 1251–1260, Apr. 2019, doi: 10.1016/j.ijrobp.2018.11.048.

[36] S. Y. Tsuji, A. Hwang, V. Weinberg, S. S. Yom, J. M. Quivey, and P. Xia, "Dosimetric Evaluation of Automatic Segmentation for Adaptive IMRT for Head-and-Neck Cancer," *Int J Radiat Oncol Biol Phys*, vol. 77, no. 3, pp. 707–714, 2010, doi: 10.1016/j.ijrobp.2009.06.012.

[37] B. E. Nelms *et al.*, "Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems," *Pract Radiat Oncol*, vol. 2, no. 4, pp. 296–305, Oct. 2012, doi: 10.1016/J.PRRO.2011.11.012.

[38] V. Batumalai, M. G. Jameson, D. F. Forstner, P. Vial, and L. C. Holloway, "How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case," *Pract Radiat Oncol*, vol. 3, no. 3, pp. e99–e106, Jul. 2013, doi: 10.1016/J.PRRO.2012.06.009.

[39] M. Clements, N. Schupp, M. Tattersall, A. Brown, and R. Larson, "Monaco treatment planning system tools and optimization processes," *Medical Dosimetry*, vol. 43, no. 2, pp. 106–117, Jun. 2018, doi: 10.1016/J.MEDDOS.2018.02.005.

[40] P. Coupé, J. v. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, "Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation," *Neuroimage*, vol. 54, no. 2, pp. 940–954, Jan. 2011, doi: 10.1016/J.NEUROIMAGE.2010.09.018.

[41] H. Lee *et al.*, "Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region," *Front Oncol*, vol. 9, no. APR, pp. 1–9, 2019, doi: 10.3389/fonc.2019.00239.

[42] A. Amjad *et al.*, "General and custom deep learning autosegmentation models for organs in head and neck, abdomen, and male pelvis," *Med Phys*, vol. 49, no. 3, pp. 1686–1700, 2022, doi: 10.1002/mp.15507.

[43] J. Yang *et al.*, "Autosegmentation for thoracic radiation treatment planning: A grand challenge at AAPM 2017," *Med Phys*, vol. 45, no. 10, pp. 4568–4581, Oct. 2018, doi: 10.1002/mp.13141.

[44] L. Maier-Hein *et al.*, "Metrics reloaded: Pitfalls and recommendations for image analysis validation," Jun. 2022, doi: 10.48550/arxiv.2206.01653.

[45] M.-C. Biston *et al.*, "Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer," *Physica Medica*, vol. 87, pp. 31–38, Jul. 2021, doi: 10.1016/j.ejmp.2021.05.037.

[46] V. Grégoire, S. Boisbouvier, P. Giraud, P. Maingon, Y. Pointreau, and L. Vieillevigne, "Management and work-up procedures of patients with head and neck malignancies treated by radiation," *Cancer/Radiotherapie*, vol. 26, no. 1–2, pp. 147–155, 2022, doi: 10.1016/j.canrad.2021.10.005.

[47] S. Willems *et al.*, "Clinical implementation of deepvoxnet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11041 LNCS, pp. 223–232, 2018, doi: 10.1007/978-3-030-01201-4_24.

[48] van der V. J *et al.*, "Benefits of deep learning for delineation of organs at risk in head and neck cancer," *Radiother Oncol*, vol. 138, pp. 68–74, Sep. 2019, doi: 10.1016/J.RADONC.2019.05.010.

[49] W. Chen *et al.*, "Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images," *Radiation Oncology*, vol. 15, no. 1, Jul. 2020, doi: 10.1186/s13014-020-01617-0.

[50] T. Nemoto *et al.*, "Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi," *J Radiat Res*, 2020, doi: 10.1093/jrr/rrz086.

[51] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks:," *Med Phys*, 2017, doi: 10.1002/mp.12045.

[52] Y. Fang, J. Wang, S. Chen, Y. Guo, Z. Zhang, and W. Hu, "The Impact Of Training Sample Size On Deep Learning Based Organ Auto Segmentation For Head Neck," *International Journal of Radiation Oncology*Biology*Physics*, 2020, doi: 10.1016/j.ijrobp.2020.07.228.

[53] A. Hänsch *et al.*, "PV-0530: Parotid gland segmentation with deep learning using clinical vs. curated training data," *Radiotherapy and Oncology*, 2018, doi: 10.1016/s0167-8140(18)30840-5.

Abstract

Background and purpose: To investigate the performance of head-and-neck (HN) organs-at-risk (OARs) contouring using three atlas-based (ABAS) solutions (STAPLE and Patch Fusions (PF) (ADMIREv3.41, Elekta AB; Sweden), majority voting fusion (MIM-Maestro, MIM Software; USA)), two Deep Learning (DL) solutions (ART-Plan Annotate, TheraPanacea; France) and ADMIRE-DL (ADMIREv3.41) and one hybrid ABAS solution (Random Forest (RF) (ADMIREv3.41)).

Material and Methods: All patients underwent iodine contrast-enhanced planning CT. Fourteen OARs were manually delineated. DL models were trained with 63 mono-centric patients and >1000 multi-centric patients, for ADMIRE-DL and ART-plan, respectively. Ten and 15 patients with varied anatomies were selected for the atlas library and for testing, respectively. The evaluation was based on geometric indices (DICE coefficient and 95th percentile-Hausdorff Distance), time needed for manual corrections and clinical dosimetric endpoints obtained using automated treatment planning.

Results: Overall, DL algorithms were more performing than ABAS methods for segmentation of HN OARs, especially when considering the computational time and the time spent on manual corrections. ADMIRE-DL had superior results compared to ART-plan with a DICE average of 0.85±0.32 vs 0.82±0.06. Regarding ABAS solutions, ADMIRE algorithms had better performances than MIM. While all the ADMIRE solutions had DICE results ≥0.8 for 7 OARs, generally better results were obtained with the RF algorithm compared to PF and STAPLE, which was the least performing. For ADMIRE-DL, ART-plan and RF solutions, a small dosimetric impact was observed when generating plans with corrected or non-corrected contours.

Conclusion: Automatic segmentation solutions can reach high accuracy results and allow reducing inter-observer variability.

Introduction

Manual contouring of organs-at-risk (OARs) is a time-consuming task that may suffer from large intra- and inter-observer variations, especially for head-and-neck (HN) cancer patients, because of the complex anatomy and the number of OARs at risk [1-4]. Contour variations may also result in important dosimetric differences[5]. Therefore, automatic segmentation (AS) methods are strongly sought after to increase contouring accuracy, improve the inter-observer variability, reduce delineation time, and facilitate treatment plan adaptation[6;7].

Among the different methods, atlas-based auto-segmentation (ABAS) uses one or more representative patients with carefully delineated OARs as a reference atlas for contouring new patients[8]. Those methods are widely spread due to their implementation convenience, but they do have several drawbacks: atlas selection strategy (single vs multi-atlas)[8]; performance plateau reached after 10-20 atlases[9]; poor performance for small and low contrast soft tissue structures[10]; increased computational time with each added atlas[11].

ABAS can be performed using a standard fusion between the atlases and the new patient [12], or by using a hybrid algorithm combining multi-ABAS with machine learning [13-17]. Despite a higher computational time, multi-ABAS studies have consistently demonstrated improved segmentation accuracy over the single atlas methods, with consequent reduction of the post-editing time [18;19]. By adding image intensity information, other studies have shown higher accuracy for model-based methods particularly on large organs such as brainstem, and spinal cord but poorer performances for tiny structures like cochlea [13-15].

Another method issued from artificial intelligence (AI) research and challenging ABAS is the use of deep learning (DL) techniques [6;7;20-25]. DL contouring typically implies the training of a convolutional neural network (CNN) directly from a set of annotated reference data. Although the training phase requires extensive GPU computing power and work in data gathering and curation, once trained, the segmentation is very fast. Most studies showed that DL methods outperformed ABAS methods [23-25]. Another few studies have compared different DL algorithms trained on the same data sets. The WBNetwas shown to provide superior results when compared to AnatomyNet and nnU-Net for 50 OARs on CT images [25-27]. However, neither of the DL methods have solved yet the problem of accurately segmenting small and low contrast structures.

Often, AS methods comparisons are based on geometric indices calculations only (DICE; Hausdorff distance (HD)) to compare the volume overlap between the reference and the automatically generated

contour[28]. However, it is highly recommended to perform additionally a dosimetric evaluation by generating treatment plans created with the AS contours [29-32]. This involves extensive time in generating treatment plans, and may also introduce inter or intra-planner variability [33;34].

In this context, the objective of the present study was to evaluate the performances 6 different algorithms for the auto-segmentation of 14 HN OARs:

- 3 commercial ABAS solutions (STAPLE[8] and Patch Fusion[35] (ADMIREv3.41, Elekta AB; Stockholm, Sweden), majority voting fusion[36] (MIM-Maestro, MIM Software; Cleveland, USA))
- 1 commercial DL solution (ART-Plan Annotate [37], TheraPanacea; Paris, France))
- 1 non-commercial hybrid ABAS solution (Random Forest (ADMIREv3.41))
- 1 non-commercial DL solution (ADMIRE-DL (ADMIREv3.41))

This study is original since the performances of ART-plan, ADMIRE-RF and ADMIRE-DL have never been investigated yet. Comparisons were first based on geometrical indices (DICE and HD), and on the time spent for correcting the contours. Finally, an *a priori* multi-criterial optimization (MCO) algorithm for automatic treatment planning was used, for the first time, to derive doses from AS contours with and without manual correction.

Materials and methods

Patient data

Seventy-eight non-operated HN cancer patients treated with radiation therapy between 2018 and 2021 and who underwent iodine contrast-enhanced planning computed tomography (CT), were selected for this study, which was approved by the hospital ethics committee. Fourteen OARs (i.e. parotids, submandibular glands, oral cavity, constrictor muscle, larynx, esophagus, trachea, thyroid, eyes, optical nerves, cochlea, brainstem, spinal cord, mandible) were manually delineated by a single expert physician (>30 years of experience), on 512x512 and 2mm-tick CT slices following HN delineation guidelines[3]. For the multi-ABAS approach, 10 patients from this database were carefully selected based on their body mass index (BMI) (from 18.9 to 30.7) to form a heterogeneous library of atlases. The same 10 atlases were used to create a library in MIM-Maestro and in ADMIRE (Fig.1). For ADMIRE-DL model training, 63 patients were used with the same set of OARs contours excluding optical nerves and cochlea. Conversely, ART-plan was trained with a large number of patients collected from multiple centers including ours. Fifteen patients having a BMI range from 12.1 to 34.7 were reserved for the testing phase (Fig.1).

Multi-ABAS and DL methodologies

Three ABAS solutions available in the research version of Monaco TPS (Monaco 5.59.11 with ADMIREv.3.41) and another one available in MIM-Maestro were investigated:

- STAPLE (Simultaneous Truth And Performance Level Estimation) consists in estimating the optimal combination of the segmentations by weighting each segmentation upon the estimated performance level [12].
- Patch Fusion (PF) algorithm computes the final probability of a voxel to belong to a structure as a weighted average of the atlases' contours based on voxel intensity information[38].
- Random Forest (RF) is a supervised learning algorithm which constructs a voxel classifier for each structure using the registered atlases as training data[16].
- MIM uses majority voting as fusion algorithm[36].

For the ADMIRE software, out of the 10 atlases used, a reference patient was selected for each test patient based on the closest BMI of the atlas and the underlying patient. No individual atlas selection was required for MIM, but a general *template scan* (patient having an anatomy close to the mean BMI of the atlas cohort) was registered with all the atlases in the library.

Two DL models were investigated:

- ADMIRE-DL model (N=63 patients from one center) is a fully connected deep convolutional neural network (DCNN) with 3D U-net architecture and added residue connections from ResNet [39;40]. It allows to obtain the segmentation of an entire whole 3D image in a single pass instead of classifying the image patch each time.
- ART-plan automated annotations (ART-plan, N>1000 patients per structure from multiple centers) rely on a local organ-specific three levels 3D-CNN (ART-Net) that uses local linear mapping for embedding the target volume and the reference anatomies with annotations. The full annotation is achieved by combining multiple networks with statistically preserving coherent principles.

Geometric evaluation of auto-segmentation solutions

To quantitatively evaluate the segmentation results, we used two geometric indices: volumetric DICE and 95thpercentile-HD ($HD_{95\%}$) together with dosimetric endpoints ($V_{95\%}$, $D_{mean,}$ $D_{5\%}$ and $D_{2\%}$) [28;30]. DICE coefficient quantifies the overlap between two volumes leading to a value between 0 (no overlap) and 1 (perfect overlap). The HD measures the distance from each point of a surface to the nearest point on the other surface. We used 95th percentile of the distances between boundary points to eliminate the possible

outliers. For the elongated organs (i.e., esophagus, trachea, constrictor muscle and spinal cord) the results were calculated only on the slices where both contours were present.

Time needed for manual corrections

Three of the automatic solutions (RF, ADMIRE-DL, ART-plan) were clinically reviewed and corrected by a dosimetrist and validated by a skilled physician on MONACO contouring workstation. The time spent on correcting and validating each structure was recorded independently.

Dosimetric evaluation – automatic treatment plans

For each patient, and for RF, ADMIRE-DL and ART-plan solutions, 2 different plans were generated: one using AS contours and another one using AS+manually corrected contours. The differences in dose distributions were then evaluated on the corrected contours. In total, 90 VMAT treatment plans were computed with mCycle auto-planning solution (Monaco 5.59.11, Elekta AB, Stockholm). The software uses a lexicographic MCO which has been extensively described before[41]. All plans were performed using two 360° arcs. A simultaneous integrated boost technique was used for delivering 70Gy to the PTV associated to the primary tumor or node therapeutic CTV and 54.25Gy to the PTV associated to prophylactic nodal CTV, in 35 fractions of 2Gy.

Statistical analysis

Statistical differences between the methods were assessed using Wilcoxon signed rank test. P-values under <0.05 were considered significant.

Results

Computational time per patient was in average 10.3±1.6min, 10.5±0.6min and 12.1±0.6min for STAPLE, PF and RF, respectively. For MIM, it was under 1min while the atlas registration took approximately 6min for a library of 10 atlases. ADMIRE-DL and ART-plan provided a solution in less than 1min and 2min, respectively. DICE scores and $HD_{95\%}$ distance results of all solutions are summarized in Fig.2.

Overall, both DICE and $HD_{95\%}$ results indicated that DL algorithms had better performances than ABAS algorithms for AS segmentation of HN OARs. With 11 common OARs, ADMIRE-DL had overall superior results compared to ART-plan with a DICE average of 0.85±0.32 vs 0.82±0.06 and 11 vs 9 OARs having DICEs ≥0.8. ART-plan had significantly better DICE results only for the thyroid ($p=0.017$). Meanwhile, ADMIRE-DL provided significantly better DICEs for trachea ($p=0.004$) and larynx ($p=0.001$), and significantly better $HD_{95\%}$ for mandible ($p=0.04$) and larynx ($p=0.001$).

Regarding ABAS solutions, ADMIRE algorithms had, overall, better DICEs and $HD_{95\%}$ than MIM, which had the lowest DICE results for 7 out of 14 OARs. While all the ADMIRE solutions had DICE results ≥0.8 for 7 OARs, generally better results were obtained with the RF algorithm compared to PF and STAPLE, which was the least performing. Compared to PF, the RF algorithm had significantly superior DICE for the eyes ($p<0.001$), mandible ($p<0.001$) and thyroid ($p=0.002$) and significantly better $HD_{95\%}$ for the thyroid ($p=0.002$). Note that RF did not segment optic nerves and cochlea owing the algorithm's limitation for such small structures.

Compared to the ADMIRE-DL, the RF algorithm provided significantly better DICE and $HD_{95\%}$ results for oral cavity ($p\leq0.007$) and eyes ($p\leq0.01$), and significantly better DICE for mandible ($p=0.004$) and constrictor muscle ($p=0.007$). Conversely, ADMIRE-DL had significantly better DICE and $HD_{95\%}$ results for thyroid ($p\leq0.02$) and esophagus ($p\leq0.004$), and significantly better DICE for parotids ($p=0.004$).

An example of segmentation from PF, ART-plan and ADMIRE-DL compared to the physicians' manual delineations is provided in Fig.3.

Ten of the OARs obtained with RF, ADMIRE-DL and ART-plan were thereafter carefully corrected by a dosimetrist and checked by a physician. The manual correction time per patient was in average 36min34sec, 26min57sec and 17min54sec for RF, ART-plan, and ADMIRE-DL, respectively. The contours generated by ADMIRE-DL model, were the fastest to correct. In general, manual corrections of eyes, spinal cord and brainstem were <2min for the 3 solutions while for oral cavity and esophagus correction times were >3min depending on the AS algorithm used (Fig.4).

After manual corrections, the DICE scores of all OARs increased, except for the oral cavity on all 3 solutions, and for the spinal cord with ART-plan (Table 1), thus highlighting inter-observer variability in oral cavity contouring between the expert physician providing the reference contours, and other physician performing manual corrections. At the same time, the $HD_{95\%}$ did not decreased consistently for all the structures after the manual corrections.

The differences in doses on corrected OARs, between treatment plans generated using the AS contours, with or without manual corrections are presented in Table 2. Target coverage differences were not significant, except for the $V_{95\%}$ for PTV70Gy, with ADMIRE-DL contours ($p=0.03$). However, coverage differences were ≤1.4%. The OARs doses were significantly different for oral cavity ($p=0.03$), right submandibular gland ($p=0.04$) and esophagus ($p=0.04$) in the plans created with ADMIRE-DL contours, but with average dose differences ≤0.65Gy. With ART-plan generated contours a significant dose difference

was found for the left parotid (p=0.01). Conversely, no significant dose differences were identified in the plans generated with the RF contours, but the largest dose differences ($D_{2\%}$) to brainstem and spinal cord.

Discussion

We showed in this study that, overall, both DICE and $HD_{95\%}$ results indicated that DL algorithms had better performances than ABAS algorithms for automatic segmentation of HN OARs. Concerning the 2 DL solutions, out of 12 contours, ADMIRE-DL outperformed ART-plan in terms of DICE for 7 OARs, with significant difference for larynx and trachea. $HD_{95\%}$ were also significantly smaller for mandible and larynx for ADMIRE-DL compared to ART-plan. Contrarily to ART-plan, ADMIRE-DL was not tailored to automatically contour optic nerves and cochlea. Nevertheless, the correction of the AS contour of small organs generally takes more time than starting from scratch [42;43]. Conversely, ART-plan was not trained to contour the constrictor muscle. However, ADMIRE-DL results were poor, showing the difficulty to get satisfying results for such organs with high anatomical variations and limited image contrast. Consistent with the literature, poor performances were observed on small organs for both ABAS and DL solutions [23-26;44;45].

Before this study, ADMIRE-DL and ART-plan algorithms had not been explored on HN site. The ADMIRE-DL algorithm was trained exclusively with manual delineations by one expert physician, providing uniformity of the training data. Ideally, there should be a consortium for the contour delineation between physicians working in a radiotherapy department, which should rely on internationally published guidelines [3]. In this study, with a limited training dataset (N=63), we showed that a model can achieve consistent results for most of the structures in HN. Hence, with a minimum of work, centers can adapt a model to their standard delineation's practices. Similarly, high accuracy segmentation results were obtained with the DeepVoxNet and another CNN with networks trained on N=70 and N=50 samples, respectively [42;46]. Other studies demonstrated that organs pattern depend on the training sample size and yet similar results can be obtained when training on a small set of carefully curated data compared to a larger set of more easily available routine-level clinical annotations [47;48]. On the opposite, ART-plan solution was trained with more than 1000 training samples per organ collected from multiple centers and can segment 50 OARs and target volumes in HN. Despite this, high performance segmentations were obtained in this study. Proving that a multi-center study approach includes combination of manual contours from different physicians (easier to obtain), ART-plan results presented good conformity to new datasets and comparable performance to a model train with data from a single center.

We also showed that, using a carefully selected atlas of patients, STAPLE, PF and RF methods provide very good agreement with manual contours (DICE≥0.8) [8;18] and, for some organs, comparable or better results than DL models (i.e. oral cavity, mandible, eyes). Conversely, MIM solution had overall inferior performance. Among these algorithms, RF, which had not been explored before, was the most performing and had significantly better results than the two DL models for oral cavity, mandible, and eyes. Therefore, with only 10 carefully selected atlases, RF algorithm may serve as an AS solution easy to implement clinically. Note that using an enlarged library of 20 patients (data not shown) did not considerably improved the performances of RF but drastically increased the computation time, demonstrating that the performance plateau phenomena still exists with this new ABAS method.

Many studies have reported the performances of different algorithms for HN OARs segmentation on CT images (Table 1 of supplementary data). All studies underline limited performance on small organs, and the importance of both manual contours' quality and training data size to get results closer to clinical acceptability. Moreover, AS has shown to reduce inter-observer variability when observers performed manual editing on the automatically generated contours, which improves the consistency of manual delineation[5].

According to the recently published guidelines, together with geometric accuracy, studies should ideally report benefit in time saving and clinical acceptability in terms of patient dose evaluation, for assessing the benefit of an automatic segmentation method[28]. Both tasks are challenging and often put aside first because of the time requested to be completed, and because of the intra-planner variability, which could introduce a bias in the observations. In this study, both tasks were completed for the three best algorithms, and an auto-planning solution was used to perform treatment plans based on AS contours with or without corrections. This was a strength of this study, and an efficient way to isolate the consequences of contour variations on the radiotherapy doses and reveal more precisely which contours require greater attention [29;30]. Among other methods, some authors proposed to superpose the original clinical plan onto the automatically delineated contours [5;23], to use automated planning strategies such as knowledge-based planning (KBP) [29;30] or to conserve the original beam configuration parameters [31]. To our knowledge, this is the first time that an *a priori* MCO auto-planning solution is used for contour evaluation.

We observed in our study that, for most structures, the correction time for ADMIRE-DL and ART-plan solutions was <1min (e.g., eyes, brainstem, submandibular glands) and <2min (e.g., mandible, parotid glands) demonstrating significant time saving versus starting from scratch. Correcting ADMIRE-DL contours was 18min and 9min faster compared with RF and ART-plan contours, respectively. Generally, the oral

cavity and esophagus took more time to be corrected. For the oral cavity, this may be correlated with the inter-observer variability since the DICE results were consistently smaller for all 3 solutions after the manual corrections. We finally observed that manual corrections of the AS contours had small impact on the dose distributions, which indicates no clinically meaningful differences between the three algorithms. The ΔD_{mean} for the investigated structures were <0.9Gy. Generally, the range of the $\Delta D_{2\%}$ were the highest for the spinal cord and for the brainstem for all the solutions, which may be an important factor in physician's decision when validating the treatment plan. However, all dose-volume constraints were respected in all the plans. Similar to other studies, for most organs, the clinical impact was not significant and thus, manual correction could be omitted [29;31].

Conclusions

DL methods were generally more performing than ABAS methods for AS segmentation of HN OARs. We showed that a DL model can reach high performances with a limited training dataset, provided that data comes from a single hospital, and if possible, only one expert physician is involved. Most ABAS methods showed consistent results but were less performing than DL algorithms especially when considering the computing time and the time spent on manual corrections. Finally, even if manual corrections are often necessary, all AS solutions allow reducing inter-observer variability when physicians perform manual editing on the automatically generated contours.

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. We are also grateful to Sophie King for her involvement in this work.

References

[1] C. L. Brouwer *et al.*, "3D Variation in delineation of head and neck organs at risk," *Radiation Oncology*, vol. 7, no. 1, pp. 1–10, Mar. 2012, doi: 10.1186/1748-717X-7-32/FIGURES/4.

[2] B. E. Nelms, W. A. Tomé, G. Robinson, and J. Wheeler, "Variations in the contouring of organs at risk: Test case from a patient with oropharyngeal cancer," *International Journal of Radiation Oncology Biology Physics*, 2012, doi: 10.1016/j.ijrobp.2010.10.019.

[3] C. L. Brouwer *et al.*, "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," *Radiotherapy and Oncology*, vol. 117, no. 1, pp. 83–90, Oct. 2015, doi: 10.1016/J.RADONC.2015.07.041.

[4] M. Awan *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," *Practical Radiation Oncology*, vol. 3, no. 3, pp. 186–193, Jul. 2013, doi: 10.1016/J.PRRO.2012.11.002.

[5] C. J. Tao *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," *Radiotherapy and Oncology*, vol. 115, no. 3, pp. 407–411, 2015, doi: 10.1016/j.radonc.2015.05.012.

[6] G. Sharp *et al.*, "Vision 20/20: Perspectives on automated image segmentation for radiotherapy," *Medical Physics*, vol. 41, no. 5, 2014, doi: 10.1118/1.4871620.

[7] T. Vrtovec, D. Močnik, P. Strojan, F. Pernuš, and B. Ibragimov, "Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods," *Medical Physics*, vol. 47, no. 9, pp. e929–e950, Sep. 2020, doi: 10.1002/MP.14320.

[8] H. X *et al.*, "Atlas-based auto-segmentation of head and neck CT images," *Med Image Comput Comput Assist Interv*, vol. 11, no. Pt 2, pp. 434–441, 2008, doi: 10.1007/978-3-540-85990-1_52.

[9] A. Larrue, D. Gujral, C. Nutting, and M. Gooding, "The impact of the number of atlases on the performance of automatic multi-atlas contouring," *Physica Medica*, 2015, doi: 10.1016/j.ejmp.2015.10.020.

[10] M. La Macchia *et al.*, "Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer," *Radiation Oncology*, 2012, doi: 10.1186/1748-717X-7-160.

[11] S. Gresswell, P. Renz, D. Werts, and Y. Arshoun, "(P059) Impact of Increasing Atlas Size on Accuracy of an Atlas-Based Auto-Segmentation Program (ABAS) for Organs-at-Risk (OARS) in Head and Neck (H&N) Cancer Patients," *International Journal of Radiation Oncology*Biology*Physics*, 2017, doi: 10.1016/j.ijrobp.2017.02.155.

[12] S. K. Warfield, K. H. Zou, and W. M. Wells, "Simultaneous truth and performance level estimation (STAPLE): An algorithm for the validation of image segmentation," *IEEE Transactions on Medical Imaging*, vol. 23, no. 7, pp. 903–921, Jul. 2004, doi: 10.1109/TMI.2004.828354.

[13] A. A. Qazi, V. Pekar, J. Kim, J. Xie, S. L. Breen, and D. A. Jaffray, “Auto-segmentation of normal and target structures in head and neck CT images: A feature-driven model-based approach,” *Medical Physics*, 2011, doi: 10.1118/1.3654160.

[14] V. Fortunati *et al.*, “Tissue segmentation of head and neck CT images for treatment planning: A multiatlas approach combined with intensity modeling,” *Medical Physics*, 2013, doi: 10.1118/1.4810971.

[15] K. D. Fritscher, M. Peroni, P. Zaffino, M. F. Spadea, R. Schubert, and G. Sharp, “Automatic segmentation of head and neck CT images for radiotherapy treatment planning using multiple atlases, statistical appearance models, and geodesic active contours,” *Medical Physics*, vol. 41, no. 5, 2014, doi: 10.1118/1.4871623.

[16] X. Han, “Learning-boosted label fusion for multi-atlas auto-segmentation,” 2013. doi: 10.1007/978-3-319-02267-3_3.

[17] G. v. Walker *et al.*, “Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer,” *Radiotherapy and Oncology*, 2014, doi: 10.1016/j.radonc.2014.08.028.

[18] D. N. Teguh *et al.*, “Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck,” *International Journal of Radiation Oncology Biology Physics*, vol. 81, no. 4, pp. 950–957, Nov. 2011, doi: 10.1016/j.ijrobp.2010.07.009.

[19] P. C. Levendag *et al.*, “Atlas Based Auto-segmentation of CT Images: Clinical Evaluation of using Auto-contouring in High-dose, High-precision Radiotherapy of Cancer in the Head and Neck,” *International Journal of Radiation Oncology*Biology*Physics*, vol. 72, no. 1, p. S401, Sep. 2008, doi: 10.1016/j.ijrobp.2008.06.1285.

[20] M. Kosmin *et al.*, “Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer,” *Radiotherapy and Oncology*, vol. 135. Elsevier Ireland Ltd, pp. 130–140, Jun. 01, 2019. doi: 10.1016/j.radonc.2019.03.004.

[21] M. Field, N. Hardcastle, M. Jameson, N. Aherne, and L. Holloway, “Machine learning applications in radiation oncology.,” *Phys Imaging Radiat Oncol*, vol. 19, pp. 13–24, Jul. 2021.

[22] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, “Image Segmentation Using Deep Learning: A Survey,” Jan. 2020.

[23] L. v. van Dijk *et al.*, “Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring,” *Radiotherapy and Oncology*, vol. 142, pp. 115–123, Jan. 2020.

[24] Y. Urago *et al.*, “Evaluation of auto-segmentation accuracy of cloud-based artificial intelligence and atlas-based models,” *Radiation Oncology*, vol. 16, no. 1, p. 175, 2021, doi: 10.1186/s13014-021-01896-1.

[25] X. Chen *et al.*, “A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy,” *Radiotherapy and Oncology*, vol. 160, pp. 175–184, 2021, doi: 10.1016/j.radonc.2021.04.019.

[26] W. Zhu *et al.*, "AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy," *Medical Physics*, vol. 46, no. 2, pp. 576–589, Feb. 2019, doi: 10.1002/MP.13300.

[27] F. Isensee, P. F. Jaeger, S. A. A. Kohl, J. Petersen, and K. H. Maier-Hein, "nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation," *Nature Methods 2020 18:2*, vol. 18, no. 2, pp. 203–211, Dec. 2020, doi: 10.1038/s41592-020-01008-z.

[28] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Medical Imaging*, 2015, doi: 10.1186/s12880-015-0068-x.

[29] W. van Rooij, M. Dahele, H. Ribeiro Brandao, A. R. Delaney, B. J. Slotman, and W. F. Verbakel, "Deep Learning-Based Delineation of Head and Neck Organs at Risk: Geometric and Dosimetric Evaluation," *Int J Radiat Oncol Biol Phys*, vol. 104, no. 3, pp. 677–684, Jul. 2019, doi: 10.1016/J.IJROBP.2019.02.040.

[30] T. Y. Lim, E. Gillespie, J. Murphy, and K. L. Moore, "Clinically Oriented Contour Evaluation Using Dosimetric Indices Generated From Automated Knowledge-Based Planning," *International Journal of Radiation Oncology*Biology*Physics*, vol. 103, no. 5, pp. 1251–1260, Apr. 2019, doi: 10.1016/j.ijrobp.2018.11.048.

[31] S. Y. Tsuji, A. Hwang, V. Weinberg, S. S. Yom, J. M. Quivey, and P. Xia, "Dosimetric Evaluation of Automatic Segmentation for Adaptive IMRT for Head-and-Neck Cancer," *International Journal of Radiation Oncology Biology Physics*, vol. 77, no. 3, pp. 707–714, 2010, doi: 10.1016/j.ijrobp.2009.06.012.

[32] T. Vrtovec, D. Močnik, P. Strojan, F. Pernuš, and B. Ibragimov, "Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods," *Medical Physics*, vol. 47, no. 9. John Wiley and Sons Ltd, pp. e929–e950, Sep. 01, 2020. doi: 10.1002/mp.14320.

[33] B. E. Nelms *et al.*, "Variation in external beam treatment plan quality: An inter-institutional study of planners and planning systems," *Practical Radiation Oncology*, vol. 2, no. 4, pp. 296–305, Oct. 2012, doi: 10.1016/J.PRRO.2011.11.012.

[34] V. Batumalai, M. G. Jameson, D. F. Forstner, P. Vial, and L. C. Holloway, "How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case," *Practical Radiation Oncology*, vol. 3, no. 3, pp. e99–e106, Jul. 2013, doi: 10.1016/J.PRRO.2012.06.009.

[35] B. A. McDonald *et al.*, "Investigation of Autosegmentation Techniques on T2-Weighted MRI for Off-line Dose Reconstruction in MR-Linac Adapt to Position Workflow for Head and Neck Cancers," *medRxiv*, p. 2021.09.30.21264327, Oct. 2021, doi: 10.1101/2021.09.30.21264327.

[36] H. Lee *et al.*, "Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region," *Frontiers in Oncology*, vol. 9, no. APR, pp. 1–9, 2019, doi: 10.3389/fonc.2019.00239.

[37] "ART-plan Annotate™, TheraPanacea, Paris; France." https://www.therapanacea.eu/our-products/annotate/

[38] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins, “Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation,” *Neuroimage*, vol. 54, no. 2, pp. 940–954, Jan. 2011, doi: 10.1016/J.NEUROIMAGE.2010.09.018.

[39] E. Shelhamer, J. Long, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017, doi: 10.1109/TPAMI.2016.2572683.

[40] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition.” [Online]. Available: http://image-net.org/challenges/LSVRC/2015/

[41] M. C. Biston *et al.*, “Evaluation of fully automated a priori MCO treatment planning in VMAT for head-and-neck cancer,” *Physica Medica*, vol. 87, pp. 31–38, Jul. 2021, doi: 10.1016/J.EJMP.2021.05.037.

[42] S. Willems *et al.*, “Clinical implementation of deepvoxnet for auto-delineation of organs at risk in head and neck cancer patients in radiotherapy,” *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11041 LNCS, pp. 223–232, 2018, doi: 10.1007/978-3-030-01201-4_24.

[43] van der V. J *et al.*, “Benefits of deep learning for delineation of organs at risk in head and neck cancer,” *Radiother Oncol*, vol. 138, pp. 68–74, Sep. 2019, doi: 10.1016/J.RADONC.2019.05.010.

[44] W. Chen *et al.*, “Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images,” *Radiation Oncology*, vol. 15, no. 1, Jul. 2020, doi: 10.1186/s13014-020-01617-0.

[45] T. Nemoto *et al.*, “Efficacy evaluation of 2D, 3D U-Net semantic segmentation and atlas-based segmentation of normal lungs excluding the trachea and main bronchi,” *Journal of Radiation Research*, 2020, doi: 10.1093/jrr/rrz086.

[46] B. Ibragimov and L. Xing, “Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks:,” *Medical Physics*, 2017, doi: 10.1002/mp.12045.

[47] Y. Fang, J. Wang, S. Chen, Y. Guo, Z. Zhang, and W. Hu, “The Impact Of Training Sample Size On Deep Learning Based Organ Auto Segmentation For Head Neck,” *International Journal of Radiation Oncology*Biology*Physics*, 2020, doi: 10.1016/j.ijrobp.2020.07.228.

[48] A. Hänsch *et al.*, “PV-0530: Parotid gland segmentation with deep learning using clinical vs. curated training data,” *Radiotherapy and Oncology*, 2018, doi: 10.1016/s0167-8140(18)30840-5.

Table. 1
Characteristics of the test cohort used for evaluation of the AS solutions

	Tumor localization	TNM	BMI	Gender	Age
Patient 1	Oral cavity	T4aN0M0	24.1	M	88Y
Patient 2	Hypopharynx	T3N2aM0	12.1	M	57Y
Patient 3	Nasal cavity	T2N0M0	31	F	83Y
Patient 4	Tonsils	T2N0M0	20.8	M	75Y
Patient 5	Hypopharynx	T0N3M0	24	M	45Y
Patient 6	Rhinopharynx	T3N1M0	19	F	58Y
Patient 7	Rhinopharynx	T3N0M0	19.8	F	69Y
Patient 8	Rhinopharynx	T2N2M0	21.4	F	71Y
Patient 9	Hypopharynx	T1N1M0	23.4	M	75Y
Patient 10	Larynx	T2N0M0	30.4	M	60Y
Patient 11	Tonsils	T2N0M0	24	F	69Y
Patient 12	Unilateral ganglion	T4N3M1	34.7	M	54Y
Patient 13	Parapharynx	T2N1M0	25.5	M	65Y
Patient 14	Subglotic	T4aN0M0	21.5	M	58Y
Patient 15	Hypopharynx	T4bN0M0	19	M	75Y

Table 2

Geometric evaluation after manual corrections of 10 OAR for the three best solutions; with * are marked the differences that are statistically significant ($p<0.05$). Abbreviations: Sub.glands = submandibular glands;

	ABAS.3		DL.1		DL.2	
DICE	without corrections	after manual corrections	without corrections	after manual corrections	without corrections	after manual corrections
Parotids	0.8 ± 0.05	0.84 ± 0.04	0.82 ± 0.04	0.84 ± 0.04	0.81 ± 0.06	0.85 ± 0.03
Oral cavity	0.87 ± 0.04	0.81 ± 0.06	0.85 ± 0.06	0.79 ± 0.08	0.84 ± 0.07	0.79 ± 0.08
Sub.glands	0.77 ± 0.13	0.83 ± 0.1	0.8 ± 0.07	0.84 ± 0.07	0.79 ± 0.13	0.82 ± 0.14
Mandible	0.92 ± 0.02	0.93 ± 0.02	0.9 ± 0.02	0.9 ± 0.02	0.89 ± 0.03	0.9 ± 0.03
Esophagus	**0.72 ± 0.1**	**0.86 ± 0.04***	0.83 ± 0.04	0.86 ± 0.03	0.84 ± 0.05	0.87 ± 0.03
Trachea	0.88 ± 0.05	0.91 ± 0.04	0.9 ± 0.02	0.9 ± 0.06	**0.87 ± 0.03**	**0.91 ± 0.04***
Thyroid	**0.74 ± 0.11**	**0.85 ± 0.03 ***	0.83 ± 0.04	0.85 ± 0.04	0.85 ± 0.04	0.86 ± 0.03
Eyes	0.91 ± 0.03	0.91 ± 0.03	0.89 ± 0.03	0.9 ± 0.03	0.87 ± 0.04	0.9 ± 0.03
Spinal cord	0.84 ± 0.05	0.84 ± 0.05	0.84 ± 0.04	0.85 ± 0.04	0.85 ± 0.03	0.84 ± 0.04
Brainstem	0.85 ± 0.04	0.86 ± 0.05	0.85 ± 0.03	0.86 ± 0.04	0.85 ± 0.06	0.86 ± 0.06
$HD_{95\%}$ (mm)	without corrections	after manual corrections	without corrections	after manual corrections	without corrections	after manual corrections
Parotids	7.2 ± 3.4	7.9 ± 7.2	6.2 ± 2.6	8.4 ± 8.1	7.8 ± 5.2	7.0 ± 7.1
Oral cavity	6.5 ± 2.6	11.2 ± 3.9	8.1 ± 3.9	11.4 ± 5.1	9.4 ± 5.1	12 ± 5.3
Sub.glands	4.5 ± 3.1	4.0 ± 2.7	3.7 ± 0.9	2.9 ± 1.3	4.7 ± 2.2	3.8 ± 2.1
Mandible	2.2 ± 1.1	1.5 ± 1	2.3 ± 0.8	2.1 ± 0.9	3.7 ± 2	2.1 ± 1.4
Esophagus	**6.1 ± 2.3**	**2.3 ± 0.9***	3 ± 1.3	2.7 ± 0.7	3.6 ± 2.6	1.9 ± 0.4
Trachea	3.2 ± 1.6	1.9 ± 0.5	2.3 ± 0.7	1.8 ± 0.7	**2.4 ± 0.5**	**1.8 ± 0.6***
Thyroid	**8 ± 8.3**	**2.5 ± 1.2 ***	4.5 ± 4.7	2.2 ± 0.6	3.0 ± 1.4	2.5 ± 1.3
Eyes	2. ± 0.5	1.9 ± 0.5	2.4 ± 0.8	2 ± 0.4	2.4 ± 0.8	2.2 ± 0.7
Spinal cord	2.1 ± 0.5	2 ± 0.5	1.8 ± 0.6	1.8 ± 0.5	1.7 ± 0.4	2 ± 0.5
Brainstem	3.9 ± 1.6	3.9 ± 1.9	4.4 ± 1.6	4.1 ± 1.7	4.1 ± 1.7	3.9 ± 1.8

Table 3

Dosimetric differences between doses generated with manually corrected contours and automatic contours, analyzed on the corrected contours; impact on the target volumes is evaluated in $V_{95\%}$ dose coverage, for the spinal cord and brainstem in $D_{2\%}$, and for the mandible in $D_{5\%}$ while for the rest of the OAR mean doses are calculated. Abbreviations: sub.glands = submandibular glands; R = right, L = left;

	ABAS.3			DL.1			DL.2		
Structure	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]	$V_{95\%}$ [%]	$\Delta V_{95\%}$ [%]	$\Delta V_{95\%}$ ranges [%]
PTV_70Gy	98.3	-0.1	[-0.9, 0.7]	98.4	0.2	[-1.4, 1.3]	98.3	-0.04	[-0.8, 2.1]
PTV_54.25Gy	99.2	-0.03	[-0.2, 0.1]	99.2	0.01	[-0.2, 0.3]	99.2	0.05	[-0.1, 0.2]
	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]	D_{mean} [Gy]	ΔD_{mean} [Gy]	ΔD_{mean} ranges [Gy]
Parotid R	16.2	-0.1	[-3.4, 3.3]	16.8	0.3	[-1.8, 3.4]	17.6	0.3	[-2.2, 3.6]
Parotid L	12.9	-0.04	[-4.3, 3.3]	15.1	0.4	[-2.4, 4]	15	1.1	[-0.8, 6]
Oral cavity	21	-0.02	[-2.7, 3.2]	21.1	-0.6	[-2.2, 0.9]	21	-0.7	[-3.8, 1.3]
Sub.gland R	32.5	-0.1	[-0.9, 1.7]	38.7	-0.5	[-2.3, 1.2]	39.1	-0.4	[-7.4, 3.1]
Sub.gland L	34.1	-0.4	[-2.8, 1.2]	40.6	-0.3	[-1.8, 1.3]	40.9	0.1	[-1.6, 1.1]
Esophagus	7.9	0.2	[-0.5, 1.7]	9.2	0.5	[-1.1, 2.4]	10.2	0.9	[-0.3, 5.5]
Trachea	10.1	-0.2	[-0.9, 0.2]	11.9	-0.1	[-0.9, 0.4]	13.5	0.2	[-1.1, 5.2]
Thyroid	29.4	-0.2	[-4.8, 1.1]	35.5	-0.5	[-3.7, 1.9]	35.4	-0.3	[-2, 1.3]
Eye R	5.2	0.5	[-1.5, 4.2]	4.9	0.4	[-4.1, 10.3]	5.1	-0.4	[-6.1, 0.8]
Eye L	4.5	-0.4	[-8.9, 2.5]	5.4	0.8	[-0.4, 8.8]	4.4	-0.8	[-8.9, 1.2]
	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]	$D_{2\%}$ [Gy]	$\Delta D_{2\%}$ [Gy]	$\Delta D_{2\%}$ ranges [Gy]
Spinal cord	25	2.5	[-4.6, 24.4]	23.7	-0.5	[-16.3, 10.1]	24.2	-2.1	[-13.6,16.2]
Brainstem	16.9	-1.5	[-20.2,13.1]	18	-1.0	[-7.1,4.9]	18.9	-0.6	[-9.7,6.9]
	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]	$D_{5\%}$ [Gy]	$\Delta D_{5\%}$ [Gy]	$\Delta D_{5\%}$ ranges [Gy]
Mandible	43.3	0.1	[-3.6, 2.1]	43.4	-0.2	[-2.9, 6.3]	44.4	0.35	[-2.5, 4.9]

Figure 1

Click here to access/download;Figure;Fig1.Overview of the study design.pdf

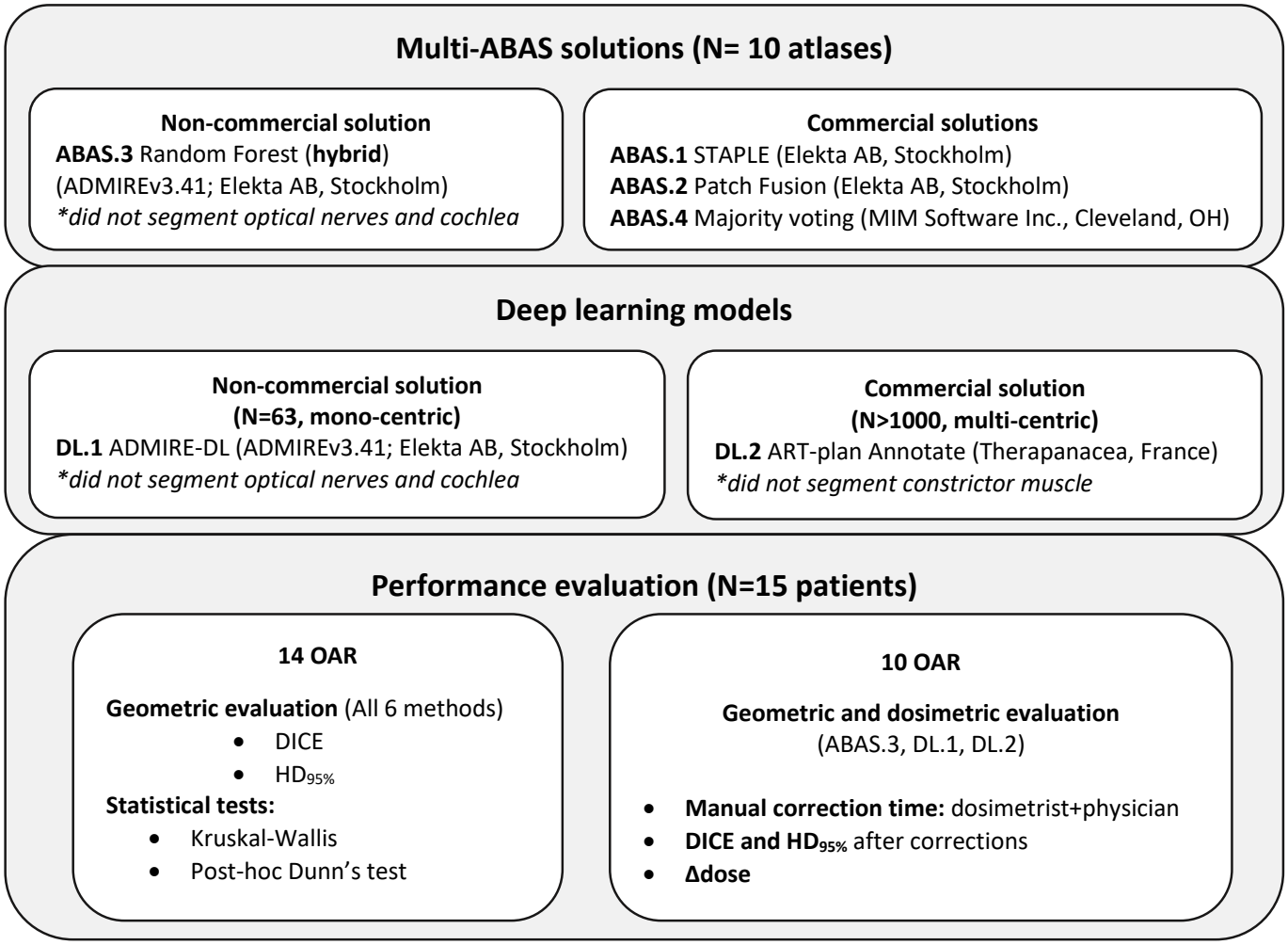

Figure 2

Click here to access/download;Figure;Fig2.Geometric evaluation.pdf

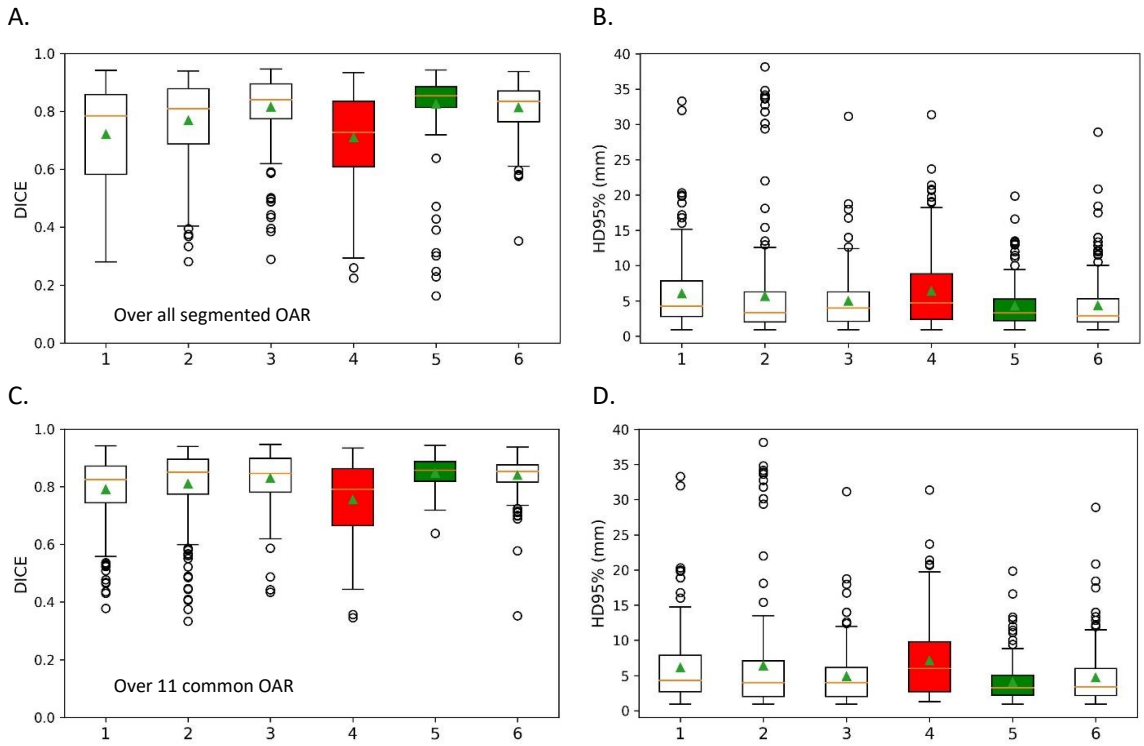

Figure 3

Click here to access/download;Figure;Fig3.Geometric evaluation per OAR.pdf

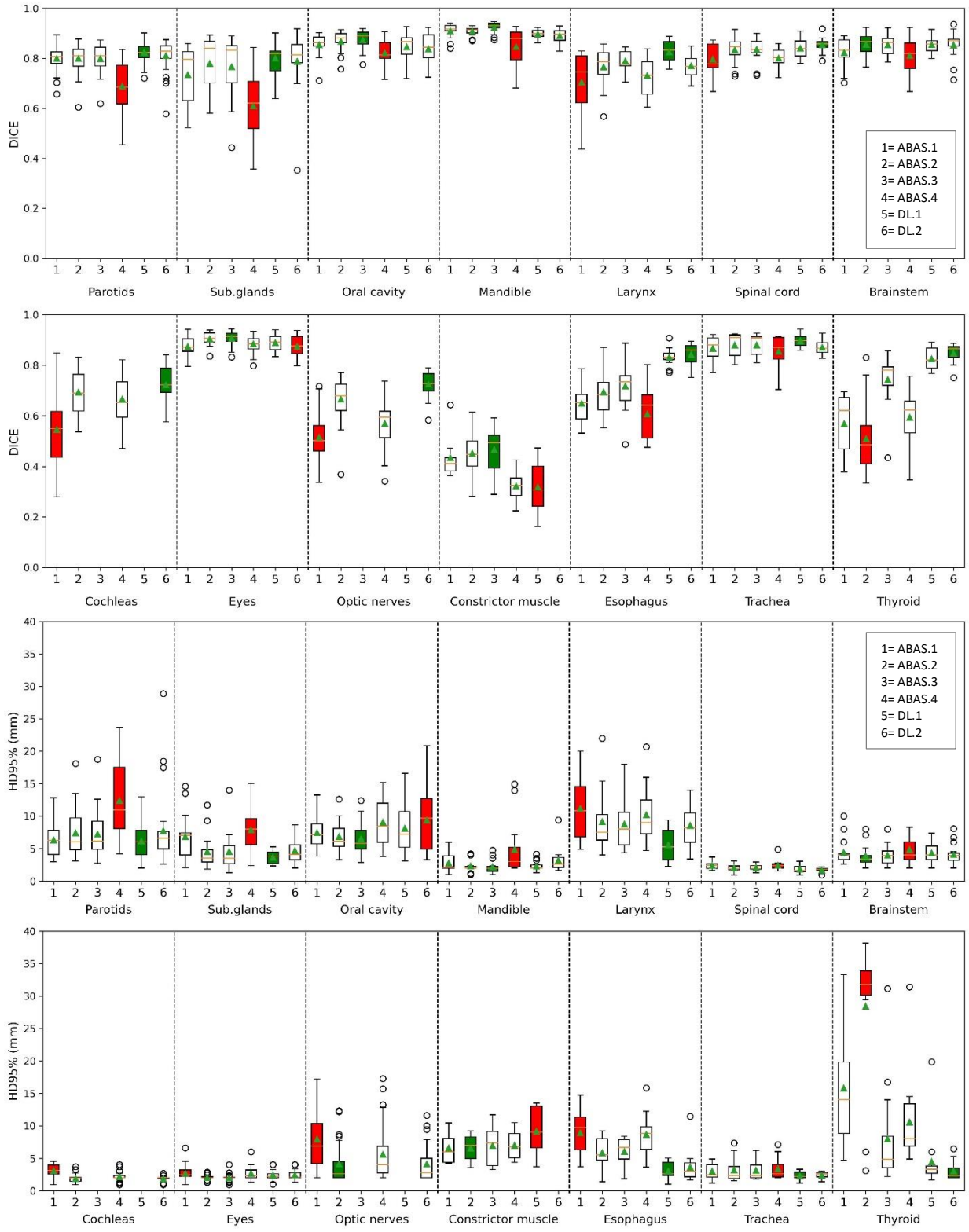

Figure 4

Click here to access/download;Figure;Fig4.Example of AS contours .pdf

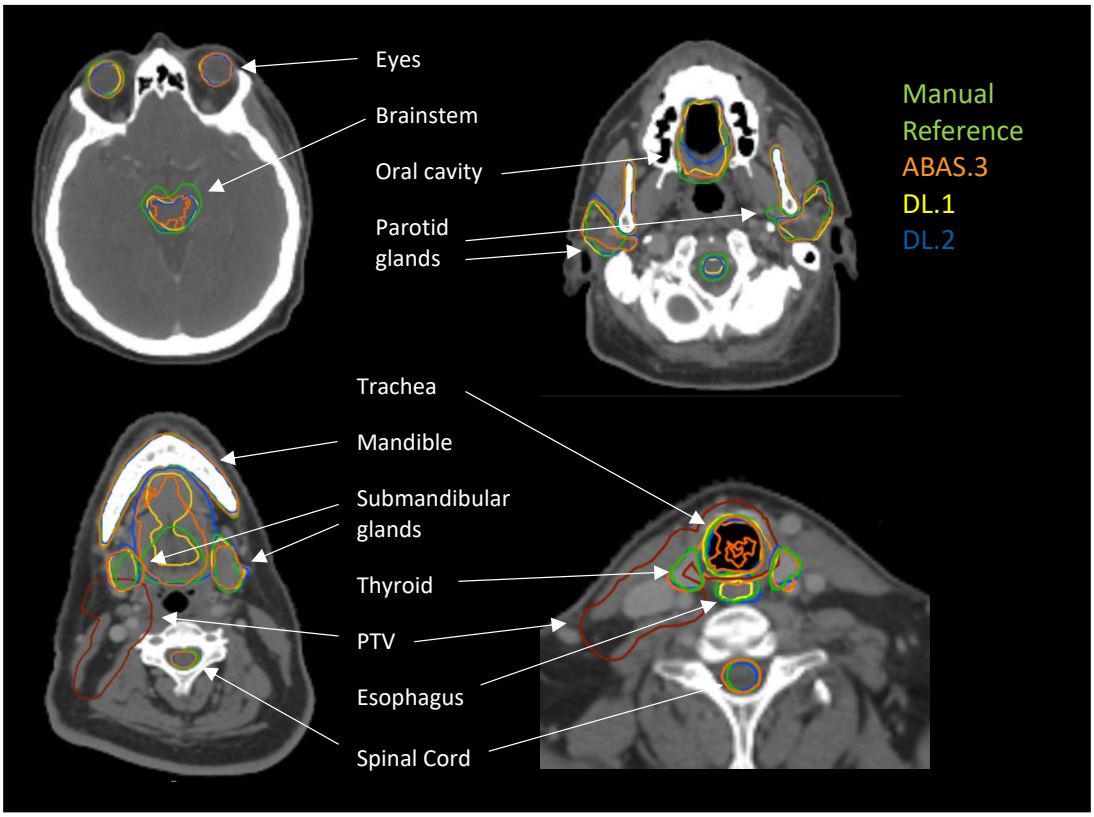

Figure 5

Click here to access/download;Figure;Fig5.Time spent on manual corrections.pdf

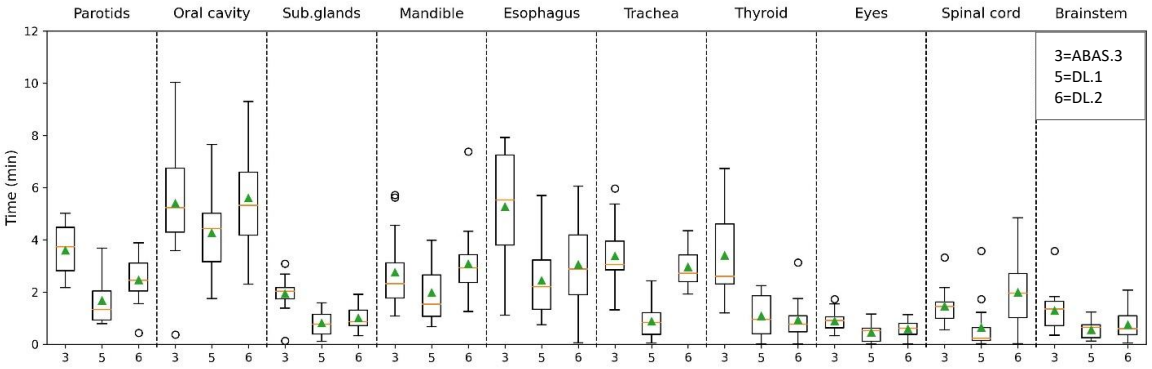

Fig. 1. Overview of the study design and performance evaluation methods; OAR = organs-at-risk, $HD_{95\%}$=95th percentile-Hausdorff Distance; * indicates the OAR that were not segmented by certain methods; Δdose=difference between the reference plan created with corrected OAR contours and the plan created with AS contours only;

Fig. 2. Geometric evaluation of the 6 automatic solutions. 1 = ABAS.1, 2 = ABAS.2, 3 = ABAS.3, 4 = ABAS.4, 5 = DL.1, 6 = DL.2; Panels A and B: analysis was performed over all the OAR. Panels C and D: analysis was performed over 11 common OAR. In red and in green are highlighted the worst and the best results, respectively determined by the mean value of DICE/$HD_{95\%}$; in the boxplots, the orange line represents the median, the green triangle indicate the mean value and the circles represent outliers.

Fig. 3. Geometric evaluation per OAR of the 4 multi-ABAS and 2 DL solutions; 1 = ABAS.1, 2 = ABAS.2, 3 = ABAS.3, 4 = ABAS.4, 5 = DL.1, 6 = DL.2; in red and in green are highlighted the worst and the best results, respectively determined by the mean value of DICE/$HD_{95\%}$; in the boxplots, the orange line represents the median, the green triangle indicate the mean value and the circles represent outliers; *Abbreviations:* Sub.glands = submandibular glands;

Fig. 4. Example of automatic segmentation uncertainties compared with manual delineations. OAR position relative to the PTV can be observed in panels C and D. A good agreement was generally observed in simple geometry structures such as eyes or spinal cord. Large contour discrepancies were noticed compared with the manual reference in the cranial and caudal slices for some structures such as oral cavity, trachea or brainstem. To illustrate OAR position relative the target, PTV volume is displayed.

Fig. 5. Time spent on manual corrections for each OAR automatically generated. 3 = ABAS.3, 5 = DL.1, and 6 = DL.2; Abbreviations: Sub.glands = submandibular glands;

Table 1 Characteristics of the test cohort used for evaluation of the AS solutions

Table 2 Geometric evaluation after manual corrections of 10 OAR for the three best solutions; with * are marked the differences that are statistically significant ($p<0.05$). Abbreviations: Sub.glands = submandibular glands;

Table 3 Dosimetric differences between doses generated with manually corrected contours and automatic contours, analyzed on the corrected contours; impact on the target volumes is evaluated in $V_{95\%}$ dose coverage, for the spinal cord and brainstem in $D_{2\%}$, and for the mandible in $D_{5\%}$ while for the rest of the OAR mean doses are calculated. Abbreviations: sub.glands = submandibular glands; R = right, L = left;

Supplementary Files

Click here to access/download

Supplementary Files

SuplementaryMaterial_revision.docx

Conflicts of interest statement

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