

Deep Learning for Automated Contouring of Primary Tumor Volumes by MRI for Nasopharyngeal Carcinoma

Li Lin, PhD* • Qi Dou, PhD* • Yue-Ming Jin, BS • Guan-Qun Zhou, PhD • Yi-Qiang Tang, MD • Wei-Lin Chen, MD • Bao-An Su, MD • Feng Liu, MD • Chang-Juan Tao, MD • Ning Jiang, PhD • Jun-Yun Li, PhD • Ling-Long Tang, PhD • Chuan-Miao Xie, MD • Shao-Min Huang, BS • Jun Ma, MD • Pheng-Ann Heng, PhD • Joseph T. S. Wee, MD • Melvin L. K. Chua, PhD • Hao Chen, PhD • Ying Sun, PhD



From the Department of Radiation Oncology (L.L., G.Q.Z., J.Y.L., L.L.T., S.M.H., J.M., Y.S.) and Imaging Diagnosis and Interventional Center (C.M.X.), Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China, Collaborative Innovation Center for Cancer Medicine, Guangdong Key Laboratory of Nasopharyngeal Carcinoma Diagnosis and Therapy, 651 Dongfeng Rd East, Guangzhou 510060, China; Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong SAR (Q.D., Y.M.J., P.A.H., H.C.); Insight Medical Technology, Shenzhen, China (H.C.); Divisions of Radiation Oncology (J.T.S.W., M.L.K.C.) and Medical Sciences (M.L.K.C.), National Cancer Center Singapore, Singapore; Oncology Academic Programme, Duke-NUS Medical School, Singapore (M.L.K.C.); Department of Radiation Oncology, Jiangxi Cancer Hospital, Nanchang, China (Y.Q.T.); Department of Radiation Oncology, Zhangzhou Affiliated Hospital of Fujian Medical University, Zhangzhou, China (W.L.C.); Department of Radiation Oncology, Quanzhou First Hospital Affiliated to Fujian Medical University, Quanzhou, China (B.A.S.); Department of Radiation Oncology, The First Affiliated Hospital of Fujian Medical University, Fuzhou, China (F.L.); Department of Radiation Oncology, Zhejiang Provincial Cancer Hospital, Key Laboratory of Radiation Oncology of Zhejiang Province, Hangzhou, China (C.J.T.); and Department of Radiation Oncology, Nanjing Medical University Affiliated Cancer Hospital, Jiangsu Cancer Hospital and Jiangsu Institute of Cancer Research, Nanjing, China (N.J.). Received August 28, 2018; revision requested October 31; revision received January 31, 2019; accepted February 6. **Address correspondence to** Y.S. (e-mail: sunying@sysucc.org.cn).

Supported by Special Support Program of Sun Yat-sen University Cancer Center (16zxtzlc06); Overseas Expertise Introduction Project for Discipline Innovation (111 Project, B14035); Health & Medical Collaborative Innovation Project of Guangzhou City, China (201604020003, 201803040003); Natural Science Foundation of Guangdong Province (2017A030312003), Sun Yat-sen University Clinical Research 5010 Program (2012011); and Innovation Team Development Plan of the Ministry of Education (IRT_17R110).

* L.L. and Q.D. contributed equally to this work.

Conflicts of interest are listed at the end of this article.

See also the editorial by Chang in this issue.

Radiology 2019; 00:1–10 • <https://doi.org/10.1148/radiol.2019182012> • Content codes:  

Background: Nasopharyngeal carcinoma (NPC) may be cured with radiation therapy. Tumor proximity to critical structures demands accuracy in tumor delineation to avoid toxicities from radiation therapy; however, tumor target contouring for head and neck radiation therapy is labor intensive and highly variable among radiation oncologists.

Purpose: To construct and validate an artificial intelligence (AI) contouring tool to automate primary gross tumor volume (GTV) contouring in patients with NPC.

Materials and Methods: In this retrospective study, MRI data sets covering the nasopharynx from 1021 patients (median age, 47 years; 751 male, 270 female) with NPC between September 2016 and September 2017 were collected and divided into training, validation, and testing cohorts of 715, 103, and 203 patients, respectively. GTV contours were delineated for 1021 patients and were defined by consensus of two experts. A three-dimensional convolutional neural network was applied to 818 training and validation MRI data sets to construct the AI tool, which was tested in 203 independent MRI data sets. Next, the AI tool was compared against eight qualified radiation oncologists in a multicenter evaluation by using a random sample of 20 test MRI examinations. The Wilcoxon matched-pairs signed rank test was used to compare the difference of Dice similarity coefficient (DSC) of pre- versus post-AI assistance.

Results: The AI-generated contours demonstrated a high level of accuracy when compared with ground truth contours at testing in 203 patients (DSC, 0.79; 2.0-mm difference in average surface distance). In multicenter evaluation, AI assistance improved contouring accuracy (five of eight oncologists had a higher median DSC after AI assistance; average median DSC, 0.74 vs 0.78; $P < .001$), reduced intra- and interobserver variation (by 36.4% and 54.5%, respectively), and reduced contouring time (by 39.4%).

Conclusion: The AI contouring tool improved primary gross tumor contouring accuracy of nasopharyngeal carcinoma, which could have a positive impact on tumor control and patient survival.

© RSNA, 2019

Online supplemental material is available for this article.

Tumor target contouring for precision head and neck radiation therapy is labor intensive and highly variable among radiation oncologists (1). Importantly, contouring inaccuracies compromise survival in patients with head and neck cancer (2). Moreover, with the progressive implementation of intensity modulated radiation therapy and proton beam therapy, contouring time of the primary gross tumor volume (GTV) has substantially increased due to the need to

consider multimodal or multiparametric imaging data sets. In tumor contouring, the manual process entails a thorough review of the tumor on diagnostic images and delineation of the GTV on the treatment-planning CT or MRI data set.

When compared with other head and neck cancers, nasopharyngeal carcinoma (NPC) is clinically distinct and exquisitely sensitive to radiation therapy; hence, the majority of these tumors are cured with radiation therapy (3). At

Abbreviations

AI = artificial intelligence, ASD = average surface distance, CNN = convolutional neural network, DSC = Dice similarity coefficient, GTV = gross tumor volume, IC = induction chemotherapy, NPC = nasopharyngeal carcinoma, 3D = three dimensional

Summary

An artificial intelligence contouring tool improved tumor target contouring accuracy for nasopharyngeal carcinoma, which could have a positive impact on tumor control and patient survival.

Key Points

- In this multicenter evaluation, artificial intelligence assistance substantially improved contouring accuracy for five of eight radiation oncologists.
- An artificial intelligence contouring tool reduced intraobserver variation by 36.4%, reduced interobserver variation by 54.5%, and reduced contouring time by 39.4%.

present, intensity modulated radiation therapy is the standard radiation therapy technique for NPC (4). GTV contouring for NPC is labor intensive and error prone, particularly due to the following factors: (a) NPC can infiltrate the adjacent skull base and neural structures, but the extent of involvement is often reflected by subtle signal changes at MRI. (b) The proximity to critical neural and other organs demands accuracy in the delineation of GTV to avoid unnecessary toxicities from radiation therapy. As such, the radiation therapy planning workflow for NPC builds on the experience of the radiation oncologist. Automation of GTV contouring by deep learning, if available, could be advantageous in this context.

Nonetheless, automation of tumor contouring for NPC by deep learning is challenging due to the substantial interpatient heterogeneity in tumor shape and the poorly defined tumor-to-normal tissue interface. More recently, deep convolutional neural networks (CNNs) have emerged as promising alternatives for volumetric medical image segmentation. Success has been achieved by performing liver, heart, and brain segmentation from three-dimensional (3D) images using 3D CNN, and this technique has yielded comparable performance to that of state-of-the-art methods (5,6).

In this study, we investigated the use of deep learning for GTV contouring of NPC. We first constructed an artificial intelligence (AI) contouring tool by applying a 3D CNN model to MRI examinations from a training cohort of 818 patients and subsequently validated its accuracy in a separate testing cohort of 203 patients. Next, the AI tool was compared against eight qualified radiation oncologists in a multicenter evaluation using 20 randomly sampled patients from the testing cohort.

Materials and Methods

One of the authors (H.C.) is an employee of a technology company (Im sight Medical Technology, Shenzhen, China), but we did not receive any financial support, equipment, or contrast agents from his company or any other industry entities. The authors had control of the data and information submitted for publication. This retrospective study was approved by the Sun Yat-sen University Cancer Center institutional review board

(approval no. YB-2017-038); the requirement to obtain informed consent was waived. The authenticity of this article has been validated by uploading the key raw data onto the Research Data Deposit public platform (www.researchdata.org.cn) (RDD no. RDDA2018000927).

MRI Examinations

We retrospectively collected MRI studies of the nasopharynx from patients with histologically proven and radiation therapy-naïve NPC between September 1, 2016, and September 30, 2017, from a single institute (Sun Yat-sen University Cancer Center). MRI examinations were performed with unenhanced T1- and T2-weighted, contrast-enhanced T1-weighted, and fat-suppressed T1-weighted sequences. Details of MRI acquisition are included in Appendix E1 (online). The exclusion criteria are shown in the study flow diagram (Fig 1). The final data set comprised 1021 patients (median age, 47 years; 751 male, 270 female), who were then randomly assigned to three cohorts: (a) a training cohort of 715 patients for 3D CNN construction, (b) a validation cohort of 103 patients for optimization of the 3D CNN hyperparameters, and (c) a testing cohort of 203 patients to test the performance of the AI contouring tool.

Human Expert Delineated GTV Contours

MRI examinations of the 1021 patients were assigned to two expert radiation oncologists (Y.S., L.L.T.; more than 15 years of experience in caring for patients with NPC) to delineate ground truth GTV via consensus. These examinations were used to train and test the AI contouring tool. A third radiologist specializing in head and neck imaging (C.M.X., 25 years of experience) was consulted in cases of disagreement. Details are described in Appendix E1 (online). Figure E1 (online) shows an example of the ground truth contour.

Network Architecture

We implemented 3D CNN to extract representative features for the complicated GTV based on four MRI pulse sequences. Specifically, we designed a full CNN architecture, which was composed of encoder and decoder paths, to conduct the segmentation task. Our network is based on the 3D CNN architecture of *VoxResNet* (6). The detailed network architecture is shown in Figure 2, and we have deposited all computer codes used for modeling and data analysis in GitHub (<https://github.com/AutoContour/NPC>). Detailed descriptions and the training and inference settings of the proposed 3D CNN are presented in Appendix E1 (online).

Performance of the AI Contouring Tool

Performance of the AI contouring tool was evaluated in the testing cohort ($n = 203$) using Dice similarity coefficient (DSC) and average surface distance (ASD). The DSC measures the spatial overlap between the AI-generated contour (A) and the ground truth contour (G), which is defined as: $DSC(A, G) = \frac{2|A \cap G|}{|A| + |G|}$ (7). The ASD counts the average distance between the surfaces of two contours (8).

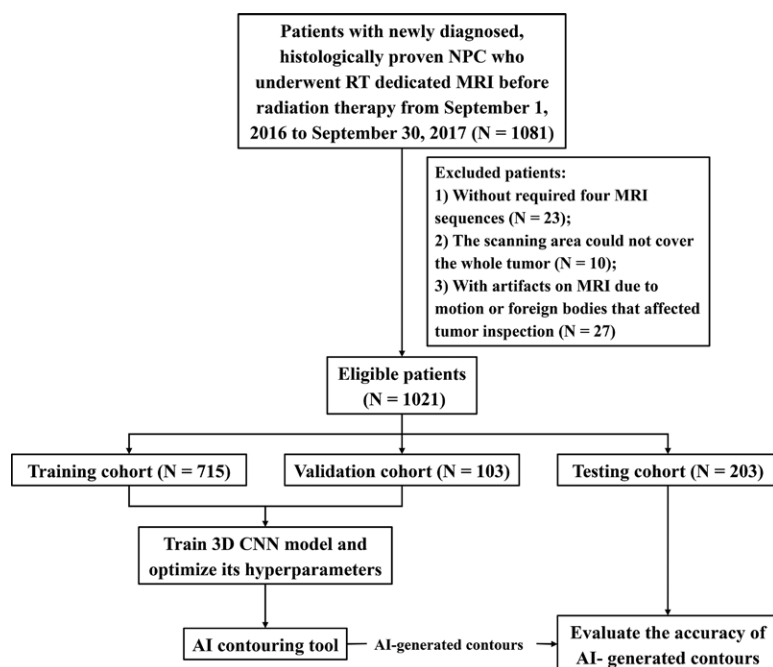


Figure 1: Study flow diagram. AI = artificial intelligence, CNN = convolutional neural network, NPC = nasopharyngeal carcinoma, RT = radiation therapy, 3D = three dimensional.

Additionally, we explored the comparison of both indexes between the following subgroups: chemotherapy-naïve versus post-induction chemotherapy (hereafter, post-IC) tumors and early T category (T1 and T2) versus advanced T category (T3 and T4) tumors. Besides volume-based indexes, section-based indexes of four labeled transverse sections—namely midcavernous sinus, skull base at the level of clivus, Eustachian cushion, and midvula—were also evaluated to determine the performance of our AI contouring tool at different anatomic locations within the tumor.

Finally, we compared the performance of GTV contours generated from our proposed 3D CNN against a 3D U-Net (9); the latter is the commonly used network architecture for medical image segmentation. When training the 3D U-Net, we retained a consistent image preprocessing, normalization, augmentation, and training strategy to ensure a neutral comparison.

Assessment of AI-generated Contours by Human Experts

Given that evaluation indexes do not provide insight into how much the contours would need to be edited to be used in clinical practice, we then asked the experts to further evaluate the applicability of AI-generated contours. Specifically, the AI-generated contours of 203 patients were assigned to the human experts (Y.S., L.L.T.) to grade their accuracy by consensus using volumetric revision magnitudes, defined as the volume needed to be edited divided by the volume of the AI-generated contour, with the result multiplied by 100. Accuracy was classified as follows: no revision required, revisions required for more than 0% to 20% of the volumetric contours, revisions required for more than 20% to 40% of the volumetric contours,

revisions required for more than 40% to 60% of the volumetric contours, revisions required for more than 60% to 80% of the volumetric contours, and revisions required for more than 80% to 100% of the volumetric contours. Next, we explored the correlation between DSC and different magnitudes of volumetric revision and compared the differences in degrees of volumetric revision in T1–T2 versus T3–T4 tumors, as well as prechemotherapy versus post-IC tumors.

Multicenter Evaluation of the AI Contouring Tool

To further evaluate the AI contouring tool, we conducted a multicenter study involving eight qualified radiation oncologists (L.L., Y.Q.T., W.L.C., B.A.S., F.L., C.J.T., N.J., and J.Y.L.) from seven high-volume (≥ 200 NPC cases per year) academic institutions. The radiation oncologists' experience with NPC was ranked according to their total number of contoured NPC targets since January 1, 2013 (Table E1 [online]). First, MRI examinations of 20 randomly sampled patients from the testing cohort, stratified by T category (T1–T2, T3–T4), were distributed

to the eight radiation oncologists for manual contouring. Next, the AI-generated contours were distributed to them for editing after a minimum interval of 2 months. The radiation oncologists were blinded to the ground truth contours, their first set of manual contours, and those by their counterparts. Contouring accuracy was assessed with DSC and ASD. Intraobserver variation was assessed with the interquartile deviation of DSC; interobserver variation was assessed with multiobserver DSC and volume coefficient of variation (the ratio between standard deviation and mean GTV). Times taken for manual, automated AI-only, and AI-assisted contouring were also reported.

Statistical Analysis

Categorical variables for the combined training-validation and testing cohorts were compared by using the χ^2 test or Fisher exact test; numeric variables were compared by using the Mann-Whitney U test. The Mann-Whitney U test was used to compare DSC and ASD between different subgroups. The Wilcoxon matched-pairs signed rank test was used to compare DSC, ASD, interquartile deviation of DSC, volume coefficient of variation, and time taken of AI tool versus manual, pre-versus post-AI assistance, and 3D CNN versus the 3D U-Net method. Kruskal-Wallis one-way analysis of variance was used to compare the median DSC among the magnitudes of volumetric revision. Correlation between the median DSCs and degrees of volumetric revision was assessed with the Spearman correlation coefficient. The χ^2 test was used to compare the difference in degrees of volumetric revision between subgroups. All analyses were performed by using Statistical Product and Service Solutions (IBM SPSS, version 21.0; New York, NY). Statistical significance was set at two-tailed $P < .05$.

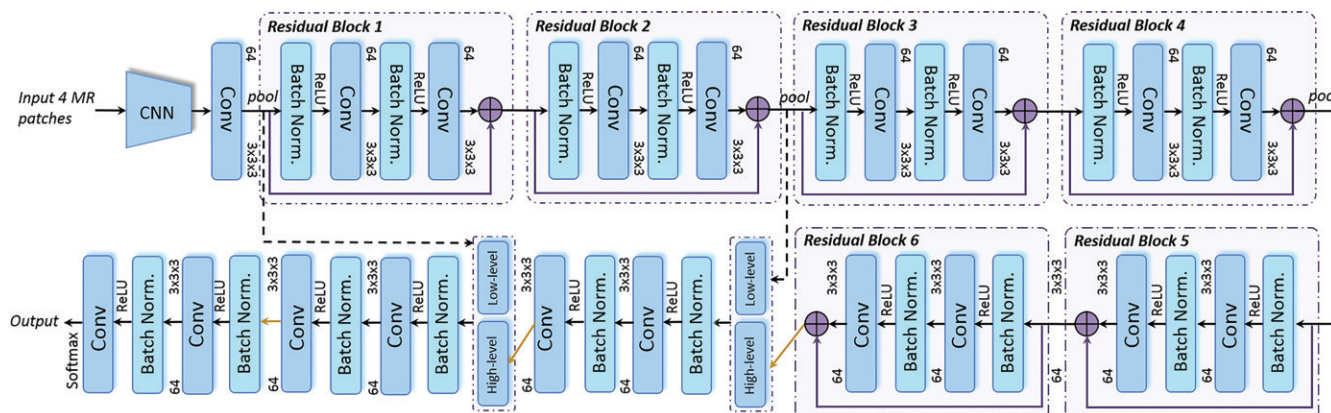


Figure 2: Network architecture of the proposed three-dimensional (3D) convolutional neural network (CNN). The network has 28 layers integrating six residual blocks. Yellow arrows indicate up-sampling operations to make dense predictions for the segmentation task. Skip connections are used to fuse low- and high-level features in the network. The batch normalization is a linear transformation of the features to reduce covariance shift, thus speeding up the training procedure. Convolution bars indicate the convolution operation, which computes the features. The number 64 indicates the number of channels in that layer, and $3 \times 3 \times 3$ denotes the size of the 3D CNN kernels.

Table 1: Clinical, imaging, and Tumor Characteristics

Characteristic	Entire Cohort ($n = 1021$)	Training-Validation Cohort ($n = 818$)	Testing Cohort ($n = 203$)	P Value
Sex81
Male	751 (73.6)	603 (73.7)	148 (72.9)	...
Female	270 (26.4)	215 (26.3)	55 (27.1)	...
Age35
Median	47	46	48	...
<18 y	3 (0.3)	2 (0.2)	1 (0.5)	...
18–60 y	872 (85.4)	706 (86.3)	166 (81.8)	...
≥ 60 y	146 (14.3)	110 (13.5)	36 (17.7)	...
T category24
T1	50 (4.9)	45 (5.5)	5 (2.5)	...
T2	123 (12.0)	94 (11.5)	29 (14.3)	...
T3	601 (58.9)	483 (59.0)	118 (58.1)	...
T4	247 (24.2)	196 (24.0)	51 (25.1)	...
Histologic finding26
WHO type I	2 (0.2)	1 (0.1)	1 (0.5)	...
WHO type II or III	1019 (99.8)	817 (99.9)	202 (99.5)	...
Imaging characteristics
MRI time point36
Chemotherapy naive	692 (67.8)	549 (67.1)	143 (70.4)	...
Post-IC	329 (32.2)	269 (32.9)	60 (29.6)	...
Tumor characteristic
No. of tumor-bearing sections per case	17 (4–41)	17 (4–41)	18 (6–36)	.16
Primary GTV (mL)	34.0 (2.1–268.6)	33.6 (2.1–268.6)	35.8 (4.4–240.5)	.50

Note.—Data are either number of patients, with the percentage in parentheses, or median, with the range in parentheses. We calculated P values by using the χ^2 or Fisher exact test for category variables and the Mann-Whitney U test for numeric variables. Two-tailed $P < .05$ indicated a significant difference. Patients were staged according to the 8th edition of American Joint Committee on Cancer staging manual. GTV = gross tumor volume, IC = induction chemotherapy, WHO = World Health Organization.

Results

Patient Characteristics

The flow diagram of this study is shown in Figure 1. A total of 4084 MRI examinations performed in 1021 patients were included. No significant differences in sex, age, T category, tumor histology,

or imaging and tumor characteristics were observed between the combined training and validation cohort and the testing cohort (Table 1). Additionally, our data set comprised an adequate number of tumors that invaded the distinct anatomic regions (Table E2 [online]). Characteristics of the 20 randomly sampled patients used in multicenter evaluation are presented in Table E3 [online].

Table 2: Accuracy of AI-generated Contours in the Testing Cohort

Volume-based Indexes	Total (<i>n</i> = 203)	Chemotherapy Condition at MRI		<i>P</i> Value	T Category		<i>P</i> Value
		Chemotherapy Naive (<i>n</i> = 143)	Post-IC (<i>n</i> = 60)		T1 and T2 (<i>n</i> = 34)	T3 and T4 (<i>n</i> = 169)	
DSC6553
Median	0.79	0.79	0.79	...	0.78	0.79	...
Interquartile range	0.76–0.81	0.76–0.81	0.75–0.81	...	0.76–0.80	0.76–0.81	...
95% CI for the median	0.78, 0.79	0.78, 0.79	0.78, 0.80	...	0.78, 0.79	0.78, 0.79	...
ASD (mm)13	<.001
Median	2.0	1.9	2.1	...	1.5	2.0	...
Interquartile range	1.6–2.4	1.5–2.4	1.7–2.5	...	1.3–2.0	1.7–2.5	...
95% CI for the median	1.9, 2.1	1.8, 2.0	1.9, 2.2	...	1.3, 1.8	1.9, 2.1	...

Note.—We calculated the *P* value by using Mann-Whitney *U* test. Two-tailed *P* < .05 indicates a significant difference. AI = artificial intelligence, ASD = average surface distance, DSC = Dice similarity coefficient, CI = confidence interval, IC = induction chemotherapy.

Performance of the AI Contouring Tool

Accuracy of the AI-generated contours is summarized in Table 2 (see also Fig E2 [online] for distribution histograms of the indexes). Figure 3 shows the level of concordance for the GTV contours between the AI tool and human experts. We observed a median DSC of 0.79 (interquartile range, 0.76–0.81; 95% confidence interval: 0.78, 0.79) and a median ASD of 2.0 mm (interquartile range, 1.6–2.4 mm; 95% confidence interval: 1.9, 2.1 mm); the latter is less than the commonly accepted 3-mm margin of systematic and random error for radiation therapy for head and neck cancers (10). These results indicate a strong concordance between our AI tool and human experts for GTV contouring.

In the subgroup analyses, the AI tool achieved comparable DSC and ASD between the post-IC and chemotherapy-naïve subgroups (median DSC, 0.79 vs 0.79; *P* = .65; median ASD, 2.1 vs 1.9 mm; *P* = .13; Table 1). For the different T categories, the AI tool achieved a significantly smaller ASD in the early T category tumors than in the advanced T category tumors (median ASD, 1.5 vs 2.0 mm; *P* < .001; Table 1), which could imply better accuracy of GTV contouring by the AI tool for smaller tumors. Nonetheless, we did not observe a difference for DSC between the different T categories. Additionally, other confounders, including age, sex, image resolution, and body mass index showed no impact on contouring accuracy (Appendix E1, Fig E3 [online]).

For section-based analysis, we observed a difference in accuracy of our AI tool at the different anatomic regions (Table E4 [online]); median section-based DSC was higher at the midvolume sections of the skull base (0.82) and the Eustachian cushion (0.83) than at the cranial-caudal sections of the cavernous sinus (0.75) and uvula (0.75), respectively. This may imply a difference in contouring accuracy within subsets of T4 tumors; for example, accuracy may be compromised for tumors that infiltrate superiorly into the cavernous sinus and inferiorly to the oropharynx and hypopharynx versus those that infiltrate laterally into the masticator space. To confirm this finding, we reviewed the AI-generated contours and observed a higher number of contouring “misses” at the oropharynx (*n* = 4), hypopharynx

(*n* = 2), temporal lobe (*n* = 2), petrous apex (*n* = 1), and pituitary fossa (*n* = 1) than at the lateral retropharynx (*n* = 2) or anterior pterygopalatine fossa (*n* = 1).

Finally, when we compared the accuracy of GTV contours generated from our 3D CNN and from 3D U-Net, median DSC and ASD of the 3D U-Net-generated contours were inferior to those generated with our 3D CNN model (median DSC, 0.72 vs 0.79; median ASD, 2.3 vs 2.0 mm; *P* < .001 for both; Table E5 [online]).

Assessment of AI-generated Contours by Human Experts

When using our grading criteria for contour accuracy, the majority (180 of 203 [88.7%]) of the AI-generated contours were deemed satisfactory by the experts (no revision required, *n* = 66; >0%–20% revision, *n* = 114). Only three contours were assessed to require >40%–60% revision, with none requiring >60% revision, thus validating the robustness of our AI contouring tool. Additionally, median DSC was correlated to the magnitudes of revision required (*R* = 0.23, *P* < .001) (Fig 4, A), indicating the reliability of using DSC as a contouring accuracy evaluation criterion.

Similarly, we also observed a higher degree of required volumetric revision in the T3–T4 tumors than in the T1–T2 tumors (13.0% vs 2.9% required >20% revision, *P* = .004) and in the post-IC subgroup than in the chemotherapy-naïve subgroup (20.0% vs 7.7% required >20% revision, *P* = .008) (Fig 4, B). The former is consistent with the larger ASD that was observed for T3–T4 tumors (Table 2).

Multicenter Evaluation of AI Contouring Tool

To further validate our findings, we tested the AI tool against eight qualified radiation oncologists. We used the GTV contours by human experts as ground truth, and our AI tool performed comparably to the assigned radiation oncologists (Fig 5, A and B, Table E6 [online]). For DSC, the AI tool outperformed four of eight radiation oncologists (median DSC: 0.79 vs 0.71, 0.71, 0.72, and 0.74; *P* < .05 for all) and was noninferior to the other four (median DSC: 0.80, 0.79, 0.78,

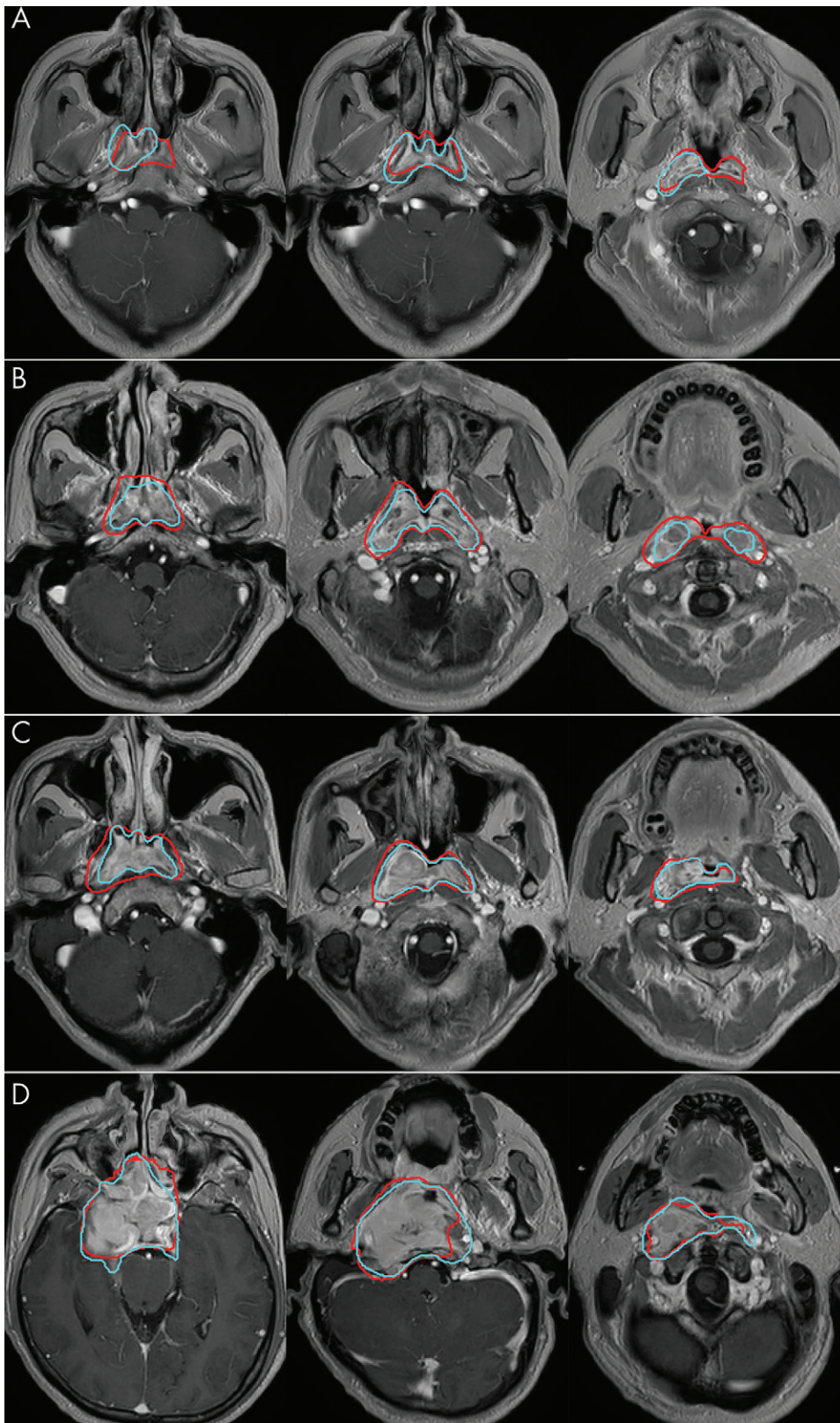


Figure 3: Example contrast-enhanced T1-weighted MRIs show the level of concordance for primary gross tumor volume contours between the artificial intelligence (AI) tool and human experts through superior, median, and inferior sections within the tumor. MRIs were obtained in patients with Dice similarity coefficients of, A, 0.67, B, 0.77, C, 0.82, and, D, 0.86. Light blue lines denote the human experts delineated ground truth, and red lines denote the AI-generated contours.

and 0.78). For ASD, our AI tool outperformed two radiation oncologists, with median ASD of 2.2 mm versus 3.3 and 2.8 mm, respectively (both $P < .05$) and performed comparably to five radiation oncologists, except for one oncologist who achieved an ASD that was smaller than that achieved with the AI tool (R1, 1.6 mm; $P = .02$). Overall, our AI tool achieved smaller interquartile deviation for DSC (0.08 vs 0.08–0.14) and ASD (0.6 mm vs 0.7–2.1 mm) compared with the manual contours of the oncologists (Fig 5, A and B; Table E6 [online]), enabling us to confirm its robustness for delineating GTV in NPC.

Next, we determined if the manual contours could be enhanced by our AI tool. We observed that AI assistance resulted in higher DSC values in five of eight radiation oncologists (Fig 5, C; Table E7 [online]) ($P < .05$ for each comparison; average median DSC, 0.74 vs 0.79; $P < .001$). This would correspond to an overall reduction in percentage of volumetric contours requiring revision (Fig E4 [online]). AI assistance also led to a reduction in intraobserver variation, with smaller interquartile deviation of DSC in seven of eight radiation oncologists after AI assistance (median interquartile deviation, 0.07 [interquartile range, 0.05–0.08] vs 0.11 [interquartile range, 0.09–0.14]; $P = .02$; Table E7 [online]; overall reduction, 36.4%). Figure 6 shows the level of variation in the manual and post-AI-assisted contours by the eight radiation oncologists.

Consistent with the increased accuracy, interobserver variation among the eight radiation oncologists was also reduced (Fig 6); median multiobserver DSC of post-AI-assisted contours was higher than that of manual

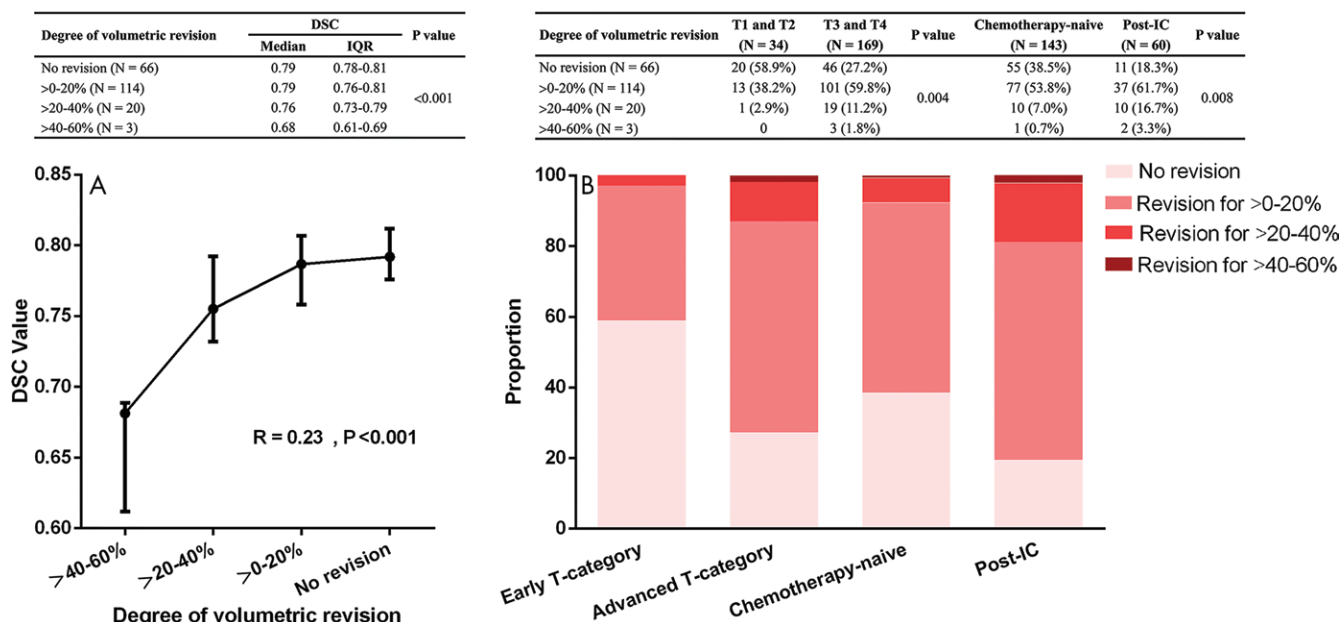


Figure 4: Expert assessment of volumetric revision magnitudes of the artificial intelligence (AI)-generated contours. **A**, Median Dice similarity coefficient (DSC) stratified by different magnitudes of volumetric revision. Kruskal-Wallis one-way analysis of variance was used to compare median DSC values among different magnitudes, and Spearman correlation coefficient showed that median DSC and degree of volumetric revision were correlated ($R = 0.23, P < .001$). AI-generated contours were assessed by the two experts who delineated the ground truth contours (Y.S., L.L.T.). Magnitude of volumetric revision was defined as the volume needed to be edited divided by the volume of the AI-generated contour, with the result multiplied by 100. **B**, Difference in volumetric revision between early (T1, T2) and advanced (T3, T4) tumors, as well as chemotherapy-naïve and post-induction chemotherapy (post-IC) tumors assessed by using the χ^2 test. Two-tailed $P < .05$ indicates a significant difference. Results indicate that by human experts' assessment, AI performed better in T1 and T2 tumors and chemotherapy-naïve tumors when compared with T3 and T4 tumors and post-IC tumors. IQR = interquartile range.

contours (0.80 vs 0.70, $P < .001$, Table E8 [online]), corresponding to a 54.5% reduction in volume coefficient of variation (0.15 vs 0.33, $P < .001$, Table E8 [online]). We observed time savings with AI intervention; median runtime of AI-only contouring was 40 seconds (range, 30–45 seconds), and average time spent editing an AI-generated contour compared with time spent with manual contouring was 18.3 vs 30.2 minutes ($P < .001$), corresponding to a time savings of 39.4% of work-hours.

Discussion

In this study, we constructed an artificial intelligence (AI) contouring tool using a large set of MRI examinations from 818 patients with nasopharyngeal carcinoma (NPC) and demonstrated the competency of our AI tool to delineate primary gross tumor volume (GTV) in the nasopharynx when compared against qualified radiation oncologists. Our AI tool was able to achieve contours comparable to those of human experts in 203 patients (median Dice similarity coefficient [DSC], 0.79; average surface distance [ASD], 2.0 mm). In addition, our AI tool performed favorably when compared against eight other experienced radiation oncologists in a separate multicenter study, outperforming half of them (median DSC, 0.79 vs 0.71, 0.71, 0.72, and 0.74; $P < .05$ for all). By allowing the radiation oncologists to edit the contours generated initially with our AI tool (AI assistance), contouring accuracy was improved in five of eight radiation oncologists. With AI assistance, there was reduction of intraobserver variation (by 36.4%), reduction of interobserver variation by 54.5%, and time savings of 39.4%.

Techniques like intensity modulated radiation therapy and proton beam therapy have led to better tumor control and reduced late-onset radiation-induced complications in patients with multiple types of tumors (11–14). However, the clinical advantages of these contemporary radiation therapy techniques are intricately linked to contouring accuracy, dose conformity, and precision of plan delivery (15). These processes are time consuming, and interobserver heterogeneity in the delineation of head and neck cancers is common (1). However, contouring accuracy is clinically important, as suboptimal tumor coverage and poor-quality radiation therapy plans are major factors for disease relapse and inferior survival (2). In this context, our AI tool simultaneously improved tumor delineation and reduced contouring variation, while it also reduced the time required for contouring.

In this study, we used a 3D CNN to automate GTV contouring on multiparametric MRIs, while a recently proposed deep deconvolutional neural network could only perform tumor segmentation for T1–T2 NPC tumors on two-dimensional axial CT images (16). Despite good accuracy (mean DSC, 0.80), they have not trained the deep deconvolutional neural network model in patients with T3–T4 disease, probably because of the lack of sensitivity to detect skull base and intracranial infiltration with CT (17,18). Although CT is the most commonly used imaging modality in treatment planning, MRI has been established as the standard modality for use in NPC staging and target contouring because of its superior soft-tissue contrast (19,20). Hence, in clinical practice, GTV contours are often generated on MRIs then registered to the treatment planning CT image (20–24).

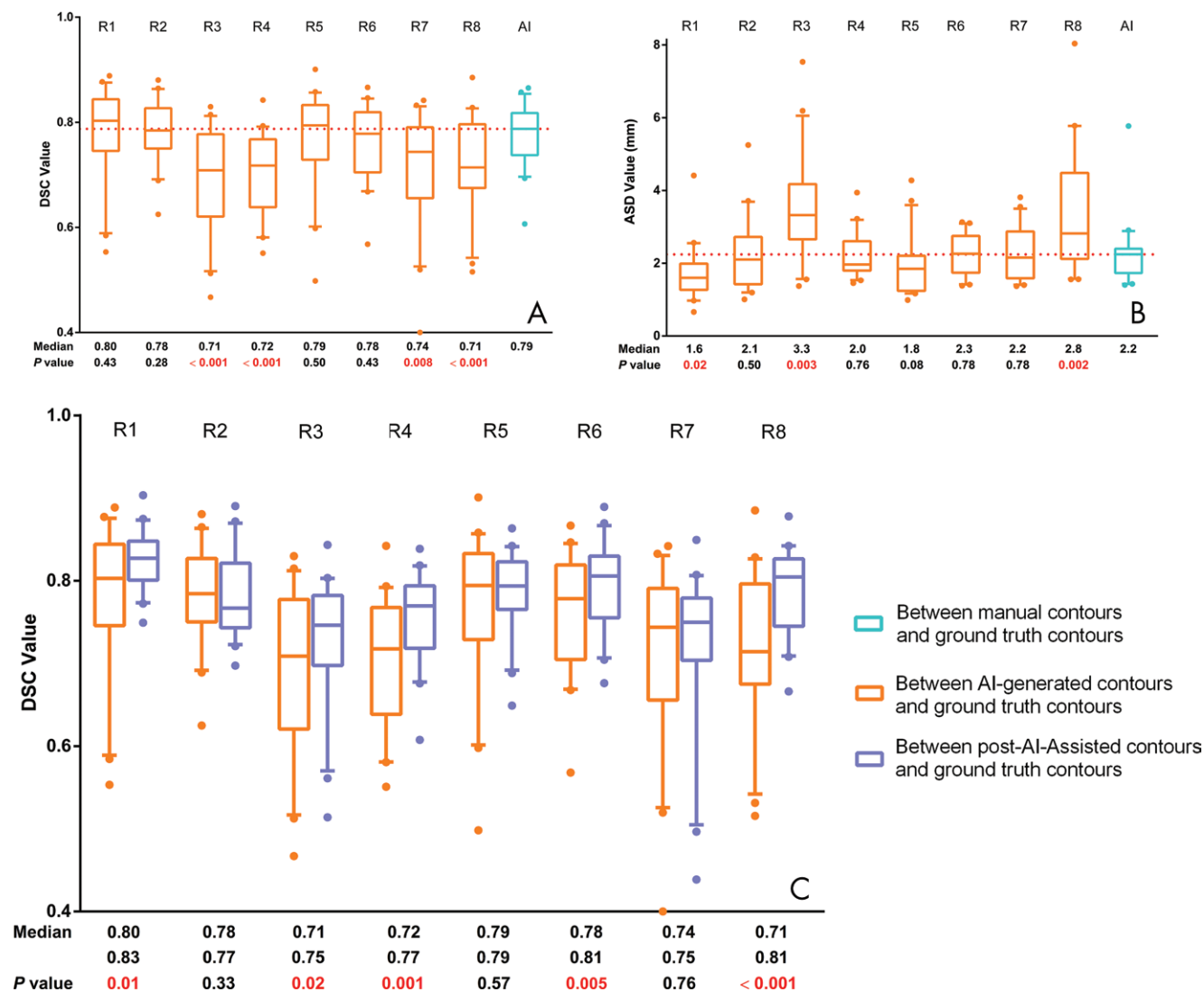


Figure 5: Results of multicenter evaluation. A, Dice similarity coefficients (DSCs) between manual contours of eight radiation oncologists and ground truth contours compared with artificial intelligence (AI)-generated contours and ground truth contours. Dotted lines indicate the median DSC of AI-generated contours. Results indicate the competency of the AI tool when compared with qualified radiation oncologists. B, Average surface distance (ASD) between manual contours of eight radiation oncologists and ground truth contours compared with AI-generated contours and ground truth contours. Dotted lines indicate the median ASD of AI-generated contours. Results indicate the competency of the AI tool when compared with qualified radiation oncologists. C, Comparison of DSC between post-AI-assisted contours and ground truth contours with DSC between manual contours and ground truth contours. Results indicate improved contouring accuracy with AI-assistance in GTV contouring for nasopharyngeal carcinoma. Error bars denote 10th and 90th percentiles. P values were calculated by using the Wilcoxon matched-pairs signed rank test. Two-tailed $P < .05$ indicates a significant difference.

However, it would be useful if our deep learning algorithm could work just as well with a CT data set.

In addition, one of the potential applications of our AI tool is its ability to facilitate the GTV recontouring process in adaptive radiation therapy. However, while we are enthusiastic to maximize its utility, we must caution the reader that our deep learning algorithm was trained on a multiparametric MRI data set acquired with a 3.0-T scanner, and we are uncertain whether it requires retraining on a CT-based data set; current adaptive replanning workflow is based on exploring the ability of an iterative artifact reduction algorithm to convert kilovoltage-based cone-beam CT images to the quality of planning helical CT images (25). Hence, much more work is needed in this space.

Several limitations of our study should be noted. First, the ground truth contours were derived from only two experienced radiation oncologists and were based on imaging modalities rather than on correlating the ground truth contours with tumor recurrence risk, given that NPC is a highly radiosensitive tumor and has very low local relapse rates (reported 5-year recurrence rates of 4.4%–11.3%) (26–28) and the majority of local recurrence occurred in field (29). Similarly, in this cohort, we observed only 33 local relapses in 1021 patients, with a median follow-up period of 20.7 months, and most of them (32 of 33) were in-field recurrence. The small number of events limits the statistical power to conduct such an analysis. Second, in the multicenter study, radiation oncologists performed

—R1—R2—R3—R4—R5—R6—R7—R8—Expert delineated ground truth contours

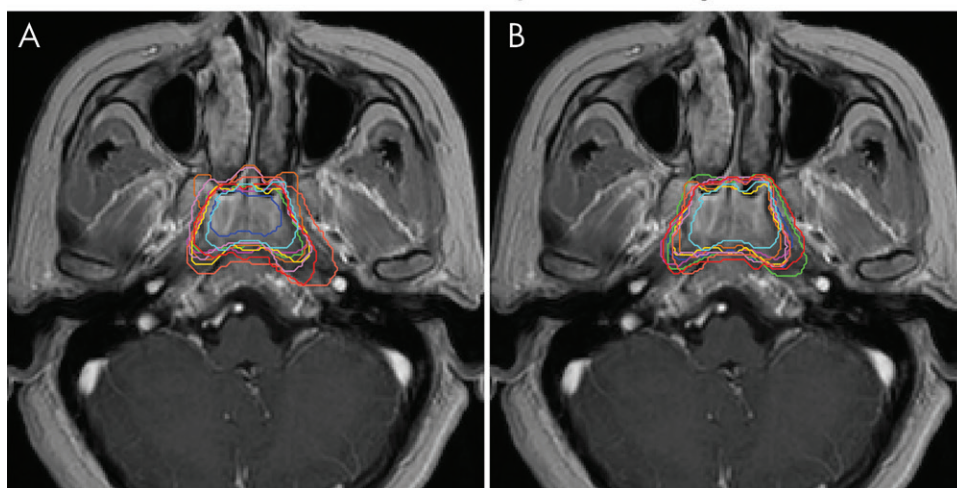


Figure 6: Example MRIs show the level of interobserver variation in the manual and post-artificial intelligence (AI)-assisted contours. Level of interobserver variation in A, manual contours and, B, post-AI-assisted contours by eight radiation oncologists.

contouring in a fixed sequence (manual contouring followed by editing AI-generated contours) rather than in a randomized sequence. Even with a 2-month interval, the unrandomized design might have introduced memory bias into evaluation of the effectiveness of AI assistance. To address more concrete real-world clinical benefits of AI assistance for tumor target and organs-at-risk (OAR) contouring in patients with NPC, we are now planning a multicenter randomized controlled trial. As superiority of CNN in OAR automatic contouring has been demonstrated by Xing et al in their recent study (30); for the purpose of clinical trial, we have codeveloped a 3D CNN to contour 43 OARs in the head and neck region. Finally, the AI tool achieved inferior accuracy at the cranial-caudal edges (ie, 0.75) compared with those at the midsections (range, 0.82–0.83) of the tumor; and this corresponded to the more frequent contouring “misses” at the oropharynx, hypopharynx, and intracranial regions than at other sites (ten vs three misses). Going forward, we aim to circumvent this deficiency by accumulating larger data sets that would include T4 tumors with more intracranial and hypopharyngeal extensions.

In summary, we implemented a deep 3D CNN to construct an AI contouring tool to automate GTV contouring for NPC. Our findings show that AI-assistance can effectively improve contouring accuracy and reduce intra- and interobserver variation and contouring time, which could have a positive impact on tumor control and patient survival.

Acknowledgments: We thank Mao-Hua Tang (Elekta Instrument [Shanghai], Shanghai, China) for supporting part of data extraction.

Author contributions: Guarantors of integrity of entire study, H.C., Y.S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agrees to ensure any questions related to the work are appropriately resolved, all authors; literature research, L.L., Q.D., Y.M.J., L.L.T., P.A.H., M.L.K.C., H.C., Y.S.; clinical studies, L.L., G.Q.Z., Y.Q.T., W.L.C., B.A.S., F.L., C.J.T., N.J., J.Y.L., C.M.X., M.L.K.C., H.C., Y.S.; statistical analysis, L.L., Q.D., P.A.H., M.L.K.C.,

H.C., Y.S.; and manuscript editing, L.L., Q.D., Y.M.J., Y.Q.T., W.L.C., B.A.S., F.L., C.J.T., J.Y.L., S.M.H., J.M., P.A.H., J.T.S.W., M.L.K.C., H.C., Y.S.

Disclosures of Conflicts of Interest: L.L. disclosed no relevant relationships. Q.D. disclosed no relevant relationships. Y.M.J. disclosed no relevant relationships. G.Q.Z. disclosed no relevant relationships. Y.Q.T. disclosed no relevant relationships. W.L.C. disclosed no relevant relationships. B.A.S. disclosed no relevant relationships. F.L. disclosed no relevant relationships. C.J.T. disclosed no relevant relationships. N.J. disclosed no relevant relationships. J.Y.L. disclosed no relevant relationships. L.L.T. disclosed no relevant relationships. C.M.X. disclosed no relevant relationships. S.M.H. disclosed no relevant relationships. J.M. disclosed no relevant relationships. P.A.H. disclosed no relevant relationships. J.T.S.W. disclosed no relevant relationships. M.L.K.C. Activities related to the present article: disclosed no relevant relationships. Activities not related to the

present article: is on the advisory board of Janssen, Astellas, and Varian; received research funding from Ferring Singapore; is on the speakers bureaus of Varian and Janssen; developed educational programs for Varian; immediate family members sold stock in Pluristem and Merrimack Pharmaceuticals; received research support from MedLever, GenomeDX, and Biosciences. Other relationships: disclosed no relevant relationships. H.C. disclosed no relevant relationships. Y.S. disclosed no relevant relationships.

References

1. Teguh DN, Levendag PC, Voet PW, 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):950–957.
2. Chen AM, Chin R, Beron P, Yoshizaki T, Mikaelian AG, Cao M. Inadequate target volume delineation and local-regional recurrence after intensity-modulated radiotherapy for human papillomavirus-positive oropharynx cancer. *Radiother Oncol* 2017;123(3):412–418.
3. Chua MLK, Wee JTS, Hui EP, Chan ATC. Nasopharyngeal carcinoma. *Lancet* 2016;387(10022):1012–1024.
4. Lee AW, Ma BB, Ng WT, Chan AT. Management of nasopharyngeal carcinoma: current practice and future perspective. *J Clin Oncol* 2015;33(29):3356–3364.
5. Dou Q, Yu L, Chen H, et al. 3D deeply supervised network for automated segmentation of volumetric medical images. *Med Image Anal* 2017;41:40–54.
6. Chen H, Dou Q, Yu L, Qin J, Heng PA. VoxResNet: deep voxelwise residual networks for brain segmentation from 3D MR images. *Neuroimage* 2018;170:446–455.
7. Carillo V, Cozzarini C, Perna L, et al. Contouring variability of the penile bulb on CT images: quantitative assessment using a generalized concordance index. *Int J Radiat Oncol Biol Phys* 2012;84(3):841–846.
8. Yousefi S, Kehrnavaz N, Gholipour A. Improved labeling of subcortical brain structures in atlas-based segmentation of magnetic resonance images. *IEEE Trans Biomed Eng* 2012;59(7):1808–1817.
9. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation. Cham, Switzerland: Springer International Publishing; 2016; 424–432.
10. Lee NY, Le QT, O’Sullivan B, Lu JJ. Nasopharyngeal Carcinoma. In: Lee NY, Lu JJ, eds. *Target Volume Delineation and Field Setup: A Practical Guide for Conformal and Intensity-Modulated Radiation Therapy*. Berlin, Germany: Springer; 2013; 1–10.
11. Zhang B, Mo Z, Du W, Wang Y, Liu L, Wei Y. Intensity-modulated radiation therapy versus 2D-RT or 3D-CRT for the treatment of nasopharyngeal carcinoma: a systematic review and meta-analysis. *Oral Oncol* 2015;51(11):1041–1046.
12. Bush DA, Cheek G, Zaheer S, et al. High-dose hypofractionated proton beam radiation therapy is safe and effective for central and peripheral early-stage non-small cell lung cancer: results of a 12-year experience at Loma Linda University Medical Center. *Int J Radiat Oncol Biol Phys* 2013;86(5):964–968.
13. McBride SM, Parambi RJ, Jang JW, Goldsmith T, Busse PM, Chan AW. Intensity-modulated versus conventional radiation therapy for oropharyngeal carcinoma: long-term dysphagia and tumor control outcomes. *Head Neck* 2014;36(4):492–498.
14. Lin SH, Wang L, Myles B, et al. Propensity score-based comparison of long-term outcomes with 3-dimensional conformal radiotherapy vs intensity-modulated radiotherapy for esophageal cancer. *Int J Radiat Oncol Biol Phys* 2012;84(5):1078–1085.
15. Jeanneret-Sozzi W, Moeckli R, Valley JF, et al. The reasons for discrepancies in target volume delineation: a SASRO study on head-and-neck and prostate cancers. *Strahlenther Onkol* 2006;182(8):450–457.

16. Men K, Chen X, Zhang Y, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. *Front Oncol* 2017;7:315.
17. Chong VF, Fan YF. Skull base erosion in nasopharyngeal carcinoma: detection by CT and MRI. *Clin Radiol* 1996;51(9):625–631.
18. Chong VF, Fan YF, Khoo JB. Nasopharyngeal carcinoma with intracranial spread: CT and MR characteristics. *J Comput Assist Tomogr* 1996;20(4):563–569.
19. Olmi P, Fallai C, Colagrande S, Giannardi G. Staging and follow-up of nasopharyngeal carcinoma: magnetic resonance imaging versus computerized tomography. *Int J Radiat Oncol Biol Phys* 1995;32(3):795–800.
20. Lee N, Xia P, Quivey JM, et al. Intensity-modulated radiotherapy in the treatment of nasopharyngeal carcinoma: an update of the UCSF experience. *Int J Radiat Oncol Biol Phys* 2002;53(1):12–22.
21. Kam MK, Teo PM, Chau RM, et al. Treatment of nasopharyngeal carcinoma with intensity-modulated radiotherapy: the Hong Kong experience. *Int J Radiat Oncol Biol Phys* 2004;60(5):1440–1450.
22. Tham IW, Hee SW, Yeo RM, et al. Treatment of nasopharyngeal carcinoma using intensity-modulated radiotherapy: the national cancer centre Singapore experience. *Int J Radiat Oncol Biol Phys* 2009;75(5):1481–1486.
23. Palazzi M, Orlandi E, Bossi P, et al. Further improvement in outcomes of nasopharyngeal carcinoma with optimized radiotherapy and induction plus concomitant chemotherapy: an update of the Milan experience. *Int J Radiat Oncol Biol Phys* 2009;74(3):774–780.
24. Lin S, Pan J, Han L, Zhang X, Liao X, Lu JJ. Nasopharyngeal carcinoma treated with reduced-volume intensity-modulated radiation therapy: report on the 3-year outcome of a prospective series. *Int J Radiat Oncol Biol Phys* 2009;75(4):1071–1078.
25. Zhang Y, Zhang L, Zhu XR, Lee AK, Chambers M, Dong L. Reducing metal artifacts in cone-beam CT images by preprocessing projection data. *Int J Radiat Oncol Biol Phys* 2007;67(3):924–932.
26. Au KH, Ngan RKC, Ng AWY, et al. Treatment outcomes of nasopharyngeal carcinoma in modern era after intensity modulated radiotherapy (IMRT) in Hong Kong: a report of 3328 patients (HKNPCSG 1301 study). *Oral Oncol* 2018;77:16–21.
27. Zhang MX, Li J, Shen GP, et al. Intensity-modulated radiotherapy prolongs the survival of patients with nasopharyngeal carcinoma compared with conventional two-dimensional radiotherapy: a 10-year experience with a large cohort and long follow-up. *Eur J Cancer* 2015;51(17):2587–2595.
28. Peng G, Wang T, Yang KY, et al. A prospective, randomized study comparing outcomes and toxicities of intensity-modulated radiotherapy vs. conventional two-dimensional radiotherapy for the treatment of nasopharyngeal carcinoma. *Radiother Oncol* 2012;104(3):286–293.
29. Li JX, Huang SM, Jiang XH, et al. Local failure patterns for patients with nasopharyngeal carcinoma after intensity-modulated radiotherapy. *Radiat Oncol* 2014;9(1):87.
30. Ibragimov B, Xing L. Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks. *Med Phys* 2017;44(2):547–557.