ACTA ONCOLOGICA
2023, VOL. 62, NO. 10, 1201–1207
https://doi.org/10.1080/0284186X.2023.2252582

Taylor & Francis
Taylor & Francis Group

ORIGINAL ARTICLE

Development of a national deep learning-based auto-segmentation model for the heart on clinical delineations from the DBCG RT nation cohort

Emma Riis Skarsø[a,b], Lasse Refsgaard[b,c], Abhilasha Saini[d], Ditte Sloth Møller[b,e], Ebbe Laugaard Lorenzen[f], Else Maae[g], Karen Andersen[h], Maja Vestmø Maraldo[i], Marie Louise Milo[j], Tine Bisballe Nyeng[e], Birgitte Vrou Offersen[a,b,c,e] and Stine Sofia Korreman[a,b,e]

[a]Danish Center for Particle Therapy, Aarhus University Hospital, Aarhus, Denmark; [b]Department of Clinical medicine, Aarhus University, Aarhus, Denmark; [c]Department of Experimental Clinical Oncology, Aarhus University Hospital, Aarhus, Denmark; [d]Department of Clinical Oncology and Palliative Care, Zealand University Hospital, Næstved, Denmark; [e]Department of Oncology, Aarhus University Hospital, Aarhus, Denmark; [f]Laboratory of Radiation Physics, Department of Oncology, Odense University Hospital, Odense, Denmark; [g]Department of Oncology, Vejle Hospital, University Hospital of Southern Denmark, Vejle, Denmark; [h]Department of Oncology, Herlev and Gentofte Hospital, Herlev, Denmark; [i]Department of Clinical Oncology, Rigshospitalet, Copenhagen University Hospital, Copenhagen, Denmark; [j]Department of Oncology, Aalborg University Hospital, Aalborg, Denmark

ABSTRACT

Background: This study aimed at investigating the feasibility of developing a deep learning-based auto-segmentation model for the heart trained on clinical delineations.

Material and methods: This study included two different datasets. The first dataset contained clinical heart delineations from the DBCG RT Nation study (1,561 patients). The second dataset was smaller (114 patients), but with corrected heart delineations. Before training the model on the clinical delineations an outlier-detection was performed, to remove cases with gross deviations from the delineation guideline. No outlier detection was performed for the dataset with corrected heart delineations. Both models were trained with a 3D full resolution nnUNet. The models were evaluated with the dice similarity coefficient (DSC), 95% Hausdorff distance (HD95) and Mean Surface Distance (MSD). The difference between the models were tested with the Mann-Whitney U-test. The balance of dataset quantity versus quality was investigated, by stepwise reducing the cohort size for the model trained on clinical delineations.

Results: During the outlier-detection 137 patients were excluded from the clinical cohort due to non-compliance with delineation guidelines. The model trained on the curated clinical cohort performed with a median DSC of 0.96 (IQR 0.94–0.96), median HD95 of 4.00 mm (IQR 3.00 mm-6.00 mm) and a median MSD of 1.49 mm (IQR 1.12 mm-2.02 mm). The model trained on the dedicated and corrected cohort performed with a median DSC of 0.95 (IQR 0.93–0.96), median HD95 of 5.65 mm (IQR 3.37 mm-8.62 mm) and median MSD of 1.63 mm (IQR 1.35 mm-2.11 mm). The difference between the two models were found non-significant for all metrics ($p > 0.05$). Reduction of cohort size showed no significant difference for all metrics ($p > 0.05$). However, with the smallest cohort size, a few outlier structures were found.

Conclusions: This study demonstrated a deep learning-based auto-segmentation model trained on curated clinical delineations which performs on par with a model trained on dedicated delineations, making it easier to develop multi-institutional auto-segmentation models.

ARTICLE HISTORY
Received 29 April 2023
Accepted 16 August 2023

KEYWORDS
Deep learning-based auto-segmentation; clinical delineations; whole heart; breast cancer; radiotherapy

Background

Delineation of target structures and organs at risk (OARs) plays a crucial role in modern radiation treatment planning. However, delineation is a manual and time-consuming process prone to large uncertainties, which may lead to unnecessary irradiation of healthy tissue [1]. In recent years, an increasing number of studies have demonstrated the promise of deep learning-based automated image segmentation for OARs [2–4], allowing for standardized and quick delineation. Training of a robust and valid deep learning-based auto-segmentation model, requires a relatively large cohort of planning CT scans with structure sets delineated consistently in accordance with well-defined guidelines [5].

Creating a high-quality dedicated dataset for training of a deep learning model can be time consuming, and for feasibility reasons, such a dedicated dataset will often be generated at a single institution. However, a deep learning model trained on planning CT scans and structure sets from a single institution might not perform well on patients from other institutions due to differences in e.g., scanners, patient positioning and level of experience. To develop a model for a

CONTACT Stine Sofia Korreman stine.korreman@oncology.au.dk Dansk Center for Partikel Terapi, Aarhus Universitetshospital, Palle Juul-Jensen Boulevard 25, Indgang B3, 8200 Aarhus, Denmark

Supplemental data for this article can be accessed online at https://doi.org/10.1080/0284186X.2023.2252582.

multi-institutional setting, the model should be robust to differences among clinics, thereby increasing the number of required delineations for a representative dataset for training the model. This makes the task of creating a dedicated cohort for deep learning development even more extensive and time consuming. A solution could be to use retrospective databases of standardized clinical delineations, in cases where such databases are available. If these clinical delineations can be used to train a model, no extra time would need to be spent on delineating a dedicated training-set. The risk of using clinical delineations is that they may contain a higher degree of uncertainties than a dedicated dataset created for the purpose of model training. Clinical delineations are created under the constraints and considerations of daily clinical workflow which may change the way guidelines are used in practice. Furthermore, guidelines may be interpreted differently in different clinics and small changes may appear over time [6].

The Danish Breast Cancer Group (DBCG) has ensured complete reporting of planning CT scans, structure sets and dose plans for all Danish patients with node-positive breast cancer initiating their radiation treatment during 2008–2016 – the DBCG RT Nation Study [7–9]. The DBCG RT Nation Study includes data from more than 8,000 patients treated at all seven radiotherapy clinics in Denmark.

In the present study, we investigated the feasibility of training a deep learning-based auto-segmentation model for the whole heart on a large, curated dataset of uncorrected clinical delineations collected from the DBCG RT Nation Study. The performance of the model was compared to that of a model trained and tested on a small, dedicated and corrected dataset. The balance of dataset quantity versus quality was further investigated for the model trained on clinical delineations by stepwise reduction of the size of the dataset.

Materials and methods

This study included two different datasets as described below. The first dataset contained clinical OAR delineations from the DBCG RT Nation study performed according to the guideline from Nielsen et al. [10]. Before training of the deep learning model, a curation procedure was performed to identify outliers exhibiting gross deviations from the delineation guideline. No corrections of the remaining delineations were performed.

The second and smaller dataset contained delineations previously generated specifically in relation to development of an auto-segmentation model (for details see supplementary S1) following published guidelines [11,12] (slightly different from the guideline used for the clinical delineations). No outlier detection was performed for this dataset before training of the deep learning model.

Data

A cohort of 1,561 high-risk breast cancer patients from the DBCG RT Nation Study who started their treatment with loco-regional adjuvant radiotherapy during 2015–2016 was identified. Delineations were made according to the ESTRO (targets) and DBCG (OAR) guidelines [10,13].

The patients were treated at seven radiotherapy clinics in Denmark, and the dataset included CT scans, delineations, and dose plans for all patients. The cohort consisted of bilateral (4), left (825)- and right (732)-sided breast cancer patients, who either received a lumpectomy (907) or mastectomy (654). All patients had routinely undergone a planning CT scan without contrast. As the patients were treated at different radiotherapy (RT) clinics and over a 2-year period, the CT scans were acquired from different scanners and with different scanning protocols. The CT slice thickness varied from 2.0 mm to 3.0 mm and the longitudinal scan-length varied from 275 mm to 524 mm. All patients were treated in the supine position with either both arms or the ipsilateral arm elevated. We chose to exclude all patients who had been a part of the DBCG Skagen Trial 1 [14] (as these patients and the deep learning model are intended for use in a different study). The DBCG RT Nation study was approved by the Danish Patient Safety Authority who waived the need for patient consent due to the nature of the study.

The second dataset included CT scans and heart delineations from 114 breast cancer patients treated from 2005–2016 and 2019–2020 in four different RT clinics in Denmark (dosimetric data was not available for this cohort). The cohort contained both left (52 patients) and right (62) sided breast cancer patients, who either received a lumpectomy (83 patients) or a mastectomy (31 patients). The heart delineations were made according to Feng et al. [11], placing the cranial border higher than in the clinical dataset. The heart delineations for this dataset were created by Milo et al. [15,16] while developing an atlas based auto-segmentation for the heart and cardiac substructures (for further details regarding this dataset, see supplementary), and were evaluated and corrected by an experienced oncologist. The CT slice thickness varied from 2.0 mm to 5.0 mm and the longitudinal scan-length varied from 255 mm to 477 mm. All patients were treated in supine position with either both arms or the ipsilateral arm elevated. This cohort was used to train a second model on this smaller, dedicated, and corrected dataset. The use of these data was approved by the Danish Data Protective Agency (No.441757). The scientific ethical committee of the Central Denmark Region waived the need for approval due to the nature of the study.

Deep learning model training and evaluation

All CT scans and RT structure files were converted into NIfTI-files with the Python package dcmrtstruct2nii [17]. The clinical cohort was randomly split into two groups with 90% for training and 10% for testing, with no stratification for treatment centre. The dedicated and corrected cohort was split into 85% for training and 15% for testing due to its smaller size (for further explanation, see supplementary). All models were trained with a 3D full resolution nnUNet with five-fold cross validation (in which the dataset was divided in five subsets, and for each fold four of the datasets were used to train the model and the fifth was used to validate) and

default parameters [18]. The number of epochs was different for the models, depending on the cohort size to avoid overfitting, varying from 1,000 epochs to 300 epochs. No early stopping was used.

The models were tested on the respective test sets and geometric evaluation was performed using the Dice similarity coefficient (DSC), 95% Hausdorff distance (HD95) and Mean Surface Distance (MSD) between model prediction and ground truth. The results were quantified using the median and interquartile range (IQR), as the results were not expected to be normally distributed. For the model trained on the clinical delineations, the model performance was furthermore evaluated in terms of dose differences between model predictions and clinical delineations. The original dose distributions were used to evaluate the mean heart dose (MHD) in both clinical delineations and model segmentations, with the dicompylercore Python package [19].

Statistical significance was tested with the Mann-Whitney U-test, when comparing unpaired data and with the Wilcoxon signed-rank test, for paired data, assuming statistical significance level at $p < 0.05$.

Model trained on clinical delineations

Outlier detection

The dataset was first curated to eliminate gross deviations from delineation guidelines (from both training and test set). Detection of gross outliers was performed by training an initial deep learning model and inspecting all cases exhibiting low performance. The initial deep learning model was trained with all patients treated from 2015–2016, with a 90%/10% split into train/test set. Inference was performed on both the test and training set and the results were compared to the clinical delineations. Potential gross outliers (in both test and training set) were identified by manual inspection of all segmentations with a DSC score lower than 0.90 and/or a HD95 above 10 mm. If the clinical delineation exhibited obvious deviation from the guideline from Nielsen et al. [10], the patient was excluded from the cohort.

Training of the model

Following the outlier detection, a new 90%/10% split into train/test set was performed. A final model was trained on the remaining uncorrected clinical delineations in the training cohort.

Cohort size reduction

The effect of cohort size was investigated by training 3 additional models with a progressively reduced cohort size until reaching the same size as the dedicated cohort. The cohorts were randomly selected from the clinical model training cohort. Model performance was geometrically evaluated in the same test set used to evaluate the original clinical model. The model performances were compared to the model trained on the dedicated and smaller cohort.

A sanity check of the robustness of the smallest model towards cohort composition was performed by training two extra models with patients randomly selected from the clinical model training cohort.

Model trained on a dedicated cohort

For the dedicated and corrected cohort, no outlier-detection was performed, as it had already been reviewed by an experienced oncologist. The dedicated model was evaluated within its own test set due to the difference in delineation guidelines for the heart. The model performance was compared with the model trained on the clinical delineations.

Results

In total, 162 patients out of the 1,561 in the clinical cohort had no heart delineation and their CT scans and structure sets were excluded from the cohort. Furthermore, 308 patients were identified as being part of the DBCG Skagen trial 1 [14]. A flow diagram of the exclusion and training process can be seen in Figure 1.

Outlier detection

The model for outlier detection was trained on 982 patients and tested on 109 patients. The model performed with a median DSC of 0.95 (IQR 0.93–0.96) on the test set. When inference was run on both training and test set a total of 164 delineations had a DSC score below 0.9 and/or a HD95 larger than 10 mm. During the manual checks, 137 of these were excluded from the cohort, due to non-compliance with the delineation guideline [10], leaving a total of 954 patients to train and test the model. The main causes of exclusion were deviation from the cranial border and omission of parts of the heart (see example in Figure 2).

Model trained on clinical delineations

The model for the final curated clinical delineation dataset was trained on 859 patients and tested on 95 patients. The deep learning model performed with a median DSC of 0.96 (IQR 0.94–0.96) in the test set, a median HD95 of 4.00 mm (IQR 3.00 mm-6.00 mm) and a median MSD of 1.49 mm (IQR 1.12 mm-2.02 mm). Figure 3 shows three examples of the model output. Two of the predicted heart-delineations in the test set had a DSC lower than 0.90, see Figure S1 for further details.

The centre-specific median DSC for centres 1–6 were: 0.96 (IQR 0.94–0.96), 0.96 (IQR 0.96–0.97), 0.96 (IQR 0.96–0.97), 0.95 (IQR 0.94–0.96), 0.96 (IQR 0.93–0.96) and 0.95 (IQR 0.93–0.96), respectively. Centre 7 was only represented with two patients in the test set, scoring 0.88 and 0.90. See the HD95 and MSD in Table S1.

A pair-wise dose comparison between the clinical heart delineations and the model segmentations can be seen in Figure 4. The median MHD in the clinically delineated hearts were 1.35 Gy and in the model segmentations 1.34 Gy. The difference was not found statistically significant.

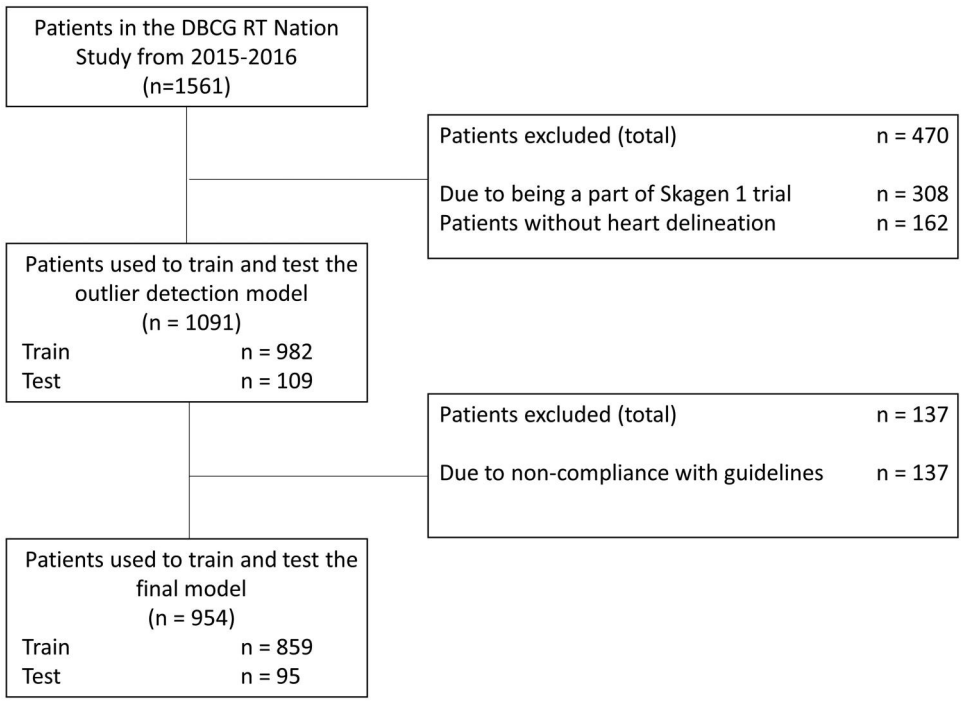

Figure 1. Overview of the dataflow process, from the total number of patients treated in 2015–2016, to the number of patients excluded before the outlier detection model, to the number of patients excluded after the outlier detection before training the model.

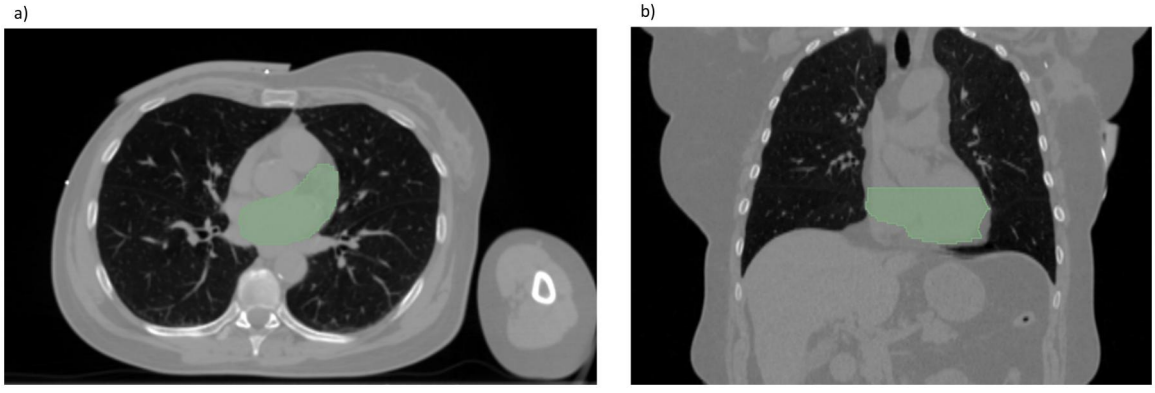

Figure 2. Examples of clinical heart delineations of patients excluded after inference with the outlier model. (a) Showing an example of a delineation omitting parts of the heart and (b) an example on the heart delineation deviating from the cranial border.

Model trained on a dedicated cohort

The dedicated model was trained on 96 patients and tested on 18 patients. The dedicated model performed with a median DSC of 0.95 (IQR 0.93–0.96) in the test set, a median HD95 of 5.65 mm (IQR 3.37 mm-8.62 mm) and median MSD of 1.63 mm (IQR 1.35 mm-2.11 mm).

The differences between the model trained on the clinical cohort and the model trained on the dedicated and corrected cohort were found to be non-significant for all three metrics ($p \gg 0.05$). See Table 1 for details.

Cohort size reduction

The reduced cohorts had 667, 381 and 95 respectively in their training cohort. The results of the models can be seen in Table 1. When comparing the model outputs with the dedicated cohort output, the differences were not found statistically significant. For the models trained on 381 and 95 patients, one and three large outliers were found in HD95, respectively, primarily due to extra volumes predicted in the arm or bowel regions of the patient. Although not statistically significant, a qualitative difference was seen in the range of all metrics for the model trained on 95 patients. Examples of this can be seen in Figure S2.

The two extra small size models, trained on 96 patients, performed with ranges of DSC 0.87–0.98/0.84–0.97, HD95 2.49 mm-24.0 mm/2.33 mm-63.0 mm and MSD 0.69 mm-4.92 mm/0.85 mm-15.2 mm. The medians were similar to the original small model for DSC, HD95 and MSD, see Table S2.

Discussion

In this study, we trained a deep learning-based model on clinical uncorrected delineations to auto-segment the heart in breast cancer patients. Our findings suggest that it is

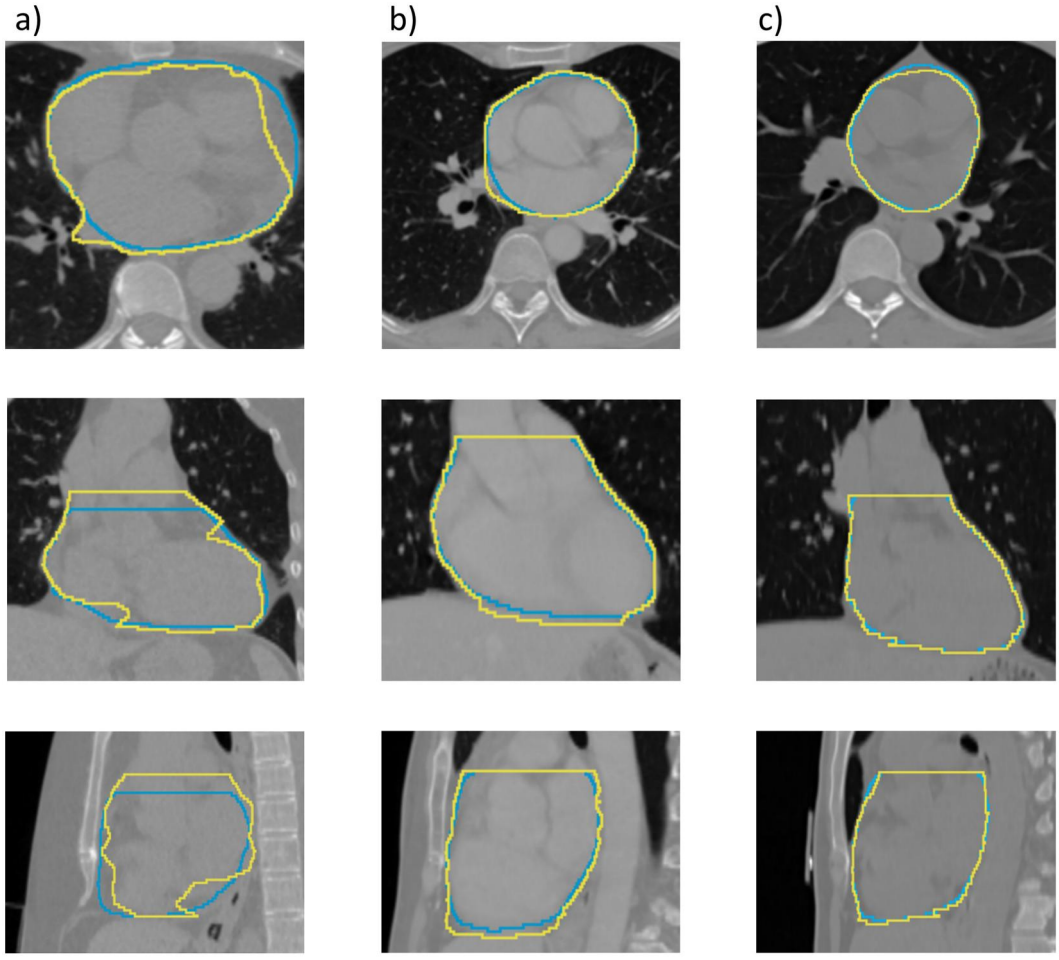

Figure 3. Examples of the lowest scored patient with the model trained on clinical delineations (a), with DSC = 0.88, HD95 = 12.0 mm and MSD = 4.21 mm, a median scored patient (b), with DSC = 0.95, HD95 = 5.46 mm and MSD = 1.50 mm and the highest scored patient (c), with DSC = 0.98, HD95 = 2.52 mm and MSD = 0.68 mm. The yellow line is showing the clinical delineation and the blue line is showing the model segmentation.

possible to collect data from multiple clinics following the same consistent guidelines and train a model performing on par with a model trained on a corrected and dedicated dataset. This eliminates the time-consuming task of generating delineations for a dedicated training dataset. We speculate that the good model performance is due to the majority of the delineations following consistent guidelines (and with curation to remove gross deviations from the guidelines), which dominates the model.

As deep learning-based auto-segmentation models are usually trained on dedicated training sets, we additionally trained a separate model on a smaller, but dedicated cohort for comparison. The dedicated model was trained on CTs of 96 patients and tested on CTs of 18 patients. It performed with a median DSC of 0.95, which is comparable to the performance of the model trained on clinical data, with a median DSC of 0.96. The same was valid for both HD95 and MSD. However, it is worth noticing, that the model performances have only been validated using quantitative geometrical performance metrics, thus no qualitative expert-assessment of the contours has been performed. In the case of implementation in clinical practice, an investigation of the clinical acceptance is warranted [20,21].

Both of the models performed on par with published auto-segmentation models for the heart [2,16,22]. Eldesoky et al. reported a mean DSC score of 0.92 with an atlas-based auto-segmentation model and Chung et al. achieved a DSC and HD95 of 0.95 and 4.56 mm respectively with a model trained on a corrected dataset with 92 cases in the training data. Additionally, the models performed on par with the inter-observer variation in manually delineated heart volumes achieved after the introduction of guidelines [10].

The model trained on the curated clinical cohort was trained using uncorrected clinical structures; thus, they represent the standard practice at the time and were created by a wide range of clinicians at different treatment centres on images from a range of scanners used in daily clinical practice. This may be considered as a strength but also a limitation. The multiple delineators introduce inter-observer variability [23] that may lead to inconsistencies in predictions. The test set is only a sample and there may be patients where the performance would be lower than in the test set. To minimize the number of delineations not following the guidelines in the training and test set, curation was performed in which potential deviations were checked by using an initial outlier detection model. When testing the

final model trained on the clinical cohort, only two patients in the test set had a DSC < 0.9. In one of these cases, the clinical delineation did not meet the cranial border as stated by the guideline and in the other, the model output did not meet the cranial border according to the delineation guideline [10].

When evaluating the MHD achieved by the model segmentation and the clinical heart delineations within the original dose plan, the difference in median MHD was found not to be statistically significant. However, for one patient, the difference in dose was 0.46 Gy, with 4.1 Gy in the clinical heart delineation and 3.64 Gy in the model segmentation.

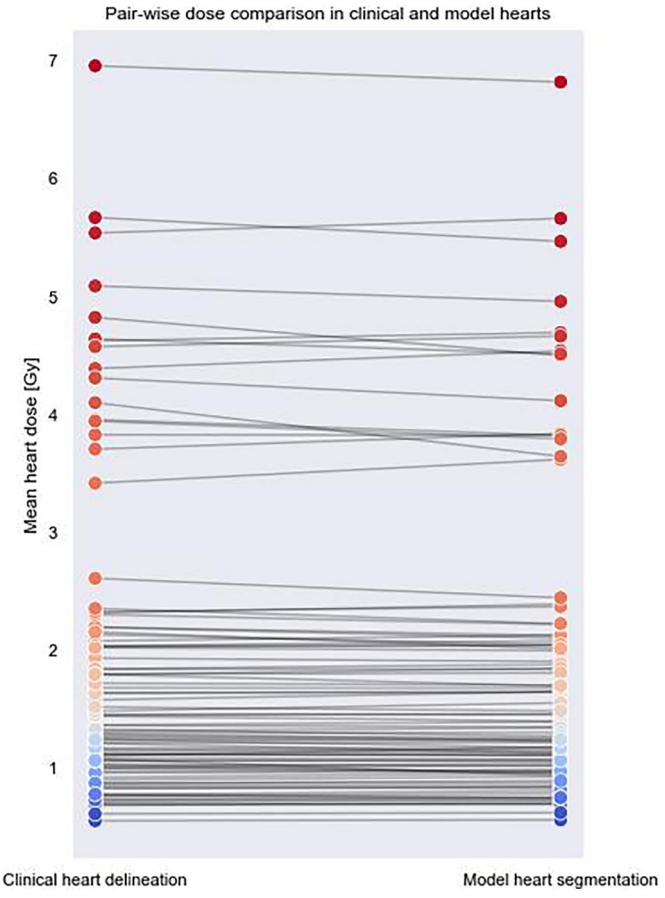

Figure 4. Pair-wise dose estimation based on the original dose distributions, in the clinical heart delineations (to the left) and in the model segmentations (to the right). The corresponding hearts are connected with a solid line, to illustrate the impact of the model segmentations.

The cohort containing clinical delineations followed the guideline created by Nielsen et al. in 2013, placing the cranial border of the heart at 'the lower part of the pulmonic trunc' [10]. The cohort with the corrected heart delineations followed the guideline created by Feng et al. with the cranial border of the heart 'just inferior to the left pulmonary artery' [11], thus the cranial border was more cranial than in the clinical cohort. The cohorts and models were therefore not mutually compatible, and each of the models was only tested within its own dataset. Comparison between the performance of the models nevertheless gave an idea of the differences in what can be achieved with a model trained on clinical delineations versus a model trained on corrected delineations.

The benefit of a model trained on a dedicated and corrected dataset is homogeneity in the data, including the low risk of delineations not following the guidelines in the training dataset. However, this may be counterbalanced by the robustness of the clinical model, where the variation in the training cohort was large in terms of patient positioning, scanner qualities and treatment centre. If the dedicated model was tested on data from another clinic, not present in the training cohort, the results might not be transferable.

When testing the cohort size impact on model performance, no statistically significant differences were found in the geometrical performance metrics. However, for both cohort size 381 and 95 a small number of outlier delineations were found, suggesting that the larger dataset produces a more robust model. When the smallest cohort size was tested with different cohort compositions, fluctuations were seen in terms of outliers, however not in the median performance.

This study demonstrated that it may be possible to train a deep learning-based auto-segmentation model on curated but uncorrected clinical delineations, which performs on par with a model trained on dedicated delineations, making it more feasible to create multi-institutional auto-segmentation models.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This study was supported by the Dansk Kræftforskningsfond; PhD scholarship funding from Department of Clinical Medicine, Aarhus University.

Table 1. Table of the geometrical performance metrics in the respective test-sets for the model trained on dedicated and corrected delineations and the model trained on curated and uncorrected clinical delineations with different cohort sizes.

Median\cohort	Dedicated cohort (size 96)	Clinical cohort (size 859)	Clinical cohort (size 667)	Clinical cohort (size 381)	Clinical cohort (size 95)
Dice similarity coefficient	0.95	0.96	0.96	0.95	0.95
	IQR 0.93–0.96	IQR 0.94–0.96	IQR 0.94–0.96	IQR 0.94–0.96	IQR 0.94–0.96
	Range 0.88–0.96	Range 0.88–0.98	Range 0.88–0.97	Range 0.87–0.97	Range 0.76–0.97
95% Hausdorff distance [mm]	5.65	4.00	4.55	4.23	4.46
	IQR 3.37–8.62	IQR 3.00–6.00	IQR 3.08–6.00	IQR 3.00–6.00	IQR 3.52–6.00
	Range 3.00–24.5	Range 2.56–18.0	Range 2.26–18.0	Range 2.22–56.4	Range 2.43–211
Mean surface distance [mm]	1.63	1.49	1.48	1.49	1.55
	IQR 1.35–2.11	IQR 1.12–2.02	IQR 1.18–2.02	IQR 1.08–2.10	IQR 1.27–2.22
	Range 0.95–3.06	Range 0.68–4.95	Range 0.78–4.96	Range 0.69–14.8	Range 0.82–29.9

The stated cohort size is the training size.

ORCID

Emma Riis Skarsø http://orcid.org/0000-0001-5408-3014
Ebbe Laugaard Lorenzen http://orcid.org/0000-0003-1895-733X

Data availability statement

Due to the nature of the research, and due to legal restrictions supporting data is not available for sharing.

References

[1] Groom N, Wilson E, Faivre-Finn C. Effect of accurate heart delineation on cardiac dose during the CONVERT trial. Br J Radiol. 2017;90(1073):20170036. doi: 10.1259/bjr.20170036.
[2] Chung SY, Chang JS, Choi MS, 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 | radiation oncology | full text. Radiat Oncol. 2021;16(1):44. doi: 10.1186/s13014-021-01771-z.
[3] Almberg SS, Lervåg C, Frengen J, et al. Training, validation, and clinical implementation of a deep-learning segmentation model for radiotherapy of loco-regional breast cancer. Radiother Oncol. 2022;173:62–68. doi: 10.1016/j.radonc.2022.05.018.
[4] Choi MS, Choi BS, Chung SY, et al. Clinical evaluation of atlas- and deep learning-based automatic segmentation of multiple organs and clinical target volumes for breast cancer. Radiother Oncol. 2020;153:139–145. doi: 10.1016/j.radonc.2020.09.045.
[5] Fang Y, Wang J, Ou X, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol. 2021;66(18):185012. doi: 10.1088/1361-6560/ac2206.
[6] van Mourik AM, Elkhuizen PHM, Minkema D, et al. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol. 2010;94(3):286–291. doi: 10.1016/j.radonc.2010.01.009.
[7] Refsgaard L, Skarsø ER, Ravkilde T, et al. End-to-end framework for automated collection of large multicentre radiotherapy datasets demonstrated in a danish breast cancer group cohort. Phys Imaging Radiat Oncol. 2023;27:100484.
[8] Refsgaard L, Skarsø ER, Ravkilde T, et al. Impact of guidelines on nationwide breast cancer treatment planning practices (DBCG RT nation study). Radiother Oncol. 2022;170: s 832–S834. doi: 10.1016/S0167-8140(22)02721-9.
[9] Refsgaard L, Ravkilde T, Skarsø ER, et al. OC-0425 dosimetric effects of national guidelines in breast cancer radiotherapy 2008-2016 (DBCG RT-Nation). Radiother Oncol. 2021;161: s 324–S325. doi: 10.1016/S0167-8140(21)06912-7.
[10] Nielsen MH, Berg M, Pedersen AN, et al. Delineation of target volumes and organs at risk in adjuvant radiotherapy of early breast cancer: national guidelines and contouring atlas by the danish breast cancer cooperative group. Acta Oncol. 2013;52(4):703–710. doi: 10.3109/0284186X.2013.765064.
[11] 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(1):10–18. doi: 10.1016/j.ijrobp.2009.10.058.
[12] Milo MLH, Offersen BV, Bechmann T, et al. Delineation of whole heart and substructures in thoracic radiation therapy: national guidelines and contouring atlas by the danish multidisciplinary cancer groups. Radiother Oncol. 2020;150:121–127. doi: 10.1016/j.radonc.2020.06.015.
[13] Offersen BV, Boersma LJ, Kirkove C, et al. ESTRO consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. Radiother Oncol. 2015;114(1):3–10. doi: 10.1016/j.radonc.2014.11.030.
[14] Francolini G, Thomsen MS, Yates ES, 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(2):282–287. doi: 10.1016/j.radonc.2017.03.011.
[15] Holm Milo ML, Slot Møller D, Bisballe Nyeng T, et al. Radiation dose to heart and cardiac substructures and risk of coronary artery disease in early breast cancer patients: a DBCG study based on modern radiation therapy techniques. Radiother Oncol. 2023;180:109453. doi: 10.1016/j.radonc.2022.109453.
[16] Milo MLH, Nyeng TB, Lorenzen EL, et al. Atlas-based auto-segmentation for delineating the heart and cardiac substructures in breast cancer radiation therapy. Acta Oncol. 2022;61(2):247–254. doi: 10.1080/0284186X.2021.1967445.
[17] Phil T, Albrecht T, Gay S, et al. Sikerdebaard/dcmrtstruct2nii: v5 [Internet]. Zenodo; 2023 [cited 2023 Mar 10]. Available from: https://zenodo.org/record/7705311
[18] Isensee F, Jaeger PF, Kohl SAA, et al. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat Methods. 2021;18(2):203–211. doi: 10.1038/s41592-020-01008-z.
[19] Panchal A, Couture G, Bot Pyup Io, et al. dicompyler/dicompyler-core v0.5.5 [Internet]. Zenodo; 2019 [cited 2022 Dec 9]. Available from: https://zenodo.org/record/3236628.
[20] Cha E, Elguindi S, Onochie I, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiother Oncol. 2021;159:1–7. doi: 10.1016/j.radonc.2021.02.040.
[21] Sherer MV, Lin D, Elguindi S, et al. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: a critical review. Radiother Oncol. 2021;160:185–191. doi: 10.1016/j.radonc.2021.05.003.
[22] Eldesoky AR, Yates ES, Nyeng TB, et al. Internal and external validation of an ESTRO delineation guideline - dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer. Radiother Oncol. 2016;121(3):424–430. doi: 10.1016/j.radonc.2016.09.005.
[23] 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(2):254–258. doi: 10.1016/j.radonc.2013.06.025.