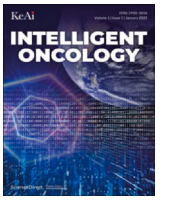
Intelligent Oncology 1 (2025) 61–71

Contents lists available at ScienceDirect

Intelligent Oncology

journal homepage: www.keaipublishing.com/en/intonc/



Full Length Article

Development and validation of AI delineation of the thoracic RTOG organs at risk with deep learning on multi-institutional datasets



Xianghua Ye[a,#], Dazhou Guo[b,#], Lujun Zhao[c], Congying Xie[d], Dandan Zheng[a], Haihua Yang[e], Xiangzhi Zhu[f], Xin Sun[a], Pingping Dong[a], Huanhuan Li[a], Weiwei Kong[g], Jianzhong Cao[h], Honglei Chen[i], Juntao Ran[j], Kai Ren[c], Hongxin Su[j], Hao Hu[k], Cuimeng Tian[l], Tianlu Wang[m], Qiang Zeng[f], Xiao Hu[n], Ping Peng[o], Junhua Zhang[p], Li Zhang[o], Tingting Zhang[q], Lue Zhou[r], Wenchao Guo[s,t], Zhanghexuan Ji[b], Puyang Wang[s,u], Hua Zhang[v], Jiali Liu[w], Le Lu[b], Senxiang Yan[a,*], Dakai Jin[b,*], Feng-Ming (Spring) Kong[w,*]

[a] *Department of Radiation Oncology, The First Affiliated Hospital, Zhejiang University, Hangzhou Zhejiang 310003, China*
[b] *DAMO Academy, Alibaba Group, New York NY 10014, USA*
[c] *Department of Radiation Oncology, Tianjin Medical University Cancer Institute and Hospital, Tianjin 300070, China*
[d] *Department of Radiation Oncology, The First Affiliated Hospital, Wenzhou Medical University, Wenzhou Zhejiang 325000, China*
[e] *Department of Radiation Oncology, Taizhou Hospital, Taizhou Zhejiang 317000, China*
[f] *Department of Radiation Oncology, Jiangsu Cancer Hospital, Nanjing Jiangsu 210019, China*
[g] *Department of Radiation Oncology, Nanjing Drum Tower Hospital, Nanjing Jiangsu 210008, China*
[h] *Department of Radiation Oncology, Shanxi Cancer Hospital, Taiyuan Shanxi 030013, China*
[i] *Department of Radiation Oncology, Jinan JunXin Med, Jinan Shandong 250001, China*
[j] *Department of Radiation Oncology, The First Hospital of Lanzhou University, Lanzhou Gansu 730000, China*
[k] *Department of Radiation Oncology, Kiang Wu Hospital, Macau SAR 999078, China*
[l] *Department of Radiation Oncology, Beijing Chest Hospital, Capital Medical University, Beijing 101149, China*
[m] *Department of Radiotherapy, Cancer Hospital & Institute; Cancer Hospital of China Medical University; Department of Radiotherapy, Cancer Hospital of Dalian University of Technology; Faculty of Medicine, Dalian University of Technology, Shenyang Liaoning 110042, China*
[n] *Department of Radiation Oncology, Zhejiang Cancer Hospital, Hangzhou Zhejiang 310022, China*
[o] *Department of Radiation Oncology, Tongji Hospital Tongji Medical College of HUST, Wuhan Hubei 430030, China*
[p] *Department of Radiation Oncology, Fudan University Shanghai Cancer Center, Department of Oncology; Shanghai Medical College, Fudan University; Shanghai Key Laboratory of Radiation Oncology, Shanghai 200032, China*
[q] *Department of Radiation Oncology, Shengli Oilfield Central Hospital, Dongying Shandong 257034, China*
[r] *Department of Radiation Oncology, Shenzhen People's Hospital (The Second Clinical Medical College , Jinan University; The First Affiliated Hospital, Southern University of Science and Technology) Shenzhen, Guangdong 518020, China*
[s] *DAMO Academy, Alibaba Group, Hangzhou Zhejiang 310023, China*
[t] *Hupan Lab, Hangzhou, Hangzhou Zhejiang 311121, China*
[u] *Johns Hopkins University, Baltimore MD 21205, USA*
[v] *LinkingMed, Beijing 100094, China*
[w] *Department of Clinical Oncology, School of Clinical Medicine, LKS Faculty of Medicine, The University of Hong Kong and University of Hong Kong–Shenzhen Hospital, Hong Kong SAR 999077, China*

* Corresponding authors.
E-mail addresses: yansenxiang@zju.edu.cn (S. Yan), dakai.jin@alibaba-inc.com (D. Jin), kong0001@hku.hk (F.-M.S. Kong).
# X. Ye and D. Guo contributed equally to this work.

https://doi.org/10.1016/j.intonc.2024.12.001
Received 5 September 2024; Received in revised form 30 November 2024; Accepted 1 December 2024
Available online 7 December 2024
2950-2616/© 2025 The Author(s). Publishing services by Elsevier B.V. on behalf of KeAi Communications Co. Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ARTICLE INFO

Keywords:
NRG-RTOG 1106
OAR segmentation
Deep learning
Partially labeled datasets

ABSTRACT

Introduction: Accurate contouring of thoracic organs at risk (OARs) is essential for minimizing complications in radiation treatment. Manual contouring of thoracic OARs is not only time-consuming but also prone to substantial user variation. To enhance the efficiency and consistency, we developed a unified deep learning (DL) OAR contouring model, DeepOAR, that was trained using multiple partially labeled datasets for segmenting a comprehensive set of thoracic OARs following the Radiation Therapy Oncology Group (RTOG)-guided OAR atlas. This DL model supports the segmentation of six required and eight optional OARs guided by the NRG-RTOG 1106 trial, providing precise and reproducible OARs contouring that are ready to be used in radiotherapy practice.

Materials and methods: Following the OAR contouring recommendation of the NRG-RTOG 1106 trial, we collected and curated three private datasets and two public datasets, comprising a total of 531 patients with partially annotated thoracic OARs. These partially annotated datasets were utilized to develop DeepOAR, which consisted of a shared encoder and 14 separate decoders, with each decoder dedicated to one specific OAR. For model training, we utilized all patients from the two public datasets and 75% of the patients from the private datasets. We reserved the remaining 25% of the private datasets for independent testing. A multi-user study involving 21 radiation oncologists was conducted on 40 randomly selected patients from the independent testing dataset to evaluate the clinical applicability of DeepOAR. The Dice coefficient score (DSC) and average surface distance (ASD) were computed to evaluate the quantitative delineation performance of the model.

Results: DeepOAR outperformed nnUNet (the benchmark medical segmentation model) across all 14 OARs, achieving mean DSC and ASD values of 88.4% and 1.0 mm, respectively, in the independent testing set. Multi-user validation demonstrated that 89.7% of DeepOAR-generated OARs were clinically acceptable or required only minor revisions. A comparison using two randomly selected patients showed that the delineation variability of DeepOAR was significantly smaller than the inter-user variation among radiation oncologists. Human editing of DeepOAR's predictions could further improve OAR delineation accuracy by an average of 3% increase in DSC and 40% reduction in ASD while significantly reducing the workload of radiation oncologists for contouring 14 thoracic OARs by an average of 77.0%.

Conclusion: We developed DeepOAR, a DL-based unified contouring model trained using multiple partially labeled datasets, to delineate a comprehensive set of 14 thoracic OARs following the RTOG-guided OAR atlas. Both qualitative and quantitative results demonstrated the strong clinical applicability of DeepOAR for the OAR delineation process in thoracic cancer radiotherapy workflows, along with improved efficiency, comprehensiveness, and quality.

1. Introduction

Precise delineation of organs at risk (OARs) is crucial to prevent surrounding healthy tissues from excessive radiation exposure and to thereby reduce the risk of severe post-treatment complications in thoracic radiotherapy.[1–3] Precise delineation also facilitates meaningful comparisons across different clinical trials. Computed tomography (CT) is the primary imaging modality used for diagnosing, staging, and assessing the treatment response of various medical conditions due to its widespread availability and capacity to provide detailed anatomical information at a relatively low cost.[4,5] Guided by the OAR atlas of the NRG-Radiation Therapy Oncology Group (RTOG) 1106 protocol,[6] currently, thoracic OAR contouring is predominantly performed via manual delineation. However, this approach demands significant clinical expertise with high labor costs and is inherently tedious. Additionally, manual delineation can exhibit substantial inter-user variability among radiation oncologists, leading to inconsistent and non-reproducible performance.[7] This study aimed to develop and validate a deep learning (DL)-based OAR contouring model capable of segmenting a comprehensive set of 14 thoracic OARs according to the NRG-RTOG 1106 guideline,[6] which has been adopted for RTOG lung radiation protocols. The targeted OARs include six required organs—the brachial plexus, esophagus, left lung, right lung, pericardium, and spinal cord—as well as eight optional organs—the aorta (A-Aorta), superior vena cava (V-Venacava-S), inferior vena cava (V-Venacava-I), proximal bronchial tree (Bronchi-Prox), pulmonary artery (A-Pulmonary), pulmonary vein (V-Pulmonary), left chest wall, and right chest wall.

Developing an accurate dl-based organ segmentation model requires large-scale datasets with high-quality and often fully labeled organ annotation, which are challenging to obtain.[8–11] Deep networks trained on sizable datasets have demonstrated promising performances comparable to human observers.[12–15] However, several existing datasets are relatively small and contain only partially labeled organs; they primarily focus on a subset of organs and a single type of cancer.[16] A possible solution is to train separate models on several smaller datasets and then combine their knowledge afterward.[17–19] Francis et al.[20] introduced a two-stage DL model with an attention mechanism to segment thoracic OARs more efficiently. Feng et al.[21] demonstrated that adding and re-training on a small number of local data can enhance the robustness of thoracic OAR segmentation models as compared to transfer learning. Shi et al.[22] implemented a cascade coarse-to-fine segmentation with an adaptive module to segment small and large whole-body organs. However, these approaches have two major drawbacks: (1) inefficiency in model training and deployment, as they require training

separate models for each dataset and combining multiple models to generate the final output, and (2) reduced accuracy and robustness, as individual models trained on smaller datasets perform less effectively compared with models trained on combined datasets. In clinical practice, there is an increasing demand for segmentation models that can utilize multiple datasets for joint training to address these challenges.

In this study, we proposed DeepOAR, a unified DL-based OAR contouring model trained using multiple partially labeled datasets to segment a comprehensive set of thoracic OARs following the RTOG-guided OAR atlas. The architecture of DeepOAR consists of a shared encoder, which extracts universal anatomical features not specific to a particular organ, and multiple decoders, each dedicated to segmenting a specific organ. Upon training DeepOAR on a union of five partially labeled datasets, it exhibited substantially improved segmentation accuracy and robustness. Two multi-user studies were conducted to validate its performance in clinical workflow.[23] The results demonstrated that the OAR contours predicted by DeepOAR required minimal revision efforts from radiation oncologists to achieve clinical acceptability. This underscores the potential of DeepOAR as a valuable and efficient tool for OAR contouring in radiotherapy planning.

2. Materials and methods

2.1. Datasets for training and testing

In this multi-institutional retrospective study, we collected 531 patients from three private and two public partially annotated datasets for DeepOAR model training and validation. The three private datasets were collected from the First Affiliated Hospital of Zhejiang University (FAH-ZU), which was approved by the Clinical Research Ethics Committee of FAH-ZU (No.: IIT20231208A). The first private dataset for model training and validation included 153 lung cancer patients with IV contrast-enhanced CT scans. Following the complete NRG-RTOG 1106 OAR atlas, an experienced radiation oncologist (X. H. Y) manually contoured 14 thoracic OARs, which were then confirmed by a senior radiation oncologist (F. M K) who developed the RTOG atlas. The second private dataset included 98 lung cancer patients with contrast-enhanced CT scans, and six thoracic OARs were annotated (i.e., the left and right lung, A-Aorta, esophagus, A-Pulmonary, and V-Pulmonary). The third private dataset contained 190 esophageal cancer patients, with each patient having paired, registered contrast-enhanced, and non-contrast-enhanced CT scans, totaling 380 scans. The same annotation process was followed for the lung cancer patients. Three OARs, including the esophagus, V-Pulmonary, and brachial plexus, were annotated. On average, the spatial resolutions of the three private datasets were $0.68 \times 0.68 \times 2.0$ mm, $0.72 \times 0.72 \times 2.5$ mm, and $0.68 \times 0.68 \times 1.8$ mm, respectively.

Two additional public datasets containing partially labeled thoracic OARs were utilized, i.e., 50 lung cancer patients with contrast-enhanced CT scans in StructSeg2019 and 40 lung cancer patients either with or without IV contrast in SegTHOR.[24] After reviewing the OAR label quality of these two public datasets, we identified four OARs (left and right lung, esophagus, and BronchiProx) from StructSeg2019 and three OARs (pericardium, esophagus, and A-Aorta) from SegTHOR that qualified for the NRG-RTOG 1106 OAR guideline. These qualified OARs were utilized to develop DeepOAR. The CT scans had an average spatial resolution of $1.17 \times 1.17 \times 5.0$ mm and $0.98 \times 0.98 \times 2.5$ mm in the StructSeg2019 and SegTHOR datasets, respectively.

For the 14 types of thoracic OARs obtained from five datasets, excluding chest walls, each type of organ was labeled in at least two datasets. Specifically, four datasets contained the esophagus, whereas three datasets contained left/right lung and A-Aorta. For model training, we used all patients from the public StructSeg2019 and SegTHOR datasets and 331 (75%) patients randomly selected from the private FAH-ZU datasets. The remaining 110 private patients, evenly distributed between contrast-enhanced CT and non-contrast CT, were kept as unseen testing data to ensure independent evaluation during the training process.

2.2. Multi-user datasets

We conducted two multi-user tests to evaluate the qualitative and quantitative performance of DeepOAR in the segmentation of thoracic OARs.

For qualitative multi-user testing, we collected 40 patients to qualitatively evaluate the DeepOAR contours, where 30 patients were curated from the FAH-ZU, and 10 non-contrast CTs were randomly selected from the non-small cell lung cancer (NSCLC) Radiomics dataset.[25] Eighteen groups (21 radiation oncologists from 14 institutes) participated in the qualitative testing. The 30 curated testing patients were stratified into three sets of equal size. The first set comprised lung cancer patients who underwent IV contrast-enhanced CT, and the second set consisted of lung cancer patients who underwent non-contrast CT. The third set comprised esophageal cancer patients who underwent IV contrast-enhanced CT. The average spatial resolutions of the three sets were $0.97 \times 0.97 \times 4.8$ mm, $0.90 \times 0.90 \times 4.0$ mm, and $0.98 \times 0.98 \times 5.0$ mm, respectively. Our objective was to qualitatively rate the DeepOAR-predicted OAR contours into the following categories: no revision, minor revision, moderate revision, major revision, and unacceptable.

To quantitatively validate the clinical applicability of DeepOAR, one lung cancer patient from the public NSCLC Radiomics dataset and one esophageal cancer patient from the internal testing dataset were randomly selected. Ten groups (11 radiation oncologists from eight institutes) participated in the quantitative testing. The lung cancer patient was imaged with contrast-enhanced CT, whereas the esophageal cancer patient underwent a non-contrast CT scan. The spatial resolutions were $0.98 \times 0.98 \times 3.0$ mm and $0.85 \times 0.85 \times 5.0$ mm for the lung and esophageal cancer patients, respectively. Following a dry-run workshop, each group independently annotated and timed the segmentation of all 14 organs from scratch. To minimize the potential bias from the heuristic methods developed during the initial annotation phase, a two-week waiting period was enforced before the groups were permitted to revise the OARs predicted by DeepOAR. Following this interval, all 14 DeepOAR-predicted OARs underwent revision and timing assessments to evaluate the efforts required to ensure the clinical acceptance of DeepOAR-predicted OAR contours.

2.3. Evaluation process

All radiation oncologists who participated in this study underwent a dry-run workshop before multi-user testing. The evaluation protocol instructions, plan assessment, and a randomly selected case from the public NSCLC Radiomics dataset, including a complete set of 14 OARs, were provided during the workshop. To ensure consistency and minimize inter-system variations in assessing and timing editing efforts, all radiation oncologists utilized a China's National Medical Products Administration-certified three-class medical device, LinkingMed intelligent cloud delineation system 3 (version 4.2.1), for OAR editing. Subsequently, the CT scans designated for testing were de-identified and uploaded to the LinkingMed system for online multi-user studies. Participating radiation oncologists submitted their manually contoured and revised OARs through the online LinkingMed platform. Only the contours were collected and downloaded to a local server for subsequent quantitative evaluation, thus maintaining the integrity and confidentiality of the patient data throughout the process. A two-tailed *t*-test was used to compare the evaluation metrics in the paired data. Statistical significance was set at two-tailed $P < 0.001$.

2.4. Baseline nnUNet models

In this study, three distinct configurations of nnUNet models were developed. (1) Single Model: A single nnUNet model trained using the internal 153-patient complete set of annotations for all 14 OARs. (2) Data-specific Ensemble: An ensemble of five nnUNet models, each associated with one of the five training datasets. In cases where different models predicted common OARs, the ensemble of nnUNet predictions utilized a majority voting approach for each voxel. (3) Organ-specific Ensemble: An ensemble of 12 organ-specific nnUNet models trained using re-organized 12 organ-wise datasets from five training sets. Paired organs, such as the left and right lung and chest walls, were included together in the same dataset. All three configurations of the nnUNet models were trained using the same standardized setup, with detailed training parameters elaborated in the Supplementary materials.

2.5. Proposed DeepOAR model

DeepOAR was developed as a unified DL segmentation model featuring a shared encoder and multiple decoders tailored to delineate the 14 thoracic OARs. This multi-organ contouring model was trained on a union of partially labeled datasets, allowing for improved accuracy and robustness. In the proposed model, the shared encoder served to extract general features that were not organ-specific, whereas each decoder was tasked to aggregate features to segment a specific organ. To supervise the training of each decoder, the model incorporated nnUNet's data augmentation scheme, training setup, and employed a combination of cross-entropy loss and Dice loss to optimize segmentation performance. The overall architecture of DeepOAR is illustrated in Figure 1, and the detailed training setup is elaborated on in the Supplementary materials.

3. Results

3.1. Quantitative performance on testing datasets

The quantitative performance of DeepOAR on the internal testing dataset is outlined in Table 1. Among the five training datasets used, the 153-patient dataset included a complete set of annotations for all 14 OARs, whereas the remaining datasets consisted of partially labeled OARs. Segmentation performance using the Single nnUNet Model trained on the 153-patient dataset yielded the lowest results, with an average DSC of 82.1% for required OARs, 86.5% for optional OARs, and 84.6% for all 14 OARs.

To maximize the utilization of all available data, the five distinct nnUNet models in the Data-specific Ensemble model were individually trained, each on one of the five datasets. The OARs predicted by each model were associated with their respective training datasets. The ensemble of predictions from the five nnUNet models yielded improved segmentation performance, achieving, on average, 84.0%, 87.3%, and 85.9% DSC for the required, optional, and all 14 OARs, respectively. In addition to the ensemble of

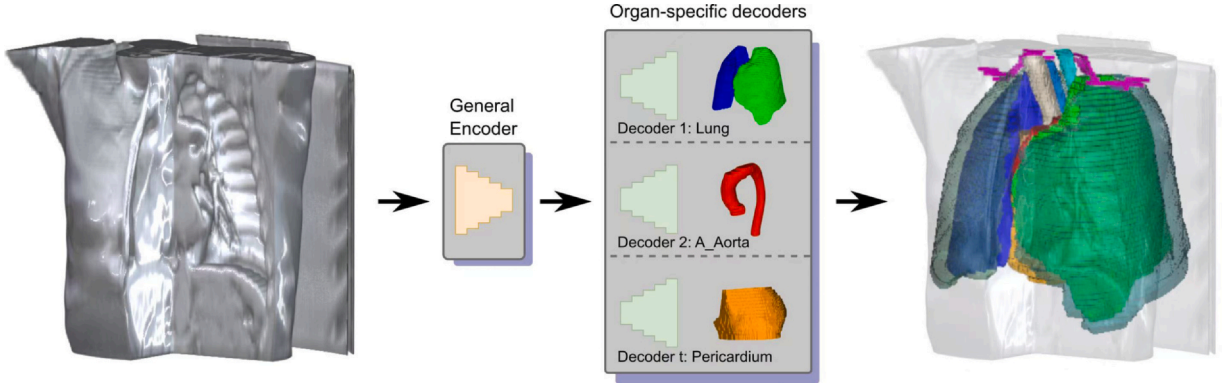


Figure 1. The overall framework of the proposed DeepOAR using a collection of multiple partially labeled datasets. It is composed of a shared encoder, multiple OAR-specific decoders, and a merging module.

Table 1
Comparisons of the accuracy between nnUNet and DeepOAR using an independent test set of 153 patients.

OARs	Single Model		Data-specific Ensemble		Organ-specific Ensemble		DeepOAR	
	DSC(%)	ASD (mm)	DSC(%)	ASD (mm)	DSC(%)	ASD (mm)	DSC(%)	ASD (mm)
Brachial Plexus	36.8 ± 8.4	3.3 ± 3.0	40.5 ± 8.3	3.2 ± 2.9	49.0 ± 8.8	2.5 ± 2.1	**52.5 ± 8.3**	**2.2 ± 1.3**
Esophagus	81.7 ± 5.5	1.2 ± 0.7	84.2 ± 5.3	0.8 ± 0.6	88.2 ± 5.8	0.7 ± 0.7	**88.8 ± 3.5**	**0.7 ± 0.3**
Lung-L	96.5 ± 0.3	0.9 ± 1.0	98.0 ± 0.3	0.8 ± 1.0	**98.2 ± 0.6**	**0.6 ± 1.4**	98.1 ± 0.2	0.7 ± 0.3
Lung-R	95.3 ± 3.3	0.7 ± 0.5	97.1 ± 3.1	0.6 ± 0.5	97.8 ± 3.7	0.5 ± 1.0	**98.8 ± 0.1**	**0.1 ± 0.1**
Pericardium	90.9 ± 1.3	1.6 ± 3.7	92.2 ± 1.3	1.6 ± 3.7	94.1 ± 1.8	1.5 ± 4.1	**93.5 ± 1.5**	**1.1 ± 0.3**
Spinal Cord	91.3 ± 2.6	1.1 ± 0.8	91.7 ± 2.4	0.9 ± 0.8	92.2 ± 2.6	0.8 ± 0.8	**92.8 ± 2.3**	**0.8 ± 0.4**
A-Aorta	91.6 ± 7.7	1.3 ± 4.4	92.8 ± 7.6	1.1 ± 4.4	92.9 ± 8.0	1.0 ± 4.6	**95.1 ± 1.9**	**1.0 ± 1.1**
A-Pulmonary	87.7 ± 3.9	1.1 ± 1.3	87.7 ± 3.7	1.1 ± 1.1	89.0 ± 3.8	0.9 ± 1.4	**88.7 ± 3.7**	**1.0 ± 0.5**
V-Pulmonary	69.6 ± 8.0	2.1 ± 1.2	70.9 ± 7.8	1.9 ± 1.1	73.2 ± 8.1	1.8 ± 1.4	**75.4 ± 6.8**	**1.3 ± 0.6**
V-Venacava-S	86.1 ± 4.0	1.1 ± 0.4	87.2 ± 3.8	0.9 ± 0.5	87.7 ± 4.5	0.8 ± 0.6	**88.7 ± 3.2**	**0.8 ± 0.6**
V-Venacava-I	85.6 ± 7.4	1.2 ± 1.2	86.3 ± 7.4	1.1 ± 1.1	**88.0 ± 8.1**	**0.9 ± 1.5**	87.2 ± 3.5	1.0 ± 0.3
Bronchi-Prox	91.0 ± 1.7	1.2 ± 0.6	91.4 ± 1.6	0.9 ± 0.7	91.0 ± 2.1	0.8 ± 0.7	**94.2 ± 1.8**	**0.8 ± 0.6**
Chest Wall-L	90.3 ± 2.0	1.4 ± 0.8	90.4 ± 2.0	1.4 ± 0.8	90.5 ± 2.5	1.2 ± 1.2	**91.2 ± 2.0**	**1.2 ± 0.6**
Chest Wall-R	91.7 ± 1.9	0.9 ± 0.5	91.8 ± 1.8	0.8 ± 0.4	92.0 ± 2.1	0.6 ± 0.5	**92.4 ± 1.5**	**0.6 ± 0.4**
Ave Required	82.1	1.45	84.0	1.3	86.8	1.0	**87.4**	**1.0**
Ave Optional	86.5	1.28	87.3	1.1	88.0	1.0	**89.1**	**1.0**
Ave All	84.6	1.35	85.9	1.2	87.5	1.0	**88.4**	**1.0**

Bold values represent statistically significant improvements. Abbreviation: 1. DSC: Dice similarity coefficient 2. ASD: average surface distance

five predictions, the five datasets were re-assembled to create 12 new datasets, each associated with a specific type of organ. Twelve nnUNet models in the Organ-specific Ensemble model were trained using the 12 re-assembled datasets, resulting in further improved performance (86.8%, 88.0%, and 87.5% DSC for the required, optional, and all 14 OARs, respectively).

DeepOAR achieved a mean DSC of 88.4% and an ASD of 1.0 mm. The qualitative results obtained from the internal testing dataset are depicted in Figure 2. For the required OARs, the mean DSC and ASD were 87.4% and 1.0 mm, respectively, whereas for the optional OARs, they were 89.1% and 1.0 mm, respectively. DeepOAR exhibited significantly superior performance, with all comparisons yielding statistically significant differences ($P < 0.001$). Compared with the Single nnUNet model, the DSC of the brachial plexus (+15.7%), esophagus (+7.1%), and V-Pulmonary (+ 5.8%) demonstrated the most significant improvements. The least improvement was observed in the case of chest walls, which showed a marginal average 0.8% increase in DSC.

3.2. Qualitative performance in multi-user evaluations

In a comprehensive evaluation conducted by 18 groups of radiation oncologists on 40 patients, the feedback revealed that the majority of assessments—9,042 out of 10,080 (14 OARs × 40 patients × 18 groups, representing 89.7%)—rated the OAR instances predicted by DeepOAR as clinically acceptable (44.9%) or requiring only minor revisions (44.8%). Figure 3 provides visual representations of the sample results for illustration purposes. A small proportion of OARs required moderate revision (8.0%) or major revision (2.2%), with only four instances categorized as unacceptable, accounting for less than 0.1% of the total evaluations. Among the instances deemed unacceptable, two cases pertained to patients with esophageal cancer. The other two unacceptable cases involved V-Venacava-I predictions that did not extend into the liver region. This occurred because the model was specifically focused on thoracic organ segmentation and intentionally cut off predictions beyond the thoracic region to reduce false positives in other parts of the body. Although most radiation oncologists considered these predictions to require only minor or no revisions, one group labeled these two instances unacceptable. In the first case, the patient underwent IV contrast-enhanced CT scanning, revealing significant dilation and tumor infiltration in the inferior esophagus. Although the model accurately identified the superior portion of the esophagus, it missed most of the extremely dilated diseased region, leading two groups to deem this prediction unacceptable. In the second case, the patient underwent non-contrast CT scanning, revealing extreme dilation in the superior esophagus. Although DeepOAR captured the inferior portion of the esophagus and the tumor, it overlooked the significantly expanded superior tumor areas (pushing toward A-Aorta), resulting in the same two groups rendering this prediction unacceptable. The remaining radiation oncologists deemed these predictions requiring major revisions.

To evaluate DeepOAR's generalizability across centers, we compared the qualitative outcomes of 30 FAH-ZU cases with those of 10 NSCLC Radiomics cases. DeepOAR's predictions for OARs in patients from the two centers were deemed clinically acceptable or required only minor revisions at rates of 89.6% and 89.9%, respectively, with clinically acceptable predictions at 43.9% vs. 47.9 % and minor revisions needed at 45.7% vs. 42.0%. The rates of OARs requiring moderate revisions were identical at 8.0% for both FAH-ZU and NSCLC Radiomics cases, whereas those requiring major revisions or rejection were minimal: 2.4% and 2.1%, respectively.

3.3. Quantitative performance in multi-user evaluations

3.3.1. Manual contouring from scratch

Table 2 provides a summary of the editing time invested by the 10 groups of radiation oncologists. The average total editing time for 14 OARs for two patients was reported as 108.7 min. Specifically, the delineation times for the required and optional OARs were

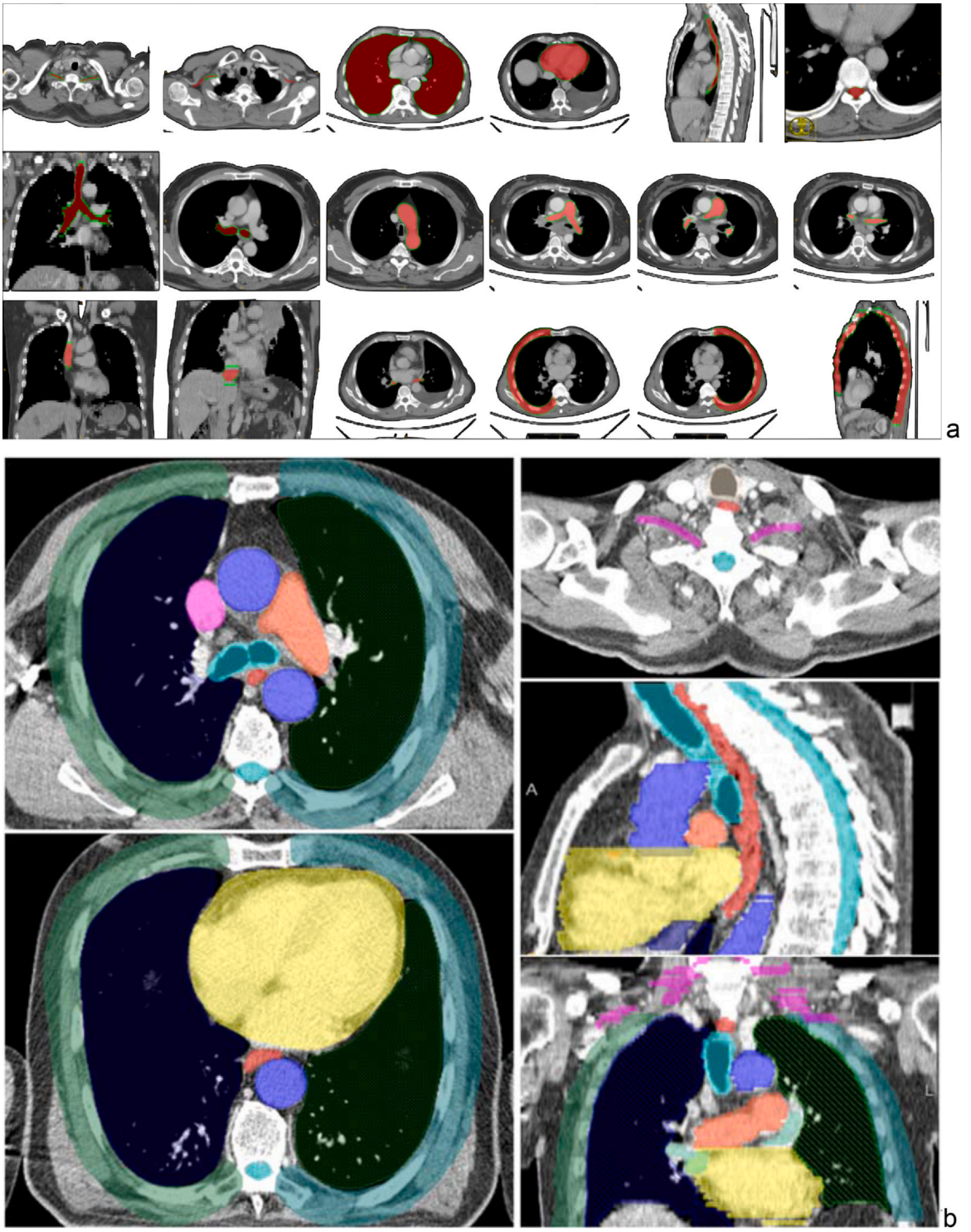


Figure 2. Examples of the proposed DeepOAR model for the segmentation of 14 thoracic OARs following the NRG-RTOG 1106 guideline.
In the top row (**a**), red masks indicate automated contouring results; green contours, the ground truth label. The bottom row (**b**) shows 3D views of DeepOAR-predicted OARs. Most auto-contoured OARs closely follow the ground truth, e.g., the lungs, pericardium, esophagus, spinal cord, A-Aorta, Bronchi-Prox, A-Pulmonary, V-Venacava-S, and V-Venacava-I.

58.9 min and 49.7 min, respectively. Large or complex OARs that required the most annotation time: the lungs (13.4 + 13.9) min, chest walls (13.6 + 12.8) min, and brachial plexus (10.1 min). Middle-sized or elongated OARs that required moderate annotation time: the spinal cord (7.6 min), pericardium (7.1 min), Bronchi-Prox (6.8 min), A-Aorta (6.8 min), and esophagus (6.8 min). Small or

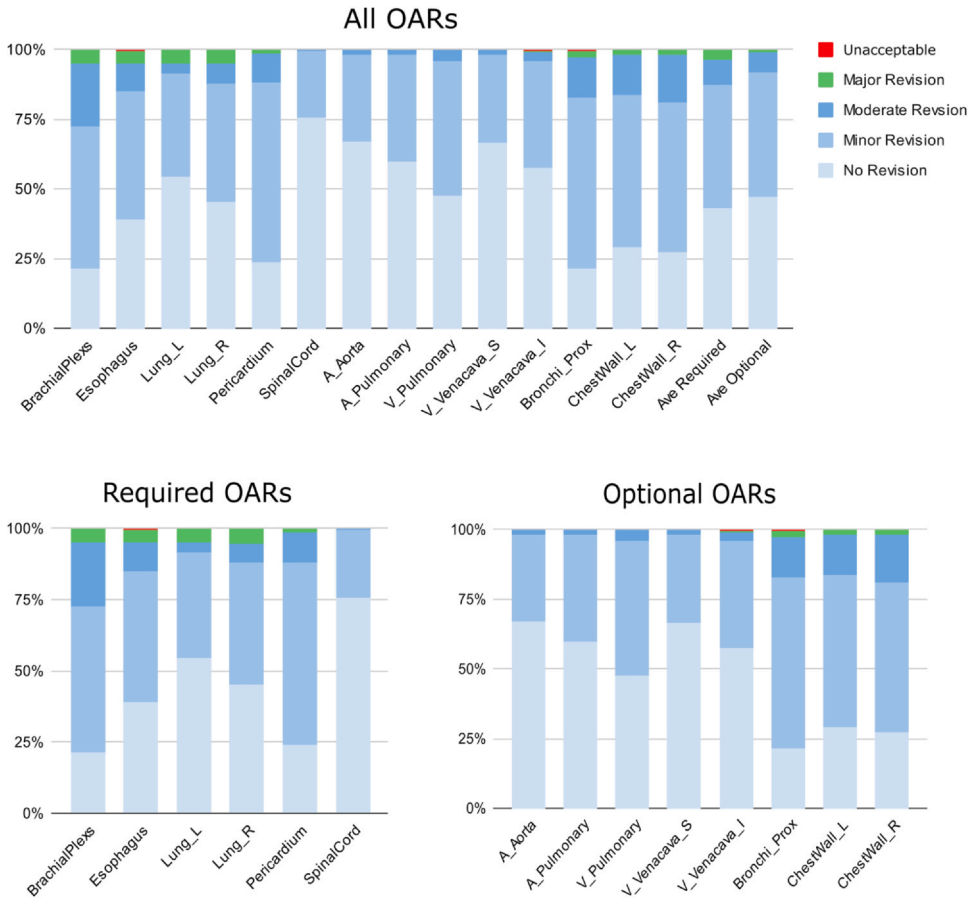


Figure 3. Summary of the qualitative assessments.
This was conducted by 18 groups of radiation oncologists on the revision efforts for DeepOAR-predicted 14 OAR contours using 40 multi-user testing patients. *The vast majority of DeepOAR-predicted OARs requiring minor revisions or no revision to be accepted clinically. Only four instances (less than 0.1%) were deemed unacceptable.*

Table 2
Contouring time taken by radiation oncologists during the assessment of manual delineation and revision for 14 OARs on one lung cancer patient and one esophageal cancer patient.

	Lung cancer patient		Esophageal cancer patient		Average	
Human	Manual Delineation (min)	Edited on DeepOAR	Manual Delineation (min)	Edited on DeepOAR	Manual Delineation (min)	Edited on DeepOAR
Brachial Plexus	10.7 ± 2.0	4.4 ± 3.4	9.6 ± 3.5	3.1 ± 3.2	10.1 ± 2.8	3.7 ± 3.3
Esophagus	6.3 ± 2.8	1.7 ± 1.5	7.3 ± 3.5	1.3 ± 1.0	6.8 ± 3.1	1.5 ± 1.2
Lung-L	14.6 ± 4.8	1.7 ± 2.7	12.4 ± 4.2	1.4 ± 1.8	13.4 ± 4.6	1.6 ± 2.3
Lung-R	14.8 ± 4.6	1.7 ± 1.7	11.9 ± 5.1	1.3 ± 1.6	13.9 ± 4.9	1.5 ± 1.6
Pericardium	8.4 ± 3.9	2.2 ± 2.2	5.9 ± 2.2	1.9 ± 1.7	7.1 ± 3.3	2.1 ± 2.0
Spinal Cord	9.0 ± 4.9	1.3 ± 2.3	6.1 ± 3.8	1.1 ± 2.1	7.6 ± 4.5	1.2 ± 2.2
A-Aorta	8.0 ± 5.3	0.8 ± 0.8	5.6 ± 2.5	1.1 ± 1.0	6.8 ± 4.2	0.9 ± 0.9
A-Pulmonary	3.4 ± 1.8	0.7 ± 0.7	3.4 ± 1.4	0.7 ± 0.6	3.4 ± 1.6	0.7 ± 0.6
V-Pulmonary	2.8 ± 1.7	0.7 ± 0.7	4.4 ± 4.1	0.8 ± 0.8	3.6 ± 3.2	0.8 ± 0.7
V-Venacava-S	1.6 ± 0.6	0.4 ± 0.3	1.3 ± 0.8	0.5 ± 0.3	1.5 ± 0.7	0.4 ± 0.3
V-Venacava-I	1.6 ± 0.4	0.5 ± 0.5	1.0 ± 0.6	0.3 ± 0.3	1.3 ± 0.6	0.4 ± 0.4
Bronchi-Prox	8.1 ± 3.8	1.6 ± 1.0	5.4 ± 2.0	1.2 ± 0.9	6.8 ± 3.3	1.4 ± 1.0
Chest Wall-L	14.9 ± 5.2	5.2 ± 4.9	12.7 ± 4.2	3.7 ± 3.9	13.6 ± 4.6	4.5 ± 4.4
Chest Wall-R	13.1 ± 4.9	4.2 ± 4.5	11.6 ± 3.8	3.7 ± 4.3	12.8 ± 4.2	3.9 ± 4.3
Total Required	62.8	13.8 (78.0%↓)	55.2	10.1 (81.7%↓)	58.9	12 (79.7%↓)
Total Optional	53.5	14.1 (73.7%↓)	45.4	12 (73.7%↓)	49.7	13 (73.8%↓)
Total All	116.3	27.9 (76.0%↓)	100.6	22.1 (78.1%↓)	108.7	25 (77.0%↓)

The percentage with a downward arrow in the parenthesis indicates the proportion of the reduced editing efforts compared with contouring from scratch. Editing time is measured in minutes.

Table 3
[a]Quantitative comparisons of contouring accuracy between manual delineation, DeepOAR, and Edited on DeepOAR in the multi-user testing study of two patients.

OARs	Manual Delineation		DeepOAR		Edited on DeepOAR	
	DSC(%)	ASD (mm)	DSC(%)	ASD (mm)	DSC(%)	ASD (mm)
Brachial Plexus	37.2 ± 10.6	3.0 ± 1.2	50.2 ± 4.4	2.9 ± 0.3	**50.8 ± 8.4**	**2.4 ± 1.0**
Esophagus	79.1 ± 8.6	0.9 ± 0.5	79.7 ± 4.8	0.9 ± 0.2	**81.1 ± 4.1**	**0.8 ± 0.2**
Lung-L	97.8 ± 0.6	1.4 ± 1.0	**99.7 ± 0.2**	**0.1 ± 0.1**	99.5 ± 0.4	0.1 ± 0.2
Lung-R	98.2 ± 0.5	1.3 ± 1.0	**99.9 ± 0.1**	**0.1 ± 0.1**	99.8 ± 0.2	0.1 ± 0.1
Pericardium	93.2 ± 2.9	2.5 ± 1.5	**95.4 ± 2.0**	**1.7 ± 1.2**	95.2 ± 1.1	1.7 ± 1.6
Spinal Cord	81.6 ± 3.9	2.8 ± 2.2	88.3 ± 0.7	0.5 ± 0.2	**88.4 ± 0.5**	**0.5 ± 0.1**
A-Aorta	89.5 ± 2.4	1.4 ± 0.5	89.6 ± 3.6	1.4 ± 0.4	**90.8 ± 2.2**	**1.2 ± 0.3**
A-Pulmonary	85.2 ± 3.4	1.2 ± 0.3	86.9 ± 2.3	1.1 ± 0.2	**87.2 ± 3.9**	**1.1 ± 0.1**
V-Pulmonary	66.3 ± 9.0	2.0 ± 0.9	69.7 ± 7.5	2.0 ± 1.2	**71.4 ± 8.1**	**1.9 ± 1.2**
V-Venacava-S	87.6 ± 2.2	0.9 ± 0.1	87.0 ± 8.9	1.0 ± 0.8	**88.2 ± 4.2**	**0.7 ± 0.3**
V-Venacava-I	89.6 ± 2.0	0.8 ± 0.2	88.6 ± 4.5	0.9 ± 0.3	**89.8 ± 2.8**	**0.7 ± 0.2**
Bronchi-Prox	82.3 ± 3.3	1.3 ± 0.3	86.3 ± 1.5	0.9 ± 0.3	**86.7 ± 1.8**	**0.9 ± 0.5**
Chest Wall-L	96.3 ± 1.5	1.1 ± 0.8	**98.5 ± 0.8**	**0.2 ± 0.2**	97.3 ± 2.3	0.6 ± 0.6
Chest Wall-R	96.8 ± 1.7	1.1 ± 0.9	**99.4 ± 0.1**	**0.2 ± 0.1**	98.0 ± 1.8	0.3 ± 0.3
Ave Required	81.2	1.8	85.5	1.0	**85.8**	**0.9**
Ave Optional	86.7	1.3	88.1	1.0	**88.5**	**0.9**
Ave All	84.3	1.5	87.1	1.0	**87.3**	**0.9**

[a] Bold values denote the best performance scores.

easy OARs that required the least annotation time: V-Pulmonary (3.6 min), A-Pulmonary (3.4 min), V-Venacava-S (1.5 min), and V-Venacava-I (1.3 min). Given the lung cancer patient's larger physical size compared with the esophageal cancer patient, this difference was reflected in the size of the organs on the CT images, resulting in a longer annotation time of 116.3 min for the lung cancer patient, as opposed to 100.6 min for the esophageal cancer patient. Despite these size disparities, we found that annotating diseased areas demanded significantly more effort than annotating healthy ones. For example, labeling the esophagus in the CT image of the esophageal cancer patient required an extra minute compared with the lung cancer patient. Furthermore, annotating the lungs in the lung cancer patient's CT image took 6.1 min longer compared with the esophageal cancer patient.

Table 3 presents a comparison between the radiation oncologists' manually delineated contours and the gold standard reference. The DSC and ASD for the required OARs were reported as 81.2% and 1.8 mm, respectively, whereas for the optional OARs, they were 86.7% and 1.3 mm, respectively. It was noted that the delineated contours of the brachial plexus demonstrated the greatest variance when compared to the gold standard reference, with the average DSC among multiple users being only 37.2%. This was due to inherent limitations in visual clarity and tissue contrast in CT images. According to the NRG-RTOG 1106 guidelines, the brachial plexus originates from spinal nerves exiting the spinal canal at levels C4–C5 (C5 nerve roots) to T1–T2 (T1 nerve roots). Although it may be possible to identify the actual roots and trunks of the brachial plexus directly through high-quality CT scanning with intravenous contrast, distinguishing it from non-contrast or low-quality contrast CT images can be extremely challenging, even for experienced radiation oncologists. Smaller organs, such as the V-Pulmonary, were generally labeled accurately in terms of their spatial locations, yet they recorded a DSC of 66.3%. The spinal cord, however, exhibited a relatively low DSC of 81.6%, coupled with a high ASD of 2.8 mm. A detailed review of the manually labeled contours revealed that the inferior performance could be traced back to three groups not labeling the spinal cord down to its lower extent at L2. Specifically, one group concluded their labeling at T8, whereas the other two groups stopped their labeling at approximately T10.

3.3.2. DeepOAR predictions

As illustrated in Table 3, DeepOAR yielded robust and accurate predictions, demonstrating superior delineation performance in comparison to manually annotated contours from scratch. The DSCs for the required, optional, and all OARs were enhanced from 81.2% to 85.5%, 86.7% to 88.1%, and 84.3% to 87.1%, respectively. This improvement was accompanied by reductions in ASD from 1.8 mm to 1.0 mm, 1.3 mm to 1.0 mm, and 1.5 mm to 1.0 mm, respectively. Notably, the delineation of the brachial plexus experienced a significant increase of 13.0% in absolute DSC. Furthermore, the ASD for the spinal cord and lungs saw considerable reductions—from 2.8 mm to 0.5 mm and from 1.3 mm to 0.1 mm, respectively—showcasing the enhanced accuracy and precision of DeepOAR.

3.3.3. Editing on DeepOAR predictions

Revising DeepOAR predictions to be clinically acceptable required an average of 12.0 min for editing the required OARs and 13.0 min for the optional OARs, as summarized in Table 2. This represents a considerable reduction in editing effort compared with manual annotation from scratch, with reductions of 80% for the required OARs and 74% for the optional OARs. The reduction in editing effort for most organs was directly correlated with improvements in segmentation accuracy. Significant time savings were observed: the lungs experienced reductions of 11.8 (−88%) and 12.4 (−89%) minutes, the spinal cord saw a decrease of 6.4 (−84%) minutes, the A-Aorta had 5.9 (−87%) minutes less editing time, the Bronchi-Prox required 5.4 (−79%) minutes less, the esophagus saw a reduction of 5.3 (−78%) minutes, and the pericardium had 5.0 (−72%) minutes less editing time. Smaller organs demonstrated an efficient revision process, requiring less time for editing—the V-Pulmonary had a reduction from 3.6 min to 0.8 min,

A-Pulmonary saw a decrease from 3.4 min to 0.7 min, V-Venacava-S had a reduction from 1.5 min to 0.4 min, and V-Venacava-I experienced a reduction from 1.3 min to 0.4 min.

Conversely, larger and more complex organs, despite showing a considerable reduction in editing effort in terms of absolute minutes, demanded more time for revisions. The most significant revision times were allocated to the chest walls, requiring a combined total of 8.4 min, and the brachial plexus, which required 3.7 min of editing. Excluding the brachial plexus, the remaining five required OARs necessitated 8.2 min of revision to meet the clinical acceptability criteria. Similarly, excluding the chest walls, the remaining six optional OARs required just 4.6 min of revision to achieve clinical acceptability. In terms of accuracy metrics, the average DSC and ASD for the required OARs further improved to 85.8% and 0.9 mm, respectively, whereas for optional OARs, they were enhanced to 88.5% and 0.9 mm, respectively.

4. Discussion

In this multi-institutional study, we proposed DeepOAR, a unified DL-based OAR contouring model, that was trained using multiple partially labeled datasets for segmenting a comprehensive set of thoracic OARs following the RTOG-guided atlas. By training it on a union of five partially labeled thoracic organ datasets, DeepOAR showed significantly improved accuracy than the state-of-the-art medical image segmentation method nnUNet[26] (DSC: 88.4 % of DeepOAR vs. 84.6 % of Single nnUNet Model; ASD: 1.0 mm of DeepOAR vs. 1.35 mm of Single nnUNet Model). Multi-user experiments demonstrated that 89.7 % of DeepOAR-generated OARs were acceptable or required only minor revisions in clinical usage. Compared with the manual delineation performance, the delineation accuracy of DeepOAR was notably lower compared with inter-user variation (DSC: 87.3 % of DeepOAR vs. 84.3 % of human reader). When editing on DeepOAR's predictions, the manual contouring time can be substantially reduced by 77.0 % (from 108.7 to 25.0 min). These results highlight the strong clinical applicability of DeepOAR in the OAR delineation process in thoracic cancer radiotherapy workflows, along with improved efficiency, comprehensiveness, and quality.

The RTOG is an internationally recognized reference standard for contouring OARs in clinical radiotherapy, and the NRG-RTOG 1106 atlas[6] has been widely used as the standard for lung cancer trials. Previous studies on thoracic OAR auto-segmentation have gained popularity in the literature; however, they only segmented a small number of OARs[20,21,24] (normally 4–5 OARs) and trained the DL network using a small single dataset (< 100 patients) that had limited generalizability to test cases from other institutions. To the best of our knowledge, this study is the first to segment a comprehensive set of 14 thoracic OARs according to the NRG-RTOG 1106 guideline. We accomplished this by developing a new unified DL segmentation model capable of learning diverse OAR features from a large collection of multiple partially labeled datasets, encompassing a total of 531 patients. Compared with previous studies,[27,28] our method shows significant improvement in accuracy and time efficiency.

A total of 20 radiation oncologists from 14 institutes participated in the qualitative multi-user evaluations, offering a diverse range of perspectives and expertise. This ensured a comprehensive assessment of the artifical intelligence (AI) model's performance from various professional viewpoints. For the quantitative multi-user evaluation, 11 radiation oncologists from eight different institutes participated, simulating real-world clinical practice scenarios. This experimental design allowed for a practical evaluation of the AI model's accuracy and usability in clinical environments, thereby validating the AI model's applicability in the workflows of radiotherapy planning.

Automated segmentation significantly reduces the subjective biases and inconsistencies that occur in manual segmentation due to variations in human expertise and interpretation. DeepOAR provides objective and consistent segmentation results, outperforming radiation oncologists by a substantial DSC margin ($+2.8\%$). When radiation oncologists edited DeepOAR's predictions, the average DSC value further improved to 87.3%, demonstrating the effectiveness and consistency of automated segmentation over manual methods.

It is worth noting that editing easy and large OARs, such as the lungs, pericardium, and chest walls, yielded slightly decreased performance compared to the fully automated predictions by DeepOAR. This slight performance drop may be attributed primarily to discrepancies in determining the starting and ending slices of these organs by radiation oncologists during the editing process. Such discrepancies become even more noticeable when radiation oncologists manually annotate these organs from scratch. For instance, discrepancies in selecting the starting and ending slices can be reflected in the distance metric. Manual annotation led to ASDs of 1.3 mm, 2.5 mm, and 1.1 mm for the lungs, pericardium, and chest walls, respectively. In comparison, when editing on DeepOAR's predictions, the ASDs for these organs were significantly reduced to 0.1 mm, 1.7 mm, and 0.5 mm, respectively. The significant reduction in ASD further demonstrates the effectiveness of DeepOAR in reducing human observers' biases and variations.

This study has several limitations. First, the segmentation accuracy for the brachial plexus was low compared with the other OARs. This was mainly due to the fact that identifying the brachial plexus in CT scans can be challenging. Radiation oncologists delineate the brachial plexus by first locating the adjacent key anatomic landmarks and then infering the contour of the brachial plexus. Therefore, one potential way to improve segmentation accuracy is to segment the brachial plexus's adjacent anatomies, such as the subclavian and axillary vessels, and use these supporting structures to help segment the brachial plexus. Second, it was observed that segmenting the chest wall required more editing efforts compared with the lungs, particularly around the areas deformed by lung tumors. Future efforts should involve curating more data on lung cancer patients to strengthen DeepOAR's ability to segment the chest wall. Finally, this is a retrospective study. DeepOAR can be integrated into the current clinical workflow to prospectively examine its role in radiotherapy treatment.

5. Conclusions

We developed DeepOAR, a unified DL-based OAR contouring model trained using a collection of multiple partially labeled datasets, to segment a comprehensive set of 14 thoracic OARs following the RTOG-guided atlas. Extensive quantitative and

qualitative validation demonstrated that DeepOAR achieved accurate and robust performance and produced higher accuracy in OAR segmentation compared with inter-user variation. Radiation oncologists needed minor revisions or no revision for DeepOAR-predicted OARs to warrant clinical acceptance. DeepOAR exhibited strong potential for adoption in the clinical radiotherapy workflow, providing a more standardized, quantitatively accurate, and efficient approach to OAR contouring.

Funding

Xianghua Ye, Zhejiang Provincial Science and Technology Project, 2024-KY1-001-105. This funding is intended to support the training of organ auto-segmentation models.

CRediT authorship contribution statement

H. H. Y: Validation, Investigation. **J. H. Z:** Validation, Investigation. **L. Z:** Validation, Investigation. **T. T. Z:** Validation, Formal analysis. **Q. Z:** Validation, Investigation. **X. H. Y:** Writing – review & editing, Writing – original draft, Data curation, Conceptualization. **L. Z:** Validation, Investigation. **D. Z. G:** Writing – original draft, Methodology, Formal analysis. **W. C. G:** Validation, Formal analysis. **L. J. Z:** Validation, Project administration, Data curation. **Z. H. X. J:** Validation, Investigation. **C. Y. X:** Visualization, Validation, Project administration. **L. L:** Software, Methodology, Data curation. **X. Z. Z:** Validation, Formal analysis. **P. Y. W:** Software, Project administration, Methodology. **D. D. Z:** Validation, Formal analysis, Data curation. **H. H. L:** Validation, Formal analysis. **P. P. D:** Validation, Formal analysis. **H. Z:** Software. **J. L. L:** Validation, Formal analysis, Data curation. **S. X. Y:** Project administration, Funding acquisition, Formal analysis, Conceptualization. **X. S:** Validation, Resources, Formal analysis. **D. K. J:** Writing – review & editing, Writing – original draft, Software, Methodology, Data curation, Conceptualization. **H. L. C:** Validation, Methodology, Formal analysis. **F. S. K:** Writing – review & editing, Writing – original draft, Project administration, Funding acquisition, Data curation, Conceptualization. **J. Z. C:** Validation, Investigation. **X. Z. Z:** Formal analysis, Methodology. **X. H:** Validation, Formal analysis. **K. R:** Formal analysis. **W. W. K:** Validation, Formal analysis. **M. E. L:** Validation, Formal analysis. **P. P:** Validation, Investigation. **J. T. R:** Validation, Investigation. **C. M. T:** Validation, Investigation. **T. L. W:** Validation, Investigation.

Ethical statement

This study was approved by the Clinical Research Ethics Committee of the First Affiliated Hospital of Zhejiang University (Approval No.: IIT20231208A).

Declaration of Competing Interest

Feng-Ming (Spring) Kong and Le Lu are editorial board members for Intelligent Oncology and was not involved in the editorial review or the decision to publish this article. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported in parts by the Shenzhen Science and Technology Program (KQTD20180411185028798) and High Level-Hospital Program, Health Commission of Guangdong Province, China (HKUSZH201901038).

Appendix A. Supplementary material

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.intonc.2024.12.001.

References

1. Kong FM, Ritter T, Quint DJ, et al. Consideration of dose limits for organs at risk of thoracic radiotherapy: atlas for lung, proximal bronchial tree, esophagus, spinal cord, ribs, and brachial plexus. *Int J Radiat Oncol Biol Phys.* 2011;81(5):1442–1457.
2. Glide-Hurst CK, Chetty IJ. Improving radiotherapy planning, delivery accuracy, and normal tissue sparing using cutting edge technologies. *J Thorac Dis.* 2014;6(4):303–318.
3. Banfill K, Giuliani M, Aznar M, et al. Cardiac toxicity of thoracic radiotherapy: existing evidence and future directions. *J Thorac Oncol.* 2021;16(2):216–227.
4. Liu F, Zhao B, Krug LM, et al. Assessment of therapy responses and prediction of survival in malignant pleural mesothelioma through computer-aided volumetric measurement on computed tomography scans. *J Thorac Oncol.* 2010;5(6):879–884.
5. Rubin GD. Computed tomography: revolutionizing the practice of medicine for 40 years. *Radiology.* 2014;273(2):S45–S74.
6. Kong FS, Hu C, Pryma DA, et al. Primary results of NRG-RTOG1106/ECOG-ACRIN 6697: a randomized phase II trial of individualized adaptive (chemo)radiotherapy using midtreatment ^{18}F-fluorodeoxyglucose positron emission tomography/computed tomography in stage III non-small cell lung cancer. *J Clin Oncol.* 2024;42(33):3935–3946.
7. Jin D, Guo D, Ge J, Ye X, Lu L. Towards automated organs at risk and target volumes contouring: defining precision radiation therapy in the modern era. *J Natl Cancer Cent.* 2022;2(4):306–313.
8. Kosmin M, Ledsam J, Romera-Paredes B, et al. Rapid advances in auto-segmentation of organs at risk and target volumes in head and neck cancer. *Radiother Oncol.* 2019;135:130–140.

9. Jin D, Chen Z, Jiang J, et al. Accurate esophageal gross tumor volume segmentation in PET/CT using two-stream chained 3D deep network fusion. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2019*. 2019:182–190.
10. Jin D, Zhang L, Liang Y, et al. Deep esophageal clinical target volume delineation using encoded 3D spatial context of tumors, lymph nodes, and organs at risk. In: *Lecture Notes in Computer Science. Medical Image Computing and Computer-Assisted Intervention – MICCAI 2019*. 2019:603–612.
11. Guo D, Jin D, Zhu Z, et al. Organ at risk segmentation for head and neck cancer using stratified learning and neural architecture search. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020:4223–4232.
12. Rios Velazquez E, Aerts HJ, Gu Y, et al. A semiautomatic CT-based ensemble segmentation of lung tumors: comparison with oncologists' delineations and with the surgical specimen. *Radiother Oncol.* 2012;105(2):167–173.
13. Tang H, Chen X, Liu Y, et al. Clinically applicable deep learning framework for organs at risk delineation in CT images. *Nat Mach Intell.* 2019;1(10):480–491.
14. Ye X, Guo D, Ge J, et al. Comprehensive and clinically accurate head and neck cancer organs-at-risk delineation on a multi-institutional study. *Nat Commun.* 2022;13(1):6137 [Published 2022 Oct 17].
15. Shi F, Hu W, Wu J, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nat Commun.* 2022;13(1):6566 [Published 2022 Nov 2].
16. Antonelli M, Reinke A, Bakas S, et al. The medical segmentation decathlon. *Nat Commun.* 2022;13(1):4128 [Published 2022 Jul 15].
17. Guo D, Ye X, Ge J, et al. DeepStationing: thoracic lymph node station parsing in CT scans using anatomical context encoding and key organ auto-search. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2021*. 2021:3–12.
18. Yan K, Cai J, Zheng Y, et al. Learning from multiple datasets with heterogeneous and partial labels for universal lesion detection in CT. *IEEE Trans Med Imaging.* 2021;40(10):2759–2770.
19. Ji Z, Guo D, Wang P, et al. Continual segment: Towards a single, unified and non-forgetting continual segmentation model of 143 whole-body organs in ct scans. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023:21140–21151.
20. Francis S, Jayaraj PB, Pournami PN, et al. ThoraxNet: a 3D U-Net based two-stage framework for OAR segmentation on thoracic CT images. *Phys Eng Sci Med.* 2022;45(1):189–203.
21. Feng X, Bernard ME, Hunter T, Chen Q. Improving accuracy and robustness of deep convolutional neural network based thoracic OAR segmentation. *Phys Med Biol.* 2020;65(7):07NT01 [Published 2020 Mar 31].
22. Shi F, Hu W, Wu J, et al. Deep learning empowered volume delineation of whole-body organs-at-risk for accelerated radiotherapy. *Nat Commun.* 2022;13(1):6566 [Published 2022 Nov 2].
23. Wong J, Huang V, Wells D, et al. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiat Oncol.* 2021;16(1):101 [Published 2021 Jun 8].
24. Lambert Z, Petitjean C, Dubray B, Kuan S. SegTHOR: Segmentation of Thoracic Organs at Risk in CT Images. arXiv (Cornell University); Published online November 9, 2020. p. 1–6.
25. Aerts HJ, Velazquez ER, Leijenaar RT, et al. Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach [published correction appears in Nat Commun. 2014;5:4644. Cavalho, Sara [corrected to Carvalho, Sara]]. *Nat Commun.* 2014;5:4006 [Published 2014 Jun 3].
26. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods.* 2021;18(2):203–211.
27. Heilemann G, Buschmann M, Lechner W, et al. Clinical implementation and evaluation of auto-segmentation tools for multi-site contouring in radiotherapy. *Phys Imaging Radiat Oncol.* 2023;28:100515 Published 2023 Nov 17.
28. Garrett Fernandes M, Bussink J, Stam B, et al. Deep learning model for automatic contouring of cardiovascular substructures on radiotherapy planning CT images: dosimetric validation and reader study based clinical acceptability testing. *Radiother Oncol.* 2021;165:52–59.