Physics and Imaging in Radiation Oncology 12 (2019) 38–43

Contents lists available at ScienceDirect

Physics and Imaging in Radiation Oncology

journal homepage: www.elsevier.com/locate/phro

Original Research Article

Geometrical and dosimetric evaluation of breast target volume auto-contouring

Rita Simões[*], Geert Wortel, Terry G. Wiersma, Tomas M. Janssen, Uulke A. van der Heide, Peter Remeijer

Department of Radiation Oncology, The Netherlands Cancer Institute, Amsterdam, The Netherlands

ARTICLE INFO

Keywords:
Auto-contouring
Contour comparison
Dosimetric differences

ABSTRACT

Background and purpose: Automatic delineations are often used as a starting point in the radiotherapy contouring workflow, after which they are manually reviewed and adapted. The purpose of this work was to quantify the geometric differences between automatic and manually edited breast clinical target volume (CTV) contours and evaluate the dosimetric impact of such differences.

Materials and methods: Eighty-seven automatically generated and manually edited contours of the left breast were retrieved from our clinical database. The automatic contours were obtained with a commercial auto-segmentation toolbox. The geometrical comparison was performed both locally and globally using the Dice score and the 95% Hausdorff distance (HD). Two treatment plans were generated for each patient and the obtained dosimetric differences were quantified using dose-volume histogram (DVH) parameters in the lungs, heart and planning target volume (PTV). An inter-observer variability study with four observers was performed on a subset of ten patients.

Results: A median Dice score of 0.95 and a median 95% HD of 9.7 mm were obtained. Larger breasts were consistently under-contoured. Cranial under-contouring resulted in more than 5% relative decrease in PTV coverage in 15% of the patients while lateroposterior over-contouring increased the lung V_{20Gy} by a maximum of 2%. The inter-observer variability of the PTV coverage was smaller than the difference between PTV coverage achieved by the automatic and the consensus contours.

Conclusions: Cranial under-contouring resulted in under-treatment, while lateroposterior over-contouring resulted in an increased lung dosage that is clinically irrelevant, showing the need to consider dose distributions to assess the clinical impact of local geometrical differences.

1. Introduction

Organ and target volume contouring is an important step in the radiotherapy workflow. In the clinical routine, the treatment target and the organs-at-risk are manually delineated by experts. This is, however, a time-consuming task that is prone to intra- and inter-observer variability [1,2].

Automatic contouring approaches are expected to help reduce the clinical workload as well as decrease this variability. In particular, atlas-based segmentation techniques are well suited for anatomical structure segmentation [2]. Currently, the automatic contours can serve as a starting point in the contouring workflow. After being generated, they are reviewed and manually edited before being sent to the treatment planning system [2].

A question that arises is whether the automatic contours are comparable to the ones used clinically. Often, contour comparison is performed at the geometric level only [1,3–5]. However, the widely used geometrical metrics do not necessarily reflect the actual clinical impact of the contour differences [6,7].

This study aims to evaluate the clinical quality of the contours that have been generated automatically by an auto-contouring software, by comparing them to the manually corrected contours that have been used clinically. We quantify the geometrical differences between the contours and the corresponding dosimetric implications in terms of both target coverage and dosage to the organs at risk.

2. Materials and methods

2.1. Contours

Eighty-seven left breast cancer patients treated in the period from

[*] Corresponding author at: Department of Radiation Oncology, The Netherlands Cancer Institute, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands.
E-mail address: r.simoes@nki.nl (R. Simões).

https://doi.org/10.1016/j.phro.2019.11.003
Received 17 September 2019; Received in revised form 15 November 2019; Accepted 20 November 2019

December 2017 to May 2018 were selected from our clinical database. For this retrospective study, written informed consent was waived by the Institutional Review Board. For each patient, the automatically generated clinical target volume (CTV, consisting of the whole breast) delineation was retrieved, together with the respective manual edits made by a radiation oncologist.

The automatic contours were obtained with Mirada Medical's Workflow Box™ (WB), an atlas-based auto-contouring software solution. Within this toolbox, deformable image registration is performed between the atlases' and the patient's planning CT scans using an adaptation of the Lucas-Kanade optic flow algorithm [9]. Subsequently, the resulting deformation fields are used to propagate the individual atlas contours into the same reference. The deformed contours are finally fused into the final patient-specific automatic contour.

For the WB, we used nine left breast atlases that had been previously manually delineated by a clinician and verified by another. The atlas patients were selected qualitatively and an attempt was made to capture anatomical variability.

The manual edits to the automatically generated contours were made following the institutional protocol for delineating breast CTV [10].

2.2. Geometrical comparison

To assess the geometrical similarity between the two contours, we used two well-known metrics: the Dice score and the 95% Hausdorff distance (95% HD). The Dice score is defined as follows:

$$Dice = \frac{2\ |M \bigcap A|}{|M| + |A|}$$

where M and A correspond to the manual and the automatic contours, respectively. The 95% HD is defined as the 95th percentile of the surface distances. We also compared the volumes of the two contours.

To better characterize the spatial location of the geometrical deviations, we analyzed the average surface distance map, obtained after projection of each patient's individual surface distances onto a mean shape. This mean breast shape was determined by aligning the centers-of-mass of the automatic contours and taking the average distance from each contour surface voxel to the common center-of-mass. Then, for each patient's contour surface voxel, the corresponding distance value was projected onto the surface voxel of the mean breast shape that lies in the same direction as the one defined by the patient's contour surface voxel and the common center-of-mass.

As it is known from the literature that the inter-observer variability is highest at the lateroposterior and the cranial parts of the breast, indicating that these areas are the most difficult to delineate [3,4,8], we further defined two local metrics that represent cranial under-contouring and lateroposterior over-contouring. We cropped 10% of the most cranial breast slices and determined the 5th percentile of the signed distances (5pSD) from the surface points in the cropped volumes. This metric represents the largest (in absolute value) segmentation errors in this part of the breast. Similarly, we extracted the 10% most posterior slices and determined the 95th percentile of the signed distances (95pSD) of the surface points in this region.

2.3. Plan comparison

For the plan comparison, we used an in-house developed framework for automatic treatment planning [11,12] to generate treatment plans within Pinnacle^3 for both the automatic and the manually edited contours. The prescribed dose was 42.56 Gy given in 16 fractions. The plans consisted of medial and lateral tangential 6 MV beams, each combining an open segment, delivering at least 75% of the dose, and a limited number of IMRT segments, delivering the additional dose. The open beam was set up such that it just includes the planning target volume (PTV) on the medial side, using an additional 7 mm margin to account for the penumbra. As we do not allow the beam to cross the patient midline, the beam was shifted and the collimator was rotated until the beam crossed the patient midline. The heart, plus an additional 5 mm margin, was blocked from the field. On the lateral side, the beam was opened outside the patient, with an additional margin in order to be robust against contour changes. No flash was used during the optimization. The PTV was generated by expanding the CTV with 5 mm and cropping it to 7 mm under the skin to allow build-up. Once the open beam was set up, a help structure PTVedit was created that consisted of only the part of PTV that lied within the beam. The plan was optimized with a fixed set of objectives on the heart (max $<$ 38 Gy), lungs (mean $<$ 5 Gy), PTVedit (min $>$ 97%, uniform 100%, max $<$ 105%) and conformity (max $<$ 100% outside PTV).

The FAST framework for completely automatic breast treatment planning has been in clinical use since 2015. Using this automated planning approach, two plans were created for each patient. The first plan was made using the automatically delineated PTV, the second plan was made on the manually corrected PTV. Both of these plans have their own beam setup-up and therefore their own optimization help structure PTVedit. For both plans, the target coverage $V_{95\%}$ was evaluated on the unedited manually corrected PTV, which is considered to be the radiation oncologist approved golden standard.

To compare the difference in PTV coverage achieved by the plans made using the manual and the automatic contours, we defined the relative $V_{95\%}$ difference as follows:

$$\text{PTV r}\Delta \text{V}_{95\%} = \frac{PTV\ \ V_{95\%}^{automatic} - PTV\ \ V_{95\%}^{manual}}{PTV\ \ V_{95\%}^{manual}}$$

where $V_{95\%}^{automatic}$ corresponds to the $V_{95\%}$ obtained by the plan generated using the automatic contour and $V_{95\%}^{manual}$is the $V_{95\%}$ obtained by the plan that was made based on the manual contour. These $V_{95\%}$ values were determined on the PTV obtained from the manual contour.

To investigate possible over-dosage of healthy tissues, we also analyzed the lung dosage. The difference in dose delivered to the lungs was quantified using the absolute difference in the V_{20Gy} between the plans generated using the automatic and the manually edited contours:

$$\text{Lung }\Delta \text{V}_{20Gy} = \text{Lung V}_{20Gy}{}^{automatic} - \text{Lung V}_{20Gy}{}^{manual}$$

Similarly, we determined the absolute difference of the following dose-volume histogram (DVH) parameters between the two plans: Heart D_{mean}, Lung D_{mean}, PTV $D_{2\%}$, PTV $D_{98\%}$, PTV D_{mean} and Heart V_{5Gy}.

2.4. Inter-observer variability study

We selected five cases in which the auto-contouring performed worst in terms of PTV coverage, and five random cases with $V_{95\%}$ values (determined in the plan made using the automatic contour) in the range between 92% and 96%.

Four experienced clinicians independently delineated the breasts of these ten cases, after which the majority vote over these four delineations plus the automatic contour was taken as the consensus delineation. Six different plans were made per patient: four using the manual delineations, one based on the consensus and one using the automatic contour. The obtained $V_{95\%}$ values on the consensus PTV were selected as the dosimetric performance metrics.

3. Results

3.1. Geometrical comparison

A median Dice score of 0.95 (range 0.80–1.0) and a median 95% HD of 9.7 mm (range 1.0–39.0 mm) were obtained. As shown in Fig. 1, the largest (in absolute value) average surface distances occurred at the most cranial and lateroposterior parts of the breast. In particular, in the

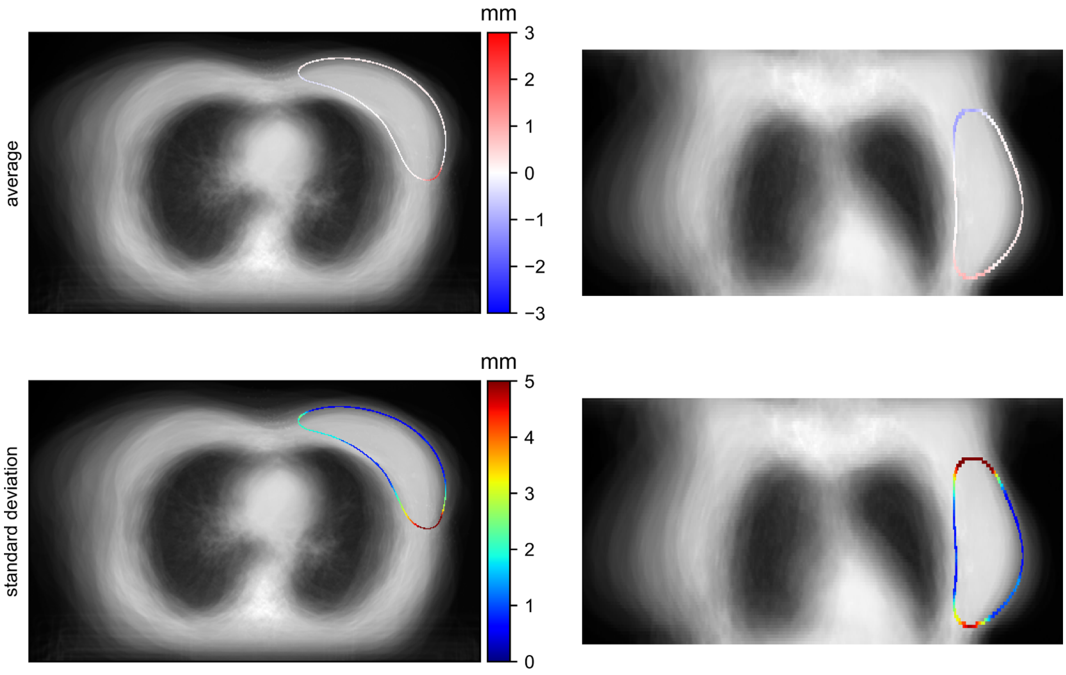

Fig. 1. Top row: axial (left) and coronal (right) midslices of the 3D average surface distance map (signed values); the standard deviations are displayed in the bottom row. The distances are overlaid on the average CT scan of all eighty-seven patients and they indicate the relative position of the contours, with negative values meaning under-contouring and positive values indicating over-contouring by the auto-contouring toolbox.

cranial area the auto-segmentation toolbox tended to under-contour the breast. On the other hand, it more often over-contoured the lateroposterior part of the breast.

Additionally, we observe from Fig. 2 that larger breasts were consistently under-contoured by the auto-contouring toolbox. For breasts with volumes in the range 500–1000 cm^3, the segmented volumes were comparable to the manual ones.

3.2. Plan comparison

In the axial view shown in Fig. 3, we can observe that, even though the delineations differed lateroposteriorly, the dose distribution was similar. However, in the sagittal view, where a clear cranial under-contouring is visible, there were marked differences in dose distribution.

A first analysis of the global geometrical metrics revealed that these correlated poorly with the PTV coverage, as represented by the PTV r$\Delta V_{95\%}$. In contrast, the local analysis on the cranial surface distances revealed a decrease of more than 5% in the $V_{95\%}$ for 15% of the patients when the automatic contours were used for plan generation (Fig. 4a)). In particular, severe under-contouring of up to 2 cm at the cranial part of the breast accounted for a reduction of more than 10% of PTV coverage. In contrast, and as would be expected, the few cases in which cranial over-contouring occurred did not have an impact on the PTV coverage.

To assess the dosimetric impact of the lateroposterior over-contouring, we investigated the relation between the 95pSD and the lung ΔV_{20Gy} (Fig. 4b)). In the worst cases of up to 4 cm of lateroposterior over-contouring there was an increase of 2% in the V_{20Gy} delivered to the lungs with respect to the clinically delivered 3%. The maximum lung V_{20Gy} obtained clinically was 8%. Also, for these patients, the clinical mean lung dose was 2.5 ± 0.5 Gy.

Additional dosimetric parameters are reported in Table 1 and a figure containing the DVH of the PTV, lungs and heart for all eighty-seven patients can be found in the Supplementary Materials (Fig. S1). It is worth noting that the PTV D_{mean} and $D_{2\%}$ are similar between the two plans, with an average difference of −37.7 cGy and −2.65 cGy, respectively. This suggests that the two plans are similar in terms of PTV dose homogeneity. The heart is blocked from the field by design of our treatment planning, meaning that there should not be relevant dose differences if the CTV contour changes. This is confirmed by the heart V_{5Gy} difference of 0%.

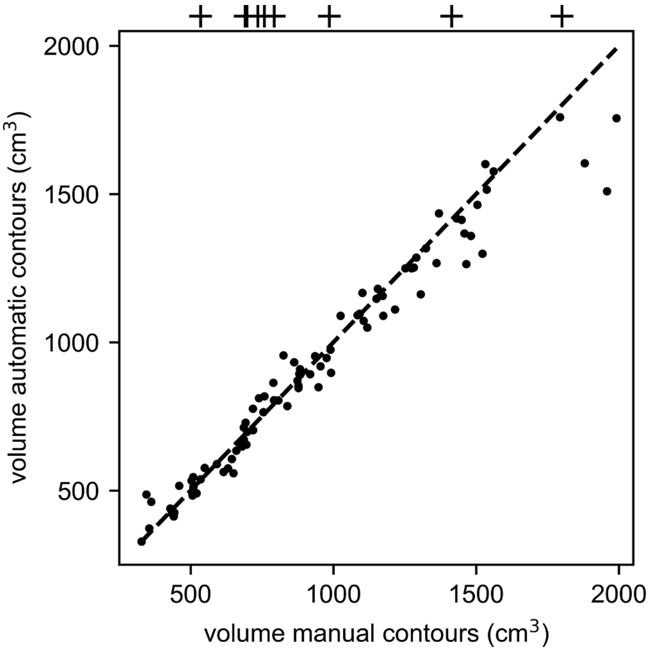

Fig. 2. Manual *vs.* automatic CTV volumes; the atlas volumes are indicated as crosses above the plot.

3.3. Inter-observer variability study

As shown in Fig. 5, the plans made using the automatic contours consistently resulted in a PTV coverage that largely differs from that which is obtained using the manual contours.

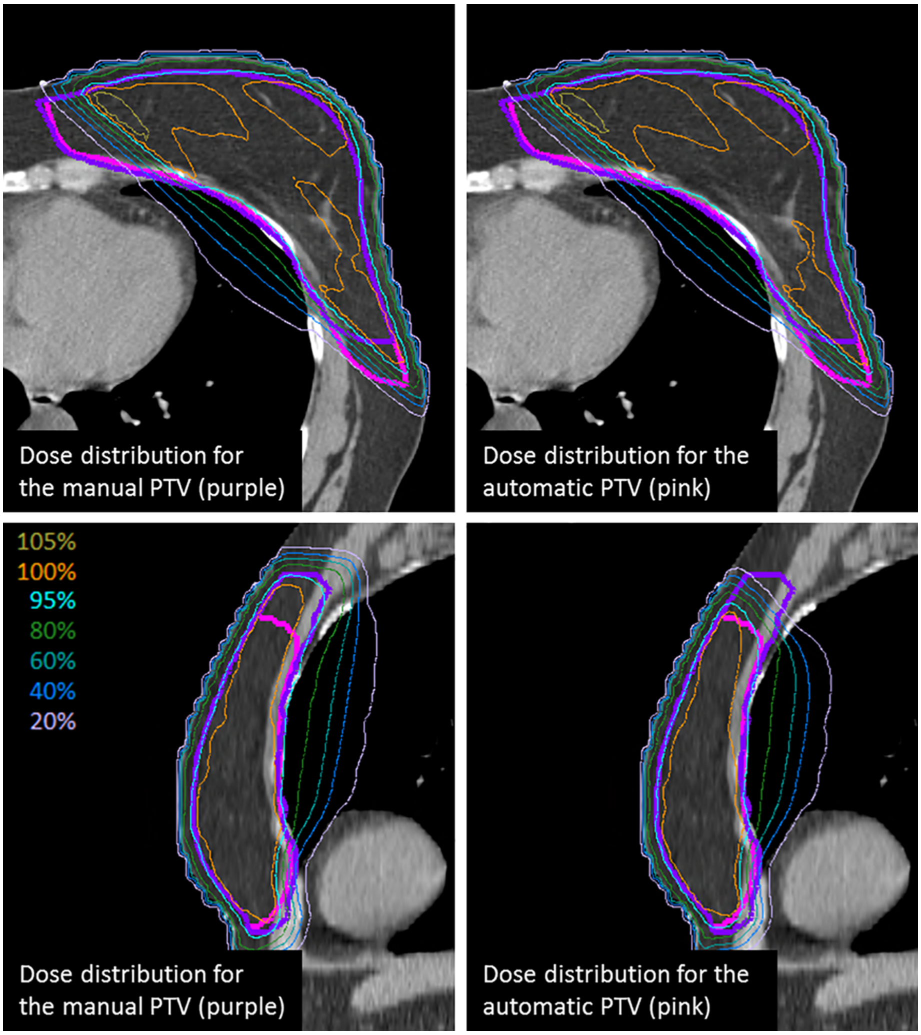
Dose distribution for the manual PTV (purple)
Dose distribution for the automatic PTV (pink)
105%
100%
95%
80%
60%
40%
20%
Dose distribution for the manual PTV (purple)
Dose distribution for the automatic PTV (pink)

Fig. 3. Manual (purple) and automatic (pink) PTVs and their respective generated plans (manual left; automatic right) in a patient for which the auto-contouring resulted in both cranial under-contouring and lateroposterior over-contouring. Top row: axial view; bottom row: sagittal view. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In particular, the mean absolute difference between the manual and the consensus $V_{95\%}$ values was lower than 1% for all patients. Also, the standard deviation of the differences over the four observers was lower than 1% for all patients.

For the automatic contours, this difference ranged from 1% to 12%, being in all cases larger than the differences between the manual and the consensus $V_{95\%}$ values, with this effect being more pronounced in the five worst-performing cases.

4. Discussion

In this study, we evaluated the clinical quality of the breast auto-contours that have been generated by an atlas-based segmentation toolbox.

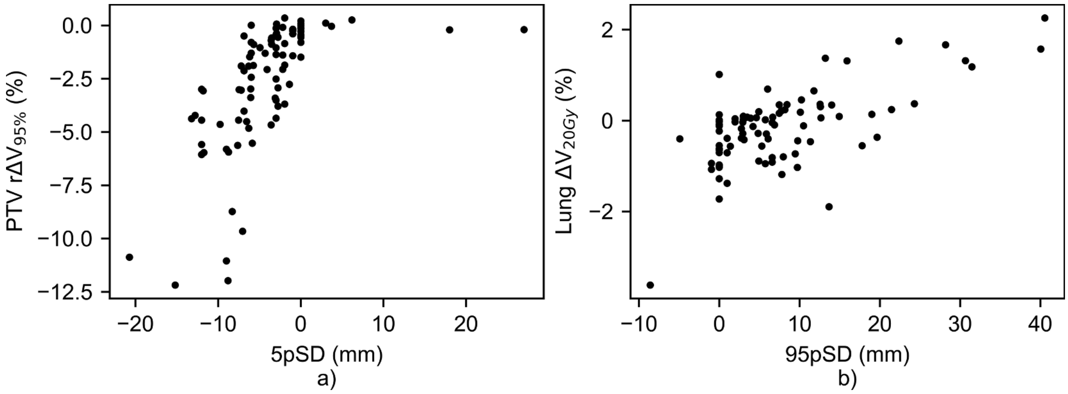

Fig. 4. a) 5th percentile of the signed distances on the most cranial 10% of the breast (5pSD) *vs.* the PTV $r\Delta V_{95\%}$; b) 95th percentile of the signed distances on the most posterior 10% of the breast (95pSD) *vs.* the ΔV_{20Gy} in both lungs.

Table 1
Absolute differences between the dosimetric parameters in the two plans (automatic – manual).

Metric	Mean	Standard deviation
Heart D_{mean} [Gy]	− 0.008	0.062
Lung D_{mean} [Gy]	− 0.072	0.342
Lung V_{20Gy} [%]	0	1
PTV $D_{2\%}$ [Gy]	− 0.027	0.240
PTV $D_{98\%}$ [Gy]	− 5.450	8.213
PTV $V_{95\%}$ [%]	− 3	3
PTV D_{mean} [Gy]	− 0.377	0.571
Heart V_{5Gy} [%]	0	0

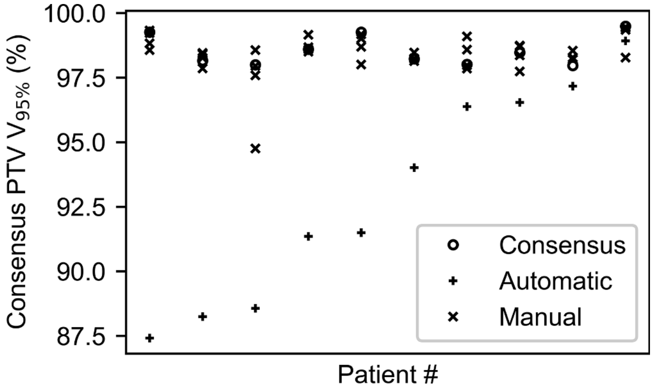

Fig. 5. Dosimetric inter-observer variability for 10 patients, ordered from left to right according to the $V_{95\%}$ (determined in the consensus PTV) obtained by the plan made based on the automatic contour.

The largest geometrical differences with respect to the manual contours occurred at the most lateroposterior and the most cranial parts of the breast. The under-contouring at the cranial level resulted in severe PTV under-coverage. In particular, we observed a decrease of more than 5% of the $V_{95\%}$ in 15% of the patients. In contrast, even extreme lateroposterior over-contouring of about 4 cm only resulted in an increase of 2% in the V_{20Gy} delivered to the lungs with respect to the clinically delivered 3%. In all cases, the V_{20Gy} of both lungs lies below 8%. This is well below the 20–30% threshold range that is typically considered when analyzing the risk for radiation pneumonitis, which is the most common dose-related complication of radiation in the thoracic cavity [13]. It is worth pointing out that, by design, our treatment plans spare the heart. We confirmed that there was no difference in the heart V_{5Gy} between the two plans. Also, the plans show similar dose homogeneity, as reflected in the average absolute difference between the plans of the PTV $D_{2\%}$ and D_{mean} values.

The posterior and cranial parts of the breast have been reported as the locations with the highest geometrical inter-observer variability, suggesting that these areas are typically more difficult to delineate. Hurkmans et al. determined the maximum distance between the PTV contours of multiple observers to be 42 mm in the posterior direction and 27 mm in the cranial direction [3]. Struikmans et al. found a smaller variability on the CTV (less than 1 cm), but also consistently higher in the posterior and cranial directions [4].

In the specific case of RT, the contours are one step in the pipeline to generate a treatment plan, which means that the actual clinical implications of possible geometrical deviations at the contour level should be evaluated downstream, at the dose level [2]. To the best of our knowledge, Li et al. were the only ones to report dosimetric differences in target and organs-at-risk delineations for breast cancer patients by comparing the dose volume histograms (DVH) obtained using the manual contours from nine radiation oncologists [8]. They found that the geometrical variability accounted for dosimetrically significant differences in the heart and the lungs, but that the PTV coverage criteria did not vary significantly.

We performed our own dosimetric inter-observer variability study on a small subset of our data, in which we evaluated the PTV coverage obtained using each of the manual delineations, the consensus delineation and the automatic one. The plans made using the automatic contour consistently and largely under-performed the plans made with the other delineations in terms of $V_{95\%}$ on the consensus PTV. This suggests that, for these patients, the issue lied in the auto-contouring itself rather than in possible anatomical particularities of the patients – in which case we would expect a large inter-observer variability. Also, our findings with respect to the dosimetric inter-observer variability of the PTV coverage are in line with those reported by Li et al. [8].

The poor performance in these cases can probably be explained by the lack of sufficient anatomical variability in the set of nine atlases. In particular, we have observed that larger breasts were significantly under-contoured by the auto-contouring toolbox, as depicted in Fig. 2. As in most atlas-based auto-segmentation solutions, for each new patient the atlas scans are non-rigidly registered to the patient's scan, after which the resulting transformation is applied to the atlas delineations. A decision rule is then applied to this set of deformed contours. When the anatomy of the patient undergoing segmentation differs considerably from that of the atlas patient, the registration outcome is likely sub-optimal, leading to erroneous contour propagations.

The most obvious recommendation for improvement of our clinical auto-contouring workflow is to add breasts with more anatomical variability to our current set of atlases. Also, and perhaps more interestingly, we could envisage incorporating a preliminary atlas selection step in which, for each new patient, only a specific subset of atlases that share common anatomical features with the patient should be used in the subsequent steps of the segmentation pipeline [5,14,15].

Additionally, it is worth pointing out that, although in a non-negligible percentage of patients the PTV coverage is considerably lower when the automatic contours are used for planning, in a large number of cases there would actually be no need for manual edits as they do not translate into dosimetrically relevant improvements. An automatic flagging system of the badly-performing auto-segmentation cases would therefore be desirable to optimize our current clinical segmentation workflow.

We believe that any clinically meaningful evaluation of auto-contouring performance should include a dosimetric assessment of geometrical differences. In that context, the described geometrical and dosimetric comparison is independent of both the delineated structures and the automatic segmentation method and could potentially be performed in future studies for a systematic assessment of the clinical impact of any update in our auto-segmentation methods.

We can conclude from this dosimetric analysis that the cranial under-contouring is likely to result in significant under-treatment while the lateroposterior over-contouring, even in the most severe cases, does not result in a clinically relevant over-dosage of the healthy tissues.

Funding statement

This research received no specific grant from any funding agency.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.phro.2019.11.003.

References

[1] Reed VK, et al. Automatic segmentation of whole breast using an atlas approach and deformable image registration. Int. J. Radiat. Oncol. Biol. Phys. 2009;73(5):1493–500.

[2] Sharp G, et al. Vision 20/20: Perspectives on automated image segmentation for radiotherapy. Med. Phys. 2014;41(5).

[3] Hurkmans CW, Borger JH, Pieters BR, Russell NS, Jansen EP, Mijnheer BJ. Variability in target volume delineation on CT scans of the breast. Int. J. Radiat. Oncol. Biol. Phys. 2001;50(5):1366–72.

[4] Struikmans H, et al. Interobserver variability of clinical target volume delineation of glandular breast tissue and of boost volume in tangential breast irradiation. Radiother. Oncol. 2005;76(3):293–9.

[5] Anders LC, Stieler F, Siebenlist K, Schäfer J, Lohr F, Wenz F. Performance of an atlas-based autosegmentation software for delineation of target volumes for radiotherapy of breast and anorectal cancer. Radiother. Oncol. 2012;102(1):68–73.

[6] Vinod SK, Jameson MG, Min M, Holloway LC. Uncertainties in volume delineation in radiation oncology: A systematic review and recommendations for future studies. Radiother. Oncol. 2016;121(2):169–79.

[7] Voet PWJ, Dirkx MLP, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJM. Does atlas-based autosegmentation of neck levels require subsequent manual contour editing to avoid risk of severe target underdosage? A dosimetric analysis. Radiother. Oncol. 2011;98(3):373–7.

[8] Li XA, et al. Variability of target and normal structure delineation for breast-cancer radiotherapy: a RTOG multi-institutional and multi-observer study. Int. J. Radiat. Oncol. Biol. Phys. 2009;73(3):944–51.

[9] Mirada Medical, “Why choose Mirada Registration ?,” 2002.

[10] Offersen BV, 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.

[11] R. De Graaf et al., Single-click automatic radiotherapy treatment planning for breast, prostate and vertebrae, in: ESTRO 35, 29 April – 3 May 2016, Turin, Italy, 2016, vol. 119, pp. S758–S759.

[12] G. Wortel et al., “Single-click generation of whole breast IMRT treatment plans,” in ESTRO 35, 29 April - 3 May 2016, Turin, Italy. Vol 119, 2016.

[13] Rodrigues G, Lock M, D’Souza D, Yu E, Van Dyk J. Prediction of radiation pneumonitis by dose–volume histogram parameters in lung cancer—a systematic review. Radiother. Oncol. 2004;71(2):127–38.

[14] Ciardo D, et al. Atlas-based segmentation in breast cancer radiotherapy: Evaluation of specific and generic-purpose atlases. The Breast 2017;32:44–52.

[15] Zaffino P, et al. Multi atlas based segmentation: should we prefer the best atlas group over the group of best atlases? Phys. Med. Biol. 2018;63(12):12NT01.