scientific data

Check for updates

OPEN

DATA DESCRIPTOR

A multicenter dataset for lymph node clinical target volume delineation of nasopharyngeal carcinoma

Xiangde Luo[1,2,3], Wenjun Liao[1]✉, Yue Zhao[1], Youjing Qiu[4], Jinfeng Xu[5], Yuan He[6], Hui Huang[7], Lu Li[1], Shichuan Zhang[1], Jia Fu[2], Guotai Wang[2,3] & Shaoting Zhang[2,3]

The deep learning (DL)-based prediction of accurate lymph node (LN) clinical target volumes (CTVs) for nasopharyngeal carcinoma (NPC) radiotherapy (RT) remains challenging. One of the main reasons is the variability of contours despite standardization processes by expert guidelines in combination with scarce data sharing in the community. Therefore, we retrospectively generated a 262-subjects dataset from four centers to develop the DL models for LN CTVs delineation. This dataset included 440 computed tomography images from different scanning phases, disease stages and treatment strategies. Three clinical expert boards, each comprising two experts (totalling six experts), manually delineated six basic LN CTVs on separate cohorts as the ground truth according to LN involvement and clinical requirements. Several state-of-the-art segmentation algorithms were evaluated on this benchmark, showing promising results for LN CTV segmentation. In conclusion, this work built a multicenter LN CTV segmentation dataset, which may be the first dataset for automatic LN CTV delineation development and evaluation, serving as a benchmark for future research.

Background & Summary

Nasopharyngeal carcinoma (NPC) is the 23rd most common cancer, and radiotherapy (RT) is the primary treatment approach for it[1,2]. In the process of RT administration, delineating lymph node (LN) clinical target volumes (CTVs) on computed tomography (CT) is a crucial step. RT planning requires accurately delineating neck lymphatic drainage regions and calculating their radiation dose, as it maximizes the chance of a cure while minimizing radiation-related toxic effects. In the past, clinicians manually performed LN CTVs delineation with slice-by-slice on CT images. It is laborious and prone to intra- and interobserver variability[3,4]. Additionally, the online adaptive RT often involves repeated manual delineation on repeated planning CT images, thus increasing much burden on radiation oncologists and prolonging patient waiting time[5,6].

Recently, deep learning (DL)-based segmentation has shown the potential to alleviate the manual delineation burden[7–11]. Furthermore, the application of DL for automated delineation has been extended to LN levels in head and neck cancer (HNC). *Van der Veen et al.*[12] examined the performance of a DL model in segmenting LN levels in HNC patients and showed that it improved time efficiency and reduced interobserver variability. Additionally, *Cardenas et al.*[13] developed a DL model to auto-delineate commonly used LN level combinations for HNC patients robustly and consistently. Those studies have investigated the accuracy of DL-based LN level segmentation in HNC. The training of those models relied on the delineation of LN levels following the 2013

[1]Department of Radiation Oncology, Radiation Oncology Key Laboratory of Sichuan Province, Sichuan Clinical Research Center for Cancer, Sichuan Cancer Hospital and Institute. Sichuan Cancer Center, Affiliated Cancer Hospital of University of Electronic Science and Technology of China, Chengdu, China. [2]School of Mechanical and Electrical Engineering, University of Electronic Science and Technology of China, Chengdu, China. [3]Shanghai AI Laboratory, Shanghai, China. [4]Department of Radiation Oncology, Daguan Hospital of Chengdu Jinjiang, Chengdu, China. [5]Department of Radiation Oncology, Nanfang Hospital, Southern Medical University, Guangzhou, China. [6]Department of Radiation Oncology, The First Affiliated Hospital of University of Science and Technology of China, Hefei, China. [7]Cancer center, Sichuan Provincial People's Hospital, University of Electronic Science and Technology of China, Chengdu, China. ✉e-mail: liao_wj2023@uestc.edu.cn

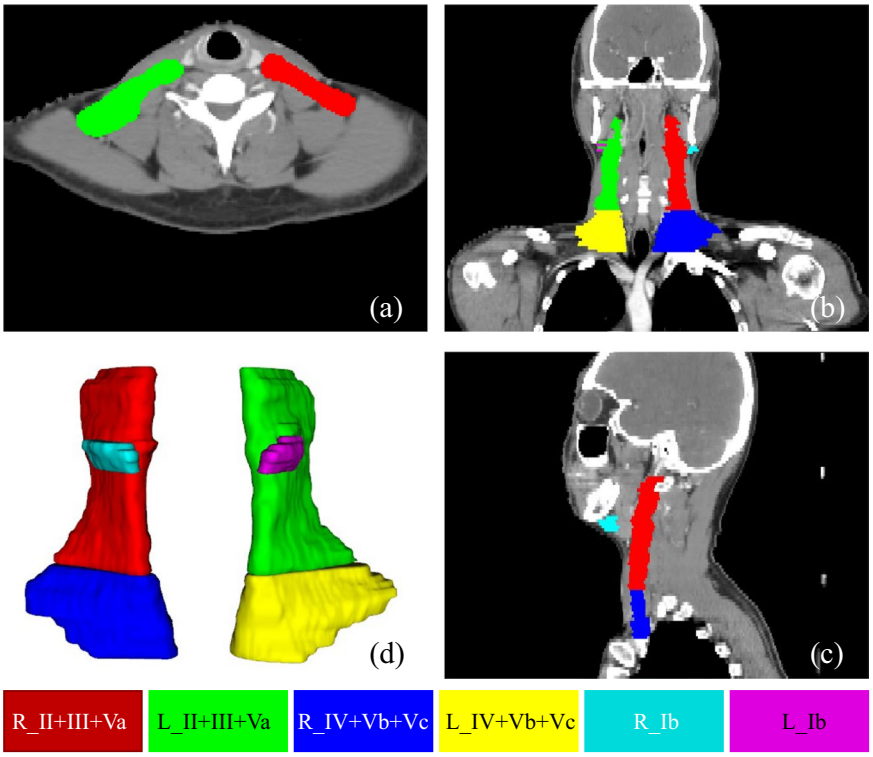


Fig. 1 An example of six annotated LN CTVs in a contrast-enhanced CT scan. The bottom table lists the annotated LN CTV categories. (**a**), (**b**), (**c**) and (**d**) represent the visualization in axial, coronal, and sagittal and annotation rendering views respectively.

updated consensus guideline[14]. However, it is important to note that those models and contours generated by them may not directly apply to NPC whose nodal metastatic characteristics differ from other HNC[15,16]. *Lin et al.* proposed a specific LN CTV delineation map to NPC based on 10,651 LNs distribution[17]. They observed that certain boundaries of levels Ib, II, IVa and V required adjustments. For instance, when delineating level Ib, the submandibular gland should be spared. Moreover, two works proposed additional modifications for levels IIb and V[18,19]. Hence, to facilitate the automated RT workflow in NPC, it is necessary to develop a DL model specifically tailored for LN CTVs delineation in this particular subtype following the most recent clinical guidelines.

In addition, specific combinations of LN levels selected according to LN metastatic risk are needed in clinical practice[20–22]. For instance, in NPC patients with N0, it is recommended to delineate a combination of bilateral levels II, III and Va. In patients with N1, the recommended delineation includes an ipsilateral combination of levels II-V and a contralateral combination of levels II, III and Va[23]. However, most previous DL models for NPC could not automatically output selective target volumes based on LN involvement[24], which does not align with the clinical requirements.

Although existing studies reported promising segmentation performance for automatic HNC organ-at-risk (OAR) and tumor segmentation[12,13,25], several potential challenges need to be noticed and handled before clinical application development. First, all of these studies developed the model and reported the performance on their collected private datasets[12,13], it's not clear what's the encouraging solution for this task due to the lack of wide and reproductive research. It's significantly different from many organ[26,27] or tumor[28] segmentation tasks which have many public datasets that can be used for model development and evaluation. However, few public HNC LN CTV datasets can be accessed for research until now[29–31]. Second, these public datasets mainly focused on survival and recurrence analysis using tumor-based radiomics. So, no study investigates CTV delineation in radiation therapy. Third, the robustness of the automatic segmentation models when dealing with patients with different stages or treatment strategies is not well evaluated. With the treatment ongoing, significant changes may occur in the location, size, and appearance of the tumor and LNs, such as reduction or disappearance[23]. It is desirable to develop a detailed LN CTV delineation dataset to investigate its potential influence in radiation therapy.

In this work, we aim to build a multicenter NPC LN CTVs segmentation benchmark for model development and evaluation. We retrospectively collected 262 patients with 440 CT images from four hospitals. To develop a specialized DL model for delineating LN CTVs in NPC, we modified the delineation of levels Ib, IIb, III, and the supraclavicular region based on previous studies[17–19] and the 2013 consensus[14]. We invited three clinical expert boards to delineate a comprehensive set of six basic LN CTVs manually as the ground truth, an example of this dataset is shown in Fig. 1. To comprehensively investigate and evaluate the performance of recent works, the dataset consists of patients with different stages, treatment strategies, imaging modalities, and raters. Here, we split the dataset into 5 cohorts according to the sources, an internal cohort for training and validation

Variable	Internal cohorts (N = 150 pts)		External cohorts with different modalities (N = 92 pts)		External cohorts with different treatment strategies (N = 20 pts)	
	Training cohort (N = 120, %)	Testing cohort (N = 30, %)	Testing cohort 1 (N = 60, %)	Testing cohort 2 (N = 32, %)	Testing cohort 3 (N = 12, %)	Testing cohort 4 (N = 8, %)
Sex						
7D2Male	85 (70.8)	22 (73.3)	45 (75.0)	23 (71.9)	11 (91.7)	7 (87.5)
7D2Female	35 (29.2)	8 (26.7)	15 (25.0)	9 (28.1)	1 (8.3)	1 (12.5)
Age (median)	48 (24-75)	47 (22-71)	49 (26-74)	49 (22-62)	48 (32-72)	52 (49-66)
T category*						
7D2T1	5 (4.2)	1 (3.3)	3 (5.0)	3 (9.4)	0 (0.0)	2 (25.0)
7D2T2	17 (14.2)	3 (10.0)	10 (16.7)	4 (12.5)	3 (25.0)	2 (25.0)
7D2T3	72 (60.0)	19 (63.3)	35 (58.3)	18 (56.3)	2 (16.7)	2 (25.0)
7D2T4	26 (21.6)	7 (23.4)	12 (20.0)	7 (21.8)	7 (58.3)	2 (25.0)
N category*						
7D2N0	1 (0.8)	1 (3.3)	2 (3.3)	1 (3.1)	0 (0.0)	0 (0.0)
7D2N1	31 (25.8)	3 (10.0)	11 (18.3)	10 (31.2)	1 (8.3)	0 (0.0)
7D2N2	63 (52.6)	17 (56.7)	26 (43.4)	11 (34.4)	9 (75.0)	4 (50.0)
7D2N3	25 (20.8)	9 (30.0)	21 (35.0)	10 (31.3)	2 (16.7)	4 (50.0)
Image parameters						
7D2Center	SCH	SCH	SPH	APH	SMU	SCH
7D2CT scan	ceCT and ueCT	ceCT and ueCT	ceCT	ueCT	ceCT	ceCT
7D2No. of CT images	240	60	60	32	24	24
7D2Thickness	3 mm	3 mm	3 mm	3 mm	3 mm	3 mm
7D2Manufacture	Philips	Philips	Siemens	Siemens	Siemens	Philips
7D2Manual delineation#	SCH board	SCH board	SPH board	APH board	SCH board	SCH board

Table 1. Clinical and image characteristics of the LNCTVSeg dataset. SCH, SPH, APH and SMU mean Sichuan Cancer Hospital, Sichuan Provincial People's Hospital, Anhui Provincial Hospital and Southern Medical University, respectively. ceCT, ueCT and pts represent contrast-enhanced computed tomography, unenhanced computed tomography and patients, respectively. * Denotes that all patients were restaged according to the eighth edition of the American Joint Committee on Cancer. # Represents that manual delineation was performed by three expert groups and each group had two senior radiation oncologists with more than ten years of experience in treating NPC.

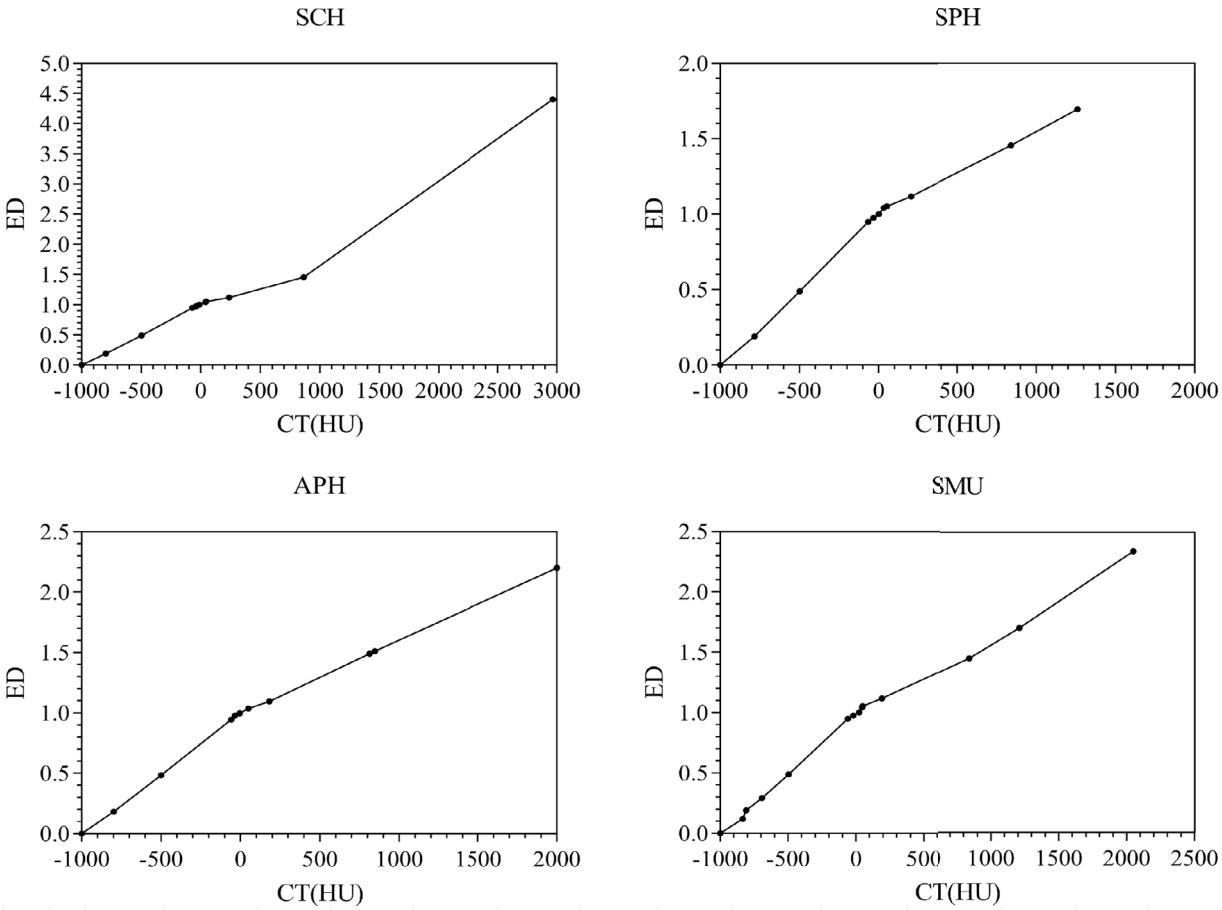
and 4 external cohorts just for evaluation. After that, we trained and evaluated several state-of-the-art (SOTA) segmentation methods on the internal training and testing sets. Moreover, we investigated the robustness and generalization ability of these SOTA methods on four external testing sets. These results provided benchmark results and potential directions for future research.

Methods

Data ethics. This retrospective study was approved by the Ethics Committee on Biomedical Research of these hospitals (approve number, SCCHEC-02-2023-005), including Sichuan Cancer Hospital (SCH), Sichuan Provincial People's Hospital (SPH), Anhui Provincial Hospital (APH), and Southern Medical University (SMU). Since this was a retrospective study with blurred face regions and all clinical and personal metadata removed, the informed consent was waived. The committee approved the open publication of this dataset.

Data collection. A total of 262 NPC patients were retrospectively collected between January 2019 and June 2022. The internal cohorts were collected from SCH and contained 300 CT images from 150 patients (Table 1). Each patient underwent an unenhanced CT (ueCT) scan followed by a contrast-enhanced CT (ceCT) scan in the same body position within several minutes. The detailed imaging protocols are that all NPC patients underwent RT and were immobilized supine using a thermoplastic head, neck, and shoulder mask. CT examinations were then performed, covering a scanning range from 3 cm above the suprasellar cistern to 2 cm below the sternal end of the clavicle. The CT images from SCH were acquired using a Brilliance CT Big Bore system from Philips Healthcare (Philips Healthcare, Best, the Netherlands), with the following scanning conditions: bulb voltage at 120 kV, current ranging from 275 to 375 mA, scan thickness of 3.0 mm, and an image matrix of 512 × 512. An injected contrast agent, iohexol, was used during the ceCT examination. Similarly, CT images from SPH, APH and SMU were acquired using a Somatom Definition AS 40 system from Siemens Healthcare (Siemens Healthcare, Forcheim, Germany), with the following conditions: bulb voltage at 120 -140 kV, 280-380 mAs current, 3.0 mm slice thickness, and an image matrix of 512 × 512. Iohexol was used as a contrast agent. In addition, to make possible dosimetric studies in the future, the electron density (ED) to Hounsfield unit (HU) of the CT conversion table for the four hospitals is present in Fig. 2.

LN CTVs delineation. During the RT treatment process, different patients need to delineate different LN CTVs according to the clinical stages for individual RT. To obtain individual LN CTVs for each NPC patient, we



SCH	HU	-1000	-796	-497	-69	-36	-10	42	45	238	865	2965			
	ED	0.000	0.190	0.489	0.949	0.976	1.000	1.043	1.052	1.117	1.456	4.400			
SPH	HU	-1000	-782.175	-497.7	-65.325	-32.112	1.000	33.913	52.913	206.162	840.362	1260.325			
	ED	0.001	0.190	0.489	0.949	0.976	1.000	1.043	1.052	1.117	1.456	1.695			
APH	HU	-1000	-797	-499	-58	-35	-4	50	182	814	847	2000			
	ED	0.001	0.183	0.484	0.944	0.979	0.996	1.037	1.094	1.49	1.512	2.2			
SMU	HU	-1000	-833	-808	-691	-495	-59	-20	23	44.5	49.6	191	840	1210	2047
	ED	0.000	0.120	0.190	0.292	0.489	0.949	0.976	1.000	1.043	1.053	1.117	1.456	1.707	2.343

Fig. 2 Curves of electronic density (ED) to Hounsfield unit (HU) of CT conversion for patients from four hospitals. SCH, SPH, APH and SMU mean Sichuan Cancer Hospital, Sichuan Provincial People's Hospital, Anhui Provincial Hospital and Southern Medical University, respectively.

Variable	LN levels included	Clinical significance in RT	Label Index
R_II+III+Va	**Right levels II, III, and Va**	**Right upper neck**	**1**
L_II+III+Va	**Left levels II, III, and Va**	**Left upper neck**	**2**
R_IV+Vb+Vc	**Right levels IV, Vb, and Vc**	**Right lower neck**	**3**
L_IV+Vb+Vc	**Left levels IV, Vb, and Vc**	**Left lower neck**	**4**
R_Ib	**Right level Ib**	**Right submandibular area**	**5**
L_Ib	**Left level Ib**	**Left submandibular area**	**6**
R_II-V	Right levels II, III, IV, and V	Right whole neck	1 & 3
R_Ib-V	Right levels Ib, II, III, IV, and V	Right whole neck and submandibular area	1 & 3 & 5
L_II-V	Left levels II, III, IV, and V	Left whole neck	2 & 4
L_Ib-V	Left levels Ib, II, III, IV, and V	Left whole neck and submandibular area	2 & 4 & 6
B_II+III+Va	Bilateral levels II, III, and Va	Bilateral upper neck	1 & 2
B_II-V	Bilateral levels II, III, IV, and V	Bilateral whole neck	1 & 2 & 3 & 4
B_Ib-V	Bilateral levels Ib, II, III, IV, and V	Bilateral whole neck and submandibular areas	1 & 2 & 3 & 4 & 5 & 6

Table 2. Clinical significance of the 13 lymph node clinical target volumes.

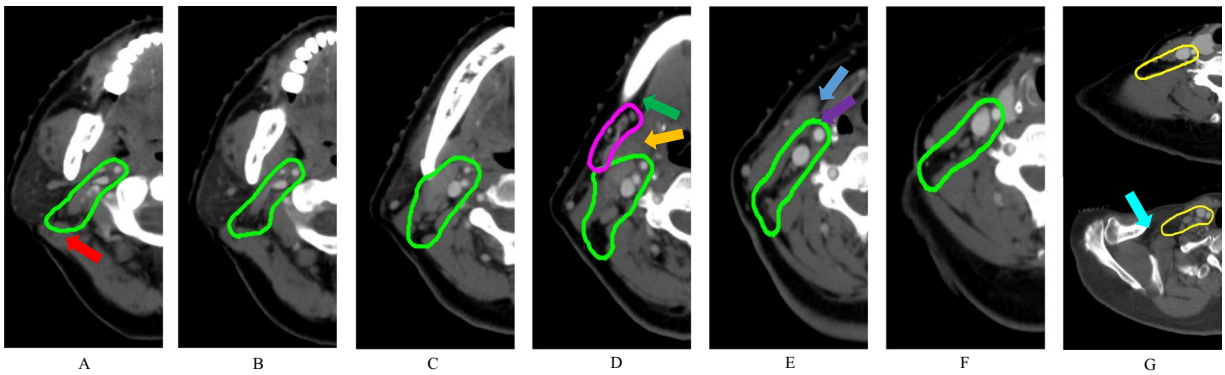


Fig. 3 Examples of experts' annotation at some key anatomical levels. The red arrow in A represents the tight muscle gap between the sternocleidomastoid and splenius capitis muscles, which can be excluded when delineating IIb. The yellow arrow and green arrow in D represent the submandibular gland (SMG) and Medial Mandibular sub-level, respectively. These two areas can be removed when delineating Ib. The purple arrow in E represents the carotid sheath. The dark blue arrow represents the space between the anterior of the sternocleidomastoid muscle and the anterior neck ribbon muscles, and this area can be spared when delineating III. The light blue in G represents the scapular hyoid muscle, which is the lateral border of level V.

Patient with IC#		Patient with adaptive RT*		
Patient	Scan interval (days)	Patient (n = 8)	Scan interval (days) (CT_{RA} to CT_{RB})	Scan interval (days) (CT_{RB} to CT_{RC})
1	48	A	28	18
2	66	B	30	16
3	67	C	21	15
4	45	D	28	20
5	67	E	30	18
6	70	F	32	22
7	48	G	24	16
8	72	H	36	18
9	42	mean days	28.6	17.9
10	72	—	—	—
11	68	—	—	—
12	66	—	—	—
mean days	60.9	—	—	—

Table 3. The scan interval of different CT images. # means 12 patients who underwent two CT scans before induction chemotherapy (IC) and after induction chemotherapy. * Denotes eight patients who underwent three repeated CT scans during adaptative radiotherapy (RT). These scans were performed after 0, 15-20, and 25-30 fractions and labelled as CT_{RA}, CT_{RB}, and CT_{RC}, respectively. - means not applicable.

Variable	R_II+III+Va	L_II+III+Va	R_IV+Vb+Vc	L_IV+Vb+Vc	R_Ib	L_Ib
Intraobserver	88.67% ± 1.85%	88.05% ± 2.59%	85.05% ± 3.40%	84.97% ± 3.31%	75.21% ± 4.28%	76.10% ± 2.92%
DL	89.45% ± 1.27%	89.14% ± 1.59%	85.87% ± 3.00%	86.10% ± 2.69%	79.95% ± 5.74%	78.84% ± 5.41%
p-value	0.220	0.137	0.384	0.289	0.062	0.204

Table 4. Quality of the DL-generated segmentation compared to intraobserver (from SCH board). *P*-values were obtained by comparing the DL-generated segmentations and intraobserver annotations.

Variable	R_II+III+Va	L_II+III+Va	R_IV+Vb+Vc	L_IV+Vb+Vc	R_Ib	L_Ib
Interobserver (SCH-SPH)	80.23% ± 9.11%	79.07% ± 7.66%	84.16% ± 3.98%	82.80% ± 5.70%	76.43% ± 11.60%	76.34% ± 7.69%
Interobserver (SCH-APH)	80.97% ± 8.95%	79.91% ± 7.45%	84.56% ± 2.60%	83.32% ± 4.32%	75.33% ± 10.49%	76.31% ± 8.03%
Interobserver (SPH-APH)	89.36% ± 1.43%	88.34% ± 1.61%	85.28% ± 3.37%	83.98% ± 4.47%	78.29% ± 7.31%	78.51% ± 6.68%

Table 5. Interobserver variabilities across three boards (SCH, SPH, APH) in the term of DSC.

Variable	DSC (%)					HD95 (mm)					NSD (%)				
	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR
Internal testing cohort															
R_II+ III+ Va	87.98 ± 3.83	86.95 ± 3.79	85.01 ± 4.94	85.42 ± 3.78	85.23 ± 4.43	4.80 ± 3.13	4.96 ± 3.84	6.35 ± 3.79	6.71 ± 3.55	6.47 ± 3.55	89.57 ± 8.26	87.75 ± 8.69	84.97 ± 10.76	85.21 ± 9.63	85.07 ± 9.84
L_II+ III+ Va	87.76 ± 3.04	86.91 ± 3.18	84.78 ± 3.46	84.74 ± 2.42	84.60 ± 3.27	4.38 ± 2.85	4.61 ± 2.89	6.00 ± 2.98	6.49 ± 3.35	6.58	90.88 ± 4.77	89.36 ± 5.44	86.45 ± 6.51	86.33 ± 5.52	86.64 ± 5.33
R_IV +Vb +Vc	84.49 ± 5.41	82.12 ± 5.35	81.18 ± 4.31	81.54 ± 4.16	80.93 ± 4.10	4.74 ± 3.85	6.22 ± 5.71	6.31 ± 3.32	6.37 ± 3.18	6.62	83.14 ± 8.58	79.19 ± 9.62	77.34 ± 7.53	78.62 ± 7.16	77.9 ± 7.19
L_IV+ Vb+ Vc	85.20 ± 3.57	83.40 ± 3.83	80.81 ± 3.40	81.04 ± 4.00	80.80 ± 3.85	4.31 ± 2.31	8.25 ± 15.27	8.91 ± 13.80	8.74 ± 13.78	8.79 ± 13.85	84.92 ± 7.1	81.87 ± 7.69	78.27 ± 6.71	79.17 ± 6.72	78.42 ± 6.77
R_Ib	76.77 ± 14.56	76.39 ± 15.14	75.57 ± 14.74	76.38 ± 14.14	76.38 ± 14.05	4.22 ± 3.94	5.16 ± 9.34	4.25 ± 4.09	4.16 ± 3.56	4.28 ± 4.45	86.59 ± 14.27	86.32 ± 16.18	85.38 ± 14.58	85.87 ± 13.9	86.77 ± 13.65
L_Ib	75.42 ± 13.00	74.75 ± 13.63	73.00 ± 13.47	73.67 ± 13.23	72.98 ± 14.39	3.73 ± 1.63	4.24 ± 2.81	3.93 ± 2.09	3.87 ± 1.81	3.91 ± 1.91	87.12 ± 8.26	86.39 ± 9.45	85.16 ± 9.52	85.82 ± 9.09	85.58 ± 10.1
R_II-V	88.39 ± 2.69	86.50 ± 3.77	86.45 ± 3.92	86.86 ± 3.21	86.58 ± 3.59	4.23 ± 2.87	5.62 ± 5.24	4.63 ± 2.76	4.69 ± 2.79	4.71 ± 2.90	90.14 ± 6.4	88.51 ± 6.82	85.63 ± 8.53	85.7 ± 7.39	85.76 ± 7.37
R_Ib-V	88.21 ± 2.78	86.28 ± 2.66	86.30 ± 4.00	86.80 ± 3.33	86.50 ± 3.66	4.01 ± 2.58	5.27 ± 4.89	4.38 ± 2.61	4.18 ± 2.35	4.17 ± 2.42	89.97 ± 6.68	86.87 ± 8.0	86.71 ± 8.9	87.43 ± 8.05	86.96 ± 8.41
L_II-V	88.43 ± 2.15	87.50 ± 2.15	86.89 ± 2.29	86.87 ± 2.26	86.85 ± 2.28	4.01 ± 2.90	4.98 ± 4.31	4.57 ± 2.99	4.54 ± 3.03	4.38 ± 3.12	91.46 ± 3.88	90.01 ± 4.4	89.11 ± 4.89	89.34 ± 4.23	89.33 ± 4.01
L_Ib-V	88.17 ± 1.93	87.21 ± 1.93	86.64 ± 1.99	86.77 ± 1.93	86.47 ± 2.10	3.56 ± 0.93	4.35 ± 3.18	3.88 ± 0.94	3.78 ± 0.86	3.83 ± 0.92	90.89 ± 4.92	88.63 ± 5.8	88.09 ± 6.57	88.56 ± 5.8	88.3 ± 5.93
B_II+ III+ Va	87.85 ± 3.18	86.91 ± 3.20	84.86 ± 4.01	85.08 ± 3.21	84.89 ± 3.34	4.52 ± 3.07	4.81 ± 3.48	6.22 ± 3.38	6.65 ± 3.30	6.46 ± 3.40	90.81 ± 4.8	88.74 ± 5.64	88.14 ± 6.46	86.62 ± 5.58	88.41 ± 5.89
B_II-V	88.46 ± 2.28	87.03 ± 2.71	86.69 ± 2.90	86.90 ± 2.53	86.74 ± 2.54	3.84 ± 1.89	5.11 ± 4.32	4.42 ± 2.27	4.35 ± 2.10	4.29 ± 2.26	90.45 ± 6.66	87.34 ± 7.77	87.38 ± 8.8	88.15 ± 7.94	87.7 ± 8.5
B_Ib-V	88.16 ± 2.17	86.78 ± 2.70	86.40 ± 2.82	86.64 ± 2.36	86.48 ± 2.48	3.64 ± 1.36	4.89 ± 4.03	4.12 ± 1.76	4.06 ± 1.45	3.98 ± 1.57	91.74 ± 3.56	90.25 ± 3.95	89.44 ± 4.54	89.93 ± 3.69	89.47 ± 3.9
Mean	85.79	84.52	83.43	83.75	83.49	4.15	5.27	5.23	5.28	5.27	89.05	87.02	85.54	86.06	85.87
External testing cohort1															
R_II+ III+ Va	88.29 ± 1.78	86.63 ± 2.84	85.57 ± 2.54	85.88 ± 2.18	85.50 ± 2.33	3.81 ± 1.17	4.95 ± 1.96	6.17 ± 2.75	6.06 ± 2.17	6.36 ± 2.67	88.7 ± 4.11	86.17 ± 5.29	84.87 ± 5.49	85.22 ± 4.75	84.85 ± 4.91
L_II+ III+ Va	86.61 ± 2.17	85.36 ± 3.30	83.43 ± 3.26	84.02 ± 2.89	82.61 ± 2.82	4.54 ± 1.61	5.64 ± 1.96	6.80 ± 2.67	6.74 ± 2.44	7.71 ± 2.81	85.86 ± 5.03	83.22 ± 6.26	81.2 ± 6.22	82.06 ± 5.8	79.95 ± 6.02
R_IV+ Vb+ Vc	83.79 ± 3.82	80.16 ± 4.30	79.97 ± 6.14	80.09 ± 5.42	79.72 ± 5.82	4.75 ± 1.86	6.07 ± 2.08	6.89 ± 3.41	6.29 ± 2.11	6.67 ± 2.50	79.0 ± 8.54	73.28 ± 8.39	73.29 ± 10.43	73.33 ± 9.64	73.22 ± 9.93
L_IV+ Vb+ Vc	83.35 ± 3.86	81.99 ± 4.54	78.85 ± 6.76	78.77 ± 6.67	78.20 ± 6.88	4.66 ± 1.85	6.60 ± 6.31	8.17 ± 7.74	7.83 ± 6.77	8.37 ± 5.91	78.49 ± 8.78	74.86 ± 8.11	72.24 ± 10.43	72.21 ± 10.46	72.17 ± 10.18
R_Ib	75.71 ± 6.14	74.12 ± 6.74	73.19 ± 6.77	73.12 ± 6.42	73.70 ± 5.73	4.02 ± 1.50	4.57 ± 1.85	4.51 ± 1.98	4.70 ± 1.89	4.39 ± 1.58	84.92 ± 6.98	84.42 ± 7.11	81.41 ± 7.88	81.53 ± 7.14	82.45 ± 6.65
L_Ib	75.82 ± 6.16	75.53 ± 6.74	73.56 ± 6.71	74.34 ± 6.06	73.64 ± 6.78	4.24 ± 2.09	4.68 ± 2.39	4.57 ± 2.05	4.28 ± 1.58	4.44 ± 1.61	84.24 ± 6.74	84.75 ± 6.38	82.45 ± 7.16	83.24 ± 6.3	82.43 ± 6.65
R_II-V	87.96 ± 1.60	86.84 ± 2.36	86.87 ± 1.96	87.06 ± 1.72	86.92 ± 1.69	4.35 ± 1.36	5.47 ± 1.56	4.81 ± 1.26	4.67 ± 1.13	4.75 ± 1.24	87.23 ± 3.96	84.64 ± 5.69	82.95 ± 5.12	83.59 ± 4.62	82.36 ± 4.88
R_Ib-V	87.39 ± 1.46	85.27 ± 2.75	86.25 ± 1.85	86.51 ± 1.48	86.39 ± 1.58	4.41 ± 1.21	5.55 ± 1.49	4.71 ± 1.26	4.55 ± 1.05	4.60 ± 1.15	87.73 ± 3.48	83.98 ± 5.84	85.92 ± 4.47	86.25 ± 4.06	86.12 ± 4.06
L_II-V	86.96 ± 1.78	84.63 ± 2.55	85.67 ± 2.08	86.07 ± 1.87	85.47 ± 1.87	4.62 ± 1.57	5.51 ± 1.58	5.24 ± 1.86	5.03 ± 1.77	5.24 ± 2.07	86.05 ± 4.77	83.28 ± 6.28	83.78 ± 5.16	84.51 ± 5.02	83.48 ± 5.47
L_Ib-V	86.34 ± 1.80	85.95 ± 1.91	85.02 ± 2.14	85.55 ± 1.93	84.77 ± 1.98	4.72 ± 1.67	5.70 ± 1.61	5.24 ± 1.67	4.96 ± 1.65	5.32 ± 1.99	87.03 ± 3.63	83.78 ± 5.06	85.01 ± 4.21	85.55 ± 4.05	84.97 ± 4.29
B_II+ III+ Va	87.45 ± 1.64	85.92 ± 1.73	84.48 ± 2.50	84.94 ± 2.10	84.07 ± 2.18	4.09 ± 1.06	5.29 ± 1.74	6.41 ± 2.30	6.42 ± 2.01	6.95 ± 2.24	86.91 ± 3.31	85.01 ± 4.7	84.85 ± 3.97	85.4 ± 3.81	84.85 ± 4.06

Continued

Variable	DSC (%)					HD95 (mm)					NSD (%)				
	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR
B_II-V	87.51 ± 1.38	86.77 ± 1.92	86.31 ± 1.70	86.61 ± 1.48	86.25 ± 1.46	4.40 ± 1.26	5.47 ± 1.34	4.91 ± 1.28	4.72 ± 1.18	4.86 ± 1.29	87.32 ± 3.21	83.64 ± 5.49	85.33 ± 4.26	85.92 ± 3.91	85.76 ± 3.88
B_Ib-V	87.14 ± 1.28	85.50 ± 2.07	85.90 ± 1.65	86.20 ± 1.45	85.58 ± 1.44	4.31 ± 1.26	5.32 ± 1.21	4.78 ± 1.10	4.64 ± 1.05	4.77 ± 1.30	85.56 ± 4.52	82.97 ± 5.79	83.39 ± 4.99	84.41 ± 4.67	83.0 ± 5.28
Mean	84.95	83.43	82.70	83.01	82.52	4.38	5.45	5.63	5.45	5.73	85.31	82.61	82.05	82.56	81.97
External testing cohort2															
R_II+ III+ Va	87.05 ± 2.15	87.06 ± 2.25	84.72 ± 2.81	84.64 ± 2.36	84.81 ± 2.55	4.85 ± 1.72	4.91 ± 1.80	6.04 ± 2.52	6.09 ± 1.81	6.44 ± 3.70	88.5 ± 4.63	88.56 ± 4.2	84.79 ± 5.63	83.92 ± 4.98	84.17 ± 5.2
L_II+ III+ Va	86.95 ± 1.63	86.60 ± 2.35	84.73 ± 2.35	84.48 ± 2.16	84.10 ± 1.78	4.34 ± 1.20	5.03 ± 1.93	14.04 ± 47.35	6.00 ± 2.12	6.11 ± 2.09	88.79 ± 3.74	88.4 ± 2.69	84.85 ± 4.34	84.7 ± 3.96	84.27 ± 4.1
R_IV+ Vb+ Vc	82.95 ± 3.20	83.05 ± 2.84	79.05 ± 5.87	80.00 ± 4.17	80.47 ± 4.29	4.76 ± 1.87	4.92 ± 1.73	13.94 ± 41.72	5.85 ± 2.25	5.73 ± 1.69	80.12 ± 7.67	80.49 ± 5.26	74.52 ± 8.59	75.78 ± 7.57	76.85 ± 7.99
L_IV+ Vb+ Vc	83.20 ± 2.87	83.07 ± 2.84	79.23 ± 5.65	79.91 ± 5.26	79.09 ± 4.95	4.45 ± 1.62	8.56 ± 14.93	33.08 ± 73.20	9.40 ± 14.35	10.04 ± 14.82	81.29 ± 6.66	80.75 ± 5.81	75.17 ± 9.64	75.7 ± 9.4	74.73 ± 9.3
R_Ib	70.60 ± 8.10	71.52 ± 8.67	69.59 ± 8.65	68.06 ± 8.64	69.89 ± 8.79	4.20 ± 2.13	4.45 ± 2.18	4.26 ± 1.86	4.57 ± 2.05	4.16 ± 2.48	81.39 ± 10.01	82.01 ± 10.27	79.82 ± 10.2	78.53 ± 10.36	80.57 ± 9.47
L_Ib	71.50 ± 8.98	72.99 ± 7.67	70.01 ± 8.13	70.87 ± 7.51	70.33 ± 7.78	3.96 ± 1.47	4.10 ± 1.52	4.27 ± 1.58	13.52 ± 51.80	4.24 ± 1.48	83.08 ± 8.02	83.51 ± 8.83	82.05 ± 7.12	81.79 ± 7.52	81.82 ± 7.46
R_II-V	86.87 ± 2.19	87.17 ± 1.75	85.53 ± 2.54	85.54 ± 2.34	85.84 ± 2.36	4.39 ± 1.24	4.29 ± 1.07	11.92 ± 40.98	4.75 ± 1.24	4.63 ± 1.28	88.59 ± 3.46	88.45 ± 3.03	84.78 ± 4.32	84.26 ± 3.84	84.17 ± 4.12
R_Ib-V	86.54 ± 1.97	86.98 ± 1.55	85.35 ± 2.31	85.38 ± 2.29	85.71 ± 2.24	4.01 ± 0.96	3.86 ± 0.79	4.33 ± 1.11	4.36 ± 1.04	4.18 ± 1.02	87.79 ± 5.4	88.28 ± 4.33	85.18 ± 5.67	85.24 ± 5.55	85.9 ± 5.59
L_II-V	86.90 ± 2.09	87.05 ± 2.10	85.33 ± 2.54	85.45 ± 2.70	85.28 ± 2.52	4.17 ± 1.39	4.45 ± 1.73	12.68 ± 41.90	4.99 ± 1.85	4.88 ± 1.95	88.37 ± 5.11	88.47 ± 4.55	85.53 ± 5.45	85.57 ± 5.62	85.31 ± 5.82
L_Ib-V	86.60 ± 2.06	86.67 ± 2.17	85.20 ± 2.38	85.33 ± 2.47	85.11 ± 2.32	3.98 ± 1.26	4.37 ± 2.18	11.92 ± 42.41	4.31 ± 1.00	4.25 ± 1.13	88.17 ± 4.78	88.65 ± 3.94	85.55 ± 4.85	85.63 ± 5.0	85.82 ± 5.18
B_II+ III+ Va	86.99 ± 1.67	86.84 ± 2.22	84.54 ± 2.41	84.56 ± 2.06	84.47 ± 1.95	4.66 ± 1.43	4.93 ± 1.71	13.55 ± 43.44	6.11 ± 1.77	6.13 ± 1.96	87.81 ± 4.56	88.26 ± 3.78	85.33 ± 4.5	85.35 ± 4.72	85.63 ± 4.88
B_II-V	86.93 ± 1.98	87.20 ± 1.78	85.51 ± 2.28	85.58 ± 2.33	85.64 ± 2.28	4.23 ± 1.28	4.19 ± 1.05	11.99 ± 41.90	4.71 ± 1.20	4.55 ± 1.26	87.86 ± 4.87	88.62 ± 3.92	85.71 ± 5.13	85.83 ± 5.23	86.45 ± 5.18
B_Ib-V	86.45 ± 1.92	86.71 ± 1.69	85.04 ± 2.19	85.08 ± 2.25	85.18 ± 2.21	4.17 ± 1.11	4.13 ± 0.93	11.90 ± 41.63	4.63 ± 1.07	4.46 ± 1.20	88.58 ± 4.91	87.54 ± 4.3	86.21 ± 4.88	86.17 ± 4.95	85.94 ± 5.16
Mean	83.81	84.07	81.83	81.91	81.99	4.32	4.78	11.84	6.10	5.37	86.18	86.38	83.04	82.96	83.2
External testing cohort3															
R_II+ III+ Va	88.64 ± 1.82	87.53 ± 2.23	85.24 ± 2.93	85.93 ± 1.94	85.49 ± 2.33	5.53 ± 5.38	5.88 ± 5.36	6.79 ± 5.11	6.39 ± 2.69	7.92 ± 5.98	90.44 ± 3.79	86.54 ± 5.7	85.3 ± 4.93	86.29 ± 3.38	85.27 ± 3.38
L_II+ III+ Va	88.14 ± 1.68	83.52 ± 2.82	84.94 ± 2.47	85.34 ± 2.38	84.62 ± 2.88	4.74 ± 1.98	4.50 ± 1.34	6.39 ± 2.88	7.02 ± 3.45	7.70 ± 4.48	89.61 ± 4.09	88.17 ± 5.66	84.8 ± 5.45	85.7 ± 4.95	85.16 ± 5.21
R_IV+ Vb+ Vc	84.28 ± 3.82	81.31 ± 4.34	79.12 ± 4.80	79.49 ± 5.28	79.94 ± 4.45	4.33 ± 1.56	5.02 ± 2.73	7.30 ± 3.26	6.38 ± 2.76	6.25 ± 2.28	82.21 ± 7.4	76.07 ± 8.05	74.44 ± 7.94	75.78 ± 7.67	76.01 ± 6.55
L_IV+ Vb+ Vc	84.12 ± 3.84	80.84 ± 7.45	78.18 ± 6.95	78.25 ± 6.36	77.94 ± 6.07	4.53 ± 1.94	4.98 ± 2.11	9.74 ± 10.52	7.58 ± 4.13	8.30 ± 4.04	83.21 ± 7.63	76.73 ± 8.8	73.74 ± 11.64	75.1 ± 10.34	74.95 ± 9.74
R_Ib	76.44 ± 5.61	77.59 ± 4.72	77.53 ± 5.13	75.17 ± 8.35	77.25 ± 5.72	3.59 ± 1.17	4.05 ± 1.26	3.36 ± 0.75	3.76 ± 1.27	3.18 ± 0.77	85.04 ± 6.94	84.07 ± 9.44	86.26 ± 6.29	84.54 ± 9.51	86.95 ± 7.54
L_Ib	76.64 ± 6.92	74.59 ± 9.65	73.77 ± 7.37	73.40 ± 7.82	71.91 ± 8.56	3.82 ± 1.74	4.96 ± 1.36	3.64 ± 0.92	3.94 ± 1.49	4.26 ± 1.68	87.13 ± 7.18	85.8 ± 7.49	84.52 ± 8.43	83.81 ± 8.95	82.8 ± 10.09
R_II-V	88.42 ± 1.42	86.71 ± 2.73	86.54 ± 2.68	86.83 ± 2.29	86.95 ± 2.16	4.25 ± 1.19	4.43 ± 1.34	4.61 ± 1.24	4.62 ± 1.40	4.42 ± 1.10	90.11 ± 3.4	85.37 ± 5.8	85.08 ± 4.25	86.01 ± 3.47	85.25 ± 3.65
R_Ib-V	87.98 ± 1.37	86.24 ± 2.53	86.26 ± 2.68	86.57 ± 2.30	86.71 ± 2.15	3.96 ± 1.04	4.68 ± 1.35	4.28 ± 0.98	4.19 ± 1.09	4.04 ± 0.89	89.91 ± 3.33	86.73 ± 4.85	86.84 ± 4.03	87.72 ± 3.74	87.51 ± 3.64
L_II-V	88.34 ± 1.59	87.33 ± 3.27	86.74 ± 2.68	86.90 ± 2.12	86.63 ± 2.61	4.23 ± 1.52	4.43 ± 1.80	4.36 ± 1.38	4.43 ± 1.31	4.47 ± 1.53	90.47 ± 3.82	88.38 ± 5.16	88.02 ± 5.34	88.15 ± 4.93	88.07 ± 5.11

Continued

Variable	DSC (%)					HD95 (mm)					NSD (%)				
	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR	nnUNet (mixup)	nnUNet (baseline)	3D UXNET	REPUXNET	SwinUNETR
L_Ib-V	87.92 ± 1.65	86.84 ± 3.23	86.26 ± 2.53	86.49 ± 2.24	86.11 ± 2.62	4.14 ± 1.28	4.23 ± 1.52	4.08 ± 1.21	4.20 ± 1.24	4.18 ± 1.30	90.26 ± 3.16	87.63 ± 5.41	87.49 ± 4.13	87.99 ± 3.85	87.86 ± 3.93
B_ II+III+Va	88.38 ± 1.59	87.99 ± 2.72	85.07 ± 2.30	85.62 ± 1.79	85.06 ± 2.34	4.67 ± 2.10	4.72 ± 1.47	6.43 ± 2.10	6.68 ± 2.74	7.50 ± 3.76	89.91 ± 3.14	87.76 ± 5.22	87.32 ± 3.97	87.67 ± 3.82	87.61 ± 3.98
B_II-V	88.38 ± 1.33	87.00 ± 1.66	86.64 ± 2.38	86.87 ± 2.01	86.80 ± 2.26	4.26 ± 1.19	5.28 ± 1.21	4.40 ± 1.05	4.51 ± 1.16	4.42 ± 1.02	89.81 ± 2.87	88.66 ± 4.5	87.14 ± 3.98	88.08 ± 3.96	87.83 ± 3.86
B_Ib-V	87.88 ± 1.27	86.61 ± 2.12	86.19 ± 2.34	86.35 ± 1.98	86.30 ± 2.27	4.10 ± 1.04	4.66 ± 1.14	4.21 ± 0.98	4.38 ± 1.05	4.22 ± 0.93	90.46 ± 3.81	88.33 ± 6.77	87.89 ± 5.2	88.19 ± 5.06	87.76 ± 5.04
Mean	85.81	84.32	83.26	83.32	83.21	4.32	4.75	5.35	5.24	5.45	88.35	85.40	84.53	85.0	84.85
External testing cohort4															
R_ II+III+Va	87.80 ± 3.59	86.69 ± 3.52	82.22 ± 5.44	84.47 ± 4.55	83.80 ± 4.98	4.36 ± 1.76	4.67 ± 1.56	17.04 ± 43.53	6.09 ± 2.45	6.66 ± 2.65	88.28 ± 5.65	87.29 ± 6.84	80.99 ± 8.84	83.92 ± 7.31	83.1 ± 7.84
L_ II+III+Va	86.65 ± 3.09	85.28 ± 4.98	79.70 ± 6.17	79.60 ± 6.87	80.77 ± 6.09	4.96 ± 2.89	6.58 ± 5.71	20.71 ± 51.40	21.19 ± 50.75	9.88 ± 8.80	87.37 ± 5.02	85.03 ± 8.16	79.0 ± 8.27	78.86 ± 9.02	79.9 ± 7.87
R_ IV+Vb+Vc	82.25 ± 5.08	80.05 ± 7.82	75.64 ± 8.06	78.46 ± 6.79	77.11 ± 8.04	4.39 ± 1.65	7.22 ± 5.33	21.43 ± 57.89	7.18 ± 4.78	8.06 ± 4.81	76.95 ± 9.61	74.99 ± 11.38	69.05 ± 11.75	73.61 ± 10.51	71.59 ± 11.28
L_ IV+Vb+Vc	81.65 ± 5.14	80.35 ± 6.12	73.92 ± 7.82	75.82 ± 7.18	76.65 ± 6.97	4.90 ± 2.49	8.90 ± 6.43	29.13 ± 65.91	13.54 ± 14.33	11.97 ± 14.11	77.56 ± 8.11	76.94 ± 11.14	67.14 ± 10.94	70.41 ± 10.9	72.02 ± 9.76
R_Ib	78.73 ± 4.25	75.00 ± 5.76	76.23 ± 6.52	68.95 ± 10.66	74.80 ± 7.13	3.05 ± 0.28	3.90 ± 1.30	3.62 ± 1.09	4.43 ± 1.84	3.51 ± 0.90	86.49 ± 6.0	84.93 ± 6.95	84.4 ± 7.37	76.37 ± 11.09	83.83 ± 8.12
L_Ib	71.19 ± 14.64	71.35 ± 8.08	68.95 ± 13.26	69.19 ± 11.15	68.88 ± 10.89	5.06 ± 2.81	4.88 ± 1.71	4.69 ± 2.10	4.50 ± 1.77	4.55 ± 1.87	79.74 ± 14.57	79.39 ± 8.3	77.63 ± 12.71	77.46 ± 11.46	77.48 ± 10.73
R_II-V	88.52 ± 3.10	86.87 ± 4.57	85.07 ± 5.11	85.98 ± 4.12	85.85 ± 4.61	3.90 ± 1.44	4.00 ± 2.01	16.37 ± 57.86	4.64 ± 1.57	4.45 ± 1.57	87.74 ± 4.16	84.07 ± 6.37	79.92 ± 7.6	81.37 ± 6.64	81.46 ± 6.16
R_Ib-V	88.26 ± 2.82	86.64 ± 4.36	84.91 ± 4.85	85.56 ± 3.84	85.65 ± 4.27	3.62 ± 1.21	5.74 ± 6.07	15.98 ± 57.39	4.19 ± 1.50	4.06 ± 1.41	89.56 ± 5.48	85.93 ± 8.57	84.07 ± 8.79	85.92 ± 6.76	85.7 ± 7.6
L_II-V	87.77 ± 2.80	84.81 ± 5.33	83.27 ± 5.65	83.52 ± 5.34	83.85 ± 5.55	4.38 ± 2.74	7.07 ± 4.79	19.93 ± 61.52	17.19 ± 46.20	7.86 ± 8.78	89.29 ± 5.29	86.64 ± 8.88	82.48 ± 9.01	83.07 ± 8.24	83.69 ± 8.32
L_Ib-V	87.03 ± 3.01	84.61 ± 5.10	82.77 ± 5.67	83.25 ± 5.18	83.30 ± 5.47	4.22 ± 2.37	6.31 ± 4.42	17.56 ± 59.49	14.94 ± 45.70	5.83 ± 4.71	89.57 ± 4.41	87.89 ± 7.49	83.36 ± 7.58	84.61 ± 6.22	84.81 ± 6.58
B_ II+III+Va	87.20 ± 2.87	86.50 ± 3.62	80.92 ± 5.37	82.11 ± 4.74	82.29 ± 4.62	4.83 ± 2.20	5.85 ± 3.84	19.74 ± 49.65	19.20 ± 50.32	8.76 ± 6.71	89.05 ± 4.04	87.36 ± 6.69	83.35 ± 6.96	84.31 ± 5.85	84.68 ± 6.06
B_II-V	88.18 ± 2.43	85.92 ± 4.26	84.18 ± 4.61	84.81 ± 3.83	84.90 ± 4.08	4.19 ± 1.90	6.07 ± 3.00	17.30 ± 58.86	14.45 ± 43.61	5.48 ± 3.20	89.89 ± 4.92	88.58 ± 6.13	84.92 ± 8.01	86.11 ± 6.16	86.43 ± 6.85
B_Ib-V	87.53 ± 2.35	85.58 ± 3.89	83.61 ± 4.52	84.13 ± 3.78	84.29 ± 4.02	4.04 ± 1.60	5.83 ± 2.60	17.12 ± 58.73	13.64 ± 40.76	5.18 ± 2.82	88.97 ± 5.1	87.87 ± 8.36	83.05 ± 8.44	83.82 ± 7.61	83.97 ± 7.72
Mean	84.83	83.05	80.11	80.45	80.93	4.30	5.92	16.97	11.17	6.63	86.19	84.38	79.95	80.76	81.44

Table 6. Segmentation results across several SOTA methods in the internal and external testing cohorts.

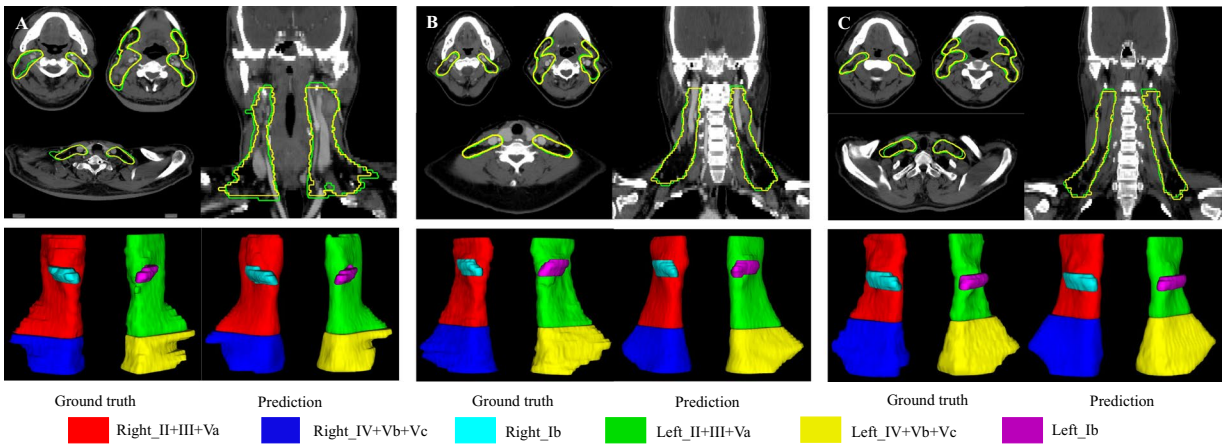


Fig. 4 Visual examples of the nnUNet(mixup) predictions in internal testing (**A**), external testing 1(**B**), and external testing 2 cohorts (**C**). We each ranked all patients in the internal testing cohort, external testing cohort 1, and external testing cohort 2 and selected the median scores of patients in the term of DSC. Yellow lines denoted the experts' delineated contours, and green lines denoted the model-predicted ones.

designed and manually delineated six basic LN CTVs based on the clinical anatomical structures, including the left (L)_Ib, L_II+III+Va, L_IV+Vb+Vc, right (R)_Ib, R_II+III+Va and R_IV+Vb+Vc. The residual seven LN CTVs can be generated by an anatomical-guided adaptive combination. For instance, combining L_II+III+Va with L_IV+Vb+Vc gives L_II-V. Hence, in addition to the six CTVs, the DL model can produce L_Ib-V, L_II-V, R_Ib-V, R_II-V, bilateral (B)_II+III+Va, B_Ib-V, and B_II-V. In total, 13 LN CTVs can automatically be produced by the DL model. The clinical and anatomical significance of 13 LN CTVs are shown in Table 2 and it provides the adaptive combination anatomical guidelines. Based on the above definitions, all radiation oncologists used ITK-SNAP[32] to delineate CTV manually.

Clinically delineation guidelines. In terms of level IIb, we adapted the ranges according to the previous work[19]. Specifically, the cranial boundary of level IIb is set at the cranial edge of C1 (the first cervical spine), the medial boundary is defined by the lateral edges of the rectus capitis lateralis, obliquus capitis superior, obliquus capitis inferior, and elevator scapulae muscles and the posterior range extends to the C1 and C2 levels, while the gap between sternocleidomastoid and splenius capitis muscles could be removed (Fig. 3A, red arrow). The caudal, lateral, and anterior boundaries remain consistent with the 2013 consensus guidelines. Regarding level Ib, we made modifications based on the previous studies[17,18]. First, we spared the submandibular gland (SMG) during delineation (Fig. 3D, yellow arrow). Second, based on the distance between all positive LNs and anatomical landmarks in level Ib, the Medial Mandibular sub-level was omitted (Fig. 3D, green arrow). For level III, we primarily modified the anterior boundary following the previous work[17]. The anterior edge of level III is defined as the region close to the anterior edge of the carotid sheath (Fig. 3E, purple arrow), and we spared the space between the anterior of the sternocleidomastoid muscle and the anterior neck ribbon muscles (Fig. 3E, dark blue arrow). Regarding level V, we used the cervical fascia anatomy-oriented nodal delineation. The lateral border of level V is defined by the scapular hyoid muscle (Fig. 3G, light blue arrow), and the medial border is close to the anterior border of the carotid sheath. To ensure the precise contours, S.C. Zhang (MD, with more than twenty years of experience in head and neck radiation therapy), Y. Zhao (MD, with more than ten years of experience in head and neck radiation therapy) and W.J. Liao (MD, with more than ten years of experience in head and neck radiation therapy), who were the main authors of the three clinical guideline papers[17–19] involved in LN levels delineation in NPC, participated in the ground truth delineation for LN CTVs in this study. These experts first manually delineated several clinical classical patients to provide annotation references for multiple expert boards to minimize the variations in annotation style. These modifications and references were applied to enhance the accuracy and clinical relevance of the DL model.

Data split. We randomly split into 120 patients for training and 30 patients for internal testing (validation cohort). External cohorts with different modalities encompassed two external testing cohorts: 1) Validation cohort 1, collected from SPH and comprised 60 ceCT images; 2) Validation cohort 2 was collected from APH and included 32 ueCT images. External cohorts with different treatment strategies consisted of two additional external testing cohorts (Validation cohorts 3 and 4), including CT images obtained before and post treatments: 1) Validation cohort 3, collected from SMU, comprised 12 patients, each with induction chemotherapy (IC)-naïve and post-IC ceCT images (referred to as CT_{preIC} and CT_{postIC}, respectively); 2) Validation cohort 4, collected from SCH, included eight patients who underwent three repeated ceCT scans during adaptative RT. These scans were performed after 0, 15-20, and 25-30 fractions and labeled as CT_{RA}, CT_{RB}, and CT_{RC}, respectively (the detailed scan intervals are presented in Table 3). Table 1 presents the detailed clinical and imaging characteristics of each cohort over the LNCTVSeg dataset. The inclusion of CT images from various centers and manufacturers enhances the heterogeneity and representativeness of our data. Meanwhile, it can be observed that the clinical and image

characteristics of these cohorts are distributed in a similar space. To protect the patients' private information, we blurred the face region and removed all metadata about clinical and personal[33].

Data Records

The dataset without private information is available on Figshare[34]. Then, all files are stored with the format NIfTI-1 (32-bit floating point). Each cohort is stored as a folder and named "Internal_Cohort", and "Testing_Cohort_X" ("X" is from 1 to 4), Raw CT images in the training or testing sets are stored in the "imagesTr" or "imagesTs" folders of each cohort with the name "lnctvseg_cx_xxxx_yyyy.nii.gz", respectively, where "cx" denotes the center number x is the digit number from 1 to 4, "xxxx" is the index number of each patient the range is 1 to the total number of patients in "cx", "yyyy" is set to "0000" (non-contrast CT) or "0001" (contrast-enhanced CT). The corresponding labels are stored in the "labelsTr" and "labelsTs" folders with the name "lnctvseg_cx_xxxx.nii.gz", respectively. The clinical baseline and CTV labelling index are presented in Tables 1 and 2.

Technical Validation

Experiment settings. To benchmark the LN CTV segmentation task, we use the nnUNet as a baseline and also investigate the results of several recent transformer-based architectures, including 1) 3D UXNET[35], a hyper-model combines large kernel size convolutions and hierarchical transformer for medical image segmentation, which can enlarge global receptive fields and reduce the model parameters and perform very well on several medical datasets; 2) REPUXNET[36], a pure 3D CNN model that performs element-wise scaling in large kernel weights to enhance the learning convergence and effectively adapt large receptive field for volumetric segmentation, which achieved the SOTA performance on abdominal organ and lesion segmentation; 3) SwinUNETR[37], a new pure transformer model that merges the merits of Swin Transformers and UNet for medical image segmentation and shows encouraging performance on several medical segmentation tasks. For a fair comparison, we used their official implementations with provided pre-trained models and trained them with the same settings. Considering the huge variations in CT imaging protocols across different centers, encompassing different vendors, parameters, and the presence or absence of contrast enhancement, we extended the baseline nnUNet with mixup-based data augmentation[38,39] to improve the generalization ability on heterogeneous images.

Evaluation metrics. Three metrics were used to evaluate segmentations, including DSC, the 95% Hausdorff distance (HD95), and normalized surface Dice (NSD) with a tolerance distance 2 mm[40]. The DSC measures the volumetric overlap between two segmentations, and the HD95 (mm) measures the boundary distance between two segmentations. Generally, a superior model should generate a higher value for DSC (maximum of 1) and a lower value for HD95 (minimum of 0). The NSD quantifies the overlap of the surfaces between two segmentations, with a perfect match scoring 1. If a target is missed, the DSC, HD95 and NSD will be set to 0, 100 and 0, respectively. We employed the DSC, HD95 and NSD implementations of the MedPy.

Intraobserver variability. In this study, twelve CT images were randomly selected from SCH and re-contoured by the same expert who had taken part in ground truth generation after two months. The DSC was calculated to assess intraobserver variability between the original and recontoured delineations for various LN CTVs. Meanwhile, we also measured the DSC between the original and the DL-generated contours (nnUNet with mixup). The results are listed in Table 4. The DSC between the original and the model-predicted contours was higher than between the original and recontoured LN CTVs, although a statistical difference was not obtained. Moreover, the variance in observed DSC for LN CTVs was lower for DL auto-segmentation than for recontouring by the expert, except for bilateral levels Ib.

Interobserver variability. we randomly selected 30 patients from SCH and invited two expert groups (SPH and APH) to independently segment the LN CTVs for these patients. Together with the initial segmentations by the SCH group, we obtained LN CTVs delineated by three expert groups (SCH, SPH, and APH). We calculated the DSC to assess interobserver variability, as shown in Table 5. For example, the DSC among the three groups around from 80.0% to 89% for four basic LN CTVs except for levels_Ib. It is indicated that although multiple expert groups delineated the target areas, the consistency of the target areas delineated by different groups was relatively good when all experts strictly followed our delineation principles and requirements.

Experimental results. Table 6 shows the quantitative segmentation results of these methods both in the internal and external cohorts. It can be found that the baseline nnUNet and the modified nnUNet perform better than the other three transformer-based models in terms of mean DSC (mDSC) on all testing cohorts. Meanwhile, the modification of the baseline also improved the baseline performance by a slight margin. Besides, the results in the term of NSD also show a similar trend compared with DSC, where the CNN-based segmentation models consistently outperform transformer-based models. (Fig. 4 shows some segmentation results.)

Limitation. Unlike previous studies including level Ia segmentation[13,25], our study did not involve it, which might limit the dataset's and model's broader applicability. That's because the patients in this dataset consist of NPC, one of the main subtypes of HNC. In NPC, LN metastases are rarely observed in this area, and thus this area is not commonly included in NPC RT.

Usage Notes

The Dataset is publicly available on Figshare under the Creative Commons 4.0 Attribution (CC-BY-4.0). All NIfTI files can be visualized via ITK-SNAP[32] and other packages.

Code availability

All code used in this work are public available, nnUNet, 3D-UXNet, REPUXNET and SwinUNETR. The pre-trained models have been released at Model. The evaluation metrics are at MedPy.

Received: 19 February 2024; Accepted: 17 September 2024;
Published online: 04 October 2024

References

1. Sun, X.-S. *et al*. The association between the development of radiation therapy, image technology, and chemotherapy, and the survival of patients with nasopharyngeal carcinoma: a cohort study from 1990 to 2012. *International Journal of Radiation Oncology* Biology* Physics* **105**, 581–590 (2019).
2. Chen, Y.-P. *et al*. Nasopharyngeal carcinoma. *The Lancet* **394**, 64–80 (2019).
3. van der Veen, J., Gulyban, A. & Nuyts, S. Interobserver variability in delineation of target volumes in head and neck cancer. *Radiotherapy and Oncology* **137**, 9–15 (2019).
4. Peng, Y.-l. *et al*. Interobserver variations in the delineation of target volumes and organs at risk and their impact on dose distribution in intensity-modulated radiation therapy for nasopharyngeal carcinoma. *Oral oncology* **82**, 1–7 (2018).
5. Ding, G. X. *et al*. A study on adaptive imrt treatment planning using kv cone-beam ct. *Radiotherapy and Oncology* **85**, 116–125 (2007).
6. Huang, H. *et al*. Determining appropriate timing of adaptive radiation therapy for nasopharyngeal carcinoma during intensity-modulated radiation therapy. *Radiation Oncology* **10**, 1–9 (2015).
7. Lin, L. *et al*. Deep learning for automated contouring of primary tumor volumes by mri for nasopharyngeal carcinoma. *Radiology* **291**, 677–686 (2019).
8. Liao, W. *et al*. Automatic delineation of gross tumor volume based on magnetic resonance imaging by performing a novel semisupervised learning framework in nasopharyngeal carcinoma. *International Journal of Radiation Oncology* Biology* Physics* **113**, 893–902 (2022).
9. Litjens, G. *et al*. A survey on deep learning in medical image analysis. *Medical image analysis* **42**, 60–88 (2017).
10. Chen, X. *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* **160**, 175–184 (2021).
11. Luo, X. *et al*. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. arXiv preprint arXiv:2312.09576 (2023).
12. Van der Veen, J., Willems, S., Bollen, H., Maes, F. & Nuyts, S. Deep learning for elective neck delineation: More consistent and time efficient. *Radiotherapy and Oncology* **153**, 180–188 (2020).
13. Cardenas, C. E. *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. *International Journal of Radiation Oncology* Biology* Physics* **109**, 801–812 (2021).
14. Grégoire, V. *et al*. Delineation of the neck node levels for head and neck tumors: a 2013 update. dahanca, eortc, hknpcsg, ncic ctg, ncri, rtog, trog consensus guidelines. *Radiotherapy and Oncology* **110**, 172–181 (2014).
15. Tang, L. *et al*. The volume to be irradiated during selective neck irradiation in nasopharyngeal carcinoma: analysis of the spread patterns in lymph nodes by magnetic resonance imaging. *Cancer* **115**, 680–688 (2009).
16. King, A. D. *et al*. Neck node metastases from nasopharyngeal carcinoma: Mr imaging of patterns of disease. *Head & Neck: Journal for the Sciences and Specialties of the Head and Neck* **22**, 275–281 (2000).
17. Lin, L. *et al*. Delineation of neck clinical target volume specific to nasopharyngeal carcinoma based on lymph node distribution and the international consensus guidelines. *International Journal of Radiation Oncology* Biology* Physics* **100**, 891–902 (2018).
18. Zhao, Y. *et al*. Level ib ctv delineation in nasopharyngeal carcinoma based on lymph node distribution and topographic anatomy. *Radiotherapy and Oncology* **172**, 10–17 (2022).
19. Zhang, J. *et al*. Level iib ctv delineation based on cervical fascia anatomy in nasopharyngeal cancer. *Radiotherapy and Oncology* **115**, 46–49 (2015).
20. Chen, J.-z. *et al*. Results of a phase 2 study examining the effects of omitting elective neck irradiation to nodal levels iv and vb in patients with n0-1 nasopharyngeal carcinoma. *International Journal of Radiation Oncology* Biology* Physics* **85**, 929–934 (2013).
21. Li, J.-G. *et al*. A randomized clinical trial comparing prophylactic upper versus whole-neck irradiation in the treatment of patients with node-negative nasopharyngeal carcinoma. *Cancer* **119**, 3170–3176 (2013).
22. Huang, C.-L. *et al*. Upper-neck versus whole-neck irradiation at the contralateral uninvolved neck in patients with unilateral n3 nasopharyngeal carcinoma. *International Journal of Radiation Oncology* Biology* Physics* **116**, 788–796 (2023).
23. Tang, L.-L. *et al*. Elective upper-neck versus whole-neck irradiation of the uninvolved neck in patients with nasopharyngeal carcinoma: an open-label, non-inferiority, multicentre, randomised phase 3 trial. *The Lancet Oncology* **23**, 479–490 (2022).
24. Men, K. *et al*. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Frontiers in oncology* **7**, 315 (2017).
25. Weissmann, T. *et al*. Deep learning for automatic head and neck lymph node level delineation provides expert-level accuracy. *Frontiers in Oncology* **13**, 1115258 (2023).
26. Luo, X. *et al*. Word: A large scale dataset, benchmark and clinical applicable study for abdominal organ segmentation from ct image. *arXiv preprint arXiv:2111.02403* (2021).
27. Liao, W. *et al*. Comprehensive evaluation of a deep learning model for automatic organs at risk segmentation on heterogeneous computed tomography images for abdominal radiotherapy. *International Journal of Radiation Oncology* Biology* Physics* (2023).
28. Menze, B. H. *et al*. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging* **34**, 1993–2024 (2014).
29. Grossberg, A. J. *et al*. Imaging and clinical data archive for head and neck squamous cell carcinoma patients treated with radiotherapy. *Scientific data* **5**, 1–10 (2018).
30. Bejarano, T., De Ornelas-Couto, M. & Mihaylov, I. B. Longitudinal fan-beam computed tomography dataset for head-and-neck squamous cell carcinoma patients. *Medical physics* **46**, 2526–2537 (2019).
31. Elhalawani, H. *et al*. Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges. *Sci. data* **4**, 170077 (2017).
32. Yushkevich, P. A. *et al*. User-guided 3d active contour segmentation of anatomical structures: significantly improved efficiency and reliability. *Neuroimage* **31**, 1116–1128 (2006).
33. Collins, S. A., Wu, J. & Bai, H. X. Facial de-identification of head ct scans. *Radiology* **296**, 22–22 (2020).
34. Luo, X. *et al*. A multicenter dataset for lymph node clinical target volume delineation of nasopharyngeal carcinoma. *figshare* https://doi.org/10.6084/m9.figshare.26793622.v3 (2019).
35. Lee, H. H., Bao, S., Huo, Y. & Landman, B. A. 3d ux-net: A large kernel volumetric convnet modernizing hierarchical transformer for medical image segmentation. *arXiv preprint arXiv:2209.15076* (2022).
36. Lee, H. H. *et al*. Scaling up 3d kernels with bayesian frequency re-parameterization for medical image segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention, 632–641 (Springer, 2023).

37. Tang, Y. *et al*. Self-supervised pre-training of swin transformers for 3d medical image analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 20730–20740 (2022).
38. Zhang, H., Cisse, M., Dauphin, Y. N. & Lopez-Paz, D. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412* (2017).
39. Zhang, L. *et al*. Generalizing deep learning for medical image segmentation to unseen domains via deep stacked transformation. *IEEE transactions on medical imaging* **39**, 2531–2540 (2020).
40. Antonelli, M. *et al*. The medical segmentation decathlon. *Nat. communications* **13**, 4128 (2022).

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grant 82203197 and 62271115, the Sichuan Science and Technology Program, China (Grant 2023NSFSC1852 and 2023NSFSC0720), and the Sichuan Provincial Cadre Health Research Project (Grant/award number: 2023-803).

Author contributions

L.W.J., L.X.D., and Z.S.T. conceived of the presented idea; L.W.J., Z.Y., Q.Y.J., H.Y., H.H., L.L., and Z.S.C. performed and confirmed the segmentations; L.X.D., and F.J. performed full data anonymization; L.W.J., X.J.F., and H.Y. collected data; L.W.J., and L.X.D. wrote the initial draft. W.G.T., Z.S.C., and Z.S.T. completed the critical review and revision of the manuscript and datasets, as well as proofreading.

Competing interests

Te authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to W.L.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

© The Author(s) 2024