applied sciences

MDPI

Article

Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring for Head and Neck Radiation Therapy

Curtise K. C. Ng [1,2,*], Vincent W. S. Leung [3] and Rico H. M. Hung [4]

[1] Curtin Medical School, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
[2] Curtin Health Innovation Research Institute (CHIRI), Faculty of Health Sciences, Curtin University, GPO Box U1987, Perth, WA 6845, Australia
[3] Department of Health Technology and Informatics, Faculty of Health and Social Sciences, The Hong Kong Polytechnic University, Hong Kong, China
[4] Department of Clinical Oncology, Pamela Youde Nethersole Eastern Hospital, Hong Kong, China
* Correspondence: curtise.ng@curtin.edu.au or curtise_ng@yahoo.com.hk; Tel.: +61-8-9266-7314; Fax: +61-8-9266-2377

Featured Application: Deep learning (DL) auto-contouring instead of atlas-based auto-contouring and manual contouring should be used for anatomy segmentation in head and neck radiation therapy for reducing contouring time, and commercial DL auto-contouring tools should be further trained by local hospital datasets for enhancing their geometric accuracy.

Abstract: Various commercial auto-contouring solutions have emerged over past few years to address labor-intensiveness, and inter- and intra-operator variabilities issues of traditional manual anatomy contouring for head and neck (H&N) radiation therapy (RT). The purpose of this study is to compare the clinical performances between RaySearch Laboratories deep learning (DL) and atlas-based auto-contouring tools for organs at risk (OARs) segmentation in the H&N RT with the manual contouring as reference. Forty-five H&N computed tomography datasets were used for the DL and atlas-based auto-contouring tools to contour 16 OARs and time required for the segmentation was measured. Dice similarity coefficient (DSC), Hausdorff distance (HD) and HD 95th-percentile (HD95) were used to evaluate geometric accuracy of OARs contoured by the DL and atlas-based auto-contouring tools. Paired sample *t*-test was employed to compare the mean DSC, HD, HD95, and contouring time values of the two groups. The DL auto-contouring approach achieved more consistent performance in OARs segmentation than its atlas-based approach, resulting in statistically significant time reduction of the whole segmentation process by 40% ($p < 0.001$). The DL auto-contouring had statistically significantly higher mean DSC and lower HD and HD95 values ($p < 0.001$–0.009) for 10 out of 16 OARs. This study proves that the RaySearch Laboratories DL auto-contouring tool has significantly better clinical performances than its atlas-based approach.

Keywords: artificial intelligence; automation; computed tomography; image segmentation; intensity-modulated radiation therapy; machine learning; nasopharyngeal cancer; organs at risk; radiotherapy; volumetric arc therapy

check for updates

Citation: Ng, C.K.C.; Leung, V.W.S.; Hung, R.H.M. Clinical Evaluation of Deep Learning and Atlas-Based Auto-Contouring for Head and Neck Radiation Therapy. *Appl. Sci.* **2022**, *12*, 11681. https://doi.org/10.3390/app122211681

Academic Editor: Yujie Huang

Received: 28 October 2022
Accepted: 16 November 2022
Published: 17 November 2022

1. Introduction

Nowadays, intensity-modulated radiation therapy (IMRT) is an important cancer treatment option as a result of its capability of reduction of toxic effects associated with radiation therapy (RT). One of the essential requirements for IMRT is anatomy contouring [1–3]. Traditionally, this requires radiation therapists and/or oncologists to manually identify and contour tumors (clinical target volumes (CTVs)) and normal tissues (organs at risk (OARs)). Labor-intensiveness, and inter- and intra-operator variabilities are two major issues of the manual contouring [1–5]. Various commercial auto-contouring solutions have emerged over past few years to address these issues. Atlas-based and deep learning (DL) approaches are used to develop these auto-contouring solutions [3–11]. The atlas-based method involves automatically registering reference patient contours (gold standard/ground truth) to new patients through deforming reference patient contours for matching new patient anatomical structures. This approach can be subdivided into three categories, namely single atlas (with use of one reference dataset), multi-atlas (using multiple datasets) and hybrid (based on outcome of statistical analysis of multiple gold standards) [3,4,12]. Over the past five years, use of DL in medical imaging has become popular [13–15]. Commercial companies such as Manteia Medical Technologies Co. (Xiamen, China) [5], Mirada Medical Limited (Oxford, UK) [7] and Carina Medical LLC (Lexington, KY, USA) [10] have applied the widely used deep convolutional neural network (DCNN) architecture to develop their auto-contouring models (software). The DL auto-contouring model development involves providing training datasets with ground truth contours to the model for learning features of tumors and normal tissues (including those not known by human) automatically. With sufficient training, the model becomes capable to automatically search for these features and locate them to achieve auto-contouring for new patients. Hence, it is believed that the DL auto-contouring approach performs better than the atlas-based method because of its capability of identifying unknown but relevant features for more accurate contouring [1–3,5,7,10–12,16–21].

It is well known that head and neck cancer segmentation is a challenging task because computed tomography (CT) images which have limited soft tissue contrast are commonly used in the RT planning process [1,2,22–26]. This results in the greatest inter- and intra-operator variabilities and takes about 2–3 h for the manual contouring in the head and neck RT [2,22–25]. Few studies compared performances between commercial atlas-based and DL auto-contouring software packages with the manual contouring as the reference [5,6,10,27]. The investigated atlas-based and DL auto-contouring software pairs included Maestro 6.6.5 (MIM Software Inc., Cleveland, OH, USA) versus AccuContour (Manteia Medical Technologies Co., Xiamen, China) [5], Maestro 6.9.6 versus INTContour (Carina Medical LLC, Lexington, KY, USA) [6], ANAtomically Constrained Deformation Algorithm (ANACONDA) (RaySearch Laboratories AB, Stockholm, Sweden) versus INTContour [10], and WorkflowBox 1.4 (Mirada Medical Limited, Oxford, UK) versus Mirada Medical Limited DLCExpert [27]. Those studies showed consistent results that the DL auto-contouring was more accurate and required less time for segmentation of OARs [5,6,10,27]. Given these promising results of the DL auto-contouring, it appears worthwhile to further investigate the potential of other unexplored DL auto-contouring software packages such as RaySearch Laboratories AB RayStation 12A RSL Head and Neck CT 2.0.0.47 DL auto-contouring model. The purpose of this study is to compare the clinical performances between RaySearch Laboratories AB RayStation DL auto-contouring model, RSL Head and Neck CT 2.0.0.47 and its atlas-based auto-contouring software, ANACONDA for the OARs segmentation in the head and neck RT with the manual contouring as the reference. It is hypothesized that the DL auto-contouring model is more accurate and requires less contouring time when compared with the atlas-based auto-contouring tool.

2. Materials and Methods

2.1. Study Design and Imaging Data

This was a retrospective study with methods based on similar studies of comparison of performances between the atlas-based and DL auto-contouring for the OARs segmentation in the head and neck RT [5,6,10,27]. Planning CT datasets of 45 head and neck cancer patients who had RT treatments between 2018 and 2021 at Pamela Youde Nethersole Eastern Hospital in Hong Kong Special Administrative Region were retrospectively collected from Eclipse treatment planning system (Varian Medical Systems, Palo Alto, CA, USA). Patient inclusion criteria were: 1. head and neck cancer histologically proven and 2. radical radiotherapy received. Patients with pre-radiotherapy surgery were excluded. Table 1 shows the patient characteristics. The study was conducted in accordance with the Declaration of Helsinki, and approved by the Human Research Ethics Committee of

Curtin University (approval number: HRE2022-0582 and date of approval: 18 October 2022), Institutional Review Board of The Hong Kong Polytechnic University (approval number: HSEARS20220815001 and date of approval: 7 October 2022), and Research Ethics Committee of Hong Kong East Cluster of Hospital Authority of Government of Hong Kong Special Administrative Region (approval number: HKECREC-2022-054 and date of approval: 6 October 2022). Patient consent was waived due to the retrospective nature. The collected CT datasets were acquired with the following parameters, slice thickness of 2 mm, 400 mAs, 120 kV and field of view of 600 mm in accordance with the routine protocol of Pamela Youde Nethersole Eastern Hospital [27]. Figure 1 shows the overview of study design.

Table 1. Patient characteristics (*n* = 45).

Characteristics	Value
Gender (n = 45)	
Male	35 (78%)
Female	10 (22%)
Age (n = 45)	
18–65 years	29 (64%)
>65 years	16 (36%)
Tumor site	
Nasopharynx	45 (100%)
Tumour classification	
T1	13 (29%)
T2	3 (6%)
T3	17 (38%)
T4	12 (27%)
Node classification	
N0	13 (29%)
N1	13 (29%)
N2	9 (20%)
N3	10 (22%)
Systemic treatment	
Yes	37 (82%)
No	8 (18%)
Treatment technique	
Volumetric arc therapy	45 (100%)
Neck irradiation	
Bilateral	45 (100%)

Figures in parentheses are proportions.

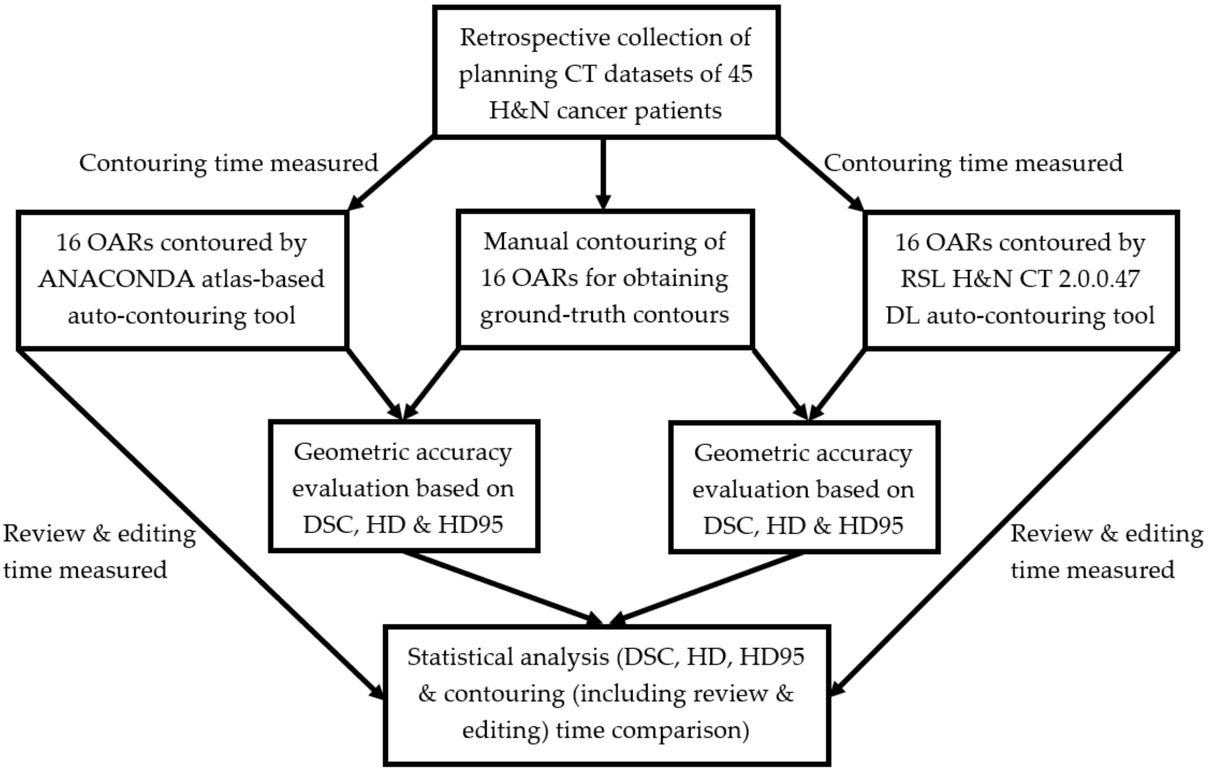
Retrospective collection of planning CT datasets of 45 H&N cancer patients
Contouring time measured
Contouring time measured
16 OARs contoured by ANACONDA atlas-based auto-contouring tool
Manual contouring of 16 OARs for obtaining ground-truth contours
16 OARs contoured by RSL H&N CT 2.0.0.47 DL auto-contouring tool
Geometric accuracy evaluation based on DSC, HD & HD95
Geometric accuracy evaluation based on DSC, HD & HD95
Review & editing time measured
Review & editing time measured
Statistical analysis (DSC, HD, HD95 & contouring (including review & editing) time comparison)

Figure 1. Study design overview. ANACONDA-ANAtomically Constrained Deformation Algorithm; CT-computed tomography; DL-deep learning; DSC-Dice similarity coefficient; H&N-head and neck; HD-Hausdorff distance; HD95-HD 95th-percentile; OARs-organs at risk.

2.2. Manual Contouring

Manual contouring for the collected 45 CT datasets was conducted by a radiation therapist with more than 10 years of experience in head and neck RT planning on the Eclipse treatment planning system as per international consensus delineation guidelines [28]. Contoured OARs included brainstem, larynx, optic chiasm, oral cavity, pituitary, spinal cord, two cochleae, two eyes, two lenses, two optic nerves and two parotid glands [27]. The manual contours were subsequently reviewed and modified by a radiation oncologist specialized in the head and neck RT for eventual clinical use, and hence considered as the ground-truth contours [10].

2.3. Atlas-Based and DL Auto-Contouring

The 45 CT datasets were exported to the RaySearch Laboratories AB RayStation 12A treatment planning system (with Intel Xeon W-10885M central processing unit (Santa Clara, CA, USA), 64 GB random access memory and NVIDIA Quadro RTX 5000 16 GB graphics card (Santa Clara, CA, USA)) in the Pamela Youde Nethersole Eastern Hospital. Its ANACONDA and RSL Head and Neck CT 2.0.0.47 were used for the atlas-based and DL auto-contouring, respectively. ANACONDA used the hybrid approach for auto-contouring. In the beginning of a new image set being contoured, multiple best-matching atlases from a pool of 40 nasopharyngeal cancer datasets within the hospital database were determined by rigid image registration. Contours from these best-matching atlases were subsequently mapped to the new image set using deformable registration and these contours eventually merged together as the segmentation outcome via a fusion algorithm [10]. The DL auto-contouring involved a pre-trained three-dimensional (3D) U-net CNN model to classify each voxel of the datasets into either specific tissue (i.e., an OAR) or non-specific one for generating labelled (segmented) datasets as the outcome. Only its pre-trained DL auto-contouring model without any finetuning was used for the DL auto-contouring. Details of its 3D U-net CNN architecture, and model training, validation and testing were available elsewhere [29].

2.4. Geometric Accuracy and Contouring Time Evaluation

Three parameters, Dice similarity coefficient (DSC) for quantification of degree of overlapping between two contours (Contour_A and Contour_B), Hausdorff distance (HD) defined as pairwise 3D point distance from one contour to another contour, and HD 95th-percentile (HD95) were employed to evaluate the geometric accuracy of the 16 OARs contoured by the atlas-based and DL auto-contouring tools with the manual contours as the reference. Equations (1)–(3) were used for DSC, HD and HD95 calculations, respectively [5,6,10,27].

$$\mathrm{DSC} = \frac{2(\mathrm{Contour_A} \cap \mathrm{Contour_B})}{\mathrm{Contour_A} + \mathrm{Contour_B}} \tag{1}$$

$$\mathrm{d_{HD}} = \max\left(\min_{a \in A} \mathrm{d(a)}, \min_{b \in B} \mathrm{d(b)}\right) \tag{2}$$

$$\vec{\mathrm{d}}_{\mathrm{HD95}}(\mathrm{A,B}) = \mathrm{k_{95}}\left(\min_{b \in B} \mathrm{d(a,b)}\right), \mathrm{d_{HD95}}(\mathrm{A,B}) \tag{3}$$

where a and b represent the points on contours A and B.

A free, open source image computing platform, 3D Slicer 5.0.3 (The Slicer Community, Boston, MA, USA) with SlicerRT extension was used to determine the DSC, HD and HD95 values [9,30,31]. The DSC values, <0.50, 0.50–0.59, 0.60–0.69, 0.70–0.79 and 0.80–1.00 indicated poor, intermediate-poor, intermediate, good-intermediate and good degree of overlapping between two contours. For HD95, <4.0 mm, 4.0–5.9 mm, 6.0–7.9 mm and ≥8.0 mm represented good, intermediate, poor and very poor geometric accuracy [27].

For the contouring time evaluation, the time required by the atlas-based and DL auto-contouring tools for the auto-contouring processes was recorded. Additionally, the same radiation therapist involved in the previous manual contouring process reviewed and edited the 16 OARs contoured by the two auto-contouring tools as per the clinical protocol of the hospital. The review and editing time was recorded as well [27]. No dataset with patient personal information was taken from the hospital.

2.5. Statistical Analysis

SPSS Statistics 28 (International Business Machines Corporation, Armonk, NY, USA) was used for statistical analysis. Paired sample *t*-test was employed to compare the mean DSC, HD, HD95, auto-contouring time, review and editing time and total segmentation process time values of the DL and atlas-based auto-contouring groups. A *p*-value less than 0.05 represented statistical significance [5,6,10].

3. Results

3.1. Geometric Accuracy

The geometric accuracy of the DL auto-contouring approach was higher than that of the atlas-based method overall. Figure 2 shows that the DL auto-contouring performed more consistently with smaller DSC, HD and HD95 variations for nearly all OARs when compared with the atlas-based auto-contouring. Additionally, its geometric accuracy (in terms of higher DSC and lower HD and HD95 values) was better for 10 out of 16 OARs. These 10 OARs were brain stem, left and right eyes, left and right lens, left and right optic nerves, left and right parotid glands, and pituitary. Except for the DSC values of left eye, the DL auto-contouring had statistically significantly higher mean DSC and lower HD and HD95 values for these 10 OARs as well (Table 2).

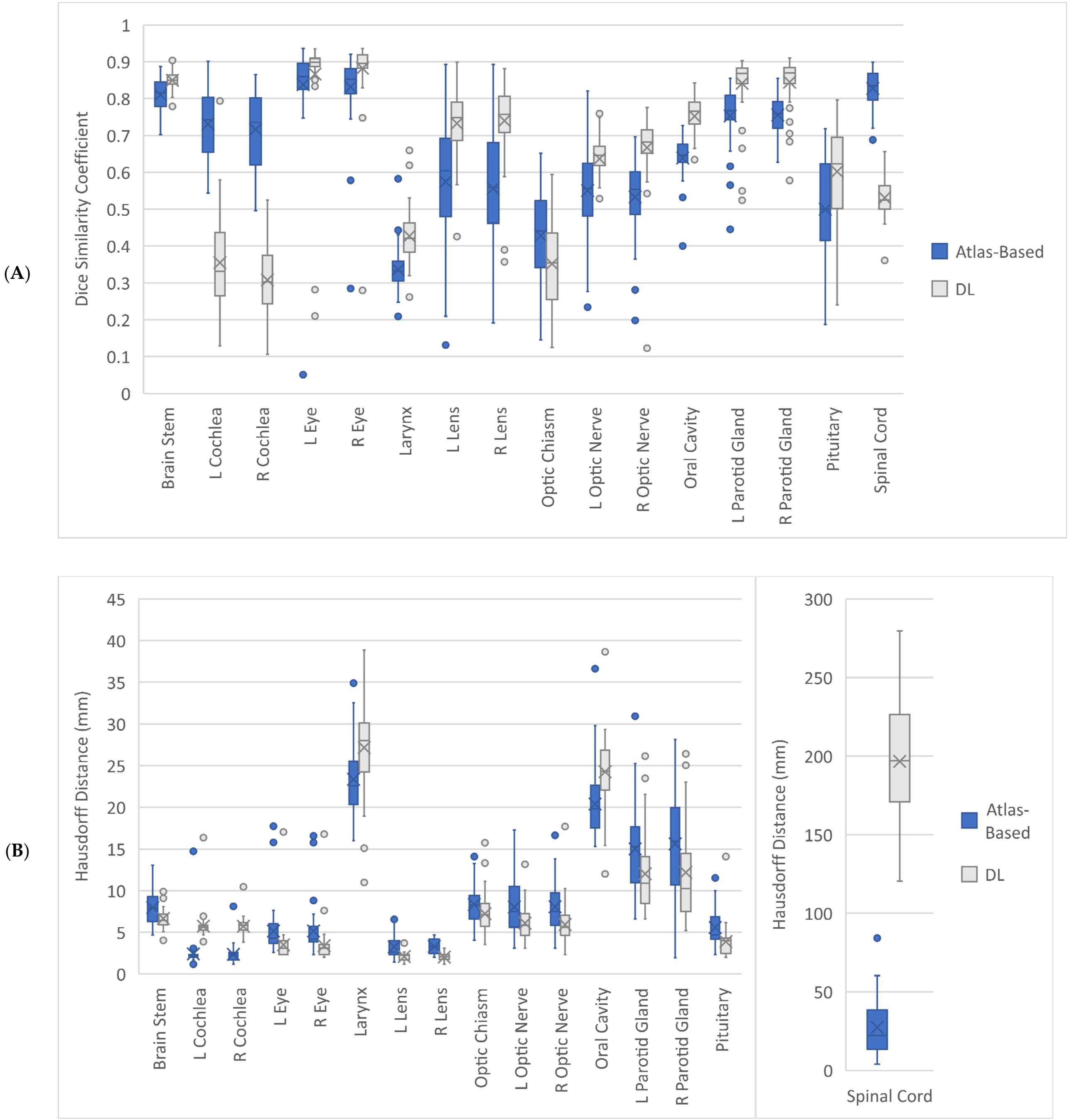

Figure 2. *Cont.*

(**C**)

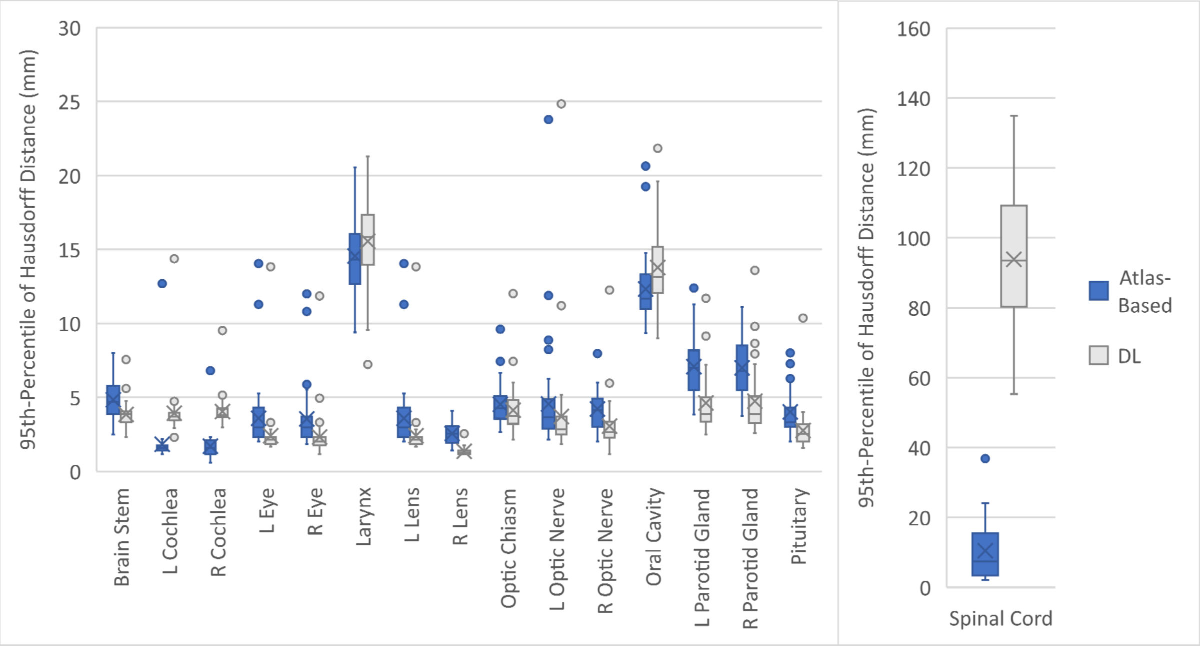

Figure 2. Boxplots of three geometric accuracy evaluation parameters, Dice similarity coefficient (**A**), Hausdorff distance (**B**), and 95th-percentile of Hausdorff distance (**C**) for organs at risk contoured by atlas-based and deep learning (DL) auto-contouring approaches. L-left; R-right.

Table 2. Comparison of geometric accuracy of contours between atlas-based and DL auto-contouring.

Organ at Risk	Geometric Accuracy Parameter	Atlas-Based Auto-Contouring (Mean (CI))	Deep Learning Auto-Contouring (Mean (CI))	*p*-Value
Brainstem	DSC	0.81 (0.80, 0.82)	0.85 (0.84, 0.86)	<0.001
	HD (mm)	7.96 (7.37, 8.55)	6.66 (6.30, 7.05)	<0.001
	HD95 (mm)	4.84 (4.42, 5.22)	3.85 (3.59, 4.11)	<0.001
Left Cochlea	DSC	0.73 (0.71, 0.75)	0.35 (0.32, 0.39)	<0.001
	HD (mm)	2.40 (2.02, 3.00)	5.73 (5.34, 6.28)	<0.001
	HD95 (mm)	1.82 (1.52, 2.34)	3.92 (3.59, 4.44)	<0.001
Right Cochlea	DSC	0.72 (0.69, 0.75)	0.31 (0.28, 0.34)	<0.001
	HD (mm)	2.36 (2.10, 2.67)	5.72 (5.45, 6.00)	<0.001
	HD95 (mm)	1.69 (1.48, 1.97)	4.04 (3.81, 4.32)	<0.001
Left Eye	DSC	0.84 (0.80, 0.87)	0.87 (0.82, 0.90)	0.339
	HD (mm)	5.15 (4.43, 6.03)	3.49 (3.03, 4.17)	<0.001
	HD95 (mm)	3.58 (3.03, 4.27)	2.40 (2.06, 2.95)	<0.001
Right Eye	DSC	0.83 (0.80, 0.86)	0.88 (0.85, 0.90)	<0.001
	HD (mm)	5.14 (4.46, 6.09)	3.39 (2.89, 4.06)	<0.001
	HD95 (mm)	3.54 (3.04, 4.18)	2.32 (1.98, 2.81)	<0.001
Larynx	DSC	0.34 (0.32, 0.35)	0.43 (0.41, 0.45)	<0.001
	HD (mm)	23.37 (22.29, 24.49)	27.17 (25.61, 28.65)	<0.001
	HD95 (mm)	14.55 (13.88, 15.24)	15.55 (14.75, 16.35)	0.021
Left Lens	DSC	0.57 (0.52, 0.63)	0.73 (0.71, 0.76)	<0.001
	HD (mm)	3.26 (2.92, 3.67)	2.08 (1.94, 2.23)	<0.001
	HD95 (mm)	2.40 (2.15, 2.67)	1.41 (1.33, 1.49)	<0.001
Right Lens	DSC	0.56 (0.51, 0.60)	0.74 (0.71, 0.77)	<0.001
	HD (mm)	3.30 (3.05, 3.58)	2.02 (1.88, 2.14)	<0.001
	HD95 (mm)	2.55 (2.35, 2.78)	1.35 (1.28, 1.43)	<0.001

Table 2. *Cont.*

Organ at Risk	Geometric Accuracy Parameter	Atlas-Based Auto-Contouring (Mean (CI))	Deep Learning Auto-Contouring (Mean (CI))	*p*-Value
Optic Chiasm	DSC	0.43 (0.39, 0.46)	0.35 (0.32, 0.38)	0.005
	HD (mm)	8.34 (7.68, 9.05)	7.23 (6.59, 7.91)	0.013
	HD95 (mm)	4.53 (4.12, 4.99)	4.13 (3.73, 4.60)	0.131
Left Optic Nerve	DSC	0.55 (0.52, 0.58)	0.64 (0.62, 0.66)	<0.001
	HD (mm)	8.03 (7.04, 8.95)	6.10 (5.51, 6.76)	0.007
	HD95 (mm)	4.55 (3.68, 5.58)	3.70 (2.92, 4.87)	0.005
Right Optic Nerve	DSC	0.53 (0.50, 0.56)	0.67 (0.64, 0.69)	<0.001
	HD (mm)	8.07 (7.16, 8.92)	5.92 (5.28, 6.71)	0.002
	HD95 (mm)	4.21 (3.80, 4.65)	3.05 (2.69, 3.55)	<0.001
Oral Cavity	DSC	0.64 (0.62, 0.65)	0.75 (0.73, 0.77)	<0.001
	HD (mm)	20.39 (19.24, 21.52)	24.26 (23.03, 25.40)	<0.001
	HD95 (mm)	12.32 (11.72, 12.97)	13.78 (13.01, 14.64)	<0.001
Left Parotid Gland	DSC	0.75 (0.73, 0.78)	0.84 (0.82, 0.86)	<0.001
	HD (mm)	15.01 (13.58, 16.47)	12.02 (10.64, 13.63)	0.006
	HD95 (mm)	7.08 (6.49, 7.60)	4.62 (3.99, 5.31)	<0.001
Right Parotid Gland	DSC	0.76 (0.74, 0.77)	0.84 (0.83, 0.86)	<0.001
	HD (mm)	15.61 (13.86, 17.30)	12.18 (10.45, 14.05)	0.009
	HD95 (mm)	7.00 (6.46, 7.57)	4.74 (4.11, 5.42)	<0.001
Pituitary	DSC	0.50 (0.46, 0.54)	0.60 (0.56, 0.64)	<0.001
	HD (mm)	5.57 (4.99, 6.17)	3.89 (3.40, 4.53)	<0.001
	HD95 (mm)	4.02 (3.55, 4.51)	2.76 (2.44, 3.19)	<0.001
Spinal Cord	DSC	0.83 (0.81, 0.84)	0.53 (0.52, 0.55)	<0.001
	HD (mm)	27.39 (22.23, 33.35)	196.68 (185.58, 208.11)	<0.001
	HD95 (mm)	10.48 (8.07, 13.21)	93.85 (88.26, 99.67)	<0.001

CI-95% confidence interval; DSC-Dice similarity coefficient; HD-Hausdorff distance; HD95-95th-percentile of Hausdorff distance.

Nonetheless, Table 2 also illustrates that both atlas-based and DL auto-contouring approaches had very poor/poor accuracy (DSC < 0.50/HD95 $\geq$ 6 mm) in contouring larynx, optic chiasm, oral cavity and spinal cord. In addition, the DL auto-contouring only had mean DSC of 0.31–0.35 for contouring left and right cochleae. Figures 3 and 4 show two examples of OARs segmentations by the three approaches, namely manual contouring (gold standard), DL and atlas-based auto-contouring.

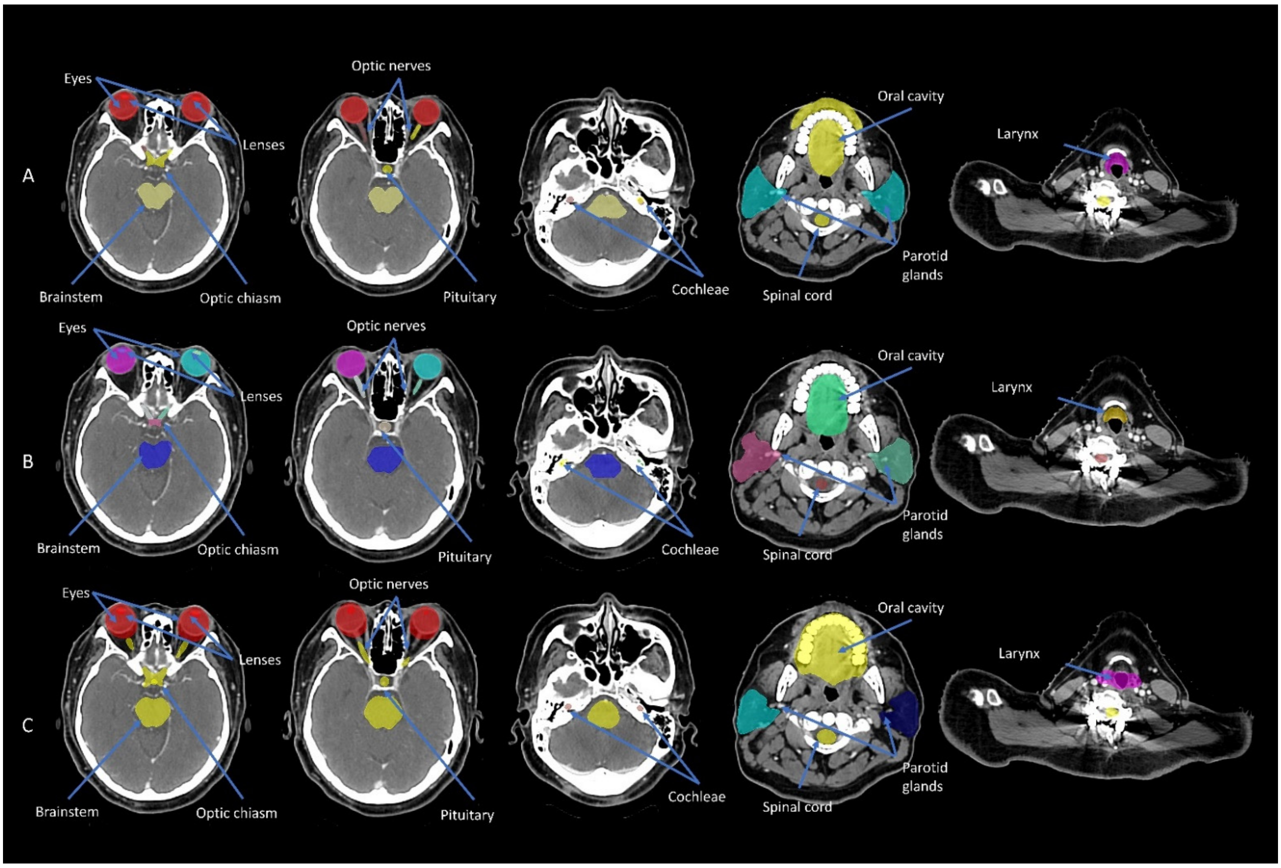

Figure 3. Organs at risk segmentation for a 75-year-old man with T3N2M0 nasopharyngeal cancer by manual contouring (**A**), DL (**B**) and atlas-based auto-contouring (**C**) with the following Dice similarity coefficient, Hausdorff distance, and 95th-percentile of Hausdorff distance values for brainstem (0.81, 7.20 mm and 4.39 mm vs. 0.83, 7.13 mm and 4.11 mm), left (L) cochlea (0.45, 5.24 mm and 2.92 mm vs. 0.81, 1.17 mm and 1.17 mm), right (R) cochlea (0.42, 6.83 mm and 3.53 mm vs. 0.79, 2.32 mm, 1.17 mm), L eye (0.85, 4.00 mm and 3.30 mm vs. 0.89, 3.30 mm, 2.33 mm), R eye (0.87, 3.52 mm and 2.33 mm vs. 0.88, 6.00 mm and 3.54 mm), larynx (0.47, 24.82 mm and 13.87 mm vs. 0.28, 18.89 mm and 12.51 mm), L lens (0.77, 2.32 mm and 1.41 mm vs. 0.53, 3.30 mm and 2.46 mm), R lens (0.67, 2.32 mm and 1.41 mm vs. 0.56, 3.30 mm and 2.46 mm), optic chiasm (0.47, 7.86 mm and 5.25 mm vs. 0.44, 9.61 mm and 4.42 mm), L optic nerve (0.66, 7.77 mm and 3.91 mm vs. 0.62, 4.33 mm and 3.31 mm), R optic nerve (0.70, 4.64 mm and 2.70 mm vs. 0.46, 8.45 mm and 3.93 mm), oral cavity (0.84, 19.31 mm and 10.24 mm vs. 0.68, 16.57 mm and 11.12 mm), L parotid gland (0.83, 11.71 mm and 4.78 mm vs. 0.57, 19.18 mm and 9.34 mm), R parotid gland (0.81, 10.73 mm and 4.74 mm vs. 0.69, 17.25 mm and 7.38 mm), pituitary (0.51, 3.71 mm and 2.54 mm vs. 0.46, 5.32 mm and 3.49 mm) and spinal cord (0.52, 220.13 mm and 104.18 mm vs. 0.79, 34.90 mm and 12.26 mm), respectively.

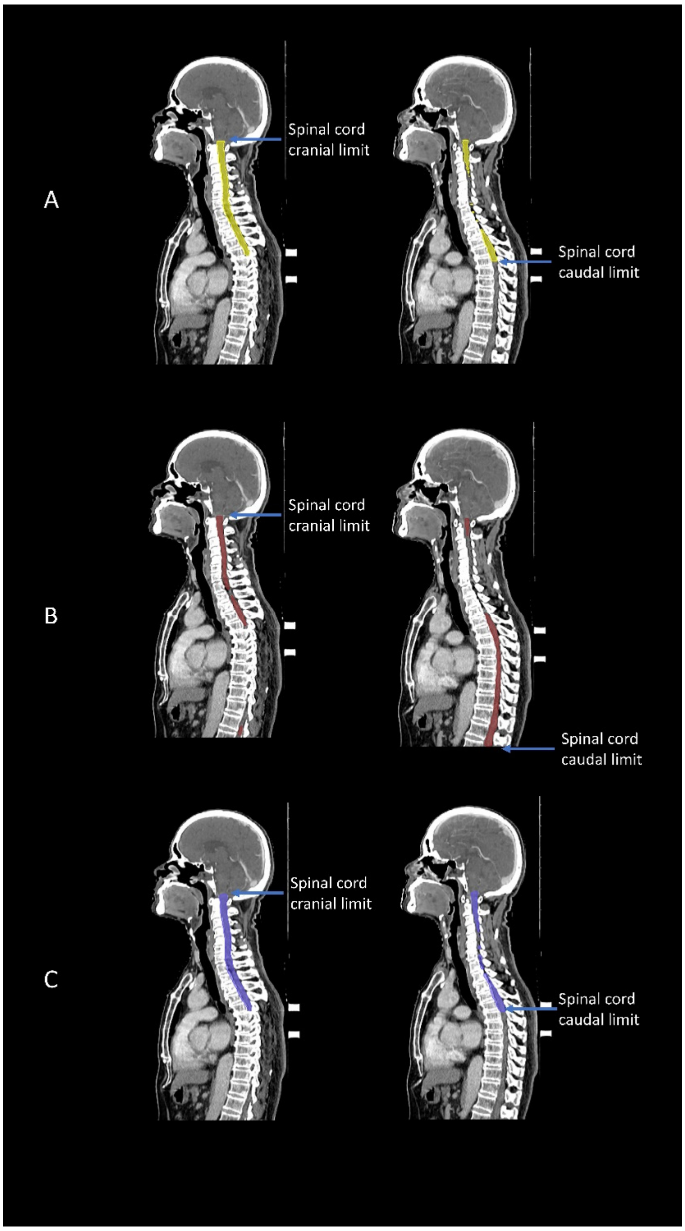
A
Spinal cord cranial limit
Spinal cord caudal limit
B
Spinal cord cranial limit
Spinal cord caudal limit
C
Spinal cord cranial limit
Spinal cord caudal limit

Figure 4. Sagittal view of spinal cord contouring for a 75-year-old man with T2N2M0 nasopharyngeal cancer by manual contouring (**A**), DL (**B**) (Dice similarity coefficient (DSC): 0.48, Hausdorff distance (HD): 188.62 mm, and 95th-percentile of Hausdorff distance (HD95): 91.27 mm) and atlas-based auto-contouring (**C**) (DSC: 0.86, HD: 6.55 mm, and HD95: 2.34 mm).

3.2. Contouring Time Evaluation

The DL auto-contouring approach required about 40% less time to complete all OARs segmentation for a dataset on average. This was contributed by 55% and 35% less time required for initial auto-contouring and subsequent review and editing processes, respectively. Table 3 illustrates the mean time required for the atlas-based and deep learning auto-contouring processes. The DL approach had statistically significantly shorter time required for auto-contouring and review and editing, and hence the whole process.

Table 3. Comparison of time required for atlas-based and deep learning auto-contouring.

Time (s)	Atlas-Based Auto-Contouring (Mean (CI))	Deep Learning Auto-Contouring (Mean (CI))	*p*-Value
Contouring	153.43 (150.63, 156.08)	69.53 (68.40, 70.58)	<0.001
Review and Editing	544.55 (495.02, 612.48)	352.98 (315.40, 397.92)	<0.001
Total	697.97 (647.55, 764.92)	422.51 (385.39, 467.14)	<0.001

CI-95% confidence interval.

4. Discussion

To the best of our knowledge, this was the first study on comparing the performances between RaySearch Laboratories AB RayStation DL auto-contouring model, RSL Head and Neck CT 2.0.0.47 and its atlas-based auto-contouring software, ANACONDA for the OARs segmentation in the head and neck RT. This study's results show that the RSL Head and Neck CT 2.0.0.47 DL auto-contouring approach could achieve more consistent performance in OARs segmentation than the ANACONDA atlas-based approach (Figure 2). These findings are in line with similar studies on comparing the INTContour DL auto-contouring approach with the Maestro 6.9.6 and ANACONDA atlas-based approaches for head and neck RT [6,10]. This consistent performance resulted in less time required for manual correction because the required adjustments became more straightforward and could be completed rapidly. Hence, this contributed to the statistically significant reduction of the review and editing time and the time required for the whole segmentation process by 35% and 40%, respectively (Table 3). For Li et al.'s [6] study, the INTContour DL and Maestro 6.9.6 atlas-based approaches required 120 s and 600 s to contour 4 head and neck OARs, respectively. Although this represented that their DL approach was able to reduce the total contouring time by 80%, the number of OARs involved in their study was only one-fourth of the number in this study. Hence, the total contouring time required for the RSL Head and Neck CT 2.0.0.47 DL auto-contouring approach appears comparable to their study. These findings are particularly important for addressing the major issues of the manual contouring, namely inter- and intra-operator variabilities and labor-intensiveness in head and neck RT [1–5].

To further address the manual contouring issue, high geometric accuracy of contours is essential. This study's results show that the RSL Head and Neck CT 2.0.0.47 DL auto-contouring had statistically significantly higher mean DSC and lower HD and HD95 values for 10 out of 16 OARs, namely brain stem, left and right eyes, left and right lens, left and right optic nerves, left and right parotid glands, and pituitary when compared with those of the ANACONDA atlas-based approach (Table 2). However, the highest mean DSC value achieved by the RSL Head and Neck CT 2.0.0.47 DL auto-contouring approach was only 0.88 for the right eye and the lowest one was 0.31 for the right cochlea. For Wang et al.'s [5] study on comparing the performances of Maestro 6.6.5 atlas-based, pre-trained and trained AccuContour DL auto-contouring tools for head and neck OARs contouring, the highest mean DSC value achieved by their pre-trained DL tool was above 0.9 but four (left and right temporomandibular joints and left and right optic nerves) out of 14 OARs contoured by their pre-trained DL model had mean DSC below 0.6. In contrast, their Maestro 6.6.5 atlas-based auto-contouring tool performed better for these four OARs. In this study, five (left and right cochleae, larynx, optic chiasm and spinal cord) out of 16 OARs contoured by the pre-trained RSL Head and Neck CT 2.0.0.47 DL model had mean DSC below 0.6. Except larynx, the ANACONDA atlas-based approach had higher mean DSC values for these OARs. The DSC findings of this study appears comparable to the corresponding tools in Wang et al.'s study [5].

According to van Dijk et al.'s [27] study on comparing the WorkflowBox 1.4 atlas-based auto-contouring tool with the Mirada Medical Limited DLCExpert DL auto-contouring model for head and neck RT, the DL model was more accurate to contour small OARs. However, this was not supported by Wang et al.'s [5] findings. For example, their Maestro 6.6.5 atlas-based auto-contouring tool was more accurate to contour the left and right optic nerves when compared with their pre-trained DL model. Although this study shows that the RSL Head and Neck CT 2.0.0.47 DL auto-contouring tool had higher mean DSC values for contouring the left and right optic nerves than its counterpart, aligning with van Dijk et al.'s findings, its performance in contouring two other small structures, left and right cochleae was the worst (mean DSC values: 0.31–0.35). Notable discrepancies between mean DSC and HD/HD95 values of these two OARs (mean HD95 of left and right cochleae: 3.92 mm (good) and 4.04 mm (intermediate), respectively) were also found. Despite that the DSC and HD/HD95 are parameters to evaluate the geometric accuracy of contours, the DSC is the indicator of shape similarity while HD/HD95 are the parameters to illustrate location similarity. HD95 is generally considered a better descriptor of location similarity than the HD due to its capability of excluding the outliers [6,27]. Similar discrepancies between the DSC and HD values were also found in Wang et al.'s [5] study. For example, their pre-built DL model had the mean DSC values of about 0.8 (good) for oral cavity and left and right parotid glands but the corresponding mean HD values were more than 17 mm (poor). In order to address these discrepancies and improve the geometric accuracy of the DL auto-contouring tool, Wang et al. [5] used 120 datasets from their local hospital database to finetune the DL model. In this way, the DL model could learn the local contouring protocols and become more capable to accurately contour the OARs as per the hospital's requirements. After further training, their AccuContour DL model was able to achieve DSC values of all 14 OARs above 0.7. This highlights finetuning of the commercial DL model is key to improve its geometric accuracy. In this study, discrepancies between the local hospital contouring protocol and DL auto-contouring tool were also found. For example, Figure 4 shows the notable difference of the spinal cord caudal limits of manual contouring protocol and the DL approach. It is expected that further training of the RSL Head and Neck CT 2.0.0.47 DL model would help to reduce these discrepancies and improve its accuracy [29].

This study had several major limitations. Only 45 nasopharyngeal cancer datasets were used for the atlas-based and DL auto-contouring tool evaluation. However, the number of evaluation datasets of this study was more than a double of the dataset numbers of the similar studies by Wang et al. [5] (20 datasets) and Li et al. [6] (22 datasets). Additionally, this study covered 16 OARs which was greater than the numbers of OARs (4–14) of other similar studies [5,6,10]. Although only one radiation therapist with more than 10 years of experience in head and neck RT planning was involved in the contouring time evaluation, the international consensus delineation guidelines were used for review and editing the contours for minimizing the inter- and intra-operator variabilities [5,28]. Besides, this study only evaluated one DL auto-contouring model and did not investigate dosimetric impact of contouring accuracy but these arrangements were consistent with the other similar studies [5,6,10,27].

5. Conclusions

This study compared the clinical performances between RaySearch Laboratories AB RayStation DL auto-contouring model, RSL Head and Neck CT 2.0.0.47 and its atlas-based auto-contouring software, ANACONDA for 16 OARs segmentation in the head and neck RT with the manual contouring as the reference. Its results show that the DL auto-contouring model was more accurate and required less contouring time when compared with the atlas-based auto-contouring tool. The DL auto-contouring approach could achieve more consistent performance in the OARs segmentation than the atlas-based approach, resulting in statistically significant reduction of the time required for the whole segmentation process by 40%. Additionally, the DL auto-contouring had statistically significantly higher mean DSC and lower HD and HD95 values for 10 out of 16 OARs, namely brain stem, left and right eyes, left and right lens, left and right optic nerves, left and right parotid glands, and pituitary. These outcomes are particularly important for addressing the major issues of the manual contouring, namely inter- and intra-operator variabilities and labor-intensiveness in head and neck RT. However, for future studies, more radiation oncologists and head and neck cancer datasets with greater varieties should be used to evaluate the performances (in terms of contouring time, geometric accuracy and associated dosimetric impact) of multiple atlas-based, pre-trained and trained DL auto-contouring tools in contouring a great number of OARs. Comparison of performances of RaySearch Laboratories DL and atlas-based auto-contouring tools in contouring both CTVs and OARs for other cancer types such as breast cancer should be conducted as well.

Author Contributions: Conceptualization, C.K.C.N., V.W.S.L. and R.H.M.H.; methodology, C.K.C.N., V.W.S.L. and R.H.M.H.; software, V.W.S.L.; validation, C.K.C.N., V.W.S.L. and R.H.M.H.; formal analysis, C.K.C.N., V.W.S.L. and R.H.M.H.; investigation, C.K.C.N., V.W.S.L. and R.H.M.H.; resources, C.K.C.N., V.W.S.L. and R.H.M.H.; data curation, R.H.M.H.; writing—original draft preparation, C.K.C.N., V.W.S.L. and R.H.M.H.; writing—review and editing, C.K.C.N., V.W.S.L. and R.H.M.H.; visualization, C.K.C.N., V.W.S.L. and R.H.M.H.; project administration, C.K.C.N.; funding acquisition, C.K.C.N. and V.W.S.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by The Hong Kong Polytechnic University Large Equipment Fund for Teaching 2020/21, grant number HTI-4.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Human Research Ethics Committee of Curtin University (approval number: HRE2022-0582 and date of approval: 18 October 2022), Institutional Review Board of The Hong Kong Polytechnic University (approval number: HSEARS20220815001 and date of approval: 7 October 2022), and Research Ethics Committee of Hong Kong East Cluster of Hospital Authority of Government of Hong Kong Special Administrative Region (approval number: HKECREC-2022-054 and date of approval: 6 October 2022).

Informed Consent Statement: Patient consent was waived due to the retrospective nature.

Data Availability Statement: The datasets used in this study are not publicly available due to strict requirements set out by the Research Ethics Committee of Hong Kong East Cluster of Hospital Authority of Government of Hong Kong Special Administrative Region.

Acknowledgments: The authors would like to thank Venus W. Ho for her assistance with the collection of data from Department of Clinical Oncology, Pamela Youde Nethersole Eastern Hospital, Hong Kong Special Administrative Region.

Conflicts of Interest: The authors declare no conflict of interest. The funder had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Oktay, O.; Nanavati, J.; Schwaighofer, A.; Carter, D.; Bristow, M.; Tanno, R.; Jena, R.; Barnett, G.; Noble, D.; Rimmer, Y.; et al. Evaluation of deep learning to augment image-guided radiotherapy for head and neck and prostate cancers. *JAMA Netw. Open* **2020**, *3*, e2027426. [CrossRef] [PubMed]
2. Cardenas, C.E.; Beadle, B.M.; Garden, A.S.; Skinner, H.D.; Yang, J.; Rhee, D.J.; McCarroll, R.E.; Netherton, T.J.; Gay, S.S.; Zhang, L.; et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. *Int. J. Radiat. Oncol. Biol. Phys.* **2021**, *109*, 801–812. [CrossRef] [PubMed]
3. Kosmin, M.; Ledsam, J.; Romera-Paredes, B.; Mendes, R.; Moinuddin, S.; de Souza, D.; Gunn, L.; Kelly, C.; Hughes, C.O.; Karthikesalingam, A.; et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother. Oncol.* **2019**, *135*, 130–140. [CrossRef] [PubMed]
4. Lee, H.; Lee, E.; Kim, N.; Kim, J.H.; Park, K.; Lee, H.; Chun, J.; Shin, J.I.; Chang, J.S.; Kim, J.S. Clinical evaluation of commercial atlas-based auto-segmentation in the head and neck region. *Front. Oncol.* **2019**, *9*, 239. [CrossRef]
5. Wang, J.; Chen, Z.; Yang, C.; Qu, B.; Ma, L.; Fan, W.; Zhou, Q.; Zheng, Q.; Xu, S. Evaluation exploration of atlas-based and deep learning-based automatic contouring for nasopharyngeal carcinoma. *Front. Oncol.* **2022**, *12*, 833816. [CrossRef] [PubMed]
6. Li, Y.; Rao, S.; Chen, W.; Azghadi, S.F.; Nguyen, K.N.B.; Moran, A.; Usera, B.M.; Dyer, B.A.; Shang, L.; Chen, Q.; et al. Evaluating automatic segmentation for swallowing-related organs for head and neck cancer. *Technol. Cancer Res. Treat.* **2022**, *21*, 15330338221105724. [CrossRef] [PubMed]
7. Brunenberg, E.J.L.; Steinseifer, I.K.; van den Bosch, S.; Kaanders, J.H.A.M.; Brouwer, C.L.; Gooding, M.J.; van Elmpt, W.; Monshouwer, R. External validation of deep learning-based contouring of head and neck organs at risk. *Phys. Imaging Radiat. Oncol.* **2020**, *15*, 8–15. [CrossRef]

8. Aliotta, E.; Nourzadeh, H.; Choi, W.; Leandro Alves, V.G.; Siebers, J.V. An automated workflow to improve efficiency in radiation therapy treatment planning by prioritizing organs at risk. *Adv. Radiat. Oncol.* **2020**, *5*, 1324–1333. [CrossRef]
9. Ayyalusamy, A.; Vellaiyan, S.; Subramanian, S.; Ilamurugu, A.; Satpathy, S.; Nauman, M.; Katta, G.; Madineni, A. Auto-segmentation of head and neck organs at risk in radiotherapy and its dependence on anatomic similarity. *Radiat. Oncol. J.* **2019**, *37*, 134–142. [CrossRef]
10. Chen, W.; Li, Y.; Dyer, B.A.; Feng, X.; Rao, S.; Benedict, S.H.; Chen, Q.; Rong, Y. Deep learning vs. atlas-based models for fast auto-segmentation of the masticatory muscles on head and neck CT images. *Radiat. Oncol.* **2020**, *15*, 176. [CrossRef]
11. Robert, C.; Munoz, A.; Moreau, D.; Mazurier, J.; Sidorski, G.; Gasnier, A.; Beldjoudi, G.; Grégoire, V.; Deutsch, E.; Meyer, P.; et al. Clinical implementation of deep-learning based auto-contouring tools-Experience of three French radiotherapy centers. *Cancer Radiother.* **2021**, *25*, 607–616. [CrossRef]
12. Karagiannis, E.; Koreas, P.; Strouthos, I.; Leczynski, A.; Grimm, M.; Zamboglou, N.; Ferentinos, K. Evaluation of an atlas-based auto-segmentation tool of target volumes and organs at risk in head and neck radiation therapy. *J. Oncol. Res. Ther.* **2021**, *6*, 10113. [CrossRef]
13. Sun, Z.; Ng, C.K.C. Artificial intelligence (enhanced super-resolution generative adversarial network) for calcium deblooming in coronary computed tomography angiography: A feasibility study. *Diagnostics* **2022**, *12*, 991. [CrossRef]
14. Ng, C.K.C. Artificial intelligence for radiation dose optimization in pediatric radiology: A systematic review. *Children* **2022**, *9*, 1044. [CrossRef]
15. Sun, Z.; Ng, C.K.C. Finetuned super-resolution generative adversarial network (artificial intelligence) model for calcium deblooming in coronary computed tomography angiography. *J. Pers. Med.* **2022**, *12*, 1354. [CrossRef]
16. Zhong, Y.; Yang, Y.; Fang, Y.; Wang, J.; Hu, W. A preliminary experience of implementing deep-learning based auto-segmentation in head and neck cancer: A study on real-world clinical cases. *Front. Oncol.* **2021**, *11*, 638197. [CrossRef]
17. Nikolov, S.; Blackwell, S.; Zverovitch, A.; Mendes, R.; Livne, M.; De Fauw, J.; Patel, Y.; Meyer, C.; Askham, H.; Romera-Paredes, B.; et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: Deep learning algorithm development and validation study. *J. Med. Internet Res.* **2021**, *23*, e26151. [CrossRef]
18. Kim, N.; Chun, J.; Chang, J.S.; Lee, C.G.; Keum, K.C.; Kim, J.S. Feasibility of continual deep learning-based segmentation for personalized adaptive radiation therapy in head and neck area. *Cancers* **2021**, *13*, 702. [CrossRef]
19. Zhou, H.; Li, Y.; Gu, Y.; Shen, Z.; Zhu, X.; Ge, Y. A deep learning based automatic segmentation approach for anatomical structures in intensity modulation radiotherapy. *Math. Biosci. Eng.* **2021**, *18*, 7506–7524. [CrossRef]
20. Iyer, A.; Thor, M.; Onochie, I.; Hesse, J.; Zakeri, K.; LoCastro, E.; Jiang, J.; Veeraraghavan, H.; Elguindi, S.; Lee, N.Y.; et al. Prospectively-validated deep learning model for segmenting swallowing and chewing structures in CT. *Phys. Med. Biol.* **2022**, *67*, 024001. [CrossRef]
21. Bilimagga, R.S.; Anchineyan, P.; Nmugam, M.S.; Thalluri, S.; Goud, P.S. Autodelineation of organ at risk in head and neck cancer radiotherapy using artificial intelligence. *J. Can. Res. Ther.* **2022**. [CrossRef]
22. Hong, T.S.; Tomé, W.A.; Harari, P.M. Heterogeneity in head and neck IMRT target design and clinical practice. *Radiother. Oncol.* **2012**, *103*, 92–98. [CrossRef] [PubMed]
23. Segedin, B.; Petric, P. Uncertainties in target volume delineation in radiotherapy—Are they relevant and what can we do about them? *Radiol. Oncol.* **2016**, *50*, 254–262. [CrossRef] [PubMed]
24. Multi-Institutional Target Delineation in Oncology Group. Human-computer interaction in radiotherapy target volume delineation: A prospective, multi-institutional comparison of user input devices. *J. Digit. Imaging* **2011**, *24*, 794–803. [CrossRef] [PubMed]
25. Kieselmann, J.P.; Kamerling, C.P.; Burgos, N.; Menten, M.J.; Fuller, C.D.; Nill, S.; Cardoso, M.J.; Oelfke, U. Geometric and dosimetric evaluations of atlas-based segmentation methods of MR images in the head and neck region. *Phys. Med. Biol.* **2018**, *63*, 145007. [CrossRef]
26. Jarrett, D.; Stride, E.; Vallis, K.; Gooding, M.J. Applications and limitations of machine learning in radiation oncology. *Br. J. Radiol.* **2019**, *92*, 20190001. [CrossRef]
27. Van Dijk, L.V.; Van den Bosch, L.; Aljabar, P.; Peressutti, D.; Both, S.; Steenbakkers, R.J.H.M.; Langendijk, J.A.; Gooding, M.J.; Brouwer, C.L. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiother. Oncol.* **2020**, *142*, 115–123. [CrossRef]
28. Brouwer, C.L.; Steenbakkers, R.J.H.M.; Bourhis, J.; Budach, W.; Grau, C.; Grégoire, V.; van Herk, M.; Lee, A.; Maingon, P.; Nuttingt, C.; 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. *Radiother. Oncol.* **2015**, *117*, 83–90. [CrossRef]
29. Machine Learning–Deep-Learning Segmentation in RayStation. Available online: https://www.raysearchlabs.com/495a00/siteassets/media/publications/white-papers/wp-pdfs/wp_ml_deeplearning_2020.03.25.pdf (accessed on 24 October 2022).
30. Delpon, G.; Escande, A.; Ruef, T.; Darréon, J.; Fontaine, J.; Noblet, C.; Supiot, S.; Lacornerie, T.; Pasquier, D. Comparison of automated atlas-based segmentation software for postoperative prostate cancer radiotherapy. *Front. Oncol.* **2016**, *6*, 178. [CrossRef]
31. Kariyawasam, L.N.; Ng, C.K.C.; Sun, Z.; Kealley, C.S. Use of three-dimensional printing in modelling an anatomical structure with a high computed tomography attenuation value: A feasibility study. *J. Med. Imaging Health Inform.* **2021**, *11*, 2149–2154. [CrossRef]