Clinical and Translational Radiation Oncology 41 (2023) 100635

Contents lists available at ScienceDirect

Clinical and Translational Radiation Oncology

journal homepage: www.sciencedirect.com/journal/clinical-and-translational-radiation-oncology

Young oncologists benefit more than experts from deep learning-based organs-at-risk contouring modeling in nasopharyngeal carcinoma radiotherapy: A multi-institution clinical study exploring working experience and institute group style factor

Ying Song [a,b], Junjie Hu [b], Qiang Wang [a,b], Chengrong Yu [b], Jiachong Su [a], Lin Chen [a], Xiaorui Jiang [c], Bo Chen [c], Lei Zhang [d], Qian Yu [d], Ping Li [a], Feng Wang [a], Sen Bai [a], Yong Luo [a,*], Zhang Yi [b,*]

[a] Cancer Center, West China Hospital, Sichuan University, No. 37 Guo Xue Alley, Chengdu 610065, PR China
[b] Machine Intelligence Laboratory, College of Computer Science, Sichuan University, No. 24, South Section 1 of the First Ring Road, Chengdu 610065, PR China
[c] Department of Oncology, First People's Hospital of Chengdu, No. 18, Wanxiang North Road, High-tech Zone, Chengdu 610041, PR China
[d] Department of Oncology, Second People's Hospital of Chengdu, Chengdu, PR China

ARTICLE INFO

Keywords:
Deep learning
Contouring
Organs-at-risk
Nasopharyngeal carcinoma radiotherapy
Working experience

ABSTRACT

Background: To comprehensively investigate the behaviors of oncologists with different working experiences and institute group styles in deep learning-based organs-at-risk (OAR) contouring.

Methods: A deep learning-based contouring system (DLCS) was modeled from 188 CT datasets of patients with nasopharyngeal carcinoma (NPC) in institute A. Three institute oncology groups, A, B, and C, were included; each contained a beginner and an expert. For each of the 28 OARs, two trials were performed with manual contouring first and post-DLCS edition later, for ten test cases. Contouring performance and group consistency were quantified by volumetric and surface Dice coefficients. A volume-based and a surface-based oncologist satisfaction rate (VOSR and SOSR) were defined to evaluate the oncologists' acceptance of DLCS.

Results: Based on DLCS, experience inconsistency was eliminated. Intra-institute consistency was eliminated for group C but still existed for group A and group B. Group C benefits most from DLCS with the highest number of improved OARs (8 for volumetric Dice and 10 for surface Dice), followed by group B. Beginners obtained more numbers of improved OARs than experts (7 v.s. 4 in volumetric Dice and 5 v.s. 4 in surface Dice). VOSR and SOSR varied for institute groups, but the rates of beginners were all significantly higher than those of experts for OARs with experience group significance. A remarkable positive linear relationship was found between VOSR and post-DLCS edition volumetric Dice with a coefficient of 0.78.

Conclusions: The DLCS was effective for various institutes and the beginners benefited more than the experts.

Introduction

Nasopharyngeal carcinoma (NPC) is a typical tumor in the head and neck [1], which is widely observed in southeast Asia and northern Africa, especially in southern China [2]. The high incidence and mortality worldwide of NPC demand enhancing control and prevention [3]. As the primary treatment modality for NPC, radiotherapy plays an important role [4–6]. NPC radiotherapy requires accurate contouring of organs at risk (OAR) for dose sparing. However, manual OAR contouring for NPC is a time-consuming and tedious procedure that heavily depends on personal experience. Significant observer variance and low group consistency have been reported [7,8].

Automatic contouring is a promising solution for improving clinical contouring efficiency and interobserver consistency [7,8]. There were traditional solutions based on image deformation and registration [9,10] with some unsolved problems, namely, no golden rules for atlas

Abbreviations: OAR, Organs-at-risk; NPC, Nasopharyngeal carcinoma; DLCS, Deep learning-based contouring system; VOSR, Volume-based oncologist satisfaction rate; SOSR, Surface-based oncologist satisfaction rate.

* Corresponding authors.
E-mail addresses: luyo119@126.com (Y. Luo), zhangyi@scu.edu.cn (Z. Yi).

https://doi.org/10.1016/j.ctro.2023.100635
Received 27 December 2022; Received in revised form 26 April 2023; Accepted 1 May 2023
Available online 8 May 2023

establishment [11], the heavy requirement for computational resources, and long computing time [12].

Deep learning methods provide solutions to automatic radiotherapy contouring from a new point of view. For NPC OAR contouring, there were algorithm design studies for deep learning-based contouring system (DLCS) performance improvement [13–25] and clinical studies for DLCS performance evaluation [26,27]. The DLCS performance ranged from 0.45 to 0.94 for Dice with various models [28]. However, the evaluation based on numerical geometric metrics does not fully reveal the clinical acceptance of the contouring target for different oncologists, and the clinical evaluation research was limited. Deep learning methods and registration algorithm contouring performance were evaluated for swallowing-related organs [27] and 11 organs of the head and neck site [26] by quantitative parameters, drawing the main conclusion that DLCS provided equivalent or better segmentation accuracy and higher efficiency. Besides contouring time reduction and segmentation accuracy, complicated clinical issues remained in DLCS application, e.g., inter- and intra-observer consistency variation, group consistency variation, oncologist performance improvement, and their acceptability, especially for oncologists with diverse experience and institute group contouring styles. These problems pose challenges to the DLCS's clinical effectiveness. Some research investigated this topic providing proof of lower inter-observer variability and improved consistency for the larynx contouring course [31] and atlas-based NPC OAR edition [29]. Some research purely focuses on manual contouring analysis of the prostate site [30] and multiple lesions and organs [31] with no intervening factors. For the aspect of different expertise, no statistical difference was found for a single institute [31]. However, manual contouring is a task with high variability both within and across institutions requiring end-to-end tests [32]. In clinical practice, DLCS is usually established by the dataset from one relatively large institute restricted by the data quality and patient resources but is used by various authorized oncologists from other institutes. In such circumstances, the DLCS-based contouring performance of oncologists with varied working experiences from multi-institutes is unknown.

Therefore, it is critical to evaluate the oncologist's acceptability towards DLCS, their performance during the contouring edition, and their improvement in clinical practice. In this study, we focus on oncologist behaviors from three institutions marked by A, B, and C, with diverse working experience and institute group contouring styles. We constructed a DLCS for NPC OAR contouring based on CT from one institute. We investigated oncologists' consistency, performance improvement, and acceptability based on two factors: working experiences and institute group styles.

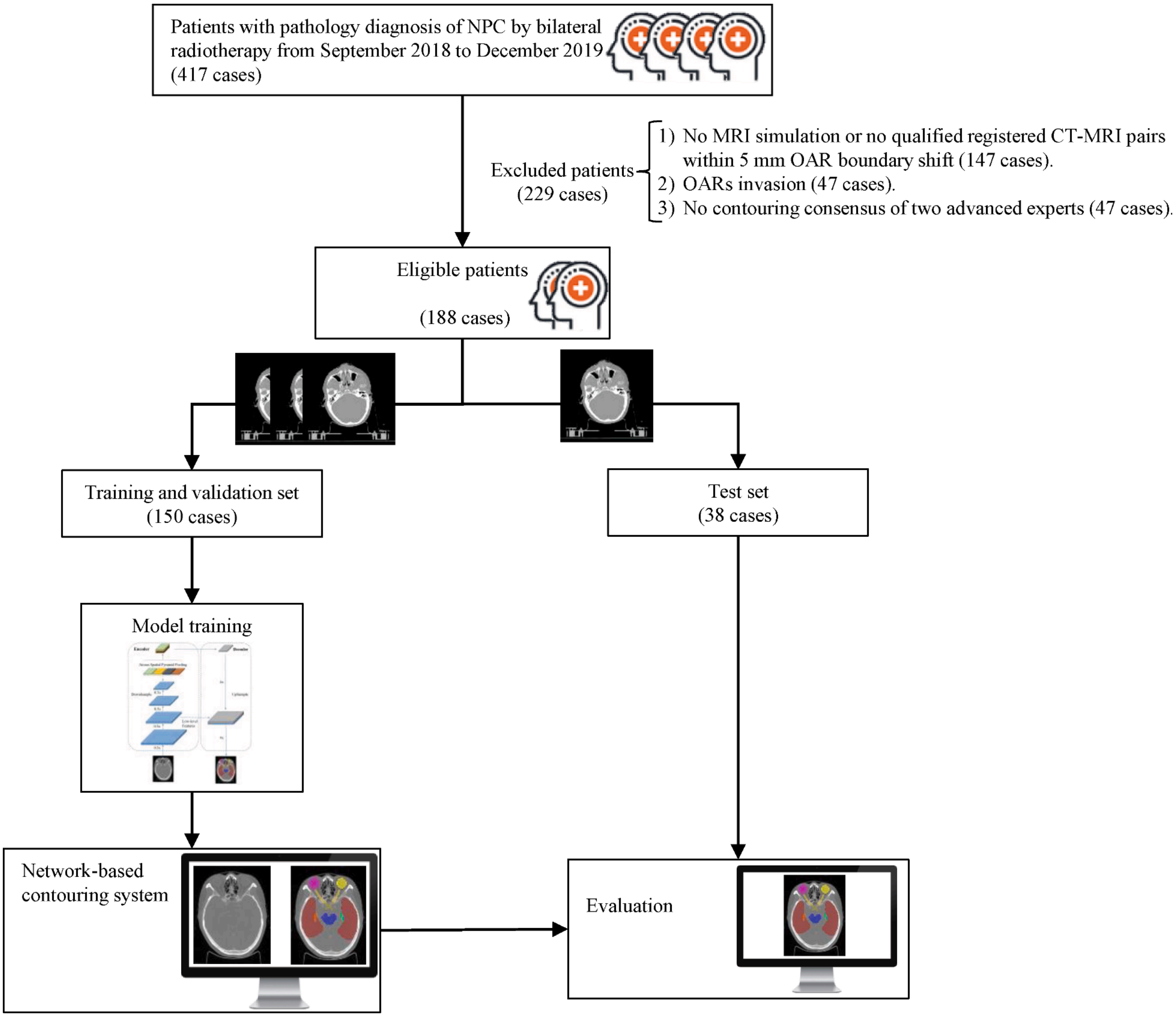
Patients with pathology diagnosis of NPC by bilateral radiotherapy from September 2018 to December 2019 (417 cases)
Excluded patients (229 cases)
1) No MRI simulation or no qualified registered CT-MRI pairs within 5 mm OAR boundary shift (147 cases).
2) OARs invasion (47 cases).
3) No contouring consensus of two advanced experts (47 cases).
Eligible patients (188 cases)
Training and validation set (150 cases)
Test set (38 cases)
Model training
Network-based contouring system
Evaluation

Fig. 1. Study flow diagram of this research.

Material and methods

2.1. Dataset

The study was approved by the ethics committee as a retrospective clinical trial with patient informed consent exempt. All the patients with the pathology diagnosis of nasopharyngeal carcinoma treated by bilateral radiotherapy from September 2018 to December 2019 were collected in the institute of group A, and 188 patients were finally included. The flow diagram of this study is shown in Fig. 1. All the patients were immobilized in the supine position by specific head-and-neck thermoplastic masks with planning CT scanning and reconstruction done by 3 mm thickness and 0.94 × 0.94 cm^2 in-plane resolution on a spiral CT (SOMATOM Definition AS+, Siemens). MRI simulation was also done for all the patients at the same immobilization position with the masks (Discovery 750 W, GE).

2.2. OAR delineation

In our clinical routine, manual contouring was done on planning CT combined with the registered MRI images by a qualified oncologist with 28 OARs. The contoured datasets are then viewed, adjusted, and approved by the corresponding responsible advanced oncologist for all patients. To avoid the oncologist personality influence, all contouring results were adjusted and approved by consensus of two advanced experts according to the anatomical boundaries in registered CT and MRI images by the international guideline [33]. The approved contours were considered as the ground truth and abbreviated as OAR_{GT}, later used for network training and the golden reference for result evaluation. The patients' characteristics are shown in Table 1.

2.3. Networks

We construct the DLCS based on an encoder-and-decoder U-Net architecture [34–36], DeepLabv3+ [37], for NPC contouring. The architecture of DeepLabv3+ is illustrated in supplemental file Fig. 1. In application, the ResNet50 network serves as the backbone for the encoder component, extracting the abstract features of multiple layers by downsampling. The decoder component mainly consists of four atrous spatial pyramid pooling modules with different rates, combined with global average pooling and other up-sampling operations. Moreover, low-level features from the encoder are re-fed to the decoder for aggregating more abstract features.

For the network training, the cross-entropy criterion was employed as the loss function, and the maximum number of the epoch was set to a constant of 100. Stochastic gradient descent was used as an optimizer, and its parameters were selected according to suggested values, with initial learning rate, momentum, and weight decay values of 0.01, 0.9, and $5*10^{-4}$, respectively. The DLCS was set on a server equipped with a Linux operation system, RAM hardware of 64 G, CPU Intel Xeon CPU E5-2620 v3 @ 2.40 GHz with 24 kernels, and NVIDIA TITAN RTX GPUs with 24G memory. 188 patients were randomly assigned to training and validation set (150 cases), and test set (38 cases). The model code can be accessed at https://github.com/hujunjiescu/DeepRT.

Table 1
Patient characteristics.

Characteristic		Entire Cohort (n = 188)	deep learning-based contouring system		
			Training-Validation (n = 150)	Test (n = 38)	P
Sex	Male	126	100	26	0.837
	Female	62	50	12	
Age	Median (IQR)	48(19–84)	48 (19–84)	45 (24–64)	
	Range	27–80	27–80	27–76	
	≤40 y	43	32	11	0.626
	40–60 y	123	100	23	
	> 60	22	18	4	
Stage	T1	33	29	4	0.563
	T2	62	49	13	
	T3	48	36	12	
	T4	45	36	9	
	N0	10	9	1	0.860
	N1	44	34	10	
	N2	104	82	22	
	N3	30	25	5	
	M0	186	148	38	1.0
	M1	2	2	0	

Note. IQR = Inter-Quartile Range.

2.4. Evaluation

Six oncologists majoring in NPC from three institutions were recruited for the study. There were three experts (over 10 years of experience labeled by O_{A10}, O_{B10}, and O_{C10}, respectively) and three beginners (1 years of experience marked by O_{A1}, O_{B1}, and O_{C1}). Two trials, independent contouring labeled as OAR_{manual}, and post-DLCS edition labeled as $OAR_{p\text{-}DLCS}$, of 28 OARs were performed for ten randomly-selected test cases on the Pinnacle3 treatment planning system (Philips Healthcare, Hamburg, Germany, version 9.1). Six oncologists were trained according to the international guideline [33], and the contouring consensus were built in advance. All the evaluations were done randomly and double-blindly, and two trials were carried out over two sessions by two months apart.

Contouring performance was evaluated by two quantitative metrics, including volumetric Dice [30,38,39] for volume overlapping evaluation and surface Dice [30,39,40] for border overlapping evaluation with a distance tolerance of 1 mm. To evaluate the subjective acceptability of OAR_{DLCS}, an objective volume-based and a surface-based oncologist satisfaction rate (VOSR and SOSR) were defined by the OAR_{DLCS} and the corresponding $OAR_{p\text{-}DLCS}$ to objectively measure the volume and surface deviation of DLCS contouring to oncologists' desired targets. The definition details are shown in Fig. 2. When no modification was needed for one oncologist, the corresponding satisfaction rate reached the maximum of 1. This way, subjective acceptability was evaluated based on the objective metric derived from the whole edition process.

Based on OAR_{GT} as a reference, volumetric and surface Dice were calculated to evaluate oncologist performance and group consistency. Based on $OAR_{p\text{-}DLCS}$ as a reference, VOSR and SOSR were calculated. To explore the oncologist acceptability impact on their post-DLCS edition performance, the correlation relationship between the satisfaction rate and post-DLCS edition performance was investigated. The Pearson linear regression method was adopted for the Dice metric, and the Spearman rank method was employed for ranked group data. Analysis of variance (ANOVA) was used for pair-wise and inter-group comparison (IBM SPSS, version 25.0, New York, NY, USA, $P < 0.05$). All P-values were from two-sided tests.

Results

3.1. Analysis of manual contouring performance and post-DLCS edition

The performance of each oncologist and experience-based intra-institute consistency is illustrated in Fig. 3. Post-DLCS edition improved intra-institute consistency significantly for most OARs except the left optical nerve, the pituitary gland, temporal lobes, and the left mandible. The intra-institute inconsistency was eliminated for group C but still existed for groups A and B. To objectively investigate the DLCS-based improvement of individual contouring performance, the post-DLCS edition improvement, defined as the post-DLCS edition metric minus the metric of manual contouring, was listed in Table 2. For those OARs with significant performance differences between manual contouring

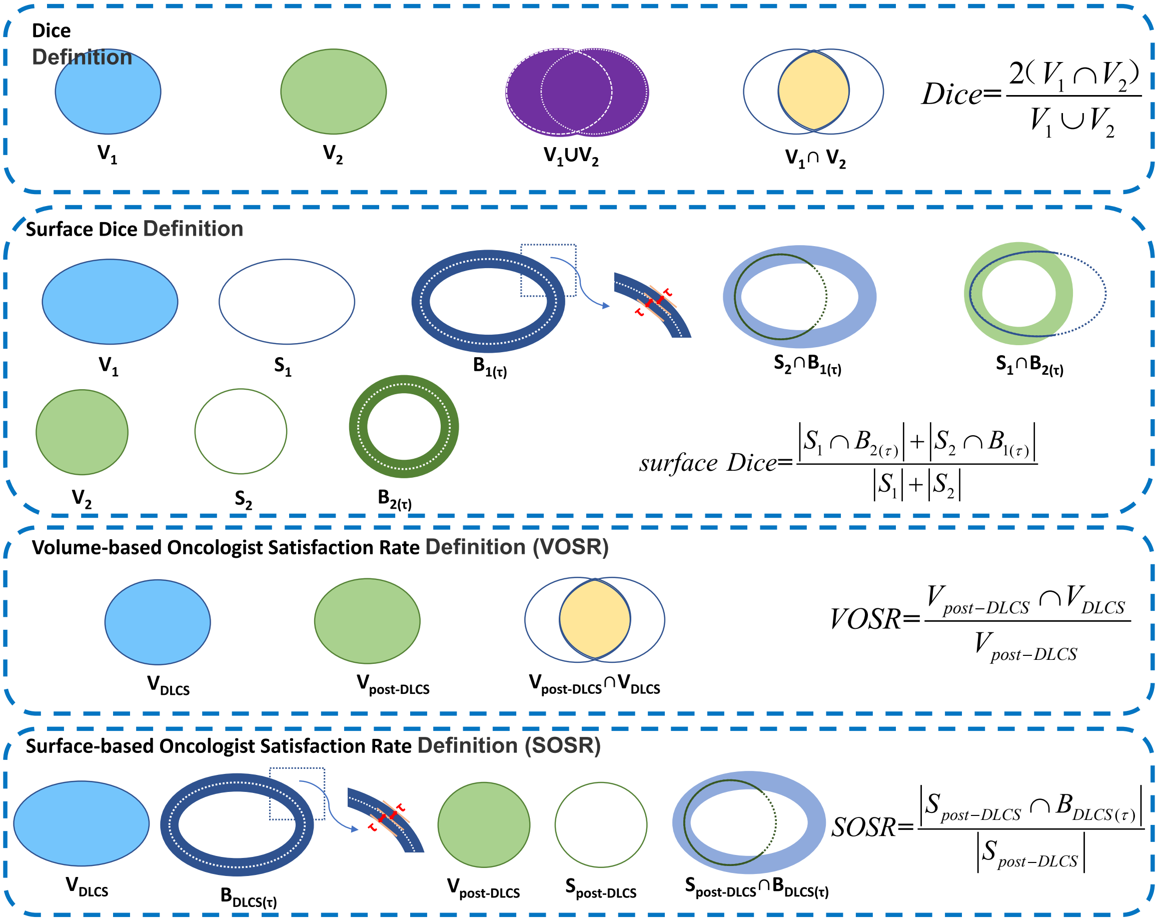
Dice Definition
V_1
V_2
$V_1 \cup V_2$
$V_1 \cap V_2$
$Dice = \frac{2(V_1 \cap V_2)}{V_1 \cup V_2}$
Surface Dice Definition
V_1
S_1
$B_{1(\tau)}$
τ
$S_2 \cap B_{1(\tau)}$
$S_1 \cap B_{2(\tau)}$
V_2
S_2
$B_{2(\tau)}$
$surface\ Dice = \frac{|S_1 \cap B_{2(\tau)}| + |S_2 \cap B_{1(\tau)}|}{|S_1| + |S_2|}$
Volume-based Oncologist Satisfaction Rate Definition (VOSR)
V_{DLCS}
$V_{post\text{-}DLCS}$
$V_{post\text{-}DLCS} \cap V_{DLCS}$
$VOSR = \frac{V_{post-DLCS} \cap V_{DLCS}}{V_{post-DLCS}}$
Surface-based Oncologist Satisfaction Rate Definition (SOSR)
V_{DLCS}
$B_{DLCS(\tau)}$
τ
$V_{post\text{-}DLCS}$
$S_{post\text{-}DLCS}$
$S_{post\text{-}DLCS} \cap B_{DLCS(\tau)}$
$SOSR = \frac{|S_{post-DLCS} \cap B_{DLCS(\tau)}|}{|S_{post-DLCS}|}$

Fig. 2. The illustration of the volumetric Dice, surface Dice, volume-based oncologist satisfaction rate, and surface-based oncologist satisfaction rate. V_1 and V_2 are the volumes of two targets, respectively. $V_1 \cup V_2$ is the union of the volumes of targets, and $V_1 \cap V_2$ is the intersection volume of targets. S_1 and S $_2$ are the surfaces of the two targets, respectively. $B_{1(\tau)}$ and $B_{2(\tau)}$ are borders extended by S_1 and S $_2$ with tolerance τ, respectively. $S_1 \cap B_{2(\tau)}$ and $S_2 \cap B_{1(\tau)}$ are intersection surfaces represented by the borders' solid lines. The operator |.| in the definition equation is the pixel summation of the corresponding surface. V_{DLCS} and $V_{post\text{-}DLCS}$ are the volumes of DLCS contouring and post-DLCS edition by oncologists, respectively. $V_{DLCS} \cap V_{post\text{-}DLCS}$ is the intersection volume. $S_{post\text{-}DLCS}$ is the surface of the post-DLCS edition by oncologists. $B_{DLCS(\tau)}$ is a border extended by the surface of V_{DLCS} with tolerance τ. $S_{post\text{-}DLCS} \cap B_{DLCS(\tau)}$ are intersection surfaces represented by the borders' solid lines.

and post-DLCS edition, all OARs obtained performance improvement without any metric deterioration for all oncologists. The DLCS-based edition performance for eyes, spinal cord, constrictor naris, and larynx was improved for all oncologists. However, no significant improvement was found for the left lens, optical nerves, and the left cochlea. For the left OARs, the performance improvement varied for different institute groups. There were more numbers of OARs with improved performance for beginners compared with experts, 13 *v.s.* 8 (volumetric Dice metric) and 10 *v.s.* 5 (surface Dice metric) for group A, and 21 *v.s.* 14 (volumetric Dice metric) and 15 *v.s.* 14 (surface Dice metric) for group B. However, this trend reversed for group C with more numbers of OARs improved for experts. Thus, we explored the two-way post-DLCS edition performance improvement analysis illustrated in Fig. 4.

3.2. Post-DLCS edition performance improvement analysis

For those OARs with performance improvement significance in Fig. 4, all OARs gained performance improvement for all institute and experienced groups. However, among OARs with experience factor significance or interaction impact, beginners obtained more performance improvement with 7 OARs in volumetric Dice and 5 OARs for surface Dice than experts with only 4 OARs for both volumetric Dice and surface Dice), indicating that beginners benefit more from the post-DLCS edition. Among OARs with institute factor significance, group C benefits most in post-DLCS performance improvement with the highest number of improved OARs (8 for the volumetric Dice and 10 for the surface Dice), followed by group B (4 for both volumetric Dice and surface Dice).

Typical two-dimensional slices and three-dimensional visualization for contouring and edition are shown in Fig. 5, demonstrated on MIM Maestro (Version 7.1, MIM Software Inc., Cleveland, OH). For ultimately the same contours, MIM only showed one color to represent the contour.

3.3. Working experience and institute contouring style analysis based on manual contouring and post-DLCS edition

Two-way variance analysis of performance metrics for working experience and institute contouring styles is shown in Fig. 6. For manual contouring, there were fourteen OARs suffering from inter-institute inconsistency, seven OARs for experience inconsistency, and ten OARs influenced by their interaction effect, qualified by both volumetric Dice and surface Dice. But the number was reduced to one (the right temporal lobe) for the institute contouring style factor and zeros for the working experience factor by post-DLCS edition, indicating that inter-institute and inter-experience consistency could be improved by DLCS. The consistency improvement also can be seen in Fig. 5 (f) and (l). The three-dimensional visualization showed clear consistency improvement with

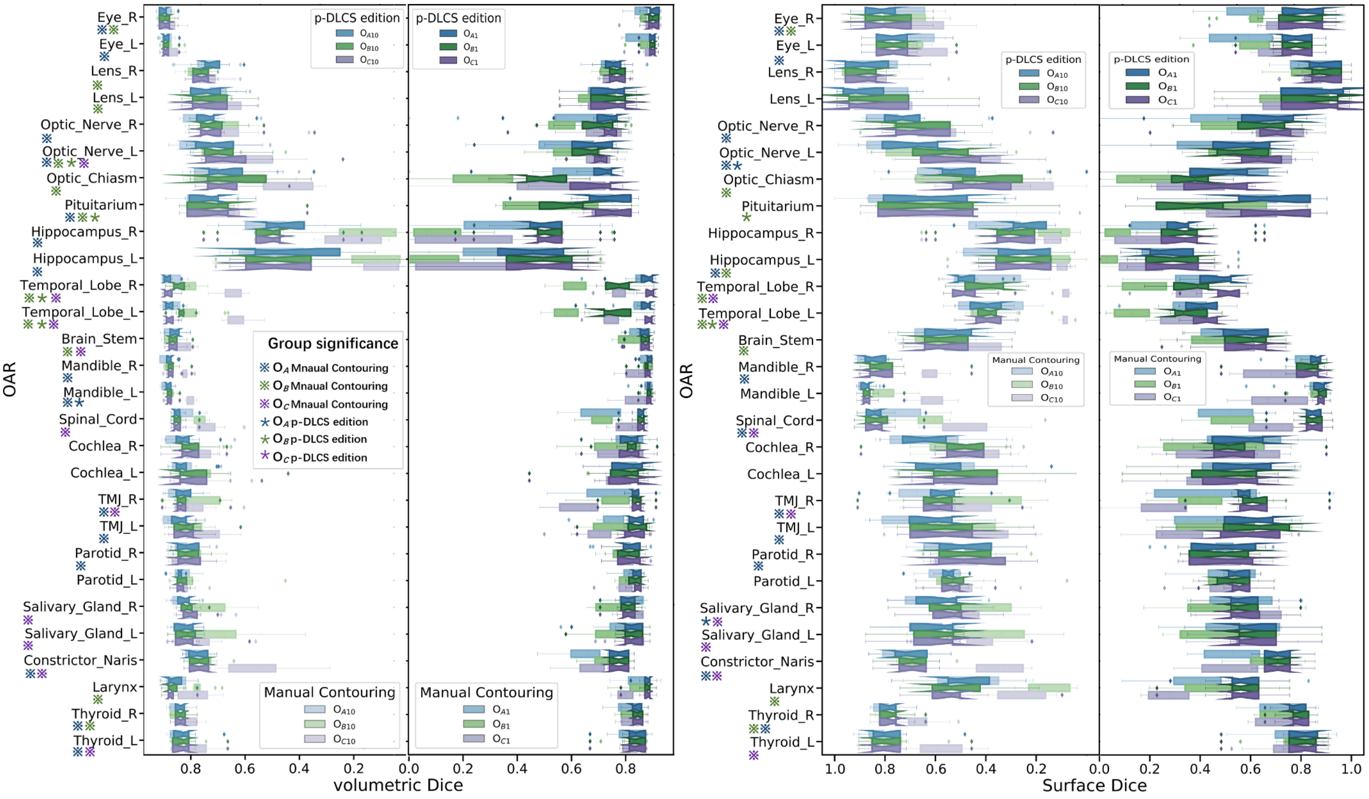

Fig. 3. Individual performance and experience-based intra-institute consistency for manual contouring and post-DLCS edition based on OAR_{GT}. Blue, green, and purple markers ※ (for manual contouring) and * (for p-DLCS edition) beside each OAR label indicate statistical significance between the beginner and the expert for groups A, B, and C, respectively. O_{A10}, O_{B10}, and O_{C10} represent three experts from three institutes, respectively. O_{A1}, O_{B1}, and O_{C1} represent three beginners from three institutes, respectively. O_A, O_B, and O_C represent groups of oncologists from three institutes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

less rendered colors. More analysis details can be found in the supplemental file table 1–28. For all those OARs except larynx with experience factor significance, the expert manual contouring and post-DLCS edition performed better than beginners both by volumetric Dice and surface Dice, indicating that working experience had a significant effect on clinical performance. For 4 OARs with institute contouring style significance in the post-DLCS edition, oncologists from group C performed slightly better than other institute groups.

3.4. *Volume-based and surface-based oncologist satisfaction analysis*

Volume-based and surface-based oncologist satisfaction rates for different groups were listed in Table 29 in the supplemental file. Most OARs got a satisfaction rate above 0.9, and some OARs got a maximum satisfaction rate of 1 with a deviation of 0, indicating that there was no manual edition for the OARs. The acceptability varied among institute groups. Pituitary gained the lowest VOSR, 0.82 from oncologists of institute A but got 0.97 and 0.96 from oncologists of institutes B and C, respectively. The same was found for optical nerves, with the lowest VOSR from institute B but 0.9 above the satisfaction rate from other oncologist groups. For experienced groups, the beginners had better acceptance of DLCS with a significantly higher satisfaction rate than the experts. A similar situation happened for SOSR. The correlation analysis between oncologist satisfaction rate and oncologist performance is listed in Table 3. All the prominence detection was under 0.05. There was a remarkable positive linear relationship between VOSR and post-DLCS edition volumetric Dice with coefficient of 0.78, and between VOSR and volumetric Dice improvement with coefficient of 0.46. This proves that oncologists' higher DLCS acceptability produced better contouring performance in clinical practice. Supplemental file Fig. 2 shows the scatter plot of the data for each group. The linear fitting line gradients of the beginners were higher than that of experts except for VOSR versus volumetric Dice improvement for group B. This also indicated that the beginners are more affected by DLCS than the experts. This was a consistent founding derived from Table 29 in the supplemental file.

3.5. *Time consuming analysis*

While the post-DLCS edition helps shorten contouring time, we focused institute contouring style and working experience impact difference. Thus, a two-way variance analysis for saving time between institute groups and working experienced groups was observed, and the result is listed in Table 4. More details of actual manual contouring time and post-DLCS edition time were listed in Table 30 and Table 31 in the supplemental file. Institute differences and the interaction impacted the saving time. However, working experience had no significant impact. It revealed that the DLCS contributes to time reduction with no difference in working experience. However, time reduction varied among institute groups, showing that institute contouring styles greatly influenced clinical applications and affected the edition process.

Discussion

In this study, we comprehensively explored the clinical impact of a deep-learning contouring system established in one institute on oncologists from three institutions with various working experience and contouring styles in nasopharyngeal carcinoma organs at risk contouring practice. Individual performance, group performance, experience-based intra-group consistency, inter-institute, and inter-experience consistency were improved, but the development varied for different

Table 2
The average improvement of post-DLCS edition performance for each oncologist.

	Volumetric Dice development gap/ Surface Dice development gap ($\tau = 1\ mm$)					
	O_{A10}	O_{A1}	O_{B10}	O_{B1}	O_{C10}	O_{C1}
Lens_L	0.07/0.08	0.07/0.04	0.01/-0.03	0.08/0.10	0.07/0.06	0.07/0.11
Lens_R	0.01/0.04	**0.05/** 0.05	0.01/0.00	**0.06/** 0.07	**0.07/** 0.09	**0.06/ 0.08**
Eye_L	**0.03/ 0.13**	**0.06/ 0.24**	**0.03/ 0.11**	**0.05/ 0.18**	**0.04/ 0.15**	**0.03/ 0.11**
Eye_R	**0.03/ 0.12**	**0.06/ 0.25**	**0.02/** 0.09	**0.05/ 0.19**	**0.04/ 0.18**	**0.04/ 0.13**
Optic_Nerve_R	0.00/-0.01	0.18/0.20	0.07/0.03	0.12/0.12	0.07/0.11	0.05/-0.01
Optic_Nerve_L	−0.04/-0.05	0.16/0.17	0.02/-0.04	0.08/0.05	0.09/0.09	0.03/0.02
Optic_Chiasm	0.00/-0.01	0.11/0.04	0.01/-0.03	**0.28/ 0.25**	**0.29/ 0.24**	0.19/0.15
Pituitarium	−0.01/-0.08	0.03/0.06	0.00/0.02	0.08/0.00	0.06/0.15	**0.07/ 0.18**
Hippocampus_R	0.00/0.01	0.20/0.16	**0.35/ 0.26**	**0.40/ 0.31**	**0.30/ 0.25**	**0.33/ 0.24**
Hippocampus_L	−0.04/0.00	0.11/0.07	**0.32/ 0.22**	**0.38/ 0.28**	**0.37/ 0.23**	**0.26/** 0.16
Temporal_Lobe_L	**0.03/** 0.12	0.05/ **0.14**	**0.07/ 0.10**	**0.21/ 0.31**	**0.25/ 0.39**	**0.16/ 0.18**
Temporal_Lobe_R	**0.04/ 0.17**	0.06/ **0.17**	**0.07/ 0.09**	**0.19/ 0.28**	**0.24/ 0.44**	**0.13/ 0.21**
Brain_Stem	0.01/0.01	0.03/0.08	0.01/0.01	**0.06/** 0.18	**0.06/ 0.17**	0.02/0.06
Spinal_Cord	**0.05/ 0.12**	**0.17/ 0.36**	**0.13/ 0.26**	**0.16/ 0.35**	**0.13/ 0.38**	**0.07/ 0.17**
Cochlea_R	0.01/0.03	0.02/0.03	0.04/0.06	**0.08/** 0.13	**0.06/ 0.13**	0.05/0.10
Cochlea_L	0.03/0.06	0.03/0.03	0.04/0.11	0.01/0.05	0.05/0.10	0.02/0.12
Mandible_L	0.01/0.00	**0.03/** 0.06	**0.03/ 0.08**	**0.02/** 0.03	**0.09/ 0.25**	**0.07/ 0.18**
Mandible_R	0.00/-0.02	**0.03/** 0.06	0.01/-0.01	0.01/-0.01	**0.06/ 0.18**	**0.06/** 0.14
TMJ_R	0.02/0.03	**0.12/ 0.20**	**0.08/ 0.21**	**0.11/ 0.24**	**0.07/ 0.17**	**0.22/ 0.39**
TMJ_L	0.00/-0.04	**0.11/ 0.23**	0.04/0.16	**0.10/** 0.20	**0.10/ 0.22**	**0.13/ 0.28**
Parotid_R	0.01/-0.02	**0.04/** 0.07	**0.04/ 0.12**	**0.05/ 0.13**	0.02/0.05	0.02/0.06
Parotid_L	**0.04/ 0.07**	**0.05/ 0.09**	0.07/0.10	**0.05/ 0.10**	**0.05/ 0.13**	**0.07/ 0.14**
Salivary_Gland_R	0.01/0.01	0.01/0.03	**0.11/ 0.24**	0.07/0.13	0.05/ **0.14**	0.00/0.01
Salivary_Gland_L	0.01/0.04	0.05/0.10	**0.15/ 0.27**	**0.09/ 0.19**	**0.08/ 0.21**	0.02/ **0.06**
Constrictor_Naris	**0.03/** 0.01	**0.16/ 0.21**	**0.07/ 0.07**	**0.10/ 0.08**	**0.25/ 0.39**	**0.14/ 0.23**
Larynx	**0.03/** 0.11	**0.05/ 0.20**	**0.12/ 0.44**	**0.05/ 0.16**	**0.10/ 0.32**	**0.09/ 0.31**
Thyroid_L	0.00/-0.01	0.03/0.05	0.01/ **0.05**	0.02/0.04	**0.07/ 0.23**	0.03/0.09
Thyroid_R	0.01/0.00	**0.05/** 0.09	0.01/0.01	**0.05/ 0.10**	**0.05/ 0.14**	**0.04/** 0.09

Note. TMJ, TemporoMandibular Joint. R, right. L, left. The bold block indicates there was a significant difference for the corresponding OAR and oncologist between manual contouring and post-DLCS edition metrics ($p < 0.05$). O_{A10}, O_{B10}, and O_{C10} represent three experts from 3 institutes, respectively. O_{A1}, O_{B1}, and O_{C1} represent three beginners from 3 institutes, respectively.

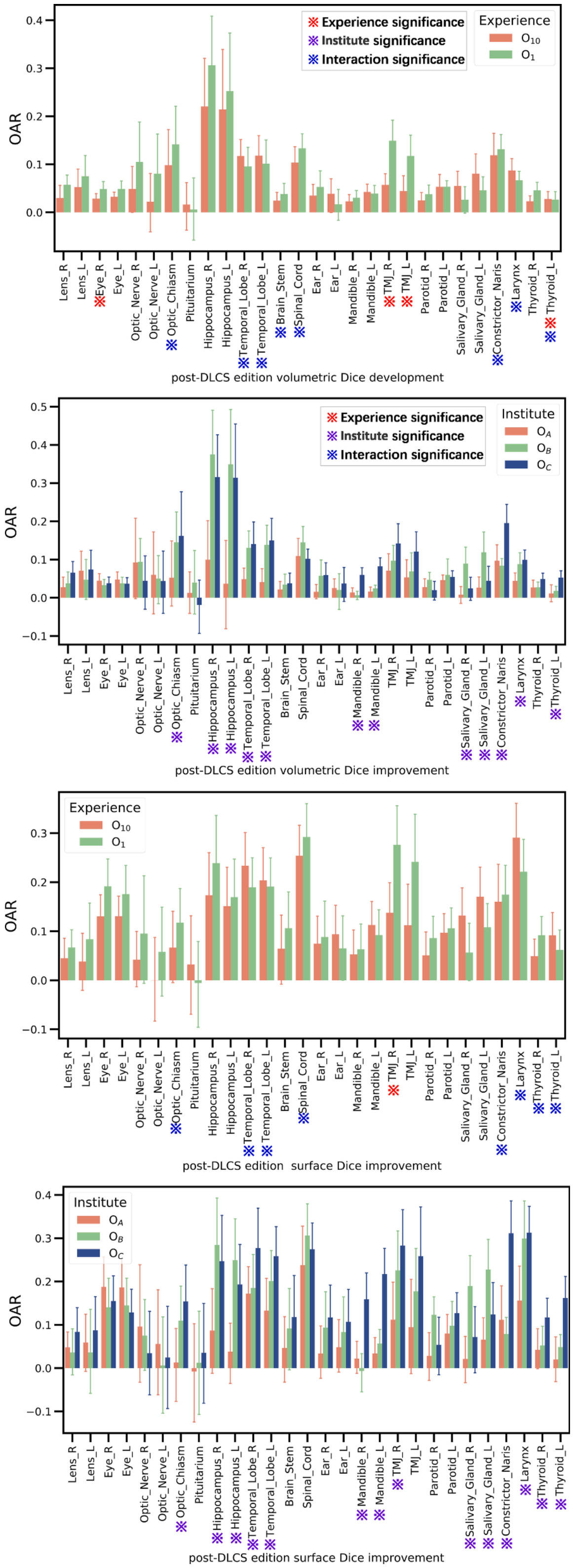

Fig. 4. Two-way post-DLCS edition performance improvement analysis by two metrics. Red, blue and purple markers ※ behind each OAR label indicate statistical significance for the institute-contouring style factor, working experience factor, and their interaction, respectively. O_A, O_B, O_C represent groups of oncologists from three institutes, respectively. O_{10} and O_1 represent experts and beginners from three institutes, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

institute groups. The DLCS established in one institute was also effective for oncologists from other institutes, with slightly better post-DLCS performance and greater improvement. All the analyses supported that beginners had better acceptance of DLCS and gained more performance improvement than experts.

We gave solid proof of DLCS impact on oncologists with different experiences and contouring styles from various institutes. However, the

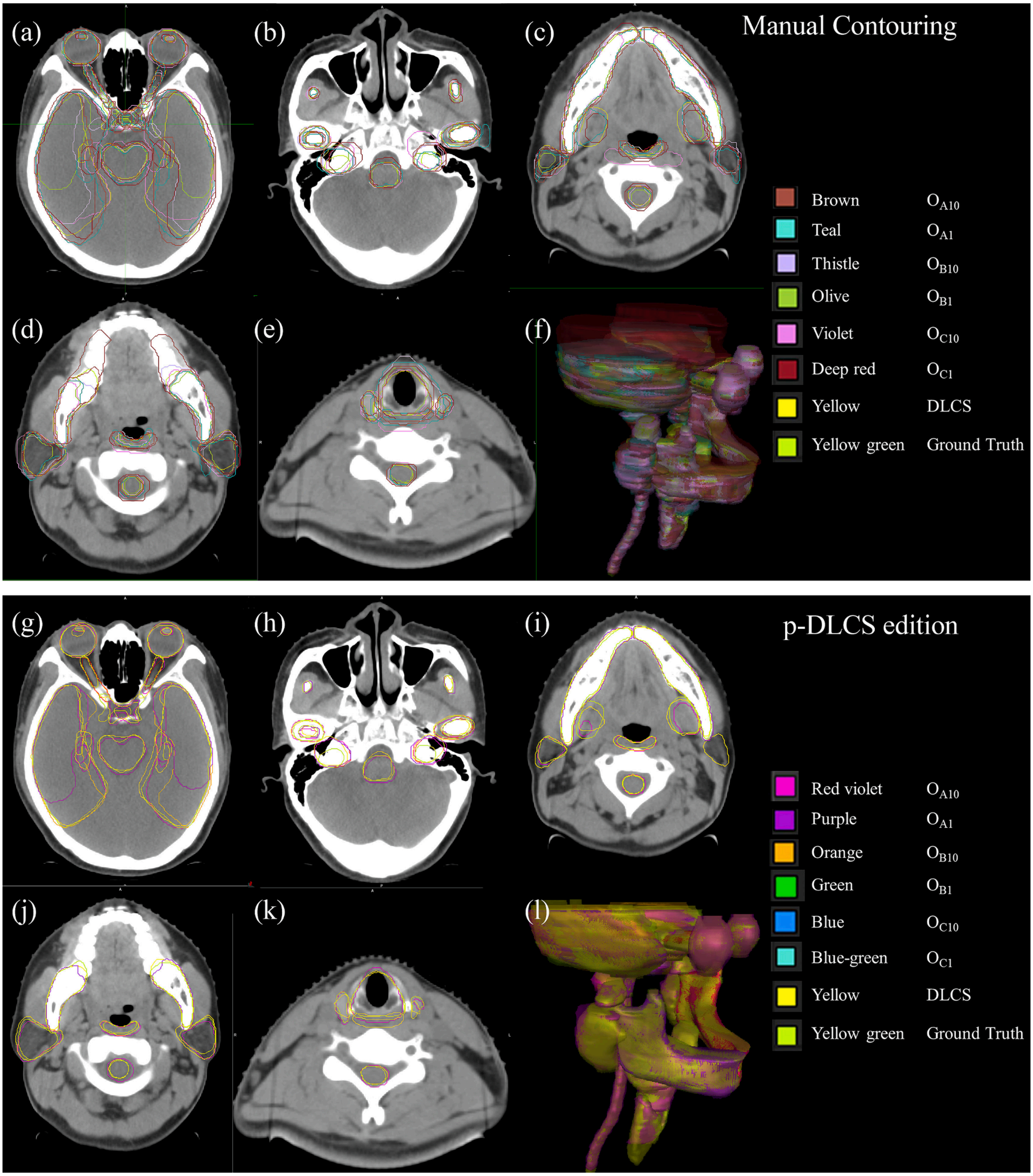

Fig. 5. Manual contouring and post-DLCS edition performance on CT of a random-selected case, created by MIM Maestro. (a)-(e), typical manual contours of each oncologist. (f) three-dimensional illustration of all manual contours. (g)-(k), corresponding edit contours of each oncologist. (l) three-dimensional illustration of all edit contours. O_{A10}, O_{B10}, O_{C10} represent three experts from 3 institutes, respectively. O_{A1}, O_{B1}, O_{C1} represent three beginners from 3 institutes, respectively.

oncologist number was limited due to the intensive manual labor for the experiments, which might cause bias due to the oncologist's personal working style. In the future, more oncologists may be invited to enhance the experiment reliability. Interestingly the DLCS failed to develop the consistency and edition performance for all OARs. Experience-based intra-group consistency improvement failed for the left optical nerve, the pituitary gland, temporal lobes, and the left mandible. The individual performance improvement failed for the left lens, optical nerves, and the left cochlea, revealing that DLCS may be not practical for every oncologist's contouring target. The possible cause may be that the DLCS was trained by CT slices, and OAR_{DLCS} was tested on CT slices. However, soft tissues, like the OARs mentioned above, had no clear boundaries on CT slices, which may cause low prediction accuracy and then lead to considerable modifications and low observer consistency. MRI can provide clear boundaries for soft tissues to solve the problem, and we are working on contouring prediction systems based on registered CT and MRI. We believe it will be a better automatic contouring framework.

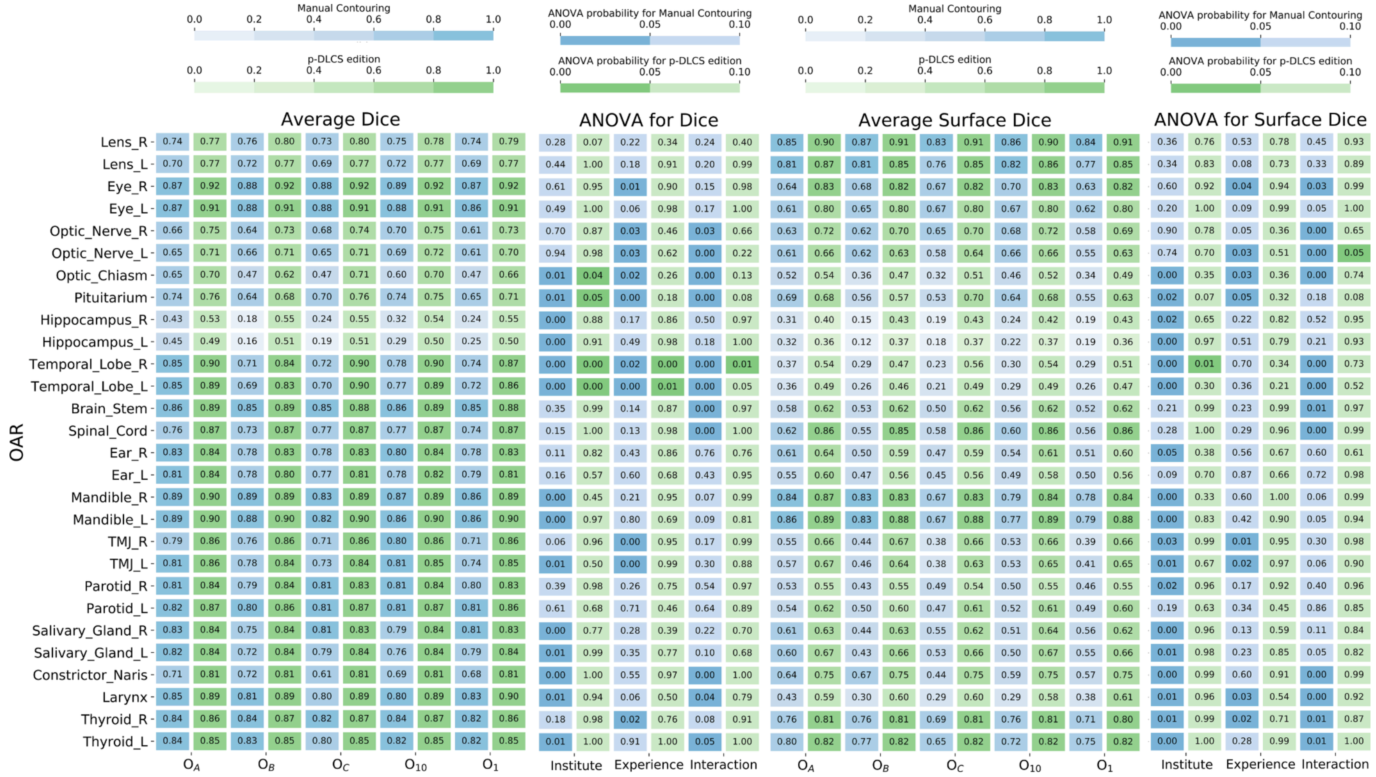

Fig. 6. Intra-institute and intra-experience consistency for manual contouring and post-DLCS edition by two-way variance performance analysis. Manual contouring and post-DLCS edition are marked by blue and green colors, respectively.). Darker blocks in the ANOVA table indicate statistical significance. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Correlation analysis between oncologist satisfaction rate and oncologist performance.

	VOSR	SOSR
post-DLCS edition volumetric Dice	0.78 ($P < 0.05$)	
post-DLCS edition surface Dice		0.09 ($P < 0.05$)
volumetric Dice development	0.46 ($P < 0.05$)	
Surface Dice development		−0.06 ($P < 0.05$)
Group	0.41 ($P < 0.05$)	0.40 ($P < 0.05$)
Experience	−0.15 ($P < 0.05$)	−0.14 ($P < 0.05$)

Note: VOSR, volume-based oncologist satisfaction rate. SOSR, surface-based oncologist satisfaction rate. DLCS, deep-learning contouring system. DLCS, deep-learning contouring system.

Table 4
Two-way variance analysis for saving time between groups and working experience (minutes).

Saving time (minutes)	O_{10}	O_1	P
O_A	17.9 ± 6.99	46.7 ± 8.26	**<0.01**
O_B	49.7 ± 8.79	30.3 ± 9.09	
O_C	28.2 ± 3.21	23.6 ± 2.55	
P	0.37		**<0.01**

Note: O_{10} and O_1 represent the experts and the beginners, respectively. O_A, O_B, and O_C represent oncologists from three institutes, respectively.

Conclusions

Individual performance, group performance, experience-based intra-group consistency, inter-institute and inter-experience consistency were improved by deep-learning organs at risk contouring system for nasopharyngeal carcinoma radiotherapy. Still, the improvement varied for different institute group. The DLCS established in one institute was also effective for oncologists from other institutes with slightly better post-DLCS performance and higher improvement. The beginners had better acceptance of DLCS and gained more performance improvement than experts, which could positively impact toxicities avoidance and complication reductions in clinical practice.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research is supported by the National Natural Science Foundation of China Grants (No. 81803056, No. 62106162, No. 81972848, and No. 2018AAA0100201).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ctro.2023.100635.

References

[1] Mohammed MA, Abd Ghani MK, Hamed RI, Ibrahim DA. Review on Nasopharyngeal Carcinoma: concepts, methods of analysis, segmentation, classification, prediction and impact: a review of the research literature. J Comput Sci 2017;21:283–98. https://doi.org/10.1016/j.jocs.2017.03.021.

[2] Tang L-L, Chen W-Q, Xue W-Q, He Y-Q, Zheng R-S, Zeng Y-X, et al. Global trends in incidence and mortality of nasopharyngeal carcinoma. Cancer Lett 2016;374(1): 22–30.

[3] Wei K-R, Zheng R-S, Zhang S-W, Liang Z-H, Li Z-M, Chen W-Q. Nasopharyngeal carcinoma incidence and mortality in China, 2013. Chin J Cancer 2017;36(1):90. https://doi.org/10.1186/s40880-017-0257-9.

[4] Katano A, Takahashi W, Yamashita H, Yamamoto K, Ando M, Yoshida M, et al. Radiotherapy alone and with concurrent chemotherapy for nasopharyngeal carcinoma: a retrospective study. Medicine (Baltimore) 2018;97(18). https://doi.org/10.1097/MD.0000000000010502.

[5] Wang R, Feng G, Li Y, et al. Concurrent chemoradiotherapy in locoregionally advanced nasopharyngeal carcinoma: treatment outcomes of a prospective multicentric clinical study. Radiotherapy Oncol J Eur Soc Therapeutic Radiol Oncol 2014.

[6] Wang F, Jiang C, Ye Z, Liu T, Sun Q, Yan F, et al. Treatment outcomes of 257 patients with locoregionally advanced nasopharyngeal carcinoma treated with nimotuzumab plus intensity-modulated radiotherapy with or without chemotherapy: a single-institution experience. Transl Oncol 2018;11(1):65–73.

[7] Walker GV, Awan M, Tao R, Koay EJ, Boehling NS, Grant JD, et al. Prospective randomized double-blind study of atlas-based organ-at-risk autosegmentation-assisted radiation planning in head and neck cancer. Radiother Oncol 2014;112(3):321–5.

[8] van Dijk LV, Van den Bosch L, Aljabar P, Peressutti D, Both S, J H M Steenbakkers R, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiother Oncol 2020;142:115–23.

[9] Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau P-Y, Malandain G, et al. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiother Oncol 2008;87(1):93–9.

[10] Sims R, Isambert A, Grégoire V, Bidault F, Fresco L, Sage J, et al. A pre-clinical assessment of an atlas-based automatic segmentation tool for the head and neck. Radiother Oncol 2009;93(3):474–8.

[11] Teguh DN, Levendag PC, Voet PWJ, Al-Mamgani A, Han X, Wolf TK, et al. Clinical validation of atlas-based auto-segmentation of multiple target volumes and normal tissue (swallowing/mastication) structures in the head and neck. Int J Radiat Oncol Biol Phys 2011;81(4):950–7.

[12] Maddah M, Zou KH, Wells WM, Kikinis R, Warfield SK. Automatic Optimization of Segmentation Algorithms Through Simultaneous Truth and Performance Level Estimation (STAPLE). In: Hellier P, Haynor DR, Barillot C, editors. Medical imaging and computer assisted intervention - MICCAI 2004: Proceedings, part I. Berlin, New York, Paris: Springer; 2004, p. 274–282.

[13] Wang T, Lei Y, Roper J, Ghavidel B, Beitler JJ, McDonald M, et al. Head and neck multi-organ segmentation on dual-energy CT using dual pyramid convolutional neural networks. Phys Med Biol 2021;66(11):115008.

[14] Chi W, Ma L, Wu J, Chen M, Lu W, Gu X. Deep learning-based medical image segmentation with limited labels. Phys Med Biol 2020;65(23):235001.

[15] Kieselmann JP, Fuller CD, Gurney-Champion OJ, Oelfke U. Cross-modality deep learning: Contouring of MRI data from annotated CT data only. Med Phys 2021;48(4):1673–84. https://doi.org/10.1002/mp.14619.

[16] Dai X, Lei Y, Wang T, Zhou J, Roper J, McDonald M, et al. Automated delineation of head and neck organs at risk using synthetic MRI-aided mask scoring regional convolutional neural network. Med Phys 2021;48(10):5862–73.

[17] Chan JW, Kearney V, Haaf S, Wu S, Bogdanov M, Reddick M, et al. A convolutional neural network algorithm for automatic segmentation of head and neck organs at risk using deep lifelong learning. Med Phys 2019;46(5):2204–13.

[18] Tong N, Gou S, Yang S, Ruan D, Sheng K. Fully automatic multi-organ segmentation for head and neck cancer radiotherapy using shape representation model constrained fully convolutional neural networks. Med Phys 2018;45(10):4558–67. https://doi.org/10.1002/mp.13147.

[19] Dai X, Lei Y, Wang T, Dhabaan AH, McDonald M, Beitler JJ, et al. Head-and-neck organs-at-risk auto-delineation using dual pyramid networks for CBCT-guided adaptive radiotherapy. Phys Med Biol 2021;66(4):045021.

[20] Liu Y, Lei Y, Fu Y, Wang T, Zhou J, Jiang X, et al. Head and neck multi-organ auto-segmentation on CT images aided by synthetic MRI. Med Phys 2020;47(9):4294–302.

[21] Iyer A, Thor M, Onochie I, Hesse J, Zakeri K, LoCastro E, et al. Prospectively-validated deep learning model for segmenting swallowing and chewing structures in CT. Phys Med Biol 2022;67(2):024001.

[22] Siciarz P, McCurdy B. U-net architecture with embedded Inception-ResNet-v2 image encoding modules for automatic segmentation of organs-at-risk in head and neck cancer radiation therapy based on computed tomography scans. Phys Med Biol 2022;67(11):115007.

[23] Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. Med Phys 2017;44(2):547–57. https://doi.org/10.1002/mp.12045.

[24] Liang S, Tang F, Huang X, Yang K, Zhong T, Hu R, et al. Deep-learning-based detection and segmentation of organs at risk in nasopharyngeal carcinoma computed tomographic images for radiotherapy planning. Eur Radiol 2019;29(4):1961–7.

[25] Tang H, Chen X, Liu Y, Lu Z, You J, Yang M, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. Nature Machine Intelligence 2019;1(10):480–91.

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

[27] Li Y, Rao S, Chen W, Azghadi SF, Nguyen KNB, Moran A, et al. Evaluating automatic segmentation for swallowing-related organs for head and neck cancer. Technol Cancer Res Treat 2022;21. https://doi.org/10.1177/15330338221105724.

[28] Fu Y, Lei Y, Wang T, Curran WJ, Liu T, Yang X. A review of deep learning based methods for medical image multi-organ segmentation. Phys Med 2021;85:107–22. https://doi.org/10.1016/j.ejmp.2021.05.003.

[29] Tao C-J, Yi J-L, Chen N-Y, Ren W, Cheng J, Tung S, et al. Multi-subject atlas-based auto-segmentation reduces interobserver variation and improves dosimetric parameter consistency for organs at risk in nasopharyngeal carcinoma: a multi-institution clinical study. Radiother Oncol 2015;115(3):407–11.

[30] Teunissen FR, Wortel RC, Wessels FJ, Claes A, van de Pol SMG, Rasing MJA, et al. Interrater agreement of contouring of the neurovascular bundles and internal pudendal arteries in neurovascular-sparing magnetic resonance-guided radiotherapy for localized prostate cancer. Clin Transl Radiat Oncol 2022;32:29–34.

[31] Joskowicz L, Cohen D, Caplan N, Sosna J. Inter-observer variability of manual contour delineation of structures in CT. Eur Radiol 2019;29(3):1391–9. https://doi.org/10.1007/s00330-018-5695-5.

[32] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, Verellen D, et al. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66.

[33] Brouwer CL, Steenbakkers RJHM, Bourhis J, Budach W, Grau C, Grégoire V, 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(1):83–90.

[34] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation. IEEE Trans Pattern Anal Mach Intell 2017;39(12):2481–95. https://doi.org/10.1109/TPAMI.2016.2644615.

[35] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41.

[36] Chen L-C, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans Pattern Anal Mach Intell 2018;40(4):834–48. https://doi.org/10.1109/TPAMI.2017.2699184.

[37] Chen L-C, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-decoder with atrous separable convolution for semantic image segmentation. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y, editors. Computer Vision – ECCV 2018. Cham: Springer International Publishing; 2018. p. 833–51.

[38] Casciaro S, Franchini R, Massoptier L, Casciaro E, Conversano F, Malvasi A, et al. Fully automatic segmentations of liver and hepatic tumors from 3-D computed tomography abdominal images: comparative evaluation of two automatic methods. IEEE Sensors J 2012;12(3):464–73.

[39] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, De Fauw J, 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(7). https://doi.org/10.2196/26151.

[40] Song Y, Hu J, Wu Q, Xu F, Nie S, Zhao Y, et al. Automatic delineation of the clinical target volume and organs at risk by deep learning for rectal cancer postoperative radiotherapy q. Radiother Oncol 2020;145:186–92.