Deep learning model for automatic contouring of cardiovascular substructures on radiotherapy planning CT images

Fernandes, Miguel Garrett; Bussink, Johan; Stam, Barbara; Wijsman, Robin; Schinagl, Dominic A. X.; Monshouwer, Rene; Teuwen, Jonas

Published in:
Radiotherapy and Oncology

DOI:
10.1016/j.radonc.2021.10.008

Publication date:
2021

Radiotherapy and Oncology 165 (2021) 52–59

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Original Article

Deep learning model for automatic contouring of cardiovascular substructures on radiotherapy planning CT images: Dosimetric validation and reader study based clinical acceptability testing

Miguel Garrett Fernandes [a,b,2,*], Johan Bussink [a], Barbara Stam [c], Robin Wijsman [d], Dominic A.X. Schinagl [a], René Monshouwer [a,1], Jonas Teuwen [b,c,1]

[a] *Department of Radiation Oncology, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands;* [b] *Department of Medical Imaging, Radboud Institute for Health Sciences, Radboud University Medical Center, Nijmegen, The Netherlands;* [c] *Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands;* [d] *Department of Radiation Oncology, University Medical Center Groningen, Groningen, The Netherlands*

ARTICLE INFO

Article history:
Received 9 June 2021
Received in revised form 27 September 2021
Accepted 11 October 2021
Available online 21 October 2021

Keywords:
Automatic cardiac contouring
Segmentation
Lung cancer
NSCLC
Surface dice
Cardiac radiotoxicity

ABSTRACT

Background and purpose: Large radiotherapy (RT) planning imaging datasets with consistently contoured cardiovascular structures are essential for robust cardiac radiotoxicity research in thoracic cancers. This study aims to develop and validate a highly accurate automatic contouring model for the heart, cardiac chambers, and great vessels for RT planning computed tomography (CT) images that can be used for dose–volume parameter estimation.

Materials and methods: A neural network model was trained using a dataset of 127 expertly contoured planning CT images from RT treatment of locally advanced non-small-cell lung cancer (NSCLC) patients. Evaluation of geometric accuracy and quality of dosimetric parameter estimation was performed on 50 independent scans with contrast and without contrast enhancement. The model was further evaluated regarding the clinical acceptability of the contours in 99 scans randomly sampled from the RTOG-0617 dataset by three experienced radiation oncologists.

Results: Median surface dice at 3 mm tolerance for all dedicated thoracic structures was 90% in the test set. Median absolute difference between mean dose computed with model contours and expert contours was 0.45 Gy averaged over all structures. The mean clinical acceptability rate by majority vote in the RTOG-0617 scans was 91%.

Conclusion: This model can be used to contour the heart, cardiac chambers, and great vessels in large datasets of RT planning thoracic CT images accurately, quickly, and consistently. Additionally, the model can be used as a time-saving tool for contouring in clinic practice.

Cardiac toxicity associated with radiotherapy of lung cancer patients has been a focus of research since correlations between cardiac dose and overall survival (OS) emerged from studies such as the RTOG-0617 [1,19,20], and after similar relationships had been found for breast cancer and lymphoma patients [2]. A 2019 systematic review [3] analyzed the association between whole-heart dosimetric parameters and outcomes in 22 studies, including 5614 unique non-small cell lung cancer (NSCLC) patients, and observed statistically significant associations in multivariable analyses only in 10 cases. Additionally, no specific parameter was consistently related with OS or cardiac events across multiple studies. Only five of the analyzed articles investigated relationships between OS and dosimetric parameters of specific cardiac substructures [4–8]. However, statistically significant relationships were found in all of these, namely, regarding dose to the pulmonary artery (PA), superior vena cava (SVC), and heart base. Later studies have continued to report findings related to dose to substructures and outcomes [9]. Other developments also suggest that sparing of the heart and other large blood reservoirs reduces dose

Abbreviations: AO, aorta; CA, clinically acceptable; CT, computed tomography; DL, Deep Learning; DVH, dose–volume histogram; GTV, gross tumor volume; HT, heart; IVC, inferior vena cava; LA, left atrium; LL, left lung; LV, left ventricle; MAE, mean absolute error; NCI, national cancer institute; NCTN, National Clinical Trials Network; NSCLC, non-small cell lung cancer; NTCP, normal tissue complication probability; OAR, organ at risk; OS, overall survival; PA, pulmonary artery; RA, right atrium; RC, requires corrections; RL, right lung; RT, radiotherapy; RV, right ventricle; SVC, superior vena cava.

* Corresponding author at: PO Box 9101, 6500 HB Nijmegen, The Netherlands.
E-mail address: miguel.fernandes@radboudumc.nl (M. Garrett Fernandes).

[1] Shared last author.

[2] Author Responsible for Statistical Analysis.

https://doi.org/10.1016/j.radonc.2021.10.008
0167-8140/© 2021 The Authors. Published by Elsevier B.V.

to the blood pool, which may decrease the probability of lymphopenia [10]. Therefore, current cardiac radiotoxicity literature remains inconclusive, with research presently shifting more towards investigating potential radiosensitive cardiac substructures and dose to the blood pool.

Cardiac toxicity research is limited in part by the availability of accurate and consistent contours for cardiovascular structures [11]. In this sense, automatic contouring tools can potentially offer a solution to contour cardiovascular structures in larger datasets [12,13]. Payer et al. [14] developed a cardiac contouring model that achieved an average of 90% volumetric Dice for the heart substructures [15]. Unfortunately, there was no evaluation of the clinical acceptability of these automatic contours. Furthermore, the algorithm was developed using Computed Tomography (CT) images acquired following a coronary CT angiography protocol that do not represent those found in radiotherapy planning regarding acquisition parameters or patient anatomy (absence of tumor). Other models such as those trained with images from breast cancer patients [16] are unlikely to generalize as well for lung cancer patient images which regularly contain tumors close to the heart. Despite the current state of the art for automatic cardiac contouring in lung cancer patients reporting high accuracies for the whole heart, the cardiac chambers and the great vessels [17,18], it still lacks in clinical validation. Specifically, model accuracy has been evaluated in datasets of only a few dozens of patients, often from a single institute and dosimetric parameter estimation and clinical acceptability has either not been evaluated [17], or only evaluated at the same scale [18]. Impact of adjacency of the tumor to the contoured structures has also not been reported. However, due to the high anatomical variability of tumors and the perturbation they can cause on the surrounding anatomy, tumor adjacency is a major confounding factor for automatic contouring and should also be evaluated.

We aim to improve automatic cardiac contouring accuracy for RT planning CTs by employing a 3D Deep Learning (DL) model with a newly introduced inductive bias, while providing extensive clinical validation results using a multi-institutional dataset and evaluating clinical acceptability in the RTOG-0617 dataset [1,19,20]. Active learning was used to improve model generalizability and an evaluation of the impact of tumor adjacency to the contoured structure on model accuracy is provided.

Methods

Data

For model development, two independent datasets, RUMC [21] and ML1 [22], were used and are described in detail in Table 1. A clinical acceptability study was then performed on a dataset sampled from the RTOG-0617 data [1,19,20].

The RUMC dataset [21] consisted of 157 RT planning 3D CT scans from 157 patients treated for irresectable advanced stage NSCLC at the Radboud University Medical Center between 2008 and 2014. The heart and the cardiac chambers of all CT scans were manually delineated by an expert radiation oncologist with more than ten years of experience (RW) following Feng et al. [23]. The great vessels were delineated (see Supplemental Material, Section 1), and independently verified by a radiation oncologist (RW). The expert contours of the heart (HT), left ventricle (LV), right ventricle (RV), left atrium (LA), right atrium (RA), aorta (AO), PA, SVC, and inferior vena cava (IVC), were used as the ground truth for our automatic contouring model. To evaluate the impact of adjacency between the tumor and these structures on contouring accuracy, tumor-structure adjacency was automatically assessed using the expert contours and the planning gross tumor volumes (GTVs) (including irradiated lymph nodes, see Table 1). The dataset was randomly split on a patient-level into a training, validation, and test subset, composed of 111, 16, and 30 patients, respectively. For the training and test subsets, scans were stratified by specific scanner types and patient sex.

The public ML1 dataset consisted of 422 non-contoured 3D CT scans from 422 stage I–IIIb NSCLC patients acquired for RT planning at the MAASTRO clinic [22]. This dataset was used to introduce multi-institutionality to the test data and to identify problematic cases to be fed back to the model for fine-tuning. On that basis, we selected 20 problematic cases by predicting the heart contour with a preliminary model on the entire dataset and visually choosing the patients for which the predictions were problematic, cases of generalized failure or with systematic errors, excluding outliers with severe artifacts, pericardial effusion, or pneumectomy. Subsequently, all structures were manually contoured and independently verified by an expert radiation oncologist (RW) (procedure shown in Fig. SM2.1). The addition of these scans to the test set biases the contour accuracy results negatively. However, these should also have the highest potential to increase the model's generalizability after fine-tuning in the test set before it being used to contour the clinical acceptability study data (see Section 2.2).

Model

For the automatic contouring algorithm, we implemented the DL architecture proposed by Nikolov et al. [24] using Pytorch 1.4 [25]. This architecture, inspired by the original U-net architecture [26], was used by Nikolov et al. to achieve expert-level contouring accuracy for multiple head-and-neck Organs at Risk (OAR)s in planning CT volumes. Compared to the conventional U-net and 3D U-net [27] architectures, there are some notable differences. Instead of 3–4 contracting and expanding levels, this architecture consists of 8 such levels and is therefore capable of more complex abstractions. This depth is made possible by the skip-connections present in each contracting and expanding module, which promote residual learning, improving training efficiency and attenuating the vanishing gradients problem. To further increase the receptive field, the lowest level consists of a fully-connected block rather than a convolutional block. A detailed description of the model architecture can be found in Nikolov et al. [24]. With respect to the approach of Nikolov et al., the model used in this study included an additional inductive bias in the output layer, as this improved training efficiency and validation accuracy. Other differences include the loss function, post processing, and the use of active learning. All of these are elaborated upon in the ensuing text and the Supplemental Material.

The model takes 3D inputs of size 21 × 512 × 512, slices, rows, and columns and outputs the segmentation map of 11 labels: HT, LV, RV, LA, RA, AO, PA, SVC, IVC, Left Lung (LL), and Right Lung (RL). Prediction of the LL and RL was included as an extra supervisory signal for regularization purposes, using the planning contours as the ground truth. A new inductive bias was introduced to drive the model towards predicting masks for the substructures, LV, RV, LA, and RA, within the predicted mask for the HT by multiplying their output probabilities with the HT probability. Additionally, all structures except the HT were constrained to not overlap with each other by using a softmax layer. Cross entropy was used as the loss function. The formal definition of the output layer and the loss function can be found in Section 2 of the Supplemental Material.

The neural network weights were optimized using RAdam [28]. To avoid overfitting, training was stopped when volumetric Dice, Dice_{vol}, peaked in the RUMC validation set. We then used this model to contour the test set (30 RUMC scans + 20 ML1 scans).

Table 1
Dataset characteristics.

			RUMC			ML1		Total
		Total	Train	Validation	Test	Total	Test	
Patients		157	111	16	30	422	20	579
Scans		157	111	16	30	422	20	579
Manually Delineated		157	111	16	30	20	20	177
Gender	Males	89	62	9	18	290	17	379
	Females	68	49	7	12	132	3	200
Contrast Enhanced		155	110	16	29	145	9	300
Scanner	Philips Brilliance Big Bore	122	88	13	21	0	0	122
	SIEMENS Emotion Duo	18	14	1	3	0	0	18
	SIEMENS SOMATOM	15	9	1	5	0	0	15
	SIEMENS Biograph40	2	0	1	1	184	7	186
	Canon XiO	0	0	0	0	97	7	97
	SIEMENS Sensation Open	0	0	0	0	111	5	111
	SIEMENS Sensation 16	0	0	0	0	25	3	25
	SIEMENS Sensation 10	0	0	0	0	4	0	4
Tumor Adjacency	HT	122	89	11	22	–	11	–
	LV	15	10	0	5	–	5	–
	RV	2	2	0	0	–	2	–
	LA	113	87	8	18	–	9	–
	RA	22	18	0	4	–	1	–
	AO	133	96	11	26	–	11	–
	PA	147	103	14	30	–	14	–
	SVC	118	88	10	20	–	8	–
	IVC	5	4	0	1	–	0	–

RUMC and ML1 dataset characteristics important for contouring accuracy. The RUMC dataset was randomly split into training, validation, and test set. The test set was stratified by sex and specific scanner types. Of all ML1 data, 20 scans were visually selected for the test set due to generalized failure or systematic errors from a preliminary cardiac contouring model, excluding those with severe artifacts, pericardial effusion, or pneumectomy. Tumor was considered adjacent to an ROI if the volume of both structures, taken from the planning gross tumor volume (GTV) and the manual contours, respectively, were in direct contact. The GTV included irradiated lymph nodes. Contrast labels in the ML1 dataset were automatically estimated from the HU distribution in the AO and visually confirmed in the 20 test set cases.

For a detailed description of the training procedure, see Section 2 of the Supplemental Material.

The quality of the automatic segmentations was evaluated against the expert contours using Dice_{vol} and surface Dice, $\text{Dice}_{\text{surface}}$ [24]. Unlike Dice_{vol}, $\text{Dice}_{\text{surface}}$ is not biased towards structures with small surface-to-volume ratios and accounts for contouring uncertainty. A tolerance of 3 mm was used for $\text{Dice}_{\text{surface}}$ computation, which corresponds to the typical longitudinal scan resolution in our datasets. To evaluate the model's usefulness for dosimetric feature computation, we compared the dose-volume parameters of each structure computed using the expert contours against those computed using the automatic contours. Namely, the difference between the Dose Volume Histograms (DVHs) computed with both contours was evaluated using the mean absolute error (MAE). This consisted in sampling both DVHs in 5 Gy intervals starting from 5 Gy up to the maximum dose and computing the absolute error for each of those points.

A fine-tuning step was performed after the aforementioned quantitative evaluation by adding the 50 test scans to the training and validation sets and continuing training from the previously optimized point (see Supplemental Material, Section 2). This fine-tuning step allowed the model to also learn from the 20 problematic cases chosen from the ML1 dataset. These included several examples of non-contrast CTs which were lacking in the training and validation sets (Table 1). The fine-tuned version of the model was used in the clinical acceptability study.

Evaluation of clinical acceptability

To evaluate the automatic contours' clinical acceptability, an internal reader study was set up using the Grand Challenge software [29]. One hundred CT images from the RTOG-0617 dataset were selected at random and automatically contoured using our model. One patient was excluded due to a postprocessing artifact in the base of the heart in the original image. The entire volumes, with the superimposed contours, were displayed one at a time in the axial view. Three expert radiation oncologists rated the contours individually (RW, JB with 30 plus years of experience, and DS with 20 plus years of experience). For each image, the readers were asked to rate each contour as (i) "Requires corrections. Large, obvious errors" (#RCLE), (ii) "Requires corrections. Minor errors" (#RCME), (iii) "Clinically acceptable. Errors not clinically significant" (#CANSE), or (iv) "Clinically acceptable. Contours are highly accurate" (#CAHA). Here, "clinically acceptable" was defined as acceptable for use in treatment planning.

The decision of providing four labels, two acceptable and two requiring corrections was driven by two goals: having each contour labeled as acceptable or not and simultaneously giving the possibility of differentiating contours within the acceptable and requiring corrections categories. At the end of the reader study, each contour had been given one evaluation by each radiation oncologist. If at least two of them were acceptable (#CANSE or #CAHA) then the contour was considered acceptable. This majority vote methodology was defined before the reader study started, hence the inclusion of three radiation oncologists in the study. Due to the nature of this methodology, some accepted contours were labeled non-acceptable by one radiation oncologist. The atlas [23] on which the expert heart contours for the RUMC and ML1 test set were based, was provided to each reader. The readers did not discuss between themselves any specifics related to the study or the contour evaluations and were aware that the contours were made by an automatic algorithm.

Results

Table 2 presents the median $\text{Dice}_{\text{surface}}$ at 3 mm tolerance, $\text{Dice}_{\text{surface|3mm}}$, and Dice_{vol} for all structures in the RUMC test set and the ML1 dataset (quartile information in Tables SM4.1 and SM4.2). The median of the average $\text{Dice}_{\text{surface|3mm}}$ for all structures

Table 2
Median surface Dice and volumetric Dice scores for all structures in the RUMC test set and the ML1 dataset.

			HT	LV	RV	LA	RA	AO	PA	SVC	IVC	Mean
Dice$_{surface\|3mm}$	RUMC	–	0.95	0.92	0.88	0.91	0.88	0.99	0.97	0.96	0.88	0.92
		CE	0.96	0.92	0.88	0.91	0.88	0.99	0.97	0.96	0.88	0.92
		NCE	0.74	0.71	0.68	0.81	0.63	0.74	0.90	0.65	0.11	0.66
	ML1	–	0.89	0.83	0.79	0.83	0.86	0.93	0.92	0.88	0.69	0.85
		CE	0.91	0.87	0.86	0.84	0.88	0.97	0.97	0.90	0.82	0.90
		NCE	0.88	0.82	0.70	0.82	0.80	0.77	0.87	0.87	0.61	0.75
	Test set (RUMC + ML1)	–	**0.94**	**0.88**	**0.88**	**0.87**	**0.87**	**0.99**	**0.96**	**0.93**	**0.82**	**0.90**
		CE	0.94	0.92	0.88	0.90	0.88	0.99	0.97	0.95	0.88	0.91
		NCE	0.87	0.77	0.69	0.82	0.77	0.76	0.89	0.86	0.59	0.75
Dice$_{vol}$	RUMC	–	0.96	0.93	0.89	0.88	0.88	0.95	0.91	0.88	0.78	0.89
		CE	0.96	0.93	0.89	0.88	0.88	0.95	0.92	0.88	0.78	0.89
		NCE	0.89	0.82	0.82	0.85	0.73	0.75	0.83	0.54	0.03	0.70
	ML1	–	0.95	0.90	0.84	0.86	0.87	0.92	0.89	0.82	0.63	0.86
		CE	0.95	0.91	0.88	0.89	0.90	0.95	0.92	0.86	0.78	0.89
		NCE	0.94	0.90	0.75	0.84	0.85	0.80	0.87	0.77	0.54	0.77
	Test set (RUMC + ML1)	–	**0.95**	**0.92**	**0.88**	**0.87**	**0.88**	**0.94**	**0.91**	**0.86**	**0.74**	**0.88**
		CE	0.96	0.92	0.89	0.89	0.88	0.95	0.92	0.88	0.78	0.89
		NCE	0.94	0.87	0.78	0.85	0.85	0.79	0.86	0.76	0.45	0.77

Median surface Dice at 3 mm tolerance, Dice$_{surface|3mm}$, and volumetric Dice, Dice$_{vol}$, scores for all 9 structures in the entire test set and independently for the 30 patients of the RUMC test set and the 20 patients of the ML1 dataset. Results for contrast enhanced (CE) and non-contrast enhanced (NCE) CTs are also given independently.

Table 3
Tumor adjacency effect on surface dice.

ROI	Tumor Adjacency to ROI			
	With	Without	p	Occurrence (%)
HT	0.957	0.955	0.981	72
LV	0.922	0.923	0.470	17
RV	–	0.893	–	0
LA	0.917	0.912	0.946	62
RA	0.898	0.882	0.548	14
AO	0.993	0.990	0.728	86
PA	0.972	–	–	100
SVC	**0.951**	**0.972**	**0.126**	69
IVC	0.910	0.878	0.676	3

Median Surface Dice at 3 mm tolerance for all structures, split by presence of tumor adjacency to the respective ROI in contrast enhanced CT scans of the RUMC test set. Difference between groups evaluated for statistical significance using a double-sided Mann-Whitney U test. Percentage of adjacency occurrence is also reported.

in the RUMC and ML1 datasets was, respectively, 0.92 and 0.85. Contrast enhancement was the largest contributor to contouring accuracy in the test set. The median average Dice$_{surface|3mm}$ in contrast enhanced scans was 0.91, while for non-contrast enhanced scans it was 0.75. The highest median Dice$_{surface|3mm}$ scores were achieved for the HT (0.94), AO (0.99), PA (0.96), and SVC (0.93).

Adjacency of tumors to the heart and vessels also affected contour accuracy. Table 3 shows how Dice$_{surface|3mm}$ was affected by tumor adjacency in contrast CTs in the RUMC dataset for all structures. Contour accuracy was considerably affected by tumor adjacency in the SVC ($p = 0.126$), but the model seemed robust against adjacency to the remaining structures.

Fig. 1 shows the predicted contours and the expert contours for the scan with median average Dice$_{surface|3mm}$ over all structures. Figure SM5.1 shows representative cases where the model did not perform as desired. For further inspection, all model predictions in the public ML1 dataset can be found at https://github.com/FernandesMG/WHS_ResUNET.

Fig. 2A shows the absolute difference between the mean dose computed with the expert contours and the mean dose computed with the model contours for each structure in the RUMC test set. The largest median absolute difference observed for the mean dose was 1.11 Gy for the LA (HT: 0.28 Gy, LV: 0.08 Gy, RV: 0.24 Gy, RA: 0.35 Gy, AO: 0.24 Gy, PA: 0.72 Gy, SVC: 1.00 Gy, IVC: 0.10 Gy). No statistically significant difference was found between the two contour groups in a Wilcoxon signed-rank test for the investigated parameters, including mean dose (lowest $p = 0.843$ for RV), 2nd percentile dose, $D_{2\%}$, (lowest $p = 0.599$ for SVC), or 98th percentile dose, $D_{98\%}$, (lowest $p = 0.715$ for SVC).

Fig. 2B shows the MAE between the DVHs computed with the expert and manual contours for each instance of the test set and ROI in the RUMC test set. The structure with the highest median MAE was the IVC (2,79%), which correlates with what was found for contouring accuracy, while for the other structures median MAE was equal to or less than 2% (HT: 0.48%, LV: 0.56%, RV: 1.64%, LA: 2.02%, RA: 1.43%, AO: 0.48%, PA: 1.09%, SVC: 1.49%).

Fig. 3 shows the clinical acceptability reader study results in the 99 scans of the RTOG-0617 dataset. A contour was regarded as acceptable if at least two of the three radiation oncologists' evaluations where either #CANSE or #CAHA. Clinical acceptability for all structures was as follows: Heart: 97%, LV: 100%, RV: 96%, LA: 98%, RA: 96%, AO: 88%, PA: 87%, SVC: 67% IVC: 90%. Also presented in Fig. 3 is the percentage of each of the four evaluations for the cases elected as clinically acceptable (CA).

Discussion

The geometric accuracy results show that our model is capable of consistently and accurately identifying and contouring the heart, cardiac chambers, and great vessels in different levels of noise, contrast, resolution, and in scans acquired in multiple institutes from stages I–IIIb NSCLC patients. We observed that most discrepancies between the model and the manual contours occurred either in contour cut-off points or when there was adjacency between a tumor and a structure. Differences in contour cut-off points were partly due to the difficulty in identifying the contour cut-off landmarks. Errors caused by the adjacency of tumors were due to low contrast between the tumors and the structures and anatomical perturbation caused by the tumor. The latter was particularly true for the SVC, which is more deformed by the tumor than the PA or the AO due to its thinner walls. Given tumor anatomic variability, this limitation could, in the future, be alleviated by performing active learning with slices where tumor adjacency is present, particularly with regards to the SVC, as the target.

Generalizability to other images used for radiotherapy planning such as 4D CT time averages was not evaluated in this study. These

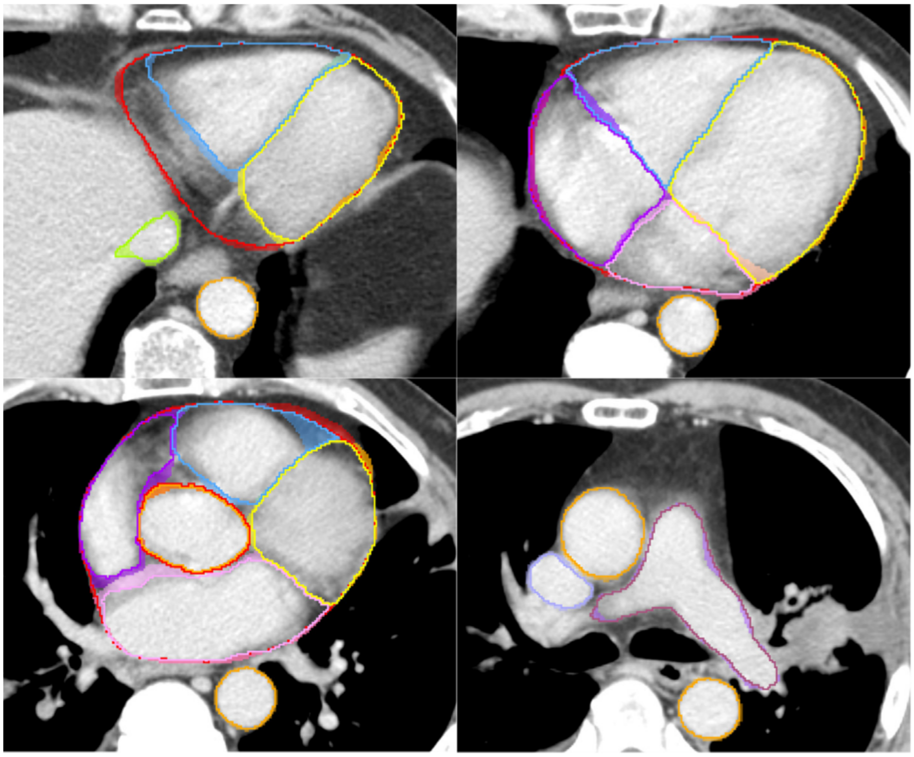

Fig. 1. Axial plane examples of the automatic contours for the patient with median surface Dice at 3 mm tolerance, averaged over all structures. Difference to the expert contours in shade.

images typically differ the most from 3D CT scans in terms of the type and amount of respiratory motion artifacts present in the lungs and tumor. Thus, the model should be able to generalize its cardiac contouring performance to 4D CT averages as it has also been observed in preliminary results of future work.

In this study, a uniform tolerance of 3 mm for $\text{Dice}_{\text{surface}}$ was used for each structure. Ideally, this tolerance would reflect the actual accepted variability over the surface of each structure such that $\text{Dice}_{\text{surface}}$ represented the fraction of the automatic contour within the acceptable limits. Estimating this tolerance in our dataset was not possible because only one manual contour was available per scan. From Lorenzen et al. [30], the average interobserver contouring variability for the heart when using guidelines is larger at the base – up to 2 cm– and in the right-posterior area – up to 1 cm – being mostly bellow 0.5 cm throughout the remaining surface area. From this literature, a uniform 3 mm tolerance seems a reasonable middle-ground between the reported ranges of interobserver contouring variability.

Compared to experts, our model appears to achieve similar or better geometric accuracy scores for the heart and cardiac chambers. For instance, our test set mean overlap results were: HT: +3%, LV: – 1%, RV: +13% when compared to those achieved by seven experts against panel reviewed gold standards in contrast CTs, following the same contouring atlas used in our study [23].

Regarding the presented dosimetric parameters in the RUMC dataset, the significance of the automatic contouring errors depends on the intended application of those dosimetric parameters. Feng et al. [23] also evaluated how differences in expert contours against a gold standard reflected in mean dose differences in four breast cancer patients with a prescribed total dose of 50 Gy: HT: 0.14 ± 0.14 Gy, LV: 0.15 ± 0.14 Gy, RV: 0.46 ± 0.37 Gy. These results are comparable to ours for NSCLC patients (Fig. 2A), which underwent treatments with a prescribed total dose of 66 Gy. Visual inspection of the high error outliers showed that the biggest rela-

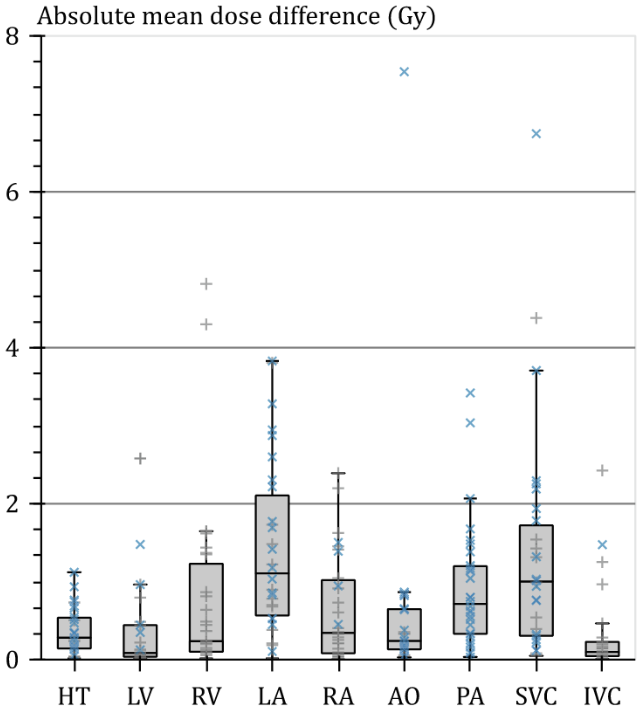

Fig. 2A. Absolute difference between the mean dose computed with the automatic contours and the expert contours for each structure in the RUMC test set. The blue cross sign represents points where there was adjacency between the tumor and the respective ROI, the gray plus sign represents no adjacency. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article).

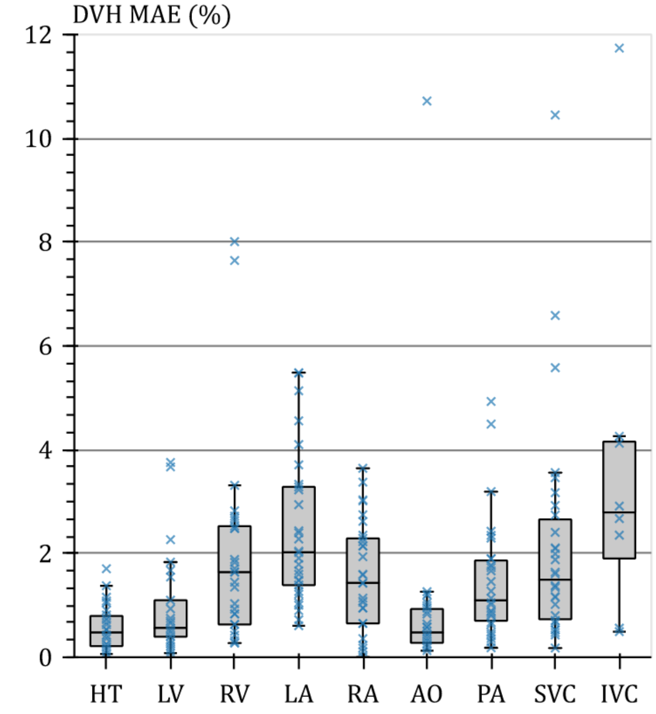

Fig. 2B. Mean Absolute Error (MAE) between the same parameters sampled from the Dose Volume Histograms (DVHs) computed with the automatic contour relative to those from the DVHs computed with the expert contours for each structure in the RUMC test set. These parameters were sampled from both DVHs in intervals of 5 Gy starting from 5 Gy until the maximum dose in both DVHs had been surpassed (i.e., V5, V10, and so on).

tive differences in dosimetric parameters resulted from contouring differences in the high dose-gradient regions of the imaged volume. Because most tumors in our dataset were located near the base of the heart, contouring errors in that region contributed disproportionally more to errors in the computed dosimetric parameters than similar contouring errors in the low gradient regions of the imaged volume, either far from the tumor or adjacent to it.

One of the advantages of using an automatic contouring model is that intra- and inter-observer variability is eliminated. This is a major benefit when compared to choosing a group of experts to contour a dataset because systematic contour differences between experts might be reflected in systematic differences in the dosimetric parameters and all other computations made using those contours [31]. Moreover, these algorithms can be easily deployed in any given institute, avoiding the time and monetary costs of manual contouring following the same atlas and contouring directions. Given this and the achieved dosimetric parameter estimation results, the model presented in this study is ideal for consistently contouring large datasets of planning CTs for cardiac toxicity research and development of robust NTCP models.

Our reader study results showed moderate acceptability rates for the SVC and high acceptability rates for all other structures. This is particularly significant since the RTOG-0617 is a multi-institutional dataset, which represents the typical scan and anatomical variability found in clinic, including a high percentage of non-contrast enhanced images. The improvement observed for non-contrast-enhanced images implies that the fine-tuning step on the test set had a noticeably positive effect on contouring. The lower acceptability rates of the SVC contours were due to cut-off errors in the cranial region of the vessel and the previously mentioned deformation caused by the tumor. Note also that the AO and PA contours, while receiving one of the highest percentages of highly accurate grades as well as highest $\text{Dice}_{\text{surface}|3\text{mm}}$ and dosimetric parameter estimation scores, did not rank as high in the majority vote for acceptability (Fig. 3). This implies that when the model commits contouring errors for the vessels, they are often large relative to those committed for the cardiac chambers and the heart (i.e., failing to identify the structure vs. failing to accurately follow an edge, respectively). Of note is also the potential bias associated with the readers not being blind to the nature of the contours. This is difficult to avoid given that the source of the contours (manual versus automatic) can frequently be derived by their intrinsic characteristics (for example high-frequency artifacts

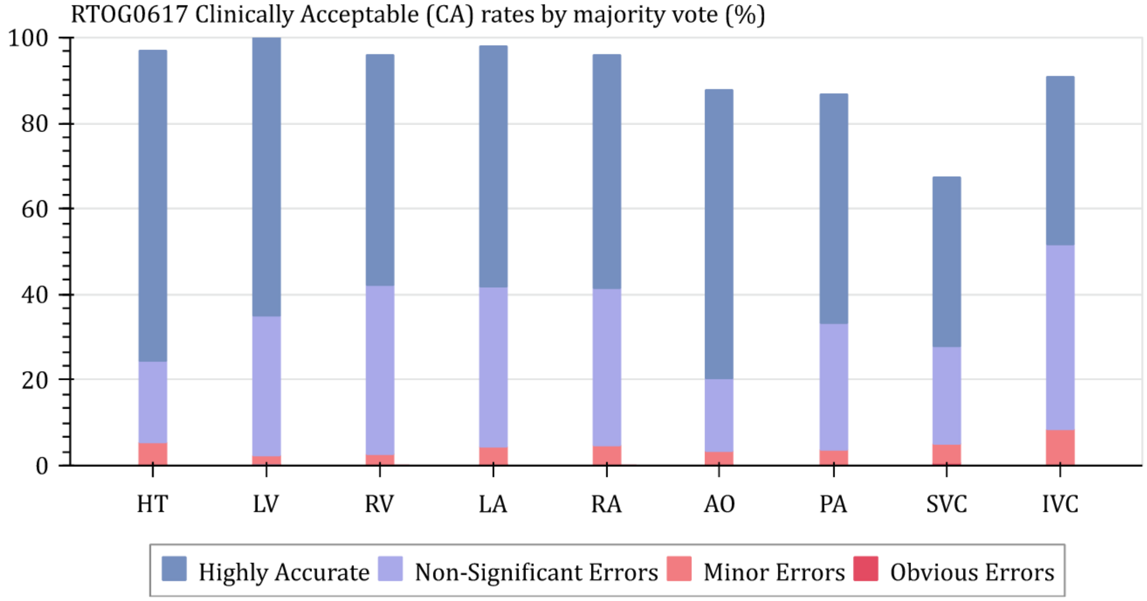

Fig. 3. Contour Clinically Acceptable (CA) rates on the 99 randomly selected scans from the RTOG-0617 dataset, evaluated by three experienced radiation oncologists. For each scan, CA rates were determined by majority vote: highly accurate (#CAHA) and non-significant errors (#CANSE) answers counted with a CA vote. Minor Errors (#RCME) and obvious errors (#RCOE) answers counted with a "Requires Corrections" vote. A contour was elected as CA if at least two of the three votes were CA. Of all contours elected as acceptable, the percentages of each of the four evaluations are stacked.

are rare in manual contours). Given the high clinical acceptability rates achieved in the reader study, our model could also potentially be used as a starting point for contouring in clinical practice as a time-saving tool.

Conclusion

In this study, we developed a DL model for automatic contouring of the heart, cardiac chambers, and great vessels. Active learning was used to take advantage of the public ML1 dataset, which increased the heterogeneity of the training set. Our extensive clinical validation showed that this model performed well in geometric contouring accuracy, dosimetric parameter estimation, and clinical acceptability evaluation in multi-institutional datasets reflective of the day-to-day scans acquired in clinic. This includes non-contrast enhanced images and cases where tumor adjacency to the cardiovascular structures is present. Thus, this model can contribute to the availability of high-quality cardiovascular contours to further cardiac radiotoxicity research in large datasets.

Funding

None.

Data Availability Statement

The model is stored in a private repository and will be shared upon request to the corresponding author. The prediction results in the public ML1 dataset are available in GIF format at: https://github.com/FernandesMG/WHS_ResUNET. The RUMC dataset is not publicly available due to legal and privacy reasons.

Conflict of Interest

Miguel Garrett Fernandes: None.
Johan Bussink: None.
Barbara Stam: None.
Robin Wijsman: None.
Dominic A. X. Schinagl: None.
Jonas Teuwen: None.
René Monshouwer: None.

Acknowledgments

This manuscript was prepared using data from Datasets NCT00533949-D1, NCT00533949-D2, and NCT00533949-D3 from the NCTN/NCORP Data Archive of the National Cancer Institute's (NCI's) National Clinical Trials Network (NCTN). Data were originally collected from clinical trial NCT number NCT00533949 RTOG-0617. All analyses and conclusions in this manuscript are the sole responsibility of the authors and do not necessarily reflect the opinions or views of the clinical trial investigators, the NCTN, the NCORP or the NCI.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.radonc.2021.10.008.

References

[1] Bradley JD, Paulus R, Komaki R, et al. Standard-dose versus high-dose conformal radiotherapy with concurrent and consolidation carboplatin plus paclitaxel with or without cetuximab for patients with stage IIIA or IIIB non-small-cell lung cancer (RTOG 0617): a randomised, two-by-two factorial p. Lancet Oncol 2015;16:187–99. https://doi.org/10.1016/S1470-2045(14)71207-0.

[2] Darby SC, Cutter DJ, Boerma M, et al. Radiation-related heart disease: current knowledge and future prospects. Int J Radiat Oncol Biol Phys 2010;76:656–65. https://doi.org/10.1016/j.ijrobp.2009.09.064.

[3] Zhang TW, Snir J, Boldt RG, et al. Is the Importance of heart dose overstated in the treatment of non-small cell lung cancer? A systematic review of the literature. Int J Radiat Oncol Biol Phys 2019;104:582–9. https://doi.org/10.1016/j.ijrobp.2018.12.044.

[4] McWilliam A, Kennedy J, Hodgson C, Vasquez Osorio E, Faivre-Finn C, van Herk M. Radiation dose to heart base linked with poorer survival in lung cancer patients. Eur J Cancer 2017;85:106–13. https://doi.org/10.1016/j.ejca.2017.07.053.

[5] Stam B, Peulen H, Guckenberger M, et al. Dose to heart substructures is associated with non-cancer death after SBRT in stage I-II NSCLC patients. Radiother Oncol 2017;123:370–5. https://doi.org/10.1016/j.radonc.2017.04.017.

[6] Vivekanandan S, Landau DB, Counsell N, et al. The impact of cardiac radiation dosimetry on survival after radiation therapy for non-small cell lung cancer. Int J Radiat Oncol 2017;99:51–60. https://doi.org/10.1016/j.ijrobp.2017.04.026.

[7] Wong OY, Yau V, Kang J, et al. Survival impact of cardiac dose following lung stereotactic body radiotherapy. Clin Lung Cancer 2018;19:e241–6. https://doi.org/10.1016/j.cllc.2017.08.002.

[8] Ma J-T, Sun L, Sun X, et al. Is pulmonary artery a dose-limiting organ at risk in non-small cell lung cancer patients treated with definitive radiotherapy? Radiat Oncol 2017;12. https://doi.org/10.1186/s13014-017-0772-5.

[9] Ghita M, Gill EK, Walls GM, et al. Cardiac sub-volume targeting demonstrates regional radiosensitivity in the mouse heart. Radiother Oncol 2020;152:216–21. https://doi.org/10.1016/j.radonc.2020.07.016.

[10] Contreras JA, Lin AJ, Weiner A, et al. Cardiac dose is associated with immunosuppression and poor survival in locally advanced non-small cell lung cancer. Radiother Oncol J Eur Soc Ther Radiol Oncol 2018;128:498–504. https://doi.org/10.1016/j.radonc.2018.05.017.

[11] Badiyan SN, Robinson CG, Bradley JD. Radiation toxicity in lung cancer patients: the heart of the problem? Int J Radiat Oncol Biol Phys 2019;104:590–2. https://doi.org/10.1016/j.ijrobp.2019.03.007.

[12] Wong J, Fong A, McVicar N, et al. Comparing deep learning-based auto-segmentation of organs at risk and clinical target volumes to expert inter-observer variability in radiotherapy planning. Radiother Oncol 2020;144:152–8. https://doi.org/10.1016/j.radonc.2019.10.019.

[13] Liu X, Li K-W, Yang R, Geng L-S. Review of Deep Learning Based Automatic Segmentation for Lung Cancer Radiotherapy. Front Oncol 2021;11:2599. https://doi.org/10.3389/fonc.2021.717039.

[14] Payer C, Stern D, Bischof H, Urschler M. Multi-label whole heart segmentation using CNNs and anatomical label configurations. STACOM@MICCAI 2017.

[15] Zhuang X, Li L, Payer C, et al. Evaluation of algorithms for Multi-Modality Whole Heart Segmentation: An open-access grand challenge. Med Image Anal 2019;58:101537. https://doi.org/10.1016/j.media.2019.101537.

[16] Morris ED, Ghanem AI, Dong M, Pantelic MV, Walker EM, Glide-Hurst CK. Cardiac substructure segmentation with deep learning for improved cardiac sparing. Med Phys 2020;47:576–86. https://doi.org/10.1002/mp.13940.

[17] Harms J, Lei Y, Tian S, et al. Automatic delineation of cardiac substructures using a region-based fully convolutional network. Med Phys 2021;48:2867–76. https://doi.org/10.1002/mp.14810.

[18] Haq R, Hotca A, Apte A, Rimner A, Deasy JO, Thor M. Cardio-pulmonary substructure segmentation of radiotherapy computed tomography images using convolutional neural networks for clinical outcomes analysis. Phys Imaging Radiat Oncol 2020;14:61–6. https://doi.org/10.1016/j.phro.2020.05.009.

[19] Clark K, Vendt B, Smith K, Freymann J, Kirby J, Koppel P, et al. The Cancer Imaging Archive (TCIA): maintaining and operating a public information repository. J Digit Imaging 2013;26:1045–57. https://doi.org/10.1007/s10278-013-9622-7.

[20] Bradley JD, Forster K. Data from NSCLC - Centuximab. The Cancer Imaging Archive 2018. https://doi.org/10.7937/TCIA.2018.jze75u7v.

[21] Wijsman R, Dankers F, Troost EGC, et al. Multivariable normal-tissue complication modeling of acute esophageal toxicity in advanced stage non-small cell lung cancer patients treated with intensity-modulated (chemo-) radiotherapy. Radiother Oncol 2015;117:49–54. https://doi.org/10.1016/j.radonc.2015.08.010.

[22] Aerts HJWL, Wee L, Velazquez ER, et al. Data From NSCLC-Radiomics. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2015.PF0M9REI.

[23] Feng M, Moran JM, Koelling T, et al. Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer. Int J Radiat Oncol Biol Phys 2011;79:10–8. https://doi.org/10.1016/j.ijrobp.2009.10.058.

[24] Nikolov S, Blackwell S, Zverovitch A, et al. Clinically applicable segmentation of head and neck anatomy for radiotherapy: deep learning algorithm development and validation study. J Med Internet Res 2021;23:e26151. https://doi.org/10.2196/26151.

[25] Paszke A, Gross S, Massa F, et al. PyTorch: An imperative style, high-performance deep learning library. In: Wallach H, Larochelle H, Beygelzimer A, d\textquotesingle Alché-Buc F, Fox E, Garnett R, eds. *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc.; 2019:8024-8035. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[26] Ronneberger O, Fischer P, Brox T. U-Net: convolutional networks for biomedical image segmentation. In: Navab N, Hornegger J, Wells WM, Frangi AF, editors. Medical image computing and computer-assisted intervention – MICCAI 2015. Cham: Springer International Publishing; 2015. p. 234–41.

[27] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: learning dense volumetric segmentation from sparse annotation. In: Ourselin S, Joskowicz L, Sabuncu MR, Unal G, Wells W, editors. Medical image computing and computer-assisted intervention – MICCAI 2016. Cham: Springer International Publishing; 2016. p. 424–32.

[28] Liu L, Jiang H, He P, et al. On the variance of the adaptive learning rate and beyond. *CoRR*. 2019;abs/1908.0. http://arxiv.org/abs/1908.03265.

[29] Meakin J, van Zeeland H, Koek M, et al. Grand-Challenge.org. doi:10.5281/zenodo.3356819

[30] Lorenzen EL, Taylor CW, Maraldo M, et al. Inter-observer variation in delineation of the heart and left anterior descending coronary artery in radiotherapy for breast cancer: A multi-centre study from Denmark and the UK. Radiother Oncol 2013;108:254–8. https://doi.org/10.1016/j.radonc.2013.06.025.

[31] Thor M, Apte A, Haq R, Iyer A, LoCastro E, Deasy JO. Using auto-segmentation to reduce contouring and dose inconsistency in clinical trials: the simulated impact on RTOG 0617. Int J Radiat Oncol Biol Phys 2021;109:1619–26. https://doi.org/10.1016/j.ijrobp.2020.11.011.