Contents lists available at ScienceDirect

Technical Innovations & Patient Support in Radiation Oncology

journal homepage: www.sciencedirect.com/journal/technical-innovations-and-patient-support-in-radiation-oncology

Comparison of the output of a deep learning segmentation model for locoregional breast cancer radiotherapy trained on 2 different datasets

Nienke Bakx [a], Maurice van der Sangen [a], Jacqueline Theuws [a], Hanneke Bluemink [a], Coen Hurkmans [a,b,*]

[a] *Catharina Hospital, Department of Radiation Oncology, 5602ZA Eindhoven, the Netherlands*
[b] *Technical University Eindhoven, Faculties of Physics and Electrical Engineering, 5600MB Eindhoven, the Netherlands*

ARTICLE INFO

Keywords:
Auto-segmentation
Loco-regional breast cancer
Deep learning
Radiotherapy
Clinical validation

ABSTRACT

Introduction: The development of deep learning (DL) models for auto-segmentation is increasing and more models become commercially available. Mostly, commercial models are trained on external data. To study the effect of using a model trained on external data, compared to the same model trained on in-house collected data, the performance of these two DL models was evaluated.

Methods: The evaluation was performed using in-house collected data of 30 breast cancer patients. Quantitative analysis was performed using Dice similarity coefficient (DSC), surface DSC (sDSC) and 95th percentile of Hausdorff Distance (95% HD). These values were compared with previously reported inter-observer variations (IOV).

Results: For a number of structures, statistically significant differences were found between the two models. For organs at risk, mean values for DSC ranged from 0.63 to 0.98 and 0.71 to 0.96 for the in-house and external model, respectively. For target volumes, mean DSC values of 0.57 to 0.94 and 0.33 to 0.92 were found. The difference of 95% HD values ranged 0.08 to 3.23 mm between the two models, except for CTVn4 with 9.95 mm. For the external model, both DSC and 95% HD are outside the range of IOV for CTVn4, whereas this is the case for the DSC found for the thyroid of the in-house model.

Conclusions: Statistically significant differences were found between both models, which were mostly within published inter-observer variations, showing clinical usefulness of both models. Our findings could encourage discussion and revision of existing guidelines, to further decrease inter-observer, but also inter-institute variability.

Introduction

In radiotherapy treatment planning, segmentation of target volumes and organs at risk (OARs) is a time-consuming process. Moreover it is prone to intra- and inter-observer variations [1]. In recent years, research on the use of deep learning (DL) models to automate this process has increased [2–6]. The primary outcome of these models is to save time, while also decreasing inter-observer variability. Most of the performed studies include only a quantitative analysis, while some others also include a qualitative analysis, such as scoring on clinical usefulness by the intended user, which is relevant with regard to the possibility of clinical use [7,8]. Also, more DL models become commercially available [2,6]. Most of these commercially available models will be trained with external data, not originating from the institute where it will be applied. However, there is limited information on the use of a DL model trained on external data. In this study, an externally trained model was compared to an in-house trained model, to examine the possible difference in performance and reflect the effect of using a commercially available model, based on the same delineation guidelines. A model is in this study defined as a DL architecture trained on a specific dataset. The models evaluated in this study consist of the same DL model architecture, but were trained on different datasets from two institutions, while evaluated on the same in-house collected dataset. Furthermore, eventual differences found in performance of the two models were further examined.

* Corresponding author at: Catharina Hospital, Department of Radiation Oncology, 5602ZA Eindhoven, the Netherlands.
E-mail address: coen.hurkmans@catharinaziekenhuis.nl (C. Hurkmans).

https://doi.org/10.1016/j.tipsro.2023.100209
Received 14 February 2023; Received in revised form 6 April 2023; Accepted 9 May 2023
Available online 13 May 2023

Materials & Methods

Patient data

The patient dataset for evaluation of the models consists of 15 left- and 15 right-sided, randomly selected breast cancer patients. All patients were treated for locally advanced breast cancer between January 2019 and February 2022. For clinical treatment, contouring of target volumes was performed by radiation oncologists (ROs) and of organs at risk (OARs) by radiotherapy technologists (RTTs) with varying experience, following ESTRO guidelines [9,10]. All clinically delineated contours were checked and adjusted if necessary by an experienced radiation oncologist (RO) before inclusion for evaluation. The RO was also involved in these tasks concerning the training data of the in-house model. In total, 11 regions of interest (ROIs) were included in the evaluation for both the left and right side, involving target volumes (breast (CTVp), axillary lymph node levels 1–3 (CTVn1, CTVn2 and CTVn3) and supraclavicular lymph nodes (CTVn4)) and organs at risk (OARs) (heart, left/right lung, esophagus, thyroid and humeral head).

DL models

In this study, the models were trained using a framework provided by RaySearch Laboratories AB in RayStation version 9B, while the evaluation was performed in version 10B-SP1 (RaySearch Laboratories AB, Sweden). This framework consists of multiple sub-models, each based on the architecture of an adapted version of U-net [11]. Model training was performed both in-house, on an in-house collected dataset, as well as externally, using data of St. Olavs Hospital and Ålesund Hospital in Norway for target structures and internally collected data of RaySearch Laboratories for OARs. All data collected and included for training followed ESTRO guidelines for targets [9,10] while OARs delineations followed atlases of Feng *et al.* and Kong *et al.* [12,13]. While the external model consists of one common model for both left- and right sided breast-cancer, two separate models were trained with the in-house data, resulting in 3 models in total. The in-house models were trained using 160 patients in total (80 for each side). However, not all patients contained delineations of all regions of interest (ROIs), with for instance 82 thyroid and 147 lung delineations. The external model was trained on a dataset containing 170 left-sided patients, with only a few cases that did not contain all ROIs. More details on training and the used datasets can be found in [14] and [15].

Evaluation

In this study a quantitative analysis was performed to compare the outcomes of the in-house and external models with the manual delineations. For a qualitative analysis of the models, we refer to previous work [14] and [15]. In our former qualitative study, it was found that the in-house models trained for left- and right-sided breast cancer patients perform equally well [15]. Therefore, the results of both models were analyzed as one cohort in this study. The Dice similarity coefficient (DSC), surface DSC (sDSC) (tolerance $\tau = 3$ mm) [16] and 95th percentile of the Hausdorff Distance (95% HD) were measured for the overlap between the automatically and manually generated contours for both models. In addition, the overlap between the automatically generated contours of the two models was measured. The Wilcoxon signed rank test was used to test for differences between both models when compared to manual delineations, with a p-value ≤ 0.05 considered significant. Besides, visual inspection was performed to check for abnormalities and investigate the origin of found differences. In addition, the values for DSC and 95% HD were compared with inter-observer variations (IOV) which are present for manual delineations of these structures, taken from various studies performed after the ESTRO guidelines were published [14,17–19].

Results

The resulting DSC scores and values of the 95% HD can be found in Table 1. For the OARs, the mean DSC values were within a range of 0.63 to 0.98 and 0.71 to 0.96 for the in-house and external model, respectively. For target volumes, these values ranged from 0.57 to 0.94 and 0.33 to 0.92. Differences of mean 95% HD values of both models, when compared with manual delineations, were within a range of 0.08 to 3.23 mm for all structures, except for the CTVn4 with a difference of 9.95 mm. Benchmarked against the IOV values, for the external model both metrics are outside the IOV for CTVn4. For the thyroid, this is the case for the DSC value found for the in-house model, and the 95% HD values of both models. The sDSC scores can be found in Table 2, and show mean values in a range of 0.70 to 0.98 and 0.65 to 0.99 for the in-house and external model, respectively. One exception is the found mean value of 0.43 for the CTVn4 for the external model. For all metrics, values were within the same range when considering the overlap between the two models as observed between the automatic and manual delineations.

Fig. 1 shows transversal slices of an example patient, visualizing delineations generated manually and by the two DL models. For a number of ROIs, a significant difference is found between the two models for all three metrics. However, not all significant differences may be clinically relevant. For example, both models show high quantitative scores on all three metrics for CTVp, heart and both lungs, in the same order of magnitude as the IOV values. Visual inspection of the delineations of heart and lungs reveals the small differences in contour mainly appear on the outermost cranial and caudal slices in a low dose region, implying a low clinical impact in most cases. Furthermore, since it only involves one or a few slices, necessary corrections are easily made. Also, significant differences were found for the CTVn3, with a relative large difference for the sDSC scores, compared to the aforementioned ROIs. However, the DSC scores and 95%HD values are within the range of the IOV values and further visual inspection did not reveal any systematic difference. In contrast, the differences for the CTVn4, esophagus and thyroid stand out more and require further investigation. First, the CTVn4 was visually inspected and it was noted that the external model always segments a larger volume than the in-house trained model (12 ± 3.2 cm^3 vs 2.4 ± 1.0 cm^3, respectively), as is shown in Fig. 2. When inspecting the esophagus, it appeared that there was always a significant difference in length of the delineated structure, which explains the high 95%HD value (Fig. 3). When only considering the overlapping part, a median DSC score of 0.78 (range 0.55 – 0.89) and 0.86 (range 0.71 – 0.92) was found for the in-house and externally developed model, respectively. The median 95%HD decreased for both models to respectively 3.24 mm (range 1.41 – 12.6) and 1.87 mm (range 1.00 – 7.21 mm). Lastly, for the thyroid it was observed that for 24 out of 30 patients the two lobes were not connected in the contours for the in-house model, in contrast to 10 patients for the external model, whereas the manual delineations always contain this connection (Fig. 4).

Discussion

Two DL models for automatic segmentation of target and OARs volumes for both right- and left-sided loco-regional breast cancer were evaluated on an in-house collected dataset. The first model was also trained on in-house collected data, while the other one was trained externally, using the same delineation guidelines. When comparing the automatically generated with manual delineations, statistically significant differences were found for most structures when comparing different metrics. However, for all structures except CTVn4 and the thyroid these values were within the found IOV.

The external model was trained and evaluated for target volumes in the study of Almberg *et al.* [14]. For some ROIs, as CTVn2-4, better quantitative results were presented in their report, when compared to the quantitative analysis of the in-house developed model [15]. However, since the in-house developed model was benchmarked against the

Table 1
Metrics (mean ± std) and corresponding inter-observer variation (IOV) for all ROIs for the in-house and external model, compared to the manual delineations and each other. Significant difference of metrics between the two models, when compared to manual delineations, is indicated with an asterisk. Inter-observer variations are acquired from 1: Almberg et al. [14], 2: Chung et al. [17], 3: Francolini et al. [18], 4: Leonardi et al. [19].

		DSC score [-]					95%HD [mm]				
		In-house vs manual	**External vs manual**		**In-house vs external**	**IOV (mean)**	**In-house vs manual**	**External vs manual**		**In-house vs external**	**IOV (mean)**
CTVp	*mean ± std*	0.94 ± 0.02	0.92 ± 0.02	*	0.93 ± 0.02	0.94^1, 0.85^2, 0.94^3	9.38 ± 9.04	8.80 ± 4.64		8.01 ± 8.94	5.71^1, 8.94^2
	median	0.95	0.92		0.93		7.81	7.77		6.47	
	(range)	(0.89, 0.97)	(0.87, 0.94)		(0.86, 0.95)		(2.45, 55.2)	(2.40, 28.2)		(3.16, 54.8)	
CTVn1	*mean ± std*	0.79 ± 0.05	0.76 ± 0.08	*	0.79 ± 0.07	0.74^1, 0.69^2, 0.72^3	11.3 ± 4.65	13.3 ± 7.26		10.1 ± 3.61	14.60^1, 13.58^2
	median	0.79	0.78		0.80		10.1	11.4		9.33	
	(range)	(0.64, 0.88)	(0.52, 0.87)		(0.52, 0.84)		(4.47, 21.9)	(5.00, 33.3)		(5.48, 20.6)	
CTVn2	*mean ± std*	0.71 ± 0.07	0.69 ± 0.07		0.67 ± 0.09	0.62^1, 0.47^2, 0.77^3, 0.55^4	9.42 ± 4.33	10.1 ± 3.61		11.3 ± 4.13	16.21^1, 18.74^2
	median	0.74	0.69		0.68		8.48	9.95		10.1	
	(range)	(0.53, 0.82)	(0.53, 0.83)		(0.46, 0.86)		(2.24, 22.2)	(4.58, 21.3)		(6.08, 23.1)	
CTVn3	*mean ± std*	0.74 ± 0.06	0.67 ± 0.10		0.72 ± 0.09	0.67^1, 0.56^2, 0.79^3, 0.58^4	7.01 ± 2.72	8.71 ± 3.07	*	7.67 ± 3.33	9.39^1, 9.87^2
	median	0.74	0.69		0.74		6.44	8.25		6.74	
	(range)	(0.58, 0.82)	(0.4, 0.82)		(0.47, 0.86)		(3.61, 14.9)	(3.00, 16.9)		(3.74, 18.8)	
CTVn4	*mean ± std*	0.57 ± 0.13	0.33 ± 0.11	*	0.30 ± 0.08	0.72^1, 0.45^2, 0.75^3, 0.69^4	6.45 ± 3.08	16.4 ± 3.88	*	16.5 ± 2.57	6.13^1, 11.82^2
	median	0.57	0.33		0.29		5.46	15.7		16.6	
	(range)	(0.22, 0.79)	(0.01, 0.54)		(0.18, 0.35)		(2.83, 18.0)	(9.43, 23.2)		(10.5, 21.8)	
Heart	*mean ± std*	0.94 ± 0.02	0.93 ± 0.02	*	0.93 ± 0.02	0.95^1, 0.91^2	7.46 ± 3.79	9.48 ± 4.67	*	11.1 ± 3.74	6.69^1, 13.00^2
	median	0.94	0.93		0.94		7.00	8.00		10.0	
	(range)	(0.90, 0.96)	(0.88, 0.96)		(0.88, 0.96)		(3.00, 23.0)	(3.00, 19.1)		(6.00, 19.0)	
Lung (left)	*mean ± std*	0.98 ± 0.01	0.96 ± 0.02	*	0.96 ± 0.02	0.98^2	2.56 ± 2.37	4.63 ± 3.55	*	4.65 ± 3.21	2.19^2
	median	0.98	0.97		0.97		1.87	2.91		3.61	
	(range)	(0.95, 0.99)	(0.91, 0.99)		(0.90, 0.98)		(1.00, 13.4)	(1.00, 10.8)		(1.00, 12.7)	
Lung (right)	*mean ± std*	0.98 ± 0.01	0.96 ± 0.03	*	0.97 ± 0.03	0.99^2	2.14 ± 1.29	5.37 ± 5.54	*	5.32 ± 5.04	2.33^2
	median	0.99	0.98		0.98		2.00	1.83		5.05	
	(range)	(0.96, 0.99)	(0.88, 0.99)		(0.90, 0.99)		(1.00, 6.16)	(1.00, 20.8)		(2.83, 9.27)	
Thyroid	*mean ± std*	0.63 ± 0.17	0.71 ± 0.17	*	0.68 ± 0.09	0.81^1, 0.72^2	8.23 ± 7.19	7.05 ± 7.31	*	5.36 ± 1.91	3.91^1, 5.37^2
	median	0.67	0.75		0.69		5.66	4.18		5.05	
	(range)	(0.00, 0.82)	(0.00, 0.87)		(0.47, 0.81)		(2.83, 41.4)	(2.00, 39.1)		(2.83, 9.27)	
Esophagus	*mean ± std*	0.70 ± 0.10	0.32 ± 0.07	*	0.27 ± 0.04	0.83^1, 0.78^2	9.58 ± 6.98	161 ± 22.4	*	161 ± 18.0	2.96^1, 7.08^2
	median	0.71	0.32		0.27		7.77	160		160	
	(range)	(0.45, 0.88)	(0.18, 0.45)		(0.18, 0.35)		(1.73, 36.8)	(116, 226)		(133, 201)	
Esophagus (overlap)	*mean ± std*	0.77 ± 0.08	0.85 ± 0.05	*	0.77 ± 0.07	0.83^1, 0.78^2	3.81 ± 2.38	2.29 ± 1.44	*	4.44 ± 2.22	2.96^1, 7.08^2
	median	0.78	0.86		0.78		3.24	1.87		3.80	
	(range)	(0.55, 0.89)	(0.71, 0.92)		(0.53, 0.87)		(1.41, 12.57)	(1.00, 7.21)		(1.41, 11.4)	
Humeral head	*mean ± std*	0.85 ± 0.06	0.85 ± 0.09		0.83 ± 0.13	–	8.18 ± 3.76	8.26 ± 4.81		9.88 ± 3.68	–
	median	0.86	0.86		0.83		7.45	7.50		10.0	
	(range)	(0.68, 0.95)	(0.52, 0.96)		(0.70, 0.94)		(1.00, 19.7)	(1.00, 21.0)		(3.00, 18.57)	

Table 2
sDSC values (mean ± std) for the in-house and external DL model for all ROIs, compared to manual delineations and each other. Significant difference of metrics between the two models, when compared to manual delineations, is indicated with an asterisk.

		In-house vs manual	External vs manual		In-house vs external
CTVp	*mean ± std*	0.87 ± 0.05	0.86 ± 0.06	*	0.89 ± 0.06
	median (range)	0.88 (0.70, 0.98)	0.87 (0.70, 0.97)		0.90 (0.63, 0.96)
CTVn1	*mean ± std*	0.70 ± 0.10	0.65 ± 0.12	*	0.68 ± 0.11
	median (range)	0.69 (0.40, 0.89)	0.63 (0.40, 0.84)		0.68 (0.40, 0.83)
CTVn2	*mean ± std*	0.82 ± 0.08	0.80 ± 0.07		0.76 ± 0.09
	median (range)	0.84 (0.60, 0.98)	0.81 (0.60, 0.89)		0.77 (0.57, 0.91)
CTVn3	*mean ± std*	0.84 ± 0.08	0.75 ± 0.10	*	0.80 ± 0.10
	median (range)	0.86 (0.70, 0.96)	0.77 (0.50, 0.96)		0.82 (0.57, 0.95)
CTVn4	*mean ± std*	0.79 ± 0.14	0.43 ± 0.11	*	0.36 ± 0.10
	median (range)	0.82 (0.40, 0.98)	0.44 (0.10, 0.69)		0.36 (0.18, 0.57)
Heart	*mean ± std*	0.85 ± 0.07	0.82 ± 0.07	*	0.94 ± 0.06
	median (range)	0.86 (0.70, 0.96)	0.82 (0.60, 0.97)		0.84 (0.65, 0.92)
Lung (left)	*mean ± std*	0.98 ± 0.02	0.93 ± 0.06	*	0.93 ± 0.05
	median (range)	0.99 (0.90, 1.00)	0.96 (0.80, 0.99)		0.95 (0.82, 0.99)
Lung (right)	*mean ± std*	0.98 ± 0.02	0.92 ± 0.09	*	0.93 ± 0.08
	median (range)	0.98 (0.90, 1.00)	0.98 (0.70, 0.99)		0.99 (0.77, 1.00)
Thyroid	*mean ± std*	0.80 ± 0.19	0.87 ± 0.18	*	0.88 ± 0.09
	median (range)	0.88 (0.10, 0.97)	0.95 (0.10, 0.99)		0.90 (0.67, 0.99)
Esophagus	*mean ± std*	0.87 ± 0.09	0.42 ± 0.07	*	0.38 ± 0.04
	median (range)	0.91 (0.70, 0.99)	0.42 (0.30, 0.55)		0.39 (0.29, 0.45)
Esophagus (overlap)	*mean ± std*	0.95 ± 0.06	0.99 ± 0.02	*	0.94 ± 0.05
	median (range)	0.97 (0.74, 1.00)	1.00 (0.92, 1.00)		0.96 (0.73, 1.00)
Humeral head	*mean ± std*	0.71 ± 0.09	0.82 ± 0.11		0.78 ± 0.09
	median (range)	0.79 (0.70, 1.00)	0.80 (0.60, 1.00)		0.75 (0.64, 0.98)

IOV values found by Chung *et al.* [17] for validation, that differ from the IOV values measured by Almberg *et al.* for their benchmark, quantitative results of both models were found sufficient in their studies. Besides, both studies already previously showed good qualitative results [14,15]. Lastly, both models resulted in a reduction of delineation time, which is the main outcome. Therefore both models were found to be clinically applicable in the institutes where they were trained and validated.

Various quantitative metrics were used in this study to assess and compare the performance of both models. Although statistically significant differences were found for most of the ROIs, the results of both models were within the range of reported IOV values, suggesting the models perform as well as an observer. The third quantitative parameter, the sDSC score, was introduced by Nikolov *et al.* [16]. It was introduced to better reflect the corrections needed by quantifying the deviation in contours rather than volumes, as is the case for the regular DSC score. Besides, a stronger correlation with time needed for correction, and relative time saved, was found for this metric, compared with the other quantitative metrics [20,21]. Since the sDSC was higher for the in-house developed model, except for the thyroid and humeral head, it could be stated that implementing the in-house developed model could reduce more delineation time than the external model. More research on actual time saving should be performed to validate this hypothesis.

By evaluating the two DL models on one dataset, it ought to reflect the reality in which an institute will use an externally trained model, which is for example the case for commercially available models. The purpose of this study is to stress the importance of thorough validation of the outcomes of the model in your own clinic. For the CTVn4, it was found that the external model delineated a larger area in cranial direction, which is remarkable since both models follow the same delineation guidelines. Apparently, there is a difference in interpretation and use of these guidelines in practice, which is an important finding. According to the ESTRO guideline, CTVn4 includes the cranial extent of the subclavian artery (i.e. 5 mm cranial of subclavian vein).

For the esophagus, it became clear that the found differences are due to the difference in the length of the region that is delineated, which stresses the importance of clear delineation guidelines and additional visual inspection, next to quantitative analysis. Lastly, for the thyroid, the difference in absence or presence of the connection between the lobes could be explained by a lack of consistency in the presence of this connection in the training dataset for the in-house model. This result was already noticed in the previous study of the in-house developed model. However, in most cases the delineation was still considered useful as a starting point for correction [15].

The study design has some limitations. Although general observations could be made based on this study, the results are intrinsically biased as we intentionally used our own manual delineations to compare with. Furthermore, no IOV values were obtained in our own institute, and the used IOV values varied for some structures, such as the lymph node levels.

Comparing the overlap between the structures resulting from both models show comparable results as when comparing to manual delineations, within the range of IOV for most structures, suggesting that the differences between the models are similar to differences between multiple observers. Furthermore, it indicates that the data used for training of both models was within this variability.

Performing this study led to practical adjustments. For example, the in-house delineation guidelines for the cranial and caudal ends of the esophagus were revised and a new consensus was reached. Furthermore, the difference in interpretation of the CTVn4 volume is seen as an important finding for both institutes, as well as for the vendor developing commercially DL segmentation models. To further investigate the nature of this difference, a multi-center study within the Netherlands will be performed, in which both the in-house as external model performance will be studied. These results could lead to a further refinement of the clinical guidelines, which will not only decrease the inter-observer but also the inter-institute variations. Therefore, studies of this nature are not only valuable to evaluate a model for clinical use in one institute, but the results can encourage discussion about and revision of existing guidelines and stimulate updates of the DL models thereafter.

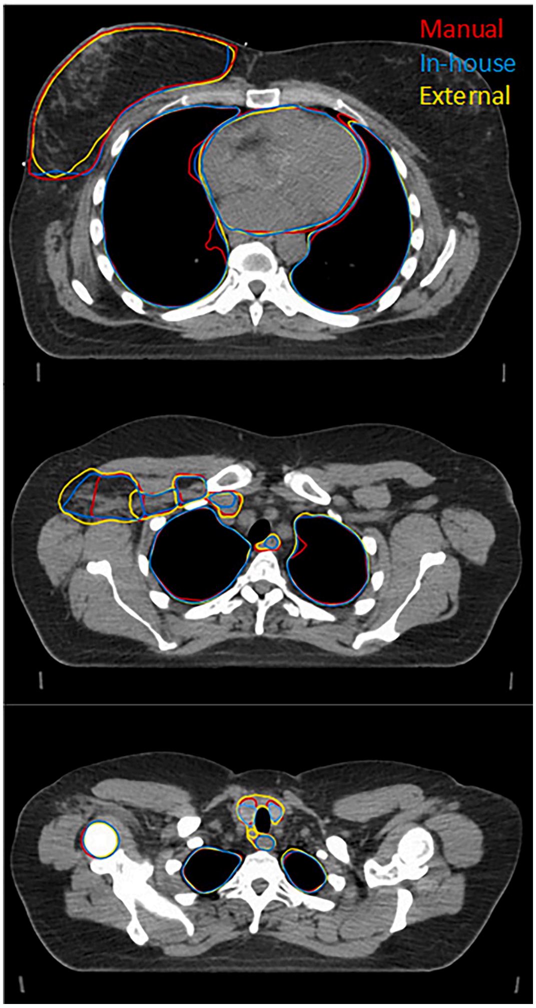

Fig. 1. Three transversal slices of an example patient of the test set, showing different structures delineated manually (red) and by the in-house (blue) and external (yellow) DL model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

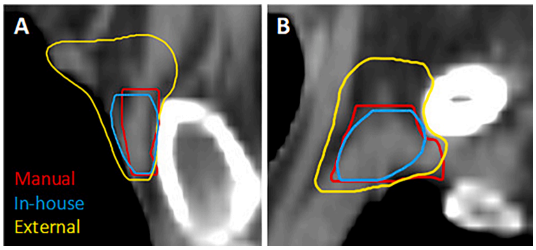

Fig. 2. Sagittal (A) and coronal (B) view of CTVn4 delineated manually (red) and by the in-house (blue) and external (yellow) DL model of a patient with left-sided breast cancer, showing a larger volume delineated by the external model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

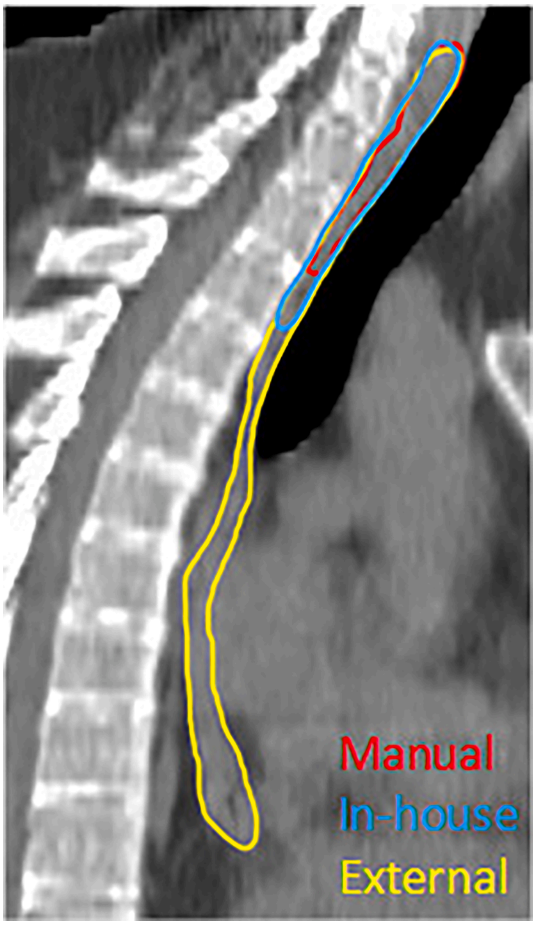

Fig. 3. Sagittal view of the esophagus delineated manually (red) and by the in-house (blue) and external (yellow) DL model, showing a difference in length for the external and in-house model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

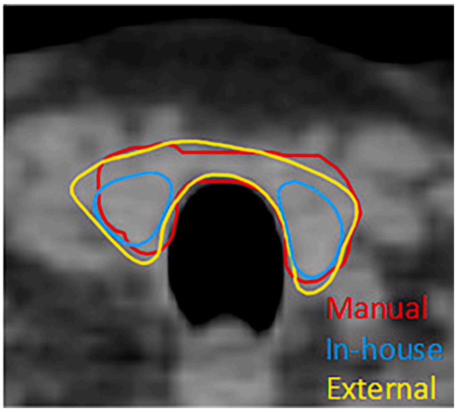

Fig. 4. Transversal view of the thyroid delineated manually (red) and by the in-house (blue) and external (yellow) DL model, showing the absence of a connection between the two lobes for the in-house model. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Nienke Bakx received funding from RaySearch Laboratories AB. RaySearch Laboratories AB had no influence on the design of the study but aided with the practical implementation and by offering advice on practical issues.

Acknowledgements

We would like to thank Fredrik Löfman, Elin Samuelsson and Jonas Söderberg from the machine learning department of RaySearch Laboratories AB for their contribution by providing a training framework and integration of the in-house model, access to the external model and many fruitful discussions. A research grant from Raysearch Laboratories AB is also acknowledged.

References

[1] Hurkmans CW, et al. Variability in target volume delineation on CT scans of the breast. Int J Radiat Oncol Biol Phys 2001;50:1366–72. https://doi.org/10.1016/S0360-3016(01)01635-2.

[2] Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med 2018;98:126–46. https://doi.org/10.1016/j.compbiomed.2018.05.018.

[3] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. Semin Radiat Oncol 2019;29:185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.

[4] Poortmans PMP, Takanen S, Marta GN, Meattini I, Kaidar-Person O. Winter is over: The use of Artificial Intelligence to individualise radiation therapy for breast cancer. Breast 2020;49:194–200. https://doi.org/10.1016/j.breast.2019.11.011.

[5] Samarasinghe G, et al. Deep learning for segmentation in radiation therapy planning: a review. J Med Imaging Radiat Oncol 2021;65:578–95. https://doi.org/10.1111/1754-9485.13286.

[6] Harrison K, et al. Machine Learning for Auto-Segmentation in Radiotherapy Planning. Clin Oncol 2022;34:74–88. https://doi.org/10.1016/j.clon.2021.12.003.

[7] Vandewinckele L, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020;153:55–66. https://doi.org/10.1016/j.radonc.2020.09.008.

[8] Mcintosh C, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. Nat Med 2021;27:999–1005. https://doi.org/10.1038/s41591-021-01359-w.

[9] Offersen BV, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. Radiother Oncol 2015;114:3–10. https://doi.org/10.1016/j.radonc.2014.11.030.

[10] Offersen BV, Boersma LJ, Kirkove C, Hol S, Aznar MC, Sola AB, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. Radiother Oncol 2016;118:205–8. https://doi.org/10.1016/j.radonc.2015.12.027.

[11] Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-net: Learning dense volumetric segmentation from sparse annotation. Conf Med Image Comput Comput-Assist Intervent 2016:424–32. https://doi.org/10.1007/978-3-319-46723-8_49.

[12] Kong F, Ritter T, Quint D. 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 2010;1:1–16. https://doi.org/10.1016/j.ijrobp.2010.07.1977.

[13] Feng M, 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.

[14] Almberg SS, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol 2022;173:62–8. https://doi.org/10.1016/j.radonc.2022.05.018.

[15] Bakx N, et al. Clinical evaluation of a deep learning segmentation model including manual adjustments afterwards for locally advanced breast cancer. Manuscript submitted for publication. 2023.

[16] Nikolov S, 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.

[17] Chung SY, et al. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. Radiat Oncol 2021;16:1–10. https://doi.org/10.1186/s13014-021-01771-z.

[18] Francolini G, et al. Quality assessment of delineation and dose planning of early breast cancer patients included in the randomized Skagen Trial 1. Radiother Oncol 2017;123:282–7. https://doi.org/10.1016/j.radonc.2017.03.011.

[19] Leonardi MC, et al. Geometric contour variation in clinical target volume of axillary lymph nodes in breast cancer radiotherapy: an AIRO multi-institutional study. Br J Radiol 2021;94:1–9. https://doi.org/10.1259/bjr.20201177.

[20] Vaassen F, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol 2020;13:1–6. https://doi.org/10.1016/j.phro.2019.12.001.

[21] Kiser KJ, Barman A, Stieb S, Fuller CD, Giancardo L. Novel Autosegmentation Spatial Similarity Metrics Capture the Time Required to Correct Segmentations Better Than Traditional Metrics in a Thoracic Cavity Segmentation Workflow. J Digit Imaging 2021;34:541–53. https://doi.org/10.1007/s10278-021-00460-3.