Physica Medica 123 (2024) 103402

Contents lists available at ScienceDirect

Physica Medica

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

Original paper

Sensitivity of automated and manual treatment planning approaches to contouring variation in early-breast cancer treatment

Michele Zeverino [a], Consiglia Piccolo [a], Maud Marguet [a], Wendy Jeanneret-Sozzi [b], Jean Bourhis [b], Francois Bochud [a], Raphaël Moeckli [a,*]

[a] Institute of Radiation Physics, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland
[b] Radiation Oncology Department, Lausanne University Hospital and University of Lausanne, Lausanne, Switzerland

ARTICLE INFO

Keywords:
Automated treatment planning
Contouring variation
Planned dose difference
Early-breast cancer treatment

ABSTRACT

Purpose: One of the advantages of integrating automated processes in treatment planning is the reduction of manual planning variability. This study aims to assess whether a deep-learning-based auto-planning solution can also reduce the contouring variation-related impact on the planned dose for early-breast cancer treatment.
Methods: Auto- and manual plans were optimized for 20 patients using both auto- and manual OARs, including both lungs, right breast, heart, and left-anterior-descending (LAD) artery. Differences in terms of recalculated dose (ΔD_{rc}^{M}, ΔD_{rc}^{A}) and reoptimized dose (ΔD_{ro}^{M}, ΔD_{ro}^{A}) for manual (M) and auto (A)-plans, were evaluated on manual structures. The correlation between several geometric similarities and dose differences was also explored (Spearman's test).
Results: Auto-contours were found slightly smaller in size than manual contours for right breast and heart and more than twice larger for LAD. Recalculated dose differences were found negligible for both planning approaches except for heart (ΔD_{rc}^{M}=-0.4 Gy, ΔD_{rc}^{A}=-0.3 Gy) and right breast (ΔD_{rc}^{M}=-1.2 Gy, ΔD_{rc}^{A}=-1.3 Gy) maximum dose. Re-optimized dose differences were considered equivalent to recalculated ones for both lungs and LAD, while they were significantly smaller for heart (ΔD_{ro}^{M}=-0.2 Gy, ΔD_{ro}^{A}=-0.2 Gy) and right breast (ΔD_{ro}^{M}=-0.3 Gy, ΔD_{ro}^{A}=-0.9 Gy) maximum dose. Twenty-one correlations were found for $\Delta D_{rc}^{M,A}$ (M=8,A=13) that reduced to four for $\Delta D_{ro}^{M,A}$ (M=3,A=1).
Conclusions: The sensitivity of auto-planning to contouring variation was found not relevant when compared to manual planning, regardless of the method used to calculate the dose differences. Nonetheless, the method employed to define the dose differences strongly affected the correlation analysis resulting highly reduced when dose was reoptimized, regardless of the planning approach.

1. Introduction

The use of automation in radiation therapy has recently rapidly increased, significantly impacting every step of the treatment process. Most efforts have been concentrated on addressing highly time-consuming tasks within the workflow, such as structures delineation and treatment planning. To achieve this, machine learning techniques, particularly knowledge- and deep learning (DL)-based approaches, have been employed to expedite these processes and enhance the overall quality of patient plans [1–3].

DL-based auto-segmentation techniques have been demonstrated to outperform other automated methods, showing good results in shortening the delineation time and approaching the accuracy of manual segmentation [4–8].

To relate the performance of automatic segmentation with dose distribution, several researchers have studied the correlation between contouring variation and dose differences. Nonetheless, the methods found in the literature used to calculate dose differences were ambiguous, and according to their formulations, they may produce different correlation results for the same case. Typically, the process starts with a plan that is first optimised for one of the two given sets of structures providing the reference dose. Then two paths can be followed to calculate the candidate dose for the concurrent structures: a simple dose-volume histogram (DVH) re-calculation [6,9–11] or a new plan re-

* Corresponding author.
E-mail address: raphael.moeckli@chuv.ch (R. Moeckli).

https://doi.org/10.1016/j.ejmp.2024.103402
Received 24 October 2023; Received in revised form 24 May 2024; Accepted 5 June 2024
Available online 12 June 2024

optimization [12–16]. Obviously, the candidate dose will be highly affected by the method used for the calculation. There is even more ambiguity by the fact that dose difference between reference and candidate doses did not always refer to a structure belonging to the same structure set [16,17]. Both approaches used to carry out the candidate dose have limitations. DVH re-calculation generally over-estimates the dose differences for structures lying near gradient regions [18]. In addition, it neglects the trade-off effects between structures deriving from the inverse planning. Dose re-optimization provides a more realistic dose difference than DVH re-calculation as geometrical variation are considered in the optimization process. However, since the optimization problem has multiple solutions, dose differences may arise due to optimization settings rather than contour geometrical variations [15].

Automated planning solutions have shown they can standardize the quality of the dose distribution for several treatment sites, reducing the planning variability [4,19,20]. For manual planning, the presence of competing objectives for target coverage and organ sparing requires multiple subjective trade-offs from the planner that are translated in the choice of specific objective functions and relative weights assigned to normal structures and targets in the definition of the optimization protocol. One of the factors influencing the dose trade-offs in the optimization problem is the mutual position of the target and organs at risk (OARs) and, in manual planning, the optimization protocol implicitly considers their spatial relationship when defining the objective functions and relative weights. If this spatial relationship no longer exists, there is no guarantee that the same optimization protocol would produce an equivalent dose distribution. Auto-planning approaches automatically introduce appropriate objectives and weights into the optimization procedure according to the prediction on the most feasible DVHs (knowledge-based) or 3D-dose distributions (DL-based) existing for the given geometry, replacing the iterative trial-and-error manual optimization process [21]. Therefore, they should be less sensitive to geometric differences and provide dose distributions that are more robust to contouring variation than manual planning.

To the best of our knowledge no one has yet compared the dosimetric impact of contouring variation between automated and manual planning approaches to assess their sensitivity to different structure sets. At the same time, no studies have evaluated the potential impact of the candidate calculation method on the correlation between contouring variation and dose difference. Therefore, the objective of this study was to assess the dosimetric impact differences between manual and DL-based auto-segmented OARs contouring variations for both manual and DL-based auto-planning approaches in left-sided early-breast cancer volumetric modulated arc therapy (VMAT) treatment under deep-inspiration breath hold (DIBH) conditions. Furthermore, for each planning approach, the correlations between contouring variations and dose differences were evaluated. Regardless of the planning approach, such correlations were evaluated against two different formulations of dose difference (e.g. recalculated vs reoptimized dose) to demonstrate how the method used may affect the results. Recalculated and reoptimized dose differences were evaluated on the manual structure set taken as reference.

2. Materials and methods

The study workflow is summarized in Fig. 1. Patients OARs were manually or DL-based automatically contoured. We first compared the accuracy of auto-segmentation to manual segmentation based on geometric similarity metrics. Second, we optimized the manual- and auto-plans by having both the manual and auto-segmented OARs generate the doses needed for dose difference computation. Third, we explored the correlation between the geometric metrics and dosimetric differences for each planning approach.

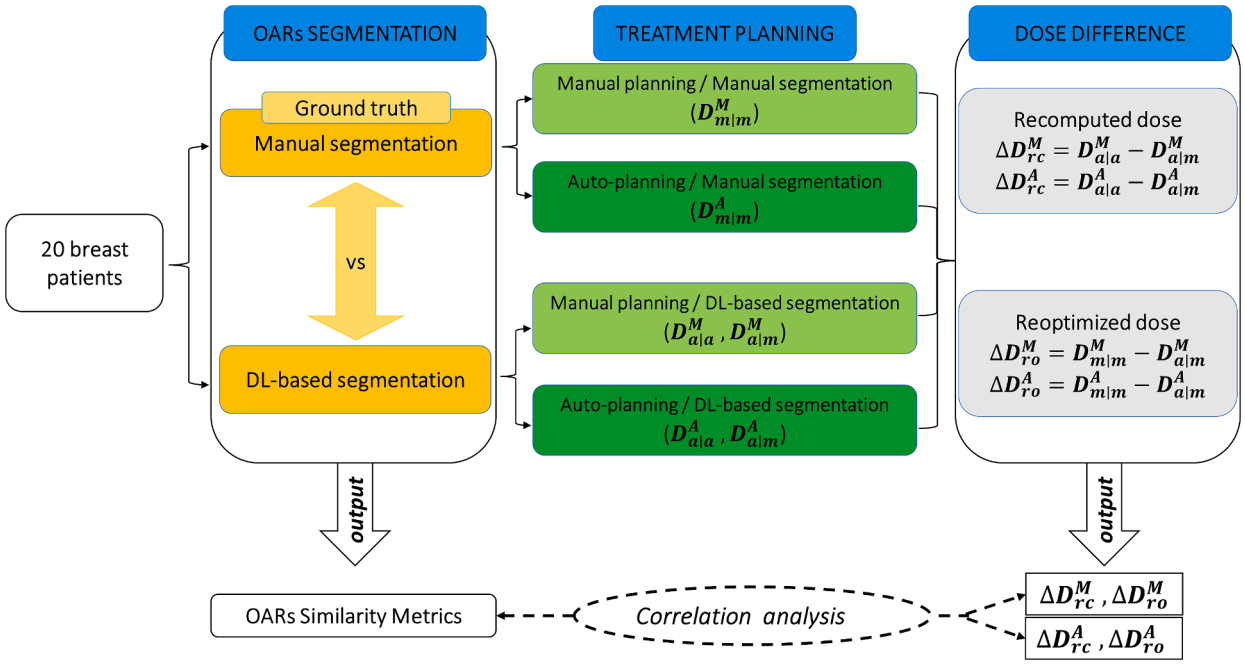

Fig. 1. Study workflow summary. OARs of twenty patients were contoured both on manual (ground truth) and automated process. For each patient, manual and auto-plans were optimized using both structure sets for a total of 4 plans providing different dose distributions named according to the following nomenclature: $D^{planningapproach}_{optimizationstructures|evaluationstructures}$, where *planning approach* = M (manual) or A (automatic) and *optimization or evaluation structures* = m (manual) or a (automatic), respectively. Regardless of the planning approach, two dose difference calculation methods were used to evaluate the correlation with these latter and OARs similarity metrics: (1) Recomputed dose difference (ΔD_{rc}) as the dose difference between automatic and manual OARs for a plan optimized with automatic structures and (2) Reoptimized dose difference (ΔD_{ro}) as the difference between plans optimized with manual and automatic OARs, respectively (see Section 2.4 for further details).

2.1. Patient contouring

Twenty left-sided early-stage breast cancer patients were randomly selected from our treatment database and involved in the study. They received a simultaneously-integrated-boost (SIB) treatment under DIBH conditions at our institute between 2021 and 2022. Dose prescription was 60 Gy and 50 Gy in 25 fractions for PTV Boost (PTV1) and PTV Whole Breast (PTV2), respectively. Targets and OARs were manually delineated by a single senior radiation oncologist according to the ESTRO and DBCG guidelines for early-stage breast cancer [22,23]. PTV1 and PTV2 were generated by expanding their respective clinical target volumes (CTVs) by 5 mm and then cropping it 3 mm under the skin. Auto-segmented structures were generated using the DL-based segmentation model available in RayStation (RS) TPS (RaySearch Laboratories, Stockholm, Sweden) and previously validated [6]. OARs evaluated in the study included contralateral breast, heart, left-anterior-descending artery (LAD) and both lungs. PTVs were only manually delineated.

We assessed the need for ethical and/or legal approval for the present study and concluded that no approval was required.

2.2. Geometric metrics for OARs

To evaluate contouring differences between manual (ground truth) and automatically generated structures, different geometrical indices were calculated for every OARs.

Given A the volume of the manual structures and B the volume of the auto-contoured structures, the indices were calculated as follows.

The volume difference (ΔV(%)) measured the percentage difference between A and B normalized to A, ΔV(cm3) was used for LAD only because of the small size of the structure:

$$\Delta V(\%) = \frac{A - B}{A} \bullet 100 \tag{1}$$

The Dice Similarity Coefficient (DSC) measured the overlap between A and B, it ranges from 0 (no overlap) to 1 (complete overlap):

$$DSC = \frac{2 \bullet |A \cap B|}{|A| + |B|} \tag{2}$$

The surface DSC (sDSC) was originally introduced by Nikolov et al. [24] to minimize the volume effect of DSC by providing a measure of the agreement between just the surfaces of two volumes above a distance threshold τ:

$$sDSC = \frac{|S_A \cap S_{B\tau}| + |S_B \cap S_{A\tau}|}{|S_A| + |S_B|} \tag{3}$$

Where S_A and S_B are the surfaces of A and B, respectively, while $S_{A\tau}$ and $S_{B\tau}$ the annuli of S_A and S_B, respectively, with τ as the difference between inner and outer radii. In this study $\tau = 3$ mm. As for the DSC, sDSC ranges from 0 (no overlap) to 1 (complete overlap).

The maximum Hausdorff distance (maxHD) measured the maximum distance from one point of A to the closest pairwise point of B:

$$maxHD(A,B) = \max\{H(A,B), H(B,A)\}, \tag{4}$$

$$H(A,B) = \max_{a \in A}\left\{ \max_{b \in B} \{d(a,b)\} \right\} \tag{5}$$

where *d(a, b)* is the HD in 3D between the point a of A and point b of B. 95HD and 99HD were the 95th and 99th percentile of the HD(A,B) distribution. Values of every HD metric, here reported in mm, tend to 0 for good overlap and increase for poor overlap.

2.3. Treatment planning

Both manual and automated treatment planning were performed with RS (v12A). The simulation CT resolution was 1x1x2 mm^3. A VMAT technique involving two reversed 6MV flattening-filter-free partial arcs was employed. For each case, both approaches shared the same treatment machine (Synergy C-arm linac equipped with Agility MLC, Elekta AB, Stockholm, Sweden) initial and final gantry angles (varying according to the patient geometry), arc span (ranging from 210° to 230°), collimator angles (5° and 355°, per arc, respectively), maximum delivery time per arc (75 s) and number of control points (segment every 3° of gantry spacing). The dose calculation grid (Collapsed Cone Convolution (CCC) algorithm) was 3x3x3 mm^3. The prescription dose was normalised to the median volume of PTV1 and plans were optimized and evaluated according to the dose-volume criteria listed in Table 1 of Supplementary Material.

Auto-planning was carried out using a DL technique based on the U-net convolutional neural network (CNN) [25]. The model was trained with 80 plans manually optimized according to the criteria listed in Table 1 of Supplementary Material. Its validation and clinical implementation are reported elsewhere [26]. The dose prediction was mimicked involving three intermediate CCC dose calculations: two over the course and one at the end of the 180 dose iterations. For each case, two auto-plans were generated by mapping the ROI required by the model to the manual and the automatic structures (see Fig. 1 of Supplementary Material for details), respectively, without any post-mimicking additional optimization.

The manual planning process started by retrieving the clinical plan with its original optimization settings. Plans that did not encompass all the mapped structures in the clinical optimization were discarded. The original optimization settings were not uniform among plans due to the subjective choice of the planner. They were used to start a new optimization and applied in turn to both manual and automatic structures to generate the manuals plans for comparison. For consistency with the auto-plans, 180 iterations were used for the optimization and no post-optimization was allowed. This approach prevented any bias related to the optimization process but, at the same time, preserved the inter-plan variability of the plan-specific set of optimization parameters.

Patients selected for this study were not involved in the training nor in the validation of the auto-plan model.

2.4. Dose comparison and dose difference evaluation

Dose-volume results from planning were extracted for every figure of merit listed in Table 1 and reported hereinafter with the nomenclature $D^{planning\ approach}_{optimization\ structures|evaluation\ structures}$. The term *planning approach* denoted the planning method using for the plan optimization, either manual (M) or automatic (A). *Optimization structures* were the structures used for the plan optimization and the *evaluation structures* were the structures used for dose-volume output, either manual (m) or automatic (a). Dose comparison was carried out for $D^{A}_{a|a}$ vs $D^{M}_{a|a}$ and $D^{A}_{m|m}$ vs $D^{M}_{m|m}$ for OARs, and $D^{A}_{a|m}$ vs $D^{M}_{a|m}$ and $D^{A}_{m|m}$ vs $D^{M}_{m|m}$ for PTVs, respectively. Paired Wilcoxon signed-rank tests were performed to assess statistically significant differences ($p < 0.05$) between planning approaches.

In addition, we defined two different dose-volume differences with the goal to assess the dosimetric impact of contouring variation when plans optimized on auto-contoured OARs were simply recalculated or fully reoptimized on the manual structure set as follows:

$$\Delta D^{M}_{rc} = D^{M}_{a|a} - D^{M}_{a|m}, \ \Delta D^{A}_{rc} = D^{A}_{a|a} - D^{A}_{a|m} \tag{6}$$

$$\Delta D^{M}_{ro} = D^{M}_{m|m} - D^{M}_{a|m}, \ \Delta D^{A}_{ro} = D^{A}_{m|m} - D^{A}_{a|m} \tag{7}$$

On the one hand, ΔD^{M}_{rc} and ΔD^{A}_{rc} represent the recalculated dose-volume differences for the manual and automatic planning approach, respec-

tively. They are defined as the difference between automatic and manual structure sets for a plan optimized with the automatic structure set. Given the previously optimized plan, such differences largely depend upon the geometric differences between manual and automatic structure sets. On the other hand, ΔD_{ro}^{M} and ΔD_{ro}^{A} indicate the reoptimized dose-volume difference for the manual and automatic planning approach, respectively. They are defined as the difference between two plans optimized on manual and automatic structure set, respectively, and are calculated according to manual structures. In this context, $D_{a|m}^{M}$ and $D_{a|m}^{A}$ correspond to the manual structure DVH output for the manual and automatic planning approach, respectively. These outputs result from a plan optimized according to the auto-contoured structures. Paired Wilcoxon signed-rank and Levene's tests were performed to assess significant differences ($p < 0.05$) between dose differences reported in Eqs. (6) and (7) median values and inter-quartile ranges (IQR), respectively.

2.5. Correlation analysis

The correlations between the previously defined geometric metrics and both dose differences reported in Equations (6) and (7) were assessed using Spearman's correlation test and evaluated with the Spearman's correlation coefficient R_S.

3. Results

3.1. Geometric metrics for OARs

For each OAR involved, scatter plots of volumes are shown in Fig. 2 of Supplementary Material and volumes similarity is reported in terms of mean values of geometric metrics in Table 1.

In terms of median values, auto-contours were found to be systematically smaller than manual contours for right breast (−4.7 %) and heart (−4.2 %) and more than twice larger for LAD. No systematic differences were observed for both lungs.

The largest variation between 95HD and HD were observed for both lungs due to the positional differences of contours lying in the mediastinal region representing a small part (<5%) of the total lung volumes. For the right breast this variation was large too, reflecting that the worst agreement occurred between contours in the medial and lateral side of the breast. HD indexes did not vary consistently for the heart indicating a minor positional variation between contours. The highest value of 95HD was found for LAD suggesting a general mismatch between auto- and manual contours.

DSC values were satisfactory ≥ 0.90 for all the OARs except the LAD. The sDSC was found to be at least equal or smaller than DSC as expected for large volumes, while it was found to be almost twice than DSC for LAD. The increment of sDSC with respect to DSC was due to the small value of the volume/surface ratio for this specific narrow and elongated organ for which the volume effect became negligible compared to an increase of 3 mm surface.

3.2. Dose comparison and dose differences

The objectives listed in Table 1 of Supplementary Material were on average fulfilled for both planning approaches except for the mean dose of the heart that was found to be slightly above the criteria (+0.1 Gy) for auto-plans.

Dose statistics are reported in Table 2. Regardless of the structure set used for optimization, no significant differences were observed for PTV1 and PTV2 coverage between auto- and manual plans, avoiding any bias in the comparison due to different target coverage. Overall, auto-plans returned significantly less dose to the left lung (−1.4 Gy and −1.1 Gy of mean dose, for manual and automated contouring, respectively), and right breast (−0.7 Gy and −-0.4 Gy of maximum dose, for manual and automated contouring, respectively) and more dose to the heart (+0.1 Gy of mean dose, independently from the planning approach) than manual plans. The homogeneity of PTV2-PTV1 volume was superior for the auto-plans as well. No significant differences were observed for right lung and LAD. OARs dose differences between auto- and manual plans were similar and independent from the contouring approach used.

Results for $\Delta D_{rc}^{A,M}$ and $\Delta D_{ro}^{A,M}$ are reported in Table 3. Regardless of the planning approach, recalculated dose differences were found to be similar and negligible except for D1% for both the heart (−0.4 Gy and −0.3 Gy for ΔD_{rc}^{M} and ΔD_{rc}^{A}, respectively) and the right breast (−1.2 Gy and −1.3 Gy for ΔD_{rc}^{M} and ΔD_{rc}^{A}, respectively). Negatives values of $\Delta D_{rc}^{A,M}$ were due to the position of the manual contours closer to dose gradient regions than the automatic contours for which the dose was originally optimized.

Although ΔD_{ro}^{A} was found to be significantly different than ΔD_{rc}^{A} for the mean dose for both lungs, the re-optimization process did not have any clinical impact if compared to the recalculated dose difference (few additional cGy for both lungs mean dose). Nonetheless, for the heart and right breast, $\Delta D_{ro}^{A,M}$ was found to be higher and closer to 0 than $\Delta D_{rc}^{A,M}$ being significant in the heart only and in both OARs for the manual and auto-planning approach, respectively. For these OARs, $D_{m|m}^{A,M}$ was higher than $D_{a|a}^{A,M}$ in particular for D1%, which suggests a smaller gap between manual OARs and PTVs. In any case, the reoptimization was effective in reducing the dose for these OARs lying adjacent to the high dose gradient. For LAD, ΔD_{ro}^{A} was significantly smaller than ΔD_{rc}^{A} while no significant difference was observed for manual planning.

When comparing re-optimized dose differences, there were minor discrepancies between ΔD_{ro}^{M} and ΔD_{ro}^{A} median values. Instead, we noted large differences in the distribution dispersions. As reported in Table 4, the inter-quartile range (IQR) and range for ΔD_{ro}^{A} were smaller than ΔD_{ro}^{M} for all OARs except the right lung and this difference was found to be significant for the left lung where IQR and range for ΔD_{ro}^{A} halved the ΔD_{ro}^{M} values. ΔD_{ro}^{M} vs ΔD_{ro}^{A} comparison is reported in Supplementary Material (Fig. 3) with boxplots for every figure of merit.

Table 1
Analytics of volumes and similarity geometric metrics. * For LAD, the volume difference units are in [cm3].

OAR		Volumes		Similarity metrics					
		manual	automatic	ΔV	95HD	99HD	HD	DSC	sDSC
		[cm^3]		[%]*		[mm]		–	
Right Breast	median	552	492	−4.7	6.3	11.7	15.5	0.90	0.90
	IQR	244	278	8.6	4.3	5.0	5.9	0.02	0.06
LAD	median	1.2	3.4	1.9*	18.2	23.1	25.2	0.39	0.73
	IQR	0.7	1.5	1.7*	14.6	16.4	16.1	0.12	0.09
Heart	median	581	561	−4.2	6.8	8.0	11.0	0.94	0.86
	IQR	187	166	3.6	2.7	3.5	3.9	0.02	0.04
Right Lung	median	2625	2619	−0.3	2.0	7.9	24.3	0.98	0.98
	IQR	376	435	3.1	0.8	9.5	14.0	0.01	0.01
Left Lung	median	2317	2328	0.2	2.1	7.2	23.5	0.98	0.98
	IQR	350	351	4.9	0.4	6.4	15.1	0.01	0.01

Table 2
Plan comparison between manual (M) and auto (A)-plans optimized with manual ($D^{M}_{m|m}$ vs $D^{A}_{m|m}$) and automatic ($D^{M}_{a|a}$ vs $D^{A}_{a|a}$) −based OARs contouring. The two rightmost columns report $D^{M}_{a|a}$ and $D^{A}_{a|a}$ evaluated on the manual structure set ($D^{M}_{a|m}$ and $D^{A}_{a|m}$). Significant p-values (<0.05) are reported in bold.

ROI			$D^{M}_{m\|m}$	$D^{A}_{m\|m}$	*p-value*	$D^{M}_{a\|a}$	$D^{A}_{a\|a}$	*p-value*	$D^{M}_{a\|m}$	$D^{A}_{a\|m}$	*p-value*
			Median (IQR)			Median (IQR)			Median (IQR)		
PTV1	D98%	Gy	57.2 (0.7)	57.2 (0.2)	*0.342*	NA	NA	*NA*	57.3 (0.7)	57.2 (0.2)	*0.503*
	D2%	Gy	61.7 (0.4)	61.5 (0.5)	*0.27*	NA	NA	*NA*	61.6 (0.7)	61.5 (0.5)	*0.447*
	V57Gy	%	98.5 (1.2)	98.6 (0.6)	*1*	NA	NA	*NA*	98.6 (1.4)	98.7 (0.5)	*0.794*
	V63Gy	%	0	0	*NA*	NA	NA	*NA*	0	0	*NA*
PTV2	D96%	Gy	46.0 (0.9)	46.4 (0.8)	*0.055*	NA	NA	*NA*	46.2 (1.0)	46.5 (0.7)	*0.064*
	V47.5 Gy	%	95.7 (1.5)	96.4 (1.2)	*0.054*	NA	NA	*NA*	96.0 (1.6)	96.4 (1.1)	*0.144*
PTV2-PTV1	D1%	Gy	58.9 (0.7)	58.0 (0.5)	***<0.001***	NA	NA	*NA*	59.0 (0.6)	58.0 (0.6)	***<0.001***
	V52.5 Gy	%	11.1 (3.6)	10.8 (3.5)	***0.039***	NA	NA	*NA*	12.2 (3.6)	10.6 (3.0)	***0.009***
Left Lung	V5	%	30.6 (9.3)	24.8 (2.8)	***<0.001***	31.3 (9.9)	24.9 (3.4)	*<0.001*	31.6 (9.5)	24.9 (3.0	***<0.001***
	V10	%	18.5 (6.9)	14.3 (2.6)	***<0.001***	18.9 (6.6)	14.5 (2.8)	*<0.001*	19.0 (6.9)	14.6 (2.8)	***<0.001***
	V20	%	9.9 (4.5)	8.0 (2.9)	***0.002***	10.0 (4.3)	7.7 (2.6)	*0.004*	10.0 (4.6)	7.8 (2.5)	***0.003***
	V40	%	2.0 (1.6)	1.4 (1.3)	*0.055*	1.7 (1.7)	1.2 (1.3)	*0.011*	1.9 (1.8)	1.4 (1.5)	***0.010***
	Mean	Gy	6.6 (1.9)	5.2 (0.9)	***<0.001***	6.4 (2.1)	5.3 (0.9)	*<0.001*	6.3 (2.1)	5.3 (0.9)	***<0.001***
Right Lung	D1%	Gy	3.9 (2.0)	3.7 (1.4)	*0.81*	3.9 (1.5)	3.4 (1.8)	*0.246*	3.9 (1.6)	3.4 (1.8)	*0.253*
	Mean	Gy	1.0 (0.5)	1.0 (0.3)	*0.764*	1.0 (0.5)	0.9 (0.43)	*0.184*	1.0 (0.5	0.9 (0.4)	*0.189*
Heart	D1%	Gy	4.6 (0.6)	5.0 (1.2)	*0.218*	4.5 (0.7)	4.9 (1.4)	*0.156*	4.9 (0.8)	5.1 (1.5)	*0.164*
	Mean	Gy	1.5 (0.2)	1.6 (0.4)	***0.024***	1.4 (0.2)	1.6 (0.3)	*0.004*	1.4 (0.2)	1.6 (0.3)	***0.033***
Right Breast	D1%	Gy	5.6 (5.4)	4.9 (2.2)	*0.136*	4.6 (5.5)	4.2 (2.0)	*0.007*	5.9 (5.6)	5.6 (2.7)	***<0.001***
	Mean	Gy	1.3 (1.0)	1.2 (0.5)	*0.312*	1.3 (1.1)	1.1 (0.6)	*0.035*	1.4 (1.1)	1.3 (0.6)	***0.028***
LAD	D0.03 cm^3	Gy	5.3 (2.5)	5.5 (1.5)	*0.453*	5.8 (3.0)	6.2 (1.3)	*0.795*	5.6 (3.1)	5.9 (1.6)	*0.655*

Table 3
Dosimetric differences due to the dose recalculation and dose re-optimization for both manual and automatic planning approaches. Significant p-values (<0.05) are reported in bold.

OAR			ΔD^{M}_{rc}	ΔD^{M}_{ro}	p-value	ΔD^{A}_{rc}	ΔD^{A}_{ro}	p-value
			Median (IQR)			Median (IQR)		
Left Lung	V5	%	0.02 (0.27)	0.63 (1.80)	*0.202*	−0.02 (0.20)	0.25 (0.74)	*0.165*
	V10	%	0.01 (0.14)	0.20 (1.29)	*0.756*	−0.03 (0.10)	0.13 (0.78)	*0.186*
	V20	%	−0.03 (0.19)	0.01 (1.17)	*0.784*	−0.03 (0.19)	0.05 (0.16)	*0.118*
	V40	%	−0.02 (0.25)	−0.12 (0.35)	*0.177*	−0.02 (0.25)	0.05 (0.19)	*0.114*
	Mean	Gy	0 (0.07)	0.05 (0.39)	*0.452*	−0.01 (0.08)	0.05 (0.12)	*0.036*
Right Lung	D1%	Gy	0.01 (0.04)	0 (0.34)	*0.985*	0.01 (0.04)	0.22 (0.54)	***0.001***
	Mean	Gy	0.04 (0.01)	−0.01 (0.18)	*0.430*	0.01 (0.01)	0.08 (0.13)	***0.048***
Heart	D1%	Gy	−0.32 (0.45)	−0.18 (0.30)	***< 0.001***	−0.27 (0.18)	−0.09 (0.28)	***0.048***
	Mean	Gy	−0.05 (0.04)	−0.03 (0.12)	***0.040***	−0.04 (0.03)	−0.03 (0.13)	*0.870*
Right Breast	D1%	Gy	−1.18 (0.84)	−0.87 (1.26)	*0.083*	−1.26 (0.59)	−0.34 (0.62)	***< 0.001***
	Mean	Gy	−0.13 (0.05)	−0.10 (0.10)	*0.784*	−0.13 (0.06)	−0.04 (0.13)	***0.002***
LAD	D0.03 cm^3	Gy	0.13 (0.72)	−0.09 (0.67)	*0.380*	0.13 (1.01)	−0.22 (0.59)	***0.007***

Table 4
Dispersion analysis for the re-optimized dose difference evaluated for both planning approaches. Paired Wilcoxon signed-rank (PWS) and Levene's test were used to assess statistically significant differences for median values and IQR, respectively. Significant p-values (<0.05) are reported in bold.

OAR			ΔD^{M}_{ro}	ΔD^{A}_{ro}	PWS	ΔD^{M}_{ro}	ΔD^{A}_{ro}	Levene
			Median		*p-value*	IQR (Range)		*p-value*
Left Lung	V5	%	0.63	0.25	*0.388*	1.80 (4.76)	0.74 (2.94)	***0.011***
	V10	%	0.20	0.13	*0.927*	1.29 (4.94)	0.78 (2.24)	***0.011***
	V20	%	0.01	0.05	*0.794*	1.17 (3.35)	0.16 (0.56)	***< 0.001***
	V40	%	−0.12	0.05	*0.053*	0.35 (1.38)	0.19 (0.60)	***0.017***
	Mean	Gy	0.05	0.05	*0.784*	0.39 (1.08)	0.13 (0.44)	***0.017***
Right Lung	D1%	Gy	0.00	0.22	***0.027***	0.34 (1.35)	0.54 (1.02)	*0.403*
	Mean	Gy	−0.01	0.08	*0.202*	0.18 (0.70)	0.13 (0.52)	*0.567*
Heart	D1%	Gy	−0.18	−0.09	*0.841*	0.30 (1.22)	0.28 (1.12)	*0.318*
	Mean	Gy	−0.02	−0.02	*0.674*	0.11 (0.37)	0.13 (0.46)	*0.281*
Breast	D1%	Gy	−0.86	−0.34	*0.069*	1.26 (5.05)	0.62 (4.99)	*0.191*
	Mean	Gy	−0.10	−0.04	***0.032***	0.10 (0.39)	0.13 (0.44)	*0.468*
LAD	D0.03 cm^3	Gy	−0.09	−0.22	*0.153*	0.67 (2.03)	0.59 (1.67)	*0.742*

3.3. Correlation analysis

For both planning approaches, we found moderate ($0.4<|R_S|\leq 0.6$) to strong ($0.6<|R_S|\leq 0.8$) correlations between ΔV and V20Gy, V40Gy and Dmean of the left lung and D1% of the right lung and recalculated dose differences. We observed moderate correlations between various geometric metrics (95HD, 99HD, and sDSC) and ΔD^{A}_{rc} for D1% and Dmean of the right breast and between ΔV, 95HD, 99HD, and DSC and ΔD^{A}_{rc} for D0.03 cc of the LAD. Specifically, a total number of 21 significant correlations were observed between the recalculated dose

difference and various metrics: 8 for manual plans and the remaining 13 for auto-plans. The total number of significant correlations was drastically reduced to 4 when geometric variations were evaluated against the re-optimized dose differences. In fact, we only observed 4 moderate correlations between 95HD of the left lung V20Gy, DSC of the right breast D1% and ΔD_{ro}^{M} and between sDSC of the right breast D1% and ΔD_{ro}^{A}. A moderate correlation found between DSC and the right lung mean dose for both $\Delta D_{rc}^{A,M}$ and ΔD_{ro}^{M} was rejected because of the unexpected trend (due to the negative sign of the correlation) although statistically significant. No other significant correlations were found, as summarized in Fig. 2 and in Table 2 of Supplementary Material.

4. Discussion

It is not uncommon to dosimetrically quantify contour variations due to different segmentation approaches, typically manual vs automatic, through DVH and dose distribution comparisons. Results may vary according to the method employed for the comparison depending on how many structure sets are used to optimize the plan and which one is used to evaluate the plan. In addition, dosimetric differences are subject to optimization strategies adopted during treatment planning and are difficult to avoid in manual planning approaches [12]. The aim of this study was to reduce all these uncertainties by analyzing the dosimetric impact on contour variations for both manual and automated planning strategies and by documenting the discrepancies that might arise from the use of various dose difference quantification methods. Manual plans were optimized with their original clinical settings to preserve the inter-planner variability and compared to auto-plans where subjective variations were not present. This was intentionally done to fully exploit the differences between two planning approaches.

Regardless of the structure set used for optimization, plan approaches provided a similar target coverage but differed in dose to OARs, since the left lung and the right breast were more preserved in auto-plans while the heart was more spared in the manual plans. Inter-planner variability for manual plans was confirmed by the higher SD observed for OARs dose with respect to auto-plans.

Geometric similarity for OARs results were found to be well aligned with the previous studies [6,10] that employed equivalent or similar DL auto-segmentation techniques, except for LAD whose manual contour was observed to be systematically smaller than the automatic contour (but still within the intra-observed variation (IOV) range reported by Almberg et al. [6]). Our manual contour was limited in the extent along the superior-inferior direction with respect to the automatic contour with large volume deviations occurring in anatomical regions placed out of the dose gradient. This was confirmed by the negligible D0.03 cc mean value of the recalculated dose difference for both planning approaches ($\Delta_{rc}^{A,M} = 0.06$ Gy). This example of discordance between (poor) geometric overlap and (good) dosimetric output for LAD was already observed for breast patients treated with 3DCRT followed by electron sequential boost [11].

The relationship between contouring differences and the corresponding dosimetric consequence is a challenging problem. The sensitivity of dose difference depends on the mutual position of the OAR and the dose gradient, which is ultimately plan specific. The irradiation technique, the shape and size of both target and OARs, the mutual position of the OARs and targets, optimization settings and clinical dose-objectives are all factors that influence dose gradients [13,16]. If dose gradients are fixed in dose re-calculation and variable in dose re-optimization, dose differences are then affected by the method employed for their calculation and the magnitude of the contouring differences. As expected, the right breast and the heart were the OARs that returned largest $\Delta_{rc}^{A,M}$ values for D1% as the closest to dose gradient regions. It is worth noting that dose differences were similar between planning approaches, which suggests a consistency of dose gradients extent and direction (but not in relative position) between manual and

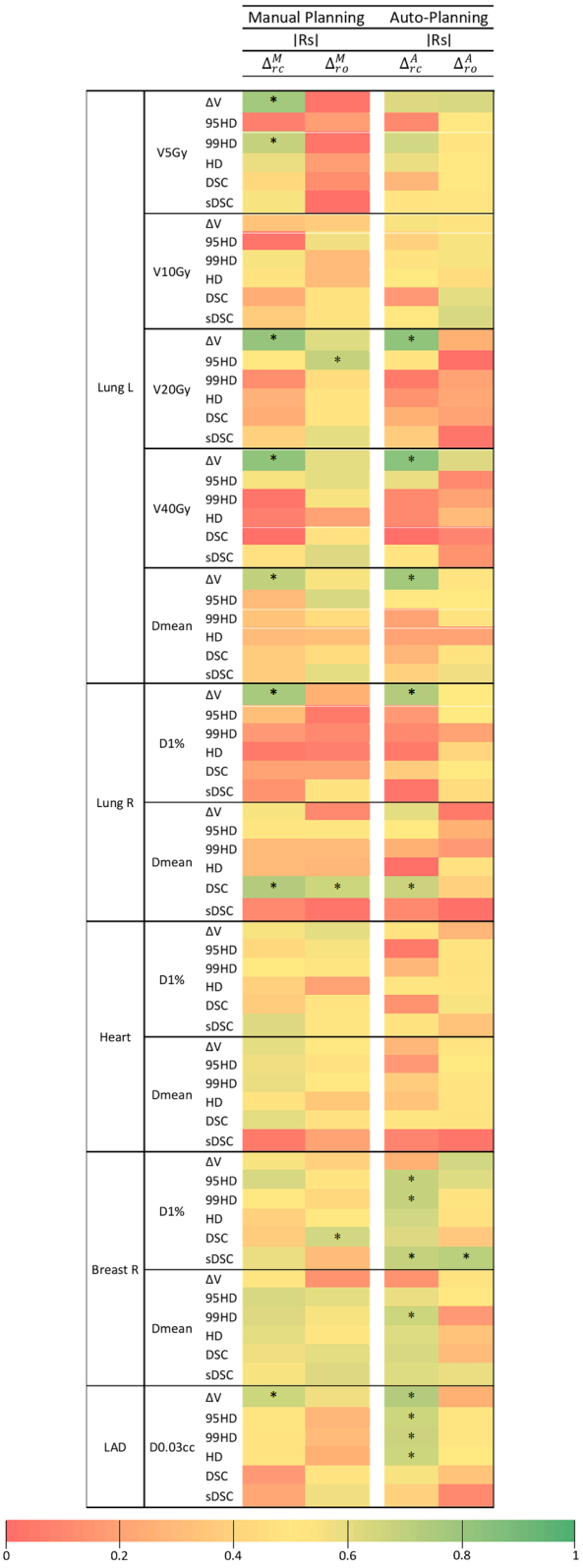

Fig. 2. Spearman correlation between geometrical metrics and dose differences for all OARs dosimetric endpoints. Recalculated and re-optimized dose differences are reported for each planning approach in the left and right column, respectively. Spearman correlation coefficient (|Rs|) is reported using an heatmap visualization in a range varying from 0 to 1. |Rs| values > 0.4 and > 0.6 were considered as moderate and strong correlations, respectively. Stars in the box represent |Rs| values as statistically significant ($p < 0.05$). Regardless of the planning approach, the number of correlations dropped sensibly when moving from recalculated to re-optimized dose differences.

auto-planning.

When plans were re-optimized, maximum dose differences increased for the heart and the right breast (D1%) and decreased for LAD (D0.03 cm^3). This was due to the difference in the optimization volumes, larger for the heart and the right breast and lower for LAD (see Fig. 1). However, we observed only minor deviations for both lungs. In general, $\Delta_{rc}^{A,M}$ was more sensitive than $\Delta_{ro}^{A,M}$ to the maximum dose difference for the OARs lying in proximity to the dose gradient, while for the same OARs both metrics returned similar values for the mean dose difference estimation.

Overall, differences between median values of Δ_{ro}^{M} and Δ_{ro}^{A} were not statistically significant or clinically relevant. However, it is interesting to note that both the IQR and range were reduced for Δ_{ro}^{A} suggesting a minor sensitivity to OARs contouring variation for auto-plans. The presence of multiple outliers for Δ_{ro}^{M} might be a consequence of the different optimization strategies adopted in manual planning including the choice of optimization functions, the relative weights assigned to objectives, and the use of optimization structures. According to our results, inter-planner optimization variability had no consequences on the mean dose differences due to contouring variation when compared to auto-planning.

This study also aimed to demonstrate how the correlation between similarity and dose metrics was highly dependent on the method used to calculate the dose difference. The strong correlation observed between ΔV and $\Delta D_{rc}^{A,M}$ for the left lung V20Gy and V40Gy disappeared when $\Delta D_{ro}^{A,M}$ was considered instead, and in general for all OARs the degree of correlation was sensibly reduced when passing from dose recalculation to dose re-optimization. This was an expected result of the study, which aimed to suggest caution when interpreting such correlations. For other treatment sites, it was demonstrated that contour variation had low significant impact on the corresponding dose evaluation metrics [12,14] and similar results were also found more recently for breast cancer [18]. In all cases, plans were re-optimized. Our findings confirmed the low degree of correlation between contour and dose variation for the evaluated OARs in breast treatments as only two moderate correlations were found for the left lung V20Gy and 95HD, and for the right breast D1% and sDSC. Although the right lung ΔDmean and its DSC was significantly correlated, we considered this a statistical random error since the trend was not as expected. Nonetheless, no significant difference was observed between auto- and manual plans although auto-planning approach provided a smaller correlation coefficient overall. The reason for such a weak correlation may be because the dose gradients produced by the VMAT technique employed in this study that were steep enough around the target to reduce any dosimetry difference due to the observed contouring variation whichever the planning approach used.

As this study represents the first attempt to evaluate the dosimetric impact of auto-planning versus manual planning on contouring variation, it has some limitations. First, the auto-segmentation of the structures was limited to OARs only, thus excluding the target volumes. This reduced the interplay effect between target and OARs contouring variation especially in the regions where the gap between normal structures and PTV was reduced (e.g., medial side of breasts, chest wall) simplifying the whole problem. Secondly, the plans were optimized using the whole OARs structure set (either manual or automatic). This could have reduced the real dosimetric impact of every single OARs as the resulting dose was affected by all other OARs. Although for this specific scenario the magnitude of the observed dose differences were not clinically relevant, it should be pointed out that the interplay effect between OARs may have a strong impact in evaluating dose differences when large deviations are observed. In addition, the results presented are valid for the specific formulation of the dose difference, the irradiation technique employed and, more importantly, for DIBH conditions only. The position of the heart under DIBH is more advantageous with respect to free breathing by reducing the dosimetric trade-offs with the PTVs. Finally, the manual structures taken as reference were delineated by a single radiation oncologist only: using multiple operators for the manual delineation would have produced more realistic geometric similarity results.

5. Conclusion

No clinically relevant dose differences due to contouring variation were observed when automated or manual planning approaches were used for early-breast cancer treatment planning under DIBH conditions. Regardless of the planning approach, the correlation between the dose differences due to contouring variation and geometric metrics was strongly affected by the method employed to calculate the dose differences. It is recommended to reoptimize the plans instead of recalculating them on the candidate structure set to carry out a more realistic scenario in terms of dose difference in the high dose gradient and correlation between contouring variation and dose difference.

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.ejmp.2024.103402.

References

[1] Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in auto-segmentation. Semin Radiat Oncol 2019;29(3):185–97. https://doi.org/10.1016/j.semradonc.2019.02.001.
[2] Wang M, Zhang Q, Lam S, Cai J, Yang R. A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning. Front Oncol 2020;10:580919. https://doi.org/10.3389/fonc.2020.580919.
[3] Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, et al. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. Phys Imag Radiat Oncol 2020;16:144–8. https://doi.org/10.1016/j.phro.2020.11.002.
[4] Vandewinckele L, Claessens M, Dinkla A, Brouwer C, Crijns W, et al. Overview of artificial intelligence-based applications in radiotherapy: Recommendations for implementation and quality assurance. Radiother Oncol 2020:156–66. https://doi.org/10.1016/j.radonc.2020.09.008.
[5] Radici L, Ferrario S, Casanova Borca V, Cante D, Paolini M, et al. Implementation of a commercial deep learning-based auto segmentation software in radiotherapy: Evaluation of effectiveness and impact on workflow. Life 2022;12:2088. https://doi.org/10.3390/life12122088.
[6] Almberg SS, Lervag C, Frengen J, Eidem M, Mikhailovna Abramova T, 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.
[7] Savenije MHF, Maspero M, Sikkes GG, van der Voort van Zyp JPN, Kotte ANTJ, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiat Oncol 2020;15:104. https://doi.org/10.1186/s13014-020-01528-0.
[8] Wong J, Fong A, McVicar N, Smith S, Giambattista J, 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.
[9] van Dijk LV, van den Bosch L, Aljabar P, Peressutti D, Both S, et al. Improving automatic delineation for head and neck organs at risk by Deep Learning Contouring. Radiother Oncol 2020;142:115–23. https://doi.org/10.1016/j.radonc.2019.09.022.
[10] Chung SY, Chang JS, Choi MS, Chang Y, Choi BS, 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;16:44. https://doi.org/10.1186/s13014-021-01771-z.
[11] Kaderka R, Gillespie EF, Mundt RC, Bryant AK, Sanudo-Thomas CB, et al. Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. Radiother Oncol 2019;131:215–20. https://doi.org/10.1016/j.radonc.2018.07.013.
[12] Guo H, Wang J, Xia X, Zhong Y, Peng J, et al. The dosimetric impact of deep learning-based auto-segmentation of organs at risk on nasopharyngeal and rectal cancer. Radiat Oncol 2021;16:113. https://doi.org/10.1186/s13014-021-01837-y.
[13] Fung NTC, Hung WM, Sze CK, Lee MCH, Ng WT. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: Time, geometrical, and

dosimetric analysis. Med Dosim 2020;45:60–5. https://doi.org/10.1016/j.meddos.2019.06.002.
[14] van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, et al. Deep learning-based delineation of head and neck organs at risk: Geometric and dosimetric evaluation. Int J Radiation Oncol Biol Phys 2019;104:677–84. https://doi.org/10.1016/j.ijrobp.2019.02.040.
[15] Kawula M, Purice D, Li M, Vivar G, Ahmadi SA, Parodi K, et al. Dosimetric impact of deep learning-based CT auto-segmentation on radiation therapy treatment planning for prostate cancer. Radiat Oncol 2022;17:21. https://doi.org/10.1186/s13014-022-01985-9.
[16] Cao M, Stiehl B, Yu VY, Sheng K, Kishan AU, et al. Analysis of geometric performance and dosimetric impact of using automatic contour segmentation for radiotherapy planning. Front Oncol 2020;10:1762. https://doi.org/10.3389/fonc.2020.01762.
[17] Simoes R, Wortel G, Wiersma TG, Janssen TM, van der Heide UA, et al. Geometric and dosimetric evaluation of breast target volume auto-contouring. Phys imaging Radiat Oncol 2019;12:38–43. https://doi.org/10.1016/j.phro.2019.11.003.
[18] Zhong Y, Guo Y, Fang Y, Wu Z, Wang J, et al. Geometric and dosimetric evaluation of deep learning based auto-segmentation for clinical target volume on breast cancer. J Appl Clin Med Phys 2023:e13951.
[19] Feng M, Valdes G, Dixit N, Solberg TD. Machine learning in radiation oncology: Opportunities, requirements, and needs. Front Oncol 2018;8:110. https://doi.org/10.3389/fonc.2018.00110.
[20] Pillai M, Adapa K, Das SK, Mazut L, Dooley J, et al. Using artificial intelligence to improce the quality and safety of radiation therapy. J Am Coll Radiol 2019;16:1267–72. https://doi.org/10.1016/j.jacr.2019.06.001.
[21] Jarrett D, Stride E, Vallis K, Gooding MJ. Applications and limitations of machine learning in radiation oncology. Br J Radiol 2019;92:20190001. https://doi.org/10.1259/bjr.20190001.
[22] Offersen BV, Boersma LJ, Hol KCS, Aznar MC, 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.
[23] Nielsen MH, Berg M, Pedersen AN, Andersen K, Glavicic V, 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;52:703–10. https://doi.org/10.3109/0284186X.2013.765064.
[24] Nikolov S, Blackwell S, Zverovitch A, Mendes R, Livne M, 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.
[25] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W., Frangi, A. (eds) Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015. MICCAI 2015. Lecture Notes in Computer Science, vol 9351. Springer, Cham. https://doi.org/10.1007/978-3-319-24574-4_28.
[26] Zeverino M, Piccolo C, Wuethrich D, Jeanneret Sozzi W, Marguet M, et al. Clinical implementation of deep learning-based automated left breast simultaneous integrated boost radiotherapy treatment planning. Phys Imaging Radiat Oncol 2023 Sep;20(28):100492. https://doi.org/10.1016/j.phro.2023.100492.