Atlas-based segmentation in breast cancer radiotherapy: Evaluation of specific and generic-purpose atlases

Delia Ciardo [a], Marianna Alessandra Gerardi [a], Sabrina Vigorito [b], Anna Morra [a], Veronica Dell'acqua [a], Federico Javier Diaz [c], Federica Cattani [b], Paolo Zaffino [d], Rosalinda Ricotti [a], Maria Francesca Spadea [d], Marco Riboldi [e, f], Roberto Orecchia [g, h], Guido Baroni [e, f], Maria Cristina Leonardi [a, *, 1], Barbara Alicja Jereczek-Fossa [a, h, 1]

[a] *Department of Radiation Oncology, European Institute of Oncology, Milan, Italy*
[b] *Unit of Medical Physics, European Institute of Oncology, Milan, Italy*
[c] *Medical Radiation Oncology, Mevaterapia, Buenos Aires, Argentina*
[d] *Department of Experimental and Clinical Medicine, Magna Graecia University, Catanzaro, Italy*
[e] *Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Milan, Italy*
[f] *Bioengineering Unit, National Center for Oncologic Hadrontherapy, CNAO Foundation, Pavia, Italy*
[g] *Scientific Direction, European Institute of Oncology, Milan, Italy*
[h] *Department of Oncology and Hemato-oncology, University of Milan, Milan, Italy*

ABSTRACT

Objectives: Atlas-based automatic segmentation (ABAS) addresses the challenges of accuracy and reli-ability in manual segmentation. We aim to evaluate the contribution of specific-purpose in ABAS of breast cancer (BC) patients with respect to generic-purpose libraries.
Materials and methods: One generic-purpose and 9 specific-purpose libraries, stratified according to type of surgery and size of thorax circumference, were obtained from the computed tomography of 200 BC patients. Keywords about contralateral breast volume and presence of breast expander/prostheses were recorded. ABAS was validated on 47 independent patients, considering manual segmentation from scratch as reference. Five ABAS datasets were obtained, testing single-ABAS and multi-ABAS with simultaneous truth and performance level estimation (STAPLE). Center of mass distance (CMD), average Hausdorff distance (AHD) and Dice similarity coefficient (DSC) between corresponding ABAS and manual structures were evaluated and statistically significant differences between different surgeries, structures and ABAS strategies were investigated.
Results: Statistically significant differences between patients who underwent different surgery were found, with superior results for conservative-surgery group, and between different structures were observed: ABAS of heart, lungs, kidneys and liver was satisfactory (median values: CMD<2 mm, DSC$\geq$0.80, AHD<1.5 mm), whereas chest wall, breast and spinal cord obtained moderate performance (median values: 2 mm $\leq$ CMD<5 mm, 0.60 $\leq$ DSC<0.80, 1.5 mm $\leq$ AHD<4 mm) and esophagus, stomach, brachial plexus and supraclavicular nodes obtained poor performance (median CMD$\geq$5 mm, DSC<0.60, AHD$\geq$4 mm). The application of STAPLE algorithm generally yields higher performance and the use of keywords improves results for breast ABAS.
Conclusion: The homogeneity in the selection of atlases based on multiple anatomical and clinical fea-tures and the use of specific-purpose libraries can improve ABAS performance with respect to generic-purpose libraries.

Abbreviations

RT	radiation therapy
ABAS	atlas-based automatic segmentation
BC	breast cancer
CTV	clinical target volume
CT	computed tomography
SCV	supraclavicular nodes
OAR	organs at risk
VF	vector field
R_{ES}	radius of the equivalent sphere
Vol_{CB}	volume of the contralateral breast
STAPLE	simultaneous truth and performance level estimation
CMD	center of mass distance
AHD	average Hausdorff distance
DSC	Dice similarity coefficient

* Corresponding author. Department of Radiation Oncology, European Institute of Oncology, Via Ripamonti 435, 20141, Milan, Italy.
E-mail address: cristina.leonardi@ieo.it (M.C. Leonardi).
[1] Co-last author.

©2017 This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/
Published Journal Article available at: http://dx.doi.org/10.1016/j.breast.2016.12.010

1. Introduction

Over the last years, the combined improvements in diagnostic imaging, segmentation techniques, dose calculation, dose delivery and quality assurance have increased the accuracy of radiation therapy (RT) [1]. Such developments usually imply increasingly complex and therefore time-consuming processes and an intensified workload for medical and non-medical staff.

As the goal of RT is to irradiate the tumor volume avoiding neighboring organs to prevent acute and late toxicity, one of the key points in the treatment planning process is the segmentation, namely the process of labeling image voxels with anatomical and biological meaningful labels [2]. In particular, the widespread practice of more conformal irradiation techniques, which maximize normal tissue sparing and improve cosmetic outcomes, requires a more accurate and time consuming segmentation of contours [3,4]. Moreover, a large number of organs at risk should be considered to take into account the low-dose bath in intensity-modulated RT and the related – still investigational – late oncogenetic effect [5,6]. Still, manual contouring is a slow time-consuming process, prone to errors and inter- and intra-operator variability, which is difficult to quantify [7–13], and this influences negatively the reliability of dose distributions and therapeutic outcomes comparison between different studies and institutions.

The recent introduction of fully or semi-automatic atlas-based segmentation techniques aims to address the challenges of accuracy and reliability of contouring.

The general impression from previous studies reporting the efficacy of atlas-based automatic segmentation (ABAS) in breast cancer (BC) patients is that the homogeneity among subjects and contours included in the library strongly influences the results of automatic contouring procedure [14]. In this frame, the aim of this study was to evaluate whether specific-purpose libraries featuring a large number of atlases and relying on homogeneity classification based on multiple selected anatomical and clinical features can improve ABAS performance in the segmentation of BC patient with respect to generic-purpose libraries. Performance of single-ABAS and multi-ABAS are also compared.

2. Materials and methods

2.1. Atlas based automatic segmentation

2.1.1. Theoretical overview

In general, an atlas is a model image segmented by an expert operator. In ABAS, a "library" of atlases is collected: the first step consists of selecting a template atlas to serve as reference for the non-rigid registration of all the remaining atlases populating the library, thus obtaining a dataset of deformation vector fields ($VF_{atlas} = \{V_1, V_2, \ldots, V_{n-1}\}$, where n is the number of atlases). When a new unlabeled computed tomography (CT) is given, it is regis-tered to the template atlas and the resulting vector field (VF_{new}) is compared to those saved in the dataset VF_{atlas}. The atlas corresponding to the vector field V_i most similar to VF_{new} is identified as the best matching atlas and its contours are therefore propagated onto the unlabeled CT scan. With other words, the similarity between the unlabeled image with all those available as atlases (or with a specific subset of the library) is calculated and deployed in order to assign a proper label to the voxels of the unlabeled image. In single-ABAS, the atlas that maximizes the similarity index is non-rigidly registered with the new image, and the contours are propagated according to the deformation vector field resulting from the registration. In multi-ABAS, the information from more than one atlas is somehow combined to generate the automatic segmentation. One possible strategy to combine information from different atlases is the Simultaneous Truth and Performance Level Estimation (STAPLE) algorithm [15], which is an expected maximization algorithm that computes a probabilistic estimate of the true segmentation by weighting each segmentation on its estimated performance level. This method is often used as a reference standard segmentation for assessing performance of different algorithms and when an improved segmentation is needed.

A manual refinement is usually required, but with much less efforts with respect to a complete manual contouring from scratch on each CT scan [2].

2.1.2. Description of the implemented ABAS strategies

The creation of the libraries was performed by means of the commercial software suite MIM 6.1.7 (MIMvista Corp., Cleveland, US-OH). One generic-purpose atlas and 9 specific-purpose sub-libraries were created.

At the building of sub-libraries, two main subgroups were stratified as function of the type of surgery, namely post-conservative surgery BC patients (hereinafter referred as "conservative-surgery" group) versus post-mastectomy BC patients (hereinafter referred as "non-conservative-surgery" group). For non-conservative-surgery patients, the side of the tumor was also considered for patient stratification. In order to describe specific-purpose atlases, some anatomical features were selected. Since the most common indicators of body size were not available in our clinical dataset, a surrogate index was derived from thoracic circumference. This was obtained as the radius of the sphere equivalent to the volume of the axial slice at the sub-mammary fold level (R_{ES}). R_{ES} thresholds were identified as the 33rd and 66th percentiles of R_{ES} distribution, corresponding to 6.0 and 6.5 cm, respectively. Three sub-groups were therefore obtained by discriminating between *small* ($R_{ES} \leq 6.0$ cm), *medium* (6.0 cm $< R_{ES} < 6.5$ cm), and *large* size ($R_{ES} \geq 6.5$ cm).

A further stratification was obtained by keywords describing the contra-lateral breast volume (Vol_{CB}) as "*small breast*" (S) and "*large breast*" (L), if it is inferior or superior to the median value of Vol_{CB} distribution (corresponding to 506 cm^3), respectively. Additional categorizing keywords were related to the presence of breast prosthesis (P) or expander (E). Therefore, there are 6 possible combinations of keywords (S, SP, SE, L, LP, LE) that the operator can chose when contouring the CT of a new patient.

2.2. Patients dataset

We collected 200 CT scans of BC patients, treated with adjuvant RT at the European Institute of Oncology (Milan, Italy) between January 2012 and December 2013. All patients gave written

informed consent for scientific research. CT scans were acquired using a GE Light Speed (GE Medical System, Fairfield, US-CT) with voxel resolution of 0.9375 × 0.9375 × 2.5 mm. In cranio-caudal direction, CT usually includes from first cervical vertebra to the second lumbar vertebra. All patients were treated supine, with arms raised above the head with TomoTherapy® (Accuray, Madison, US-WI). At the time of treatment planning, clinical target volumes (CTVs), namely breast/chest wall and supraclavicular nodes (SCV), and organs at risk (OARs) including spinal cord, thyroid, trachea, esophagus, stomach, liver, heart, contra-lateral breast, ipsilateral humeral head, ipsilateral brachial plexus, lungs, kidneys and breast implant (if present), were segmented manually for all patients by expert physicians following the institutional guidelines. Adjuvant RT was administered to the breast (in conservative-surgery patients) or to the thoracic wall with or without SCV area.

2.2.1. Atlas dataset

At the time of creation of the libraries, all contours were reviewed by a dedicated physician. The generic-purpose atlas comprises all 200 patients, whereas the 9 specific-purpose sub-li-braries are as reported in Table 1. An example of the patients selected for the creation of the libraries is shown in Fig. 1. Patients with bilateral breast implants were not included in the sub-libraries.

2.2.2. Validation dataset

The unlabeled CT scans of 47 patients (30 non-conservative-surgery patients and 17 conservative-surgery patients) not included in the libraries were used for testing purpose (Table 2).

For each CT scan, single-ABAS was performed using both the generic- and the specific-purpose libraries, thus obtaining structures sets named Gen and Spec, respectively. Moreover, multi-ABAS was obtained using the STAPLE algorithm to generate a probabi-listic estimate from the best 3 subjects of generic- and specific-purpose libraries, and corresponding structure sets were named GenS and SpecS, respectively. Finally, the single-ABAS using the specific-purpose libraries and the additional keywords resulted in the structure set SpecK. Lastly, the manual segmentation of the unlabeled test CT scan was performed from scratch, thus obtaining the ground truth structure set for ABAS quality assessment.

For the sake of clarity, it is worth summarizing that for each test patient, 6 structure sets were obtained: 1 manually segmented used as reference for the evaluation of ABAS, and 5 automatic ABAS,

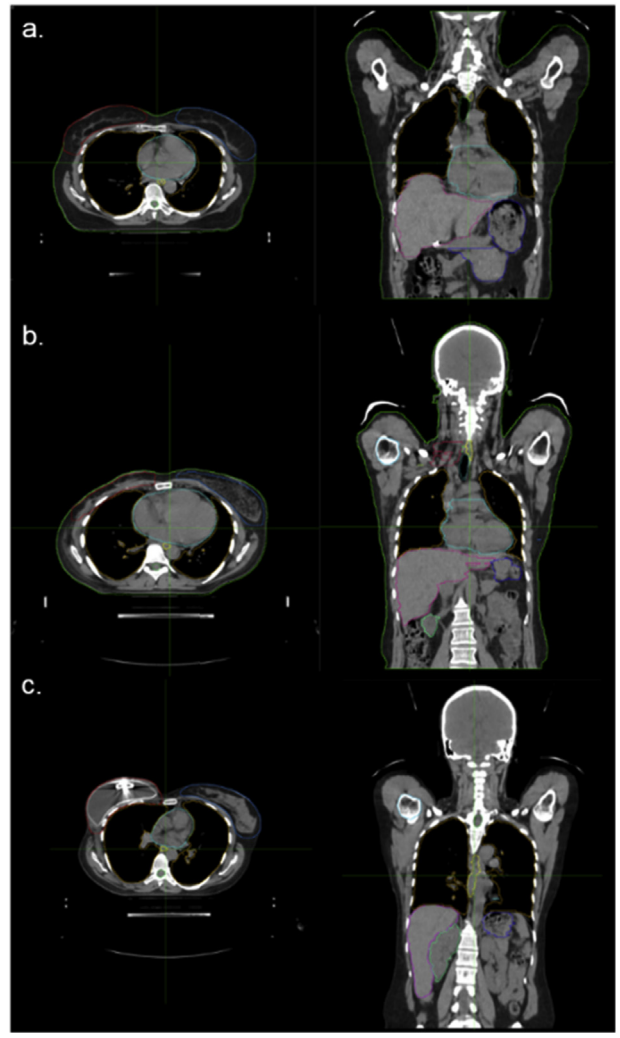

Fig. 1. Paradigms of the left breast cancer patients selected for the libraries creation. Axial and coronal slices of the CT scan of (a) a post-conservative surgery patient, (b) a post-mastectomy non-reconstructed patient and (c) a post-mastectomy surgically reconstructed patient. Segmented structures, such as breasts, heart, esophagus, trachea, liver, kidneys, stomach, lungs, brachial plexus, and supraclavicular nodes are shown.

Table 1

Characteristics of patients collected in the sub-libraries and relative keyword distribution. The column "Number of patients" is the total of patients corresponding to the columns "Surgery", "Tumor side" and "Patient size". Each of the descriptive keywords ("Contralateral breast volume", "Breast prosthesis" and "Breast expander") refers to the corresponding number of patients. Patient size was derived from thoracic circumference: the distribution of the radius of the sphere equivalent to the volume of the axial slice at the sub-mammary fold level was used to distinguish between small size (under 33rd percentile), medium size (between 33rd and 66th percentile), and large size (beyond 66th percentile). The median value of the contralateral breast volume distribution was used to distinguished between small (S) and large (L) contralateral breast volume.

Surgery	Tumor side	Patient size	Number of patients	Keywords					
				Contralateral breast volume		Breast prosthesis		Breast expander	
				S	L	yes	no	yes	no
Non-conservative breast surgery	Right	Small	21	19	2	6	15	10	11
		Medium	18	9	9	7	11	7	11
		Large	15	3	12	3	12	5	10
	Left	Small	21	20	1	3	18	12	9
		Medium	19	6	13	6	13	8	11
		Large	16	3	13	4	12	2	14
Conservative breast surgery	–	Small	19	12	7	–	–	–	–
		Medium	29	10	19	–	–	–	–
		Large	20	1	19	–	–	–	–
Patients not included in the sub-atlases			22	–	–	–	–	–	–

Table 2
Distribution of test patients. Considering the radius of the sphere equivalent to the volume of the axial slice at the sub-mammary fold level (R_{ES}), patients were classified in small (if $R_{ES} \leq 6.0$ cm), medium (6.0 cm < R_{ES} < 6.5 cm) and large size ($R_{ES} \geq 6.5$ cm). Considering the contralateral breast volume, patients were divided in small (S) and large (L). The number of patients with breast prosthesis and expander is also indicated.

Surgery	Tumor side	Patient size	Contralateral breast volume	Number of test patients	Breast prosthesis	Breast expander
Non-conservative breast surgery	Right	Small	S	3	2	1
			L	1	1	0
		Medium	S	2	0	1
			L	3	1	1
		Large	S	2	0	0
			L	3	1	1
	Left	Small	S	3	2	1
			L	3	0	2
		Medium	S	3	1	2
			L	3	2	1
		Large	S	1	0	0
			L	3	1	1
Conservative breast surgery	–	Small	S	3	–	–
			L	3	–	–
		Medium	S	3	–	–
			L	3	–	–
		Large	S	2	–	–
			L	3	–	–

2 of which obtained by applying the generic-purpose library (Gen and GenS) and 3 by applying the specific-purpose libraries (Spec, SpecS and SpecK).

For conservative-surgery patients, a structure set including heart, esophagus, liver, breasts, spinal cord, lungs, left or right kidney and stomach was considered. For the purpose of this study, CTV and contra-lateral breast were not distinguished: we will refer to them as right or left breast. For non-conservative-surgery patients, CTV (namely chest wall) and contralateral breast (distinguished in left/right), brachial plexus and SCV ipsilateral to the operated breast were included in the structure set. The thyroid was excluded from the evaluation, since it was outlined for only few patients included in the atlases.

2.3. Comparisons, metrics and statistics

According to Peroni et al. [16], a set of quantitative metrics related to position and shape of manually versus automatically segmented structure sets were used to assess the performance of ABAS.

The Euclidean distance between the centers of mass (CMD) of corresponding structures was used as an index of differences in the position. Shape discrepancies were expressed as contour distance and contour overlap between corresponding structures. Average Hausdorff distance (AHD) was chosen as measure of contour distance [17] and was computed relying on the implementation provided by the Insight Toolkit (ITK, www.itk.org, [18]). The value of AHD is equal to 0 if the contours coincide perfectly. Contour overlap was expressed by the Dice Similarity Coefficient (DSC) [19]. It is computed as the intersection volume of two structures normalized for the sum of their volumes, in a scale from 0 (no overlap) to 1 (perfect overlap). Since structures of different size are evaluated and a dependency of the DSC from structure volumes has been demonstrated [20], we put the considered metrics into relation with the volumes of the reference structure set in order to give a more accurate estimation of performance scoring.

The normality of indices distributions was assessed applying the Jarque-Bera test and the non-parametric Kruskal-Wallis test was used to perform statistical analysis. The way in which different factors (namely surgical modality, organ and ABAS library) contributed to the variability in the distribution of different metrics (CMD, DSC and AHD) was assessed by Kruskal-Wallis test. In particular, the statistically significant contribution of the factors "surgery" and "structure" is propaedeutic for the subsequent analysis of different atlases, in which the contribution of different surgery and different structures is taken into account. Post-hoc analysis with level of significance set at 0.05 was performed using the Dunn & Sidák's approach to highlight significant differences among the tested conditions. All evaluation and tests were performed in the MATLAB environment (MathWorks, Inc., US-MA).

2.4. Editing time evaluation

On a subset of 8 patients, the time required for delineation of representative structures, basing on the reached level of agreement (i.e. poor, moderate and satisfactory levels), from scratch and by editing the ABAS was recorded. For this purpose, a slice-by-slice delineation was performed with MIM 6.1.7, with the aid of the interpolation between slices in case of delineation breast and heart from scratch.

3. Results

Overall, 235 complete structure sets were obtained using ABAS, and 2650 contour comparisons were performed. The distribution of CMD, AHD and DSC for the different structures and ABAS strategies considered was not normal in 85%, 95% and 60% of the cases, thus non-parametric test was adopted.

3.1. Test on surgical modality

The analysis of CMD, AHD and DSC, considering all structures, showed statistically significant differences between groups of patients who underwent different surgery. In particular, overall statistically significant higher performance of ABAS was obtained for patients who underwent conservative-surgery with respect to the non-conservative-surgery group: median CMD decreases from 4.60 mm to 3.71 mm, median AHD decreases from 1.41 mm to 0.71 mm and median DSC increases from 0.76 to 0.85 (p-value$\ll$ 0.05 in all cases). In Fig. 2, the distributions of CMD, DSC and AHD with corresponding p-values (Kruskal-Wallis test) are shown.

3.2. Test on structures

Considering all the different ABAS strategies as a whole, a large variability was observed for different structures. Statistically

significant differences were observed between structures for all the metrics considered. In particular, post-hoc two by two inter-structure comparisons of CMD, AHD and DSC highlighted statistically significant differences in 60 (76.9%), 62 (79.5%) and 64 (82.1%) of the 78 possible combinations. DSC shows an increasing trend, whereas CMD and AHD show a decreasing trend with increasing structure volume (Fig. 3). For DSC, the logarithmic fitting curve at 95% prediction bounds has $R^2 = 0.66$, whereas the exponential fitting curve at 95% prediction bounds has $R^2 = 0.58$ and $R^2 = 0.69$ for CMD and AHD, respectively. Among all, brachial plexus and stomach have the most discordant behavior from the fitting, with DSC residuals of −0.35 and −0.17, respectively, and AHD residuals of 2.2 mm and 2.9 mm, respectively.

3.3. Test on libraries

The analysis of differences between libraries was performed considering homologous structures and homologous surgery in order to take into account the influence of the previously analyzed factors (see paragraphs 3.1. and 3.2.). Median values and corresponding p-values are reported in Tables 3 and 4 for conservative-surgery and non-conservative-surgery, respectively. Statistically significant differences are highlighted.

3.4. Editing time evaluation

The mean time required for the segmentation from scratch of esophagus, heart and breast was 2′55″, 2′42″ and 8′03″, respectively. The editing from SpecS allowed for a 12%, 41% and 44% time sparing, respectively.

4. Discussion

In this study, we reported a practical application of a commercially available automatic segmentation tool to assist the radiation oncologists in their daily work of segmentation of CTVs and OARs for BC RT. Various studies have been performed about this issue with encouraging results towards the use of ABAS: however, only few examples are available for specific application in BC. The time-sparing deriving from the use of ABAS has been recently demonstrated by Eldesoky et al. [21], who succeeded in validating specific ABAS for breast cancer. In particular, they categorized patients according to surgery and laterality but excluding patients with breast implant and comparable results as concerning structures obtaining high, moderate and poor ABAS performance. Reed et al. [22] showed that ABAS reduces the inter-operator variability when contouring whole breast CTV and the performance got worse as a function of the differences in body mass index between template and test patients. The assessment of intra/inter-observer variability is certainly a key point in segmentation and the introduced ABAS should aim to improve it. As concerning breast area segmentation, Li et al. [12] performed a multi-observer analysis to assess the variability of target and OAR delineation and its dosimetric impact. With due caution, our results can be compared with those obtained in this study. Similar results were obtained in structure overlap for heart, breast and supraclavicular nodes, which are the structures in common between the two studies. Since several studies demonstrated that the inter-observer variability decreases after modifying atlas-based segmented structures [22–24], a manual refinement will potentially further improve the inter-observer reproducibility in segmentation with advantageous effect on the reliability of intra-and inter-institution studies. Anders et al. [25] implemented spe-cific single-subject ABAS stratifying patients according to breast volume. They inferred that the results of ABAS are strongly influ-enced by the shape of the breast rather than by its volume. Un-fortunately, the differences in shape (and thus a possible shape descriptor) retrospectively identified were not described. Velker et al. [26] observed no statistically significant improvements in the segmentation of CTV for post-lumpectomy BC patients when using libraries composed by more than 12 subjects and highlighted the need of consistency and homogeneity in the structure set of atlases. In the study by Van de Velde et al. [27] cadaver scans were used to evaluate the optimal number of atlases for the multi-ABAS of the brachial plexus. Considering 3 atlases and applying the STAPLE algorithm, a moderate mean overlap of about 40% was obtained versus our corresponding median value of 15%. Probably, the good performance is ascribable to the use of magnetic resonance imaging rigidly registered to CT scan (which is highly reliable, due to the stiffness of the embalmed cadavers) that allows increasing con-sistency in the segmentation of such a poorly contrasted structure on CT images. However, best results are obtained when using 9 atlases with the STAPLE algorithm, which suggests that increasing the number of atlases and the number of subjects to include in STAPLE could be worthwhile.

The aim of our work was to determine whether patient stratification based on quantitative or qualitative features could effectively improve ABAS in clinical practice. For the chosen paradigm, namely BC treatment, several anatomical variations can be observed between patients, in terms of post-surgical outcome, anatomical characteristics and build. We decided to consider patients who underwent different surgical procedures, i.e. non-conservative-surgery and conservative-surgery. Within these large groups, we stratified patients according to the tumor side and the thorax size. Information about the breast volume and the presence of breast prosthesis and expander, the latter one partic-ularly evident in the CT scans due to its metal artifacts, were recorded for each atlas included in the sub-libraries. Our work was designed to evaluate the capabilities of the commercial software in

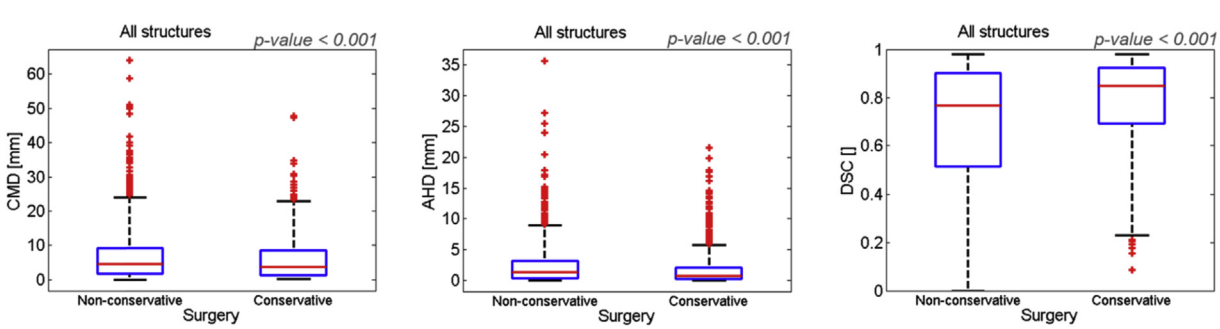

Fig. 2. Non-conservative versus conservative surgery. The boxplot of the distribution for center of mass distance (CMD), Hausdorff distance (AHD) and Dice similarity coefficient (DSC) for all structures with corresponding p-values deriving from Kruskal-Wallis test are shown.

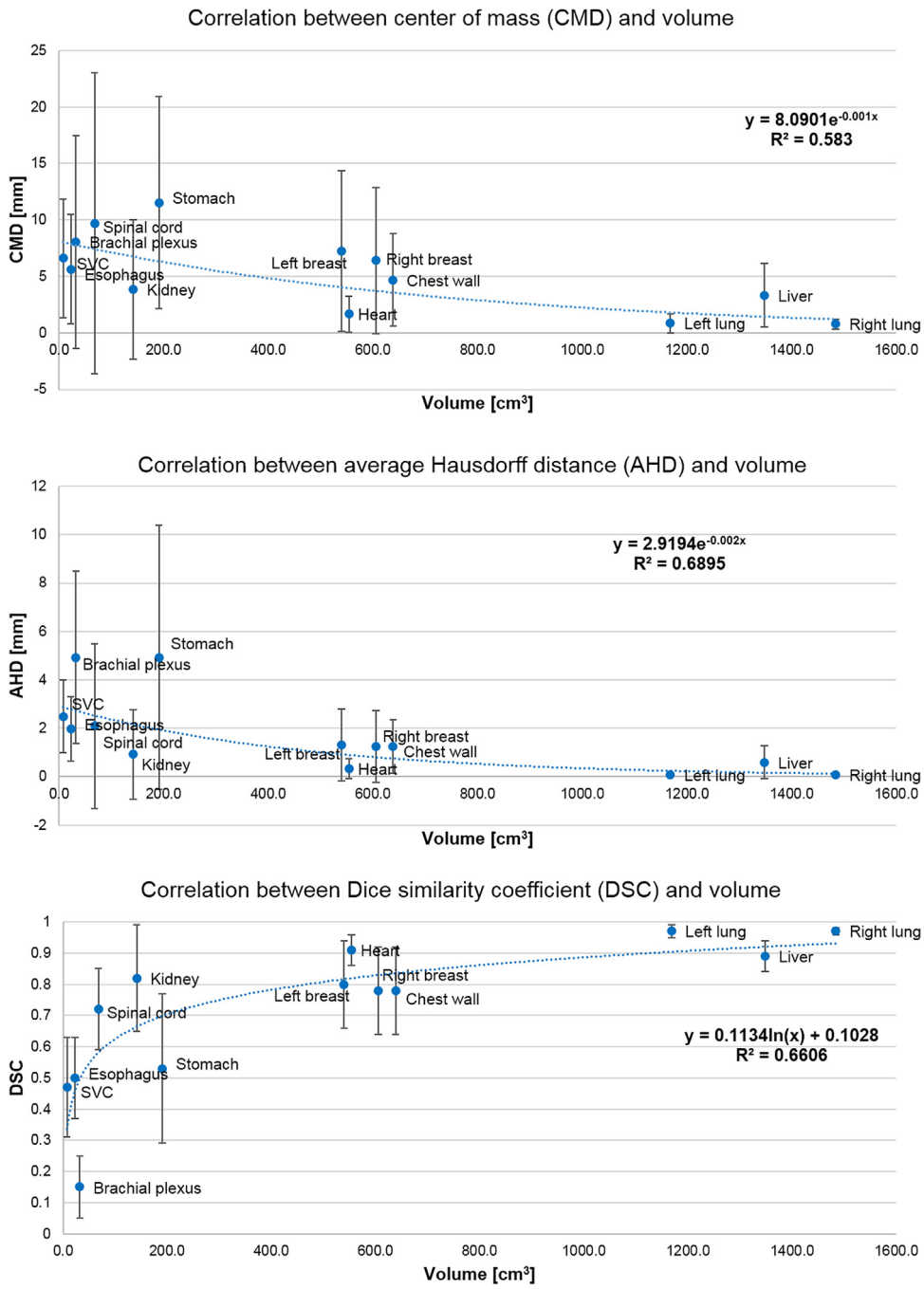

Fig. 3. Correlation between center of mass distance (CMD), Dice similarity coefficient (DSC) and average Hausdorff distance (AHD) versus structure volume. For each graph, median and interquartile range of the corresponding metric are represented in correspondence of median volume. The data comprise all the 5 implemented strategies of ABAS. Brachial plexus, supracluvicular nodes (SCV) and chest wall derive from non-conservative surgery breast cancer patients only. The dotted line is the fitting curve.

a realistic clinical situation, when some a-priori information about the anatomy of the patient is available, such as the pre-treatment surgery and the patient size.

Differences were observed between different typologies of patients: results are generally worse for non-conservative-surgery patients rather than for conservative-surgery patients (even excluding the poor results of brachial plexus and SCV, thus considering corresponding structures only). Again, this is probably due to the higher homogeneity among patients, since our non-conservative-surgery group included both surgically reconstructed and non-reconstructed breast.

Good levels of agreement were obtained for heart, liver, lungs and kidneys, with median DSC≥0.80, CMD<2 mm and AHD<1.5 mm. In particular, in the non-conservative-surgery group, slightly better statistically significant results were obtained when using STAPLE for heart and kidneys. A moderate level of agreement, with median DSC between 0.60 and 0.80, CMD<5 mm and AHD<4 mm, was obtained for spinal cord, chest wall and breasts. Interestingly, statistically significant superior results were obtained for chest wall (segmented for non-conservative-surgery cases) when specific-purpose libraries are used in combination with keywords or STAPLE algorithm, demonstrating a higher accuracy in

Table 3
Median center of mass distance (CMD), average Hausdorff distance (AHD) and Dice similarity coefficient (DSC) at varying organs at risk obtained from different atlases in conservative-surgery patients. Corresponding p-value deriving from Kruskal-Wallis test among atlases is shown. Statistically significant p-values (<0.05) are highlighted in bold. Corresponding higher quality results are highlighted in light grey.

		Conservative-surgery group									
		Heart	Esophagus	Liver	Right breast	Left breast	Spinal Cord	Right lung	Left lung	Kidney	Stomach
CMD (mm)	Gen	1.41	5.24	2.81	8.16	8.84	8.58	0.80	0.69	3.64	9.47
	GenS	1.06	4.65	2.39	6.82	9.11	8.97	0.70	0.58	2.63	13.56
	Spec	1.76	7.61	2.68	6.29	4.43	9.64	0.75	0.83	3.83	11.00
	SpecS	1.32	5.89	3.19	5.98	5.13	5.61	0.81	0.71	3.69	13.56
	SpecK	2.00	7.16	3.49	6.29	3.67	7.97	0.84	0.61	3.83	9.96
	p-value	*0.437*	*0.713*	*0.912*	*0.490*	*0.006*	*0.771*	*0.977*	*0.985*	*0.891*	*0.358*
AHD (mm)	Gen	0.26	1.77	0.51	1.78	1.13	2.22	0.07	0.07	0.82	4.30
	GenS	0.22	1.55	0.53	1.33	2.02	0.99	0.06	0.07	0.46	7.15
	Spec	0.37	2.39	0.43	1.16	0.64	2.09	0.07	0.07	1.08	5.08
	SpecS	0.26	1.53	0.49	0.96	0.62	0.82	0.07	0.06	0.98	7.36
	SpecK	0.38	2.17	0.44	0.89	0.62	1.59	0.07	0.08	1.16	5.75
	p-value	*0.087*	***0.019***	*0.495*	*0.164*	***0.020***	*0.358*	*0.920*	*0.764*	*0.204*	*0.708*
DSC	Gen	0.92	0.53	0.91	0.76	0.80	0.71	0.97	0.97	0.84	0.55
	GenS	0.93	0.57	0.89	0.81	0.79	0.82	0.97	0.97	0.88	0.41
	Spec	0.90	0.48	0.90	0.79	0.82	0.72	0.97	0.97	0.83	0.55
	SpecS	0.92	0.56	0.90	0.81	0.84	0.81	0.97	0.97	0.87	0.45
	SpecK	0.90	0.50	0.90	0.79	0.87	0.75	0.97	0.97	0.84	0.52
	p-value	***0.032***	***0.005***	*0.427*	*0.239*	***0.024***	*0.127*	*0.857*	*0.728*	*0.226*	*0.643*

Table 4
Median center of mass distance (CMD), average Hausdorff distance (AHD) and Dice similarity coefficient (DSC) at varying organs at risk obtained from different atlases in non-conservative-surgery patients. Corresponding p-value deriving from Kruskal-Wallis test among atlases is shown. Statistically significant p-values (<0.05) are highlighted in bold. Corresponding superior results are highlighted in light grey.

		Non-conservative-surgery group												
		Heart	Esophagus	Liver	Right breast	Left breast	Spinal cord	Right lung	Left lung	Kidney	Stomach	Brachial plexus	SCV	Chest wall
CMD (mm)	Gen	2.02	5.79	2.93	6.78	7.24	14.31	0.70	1.18	5.16	10.63	7.78	6.72	9.03
	GenS	1.51	4.58	4.48	4.72	7.98	9.63	0.87	0.82	3.42	10.89	6.69	7.27	6.12
	Spec	1.90	5.79	3.47	5.89	5.74	9.03	0.77	0.90	5.19	11.64	8.34	6.13	4.78
	SpecS	1.40	4.83	3.53	4.67	5.04	10.13	0.87	0.94	2.46	11.49	7.70	5.89	3.14
	SpecK	2.05	5.79	3.16	4.93	4.47	11.75	0.82	0.93	6.83	14.54	8.14	8.04	2.89
	p-value	***0.035***	*0.291*	*0.221*	*0.410*	*0.212*	*0.144*	*0.436*	*0.551*	***0.002***	*0.810*	*0.942*	*0.435*	***0.045***
AHD (mm)	Gen	0.44	2.25	0.61	1.93	1.31	3.00	0.08	0.11	1.39	4.37	4.88	2.15	2.21
	GenS	0.25	1.54	0.86	1.35	2.03	2.04	0.07	0.10	0.66	4.27	4.43	2.00	1.64
	Spec	0.43	2.43	0.62	1.31	0.92	2.46	0.08	0.10	0.96	4.50	5.68	2.77	1.61
	SpecS	0.27	1.53	0.78	1.11	0.62	1.96	0.08	0.09	0.52	4.64	5.30	2.37	1.02
	SpecK	0.47	2.07	0.52	0.90	0.61	3.08	0.08	0.11	1.86	5.55	5.01	2.81	0.91
	p-value	***0.002***	***0.000***	*0.098*	***0.026***	***0.009***	*0.286*	*0.986*	*0.575*	***0.000***	*0.771*	*0.760*	***0.005***	***0.004***
DSC	Gen	0.89	0.47	0.90	0.75	0.79	0.69	0.97	0.96	0.79	0.60	0.17	0.49	0.64
	GenS	0.92	0.53	0.87	0.82	0.77	0.73	0.97	0.97	0.84	0.56	0.16	0.53	0.73
	Spec	0.89	0.45	0.90	0.79	0.83	0.69	0.97	0.96	0.82	0.52	0.14	0.43	0.78
	SpecS	0.92	0.51	0.88	0.81	0.85	0.73	0.97	0.97	0.86	0.52	0.11	0.47	0.82
	SpecK	0.89	0.45	0.90	0.78	0.87	0.70	0.97	0.96	0.76	0.49	0.14	0.43	0.84
	p-value	***0.001***	***0.005***	*0.078*	*0.340*	***0.01***	*0.132*	*0.970*	*0.197*	***0.001***	*0.458*	*0.579*	***0.017***	***<0.001***

atlas selection. Similarly, but to a lesser extent, specific-purpose libraries showed higher performance in breast segmentation. At a qualitative inspection of breast automatic segmentation, the lateral and medial regions, followed by the inferior extent, appeared as the main sources of error. This finding is confirmed by the center of mass misalignment that, for these structures, presents a preferential direction orthogonal to the breast surface. As far as chest wall is concerned, instead, the main source of error is the presence of breast expander that hinders an accurate image registration and subsequent contour propagation due to random metal artifacts. Contour mismatches up to 30 mm can be observed, in some cases. Again, the visual inspection confirms the quantitative outcome, namely that ABAS improves with the use of specific-purpose libraries, in particular if associated with STAPLE multi-atlas combination. As far as spinal cord is concerned, ABAS was influenced by the variability affecting the manual segmentation particularly in cranio-caudal extension between test and atlas subjects. This caused a high median CMD and scattered results (even though overall acceptable), as pointed out by the large interquartile range of the computed parameters. Conversely, the ABAS of esophagus, stomach, SCV and brachial plexus resulted in poor level of agreement (median CMD$\geq$5 mm, median DSC< 0.60, median AHD$\geq$4 mm). The ABAS of SCV suffered from poor contrast of this lymph-nodal structure with respect to surroundings soft tissue. For both SCV and esophagus, higher results, but still insufficient in view of fully automatic segmentation, were obtained for generic-purpose library when STAPLE algorithm was applied. In the ABAS of the stomach, highly spread results were observed, as highlighted by the large inter-quartile range of the calculated parameters. Insufficient results were obtained for the brachial plexus, due to its poor contrast and high variability of arm setup, besides the peculiar lengthened and thin shape of this structure.

Even if a plain superior strategy did not emerge, some indications can be deduced and converted into suggestions for clinical practice. In general, our analysis demonstrated statistically significant superior results when the STAPLE algorithm was used,

since it allows smoothing the individual contribution to the final segmentation result by reducing outliers. Comparing the performance of the different libraries on the same structure, best results were obtained when using specific-purpose atlases combined with STAPLE approach or with keyword selection. This is particularly evident for breast and chest wall, for which the use of descriptive keywords reduces both median and interquartile range (results not shown). As expected, the use of keywords related to characteristics of breast/CTV affected the quality of the segmentation of these structures only. Indeed, the use of specific keywords, reducing the variety of the sample with respect to specific features of interest, limits the largest deviation from reference and leads to a reduction of the spread in the results. Nonetheless, it should be considered that the "over definition" might represent a hindrance both in the identification of a sufficient number of proper patients for the building of the sub-atlases and also for the choice of the best fitting atlas when contouring a new patient, since all the characteristics must be assessed beforehand. It is also conceivable that superior results might be obtained if the sample size of keyword-specific sub-atlases is increased. However, results are not univocal and the application of ABS should be performed carefully. In particular when considering breast, chest wall and SCV, the contour review and editing by the physician is mandatory. As demonstrated by our pilot analysis, usually a post-ABAS editing allows for time sparing with respect to contouring from scratch and a metrics-related dependence, still investigational, was observed: interestingly, higher time saving was observed for structures that obtained medium or superior ABAS results, such as heart and breast. For less reproducible structures, such as brachial plexus and stomach, instead, the use of ABAS is not recommended. This preliminary finding deserves a more in-depth investigation, in order to help the clinicians to define the most strategy in the use of ABAS.

As explained in Valentini et al. [28], the use of contour similarity indices allows for a first level evaluation of the performance and for comparison studies. Thus, the validation of the accuracy through the considered metrics (CMD, AHD and DSC) represents a necessary but not sufficient condition for the introduction of ABAS in clinical practice. It should be considered, for instance, that these metrics suffer from a dependency from structure volumes, as demonstrated by our results and confirmed by Isambert et al. and Zaffino et al. [20,29]. As a further development of this work, the clinical impact of ABAS might be examined in-depth by a comprehensive dosimetric analysis aiming to assess the differences between manual contouring, raw ABAS and manually corrected ABAS. In this sense, it would be interesting to consider the contouring of thyroid, since its precise determination is essential, especially for SCV irradiation with intensity-modulated and volumetric arc irradiation approaches.

However, Voet et al. [30] showed a statistically significant linear correlation between the reduction in target coverage and DSC, even though the large spread prevent from definitive conclusions. The observed discrepancies in breast ABAS caution against the blind introduction of ABAS in clinical practice, since the involved areas are crucial for the dosimetry optimization to neighboring vital organs at risk, such as heart, lungs and spinal cord. Improvements in the accuracy of ABAS of these areas represent one of the possible future perspectives in the implementation of more advanced ABAS software. Of course, for the different districts considered, the challenge remains the choice of anatomical and clinical factors to be considered in atlas building.

Further developments in this field are supported by the continuous efforts performed in computer science to develop new strategies and algorithms for the selection of the best subjects in multi-ABAS [14,31,32]. On the other side, several commercial software are available and new versions are continuously updated. Their clinical use requires a careful evaluation of their performance and a word of caution must be given for their introduction in clinical practice.

Conflict of interest

All the authors declare that there is no actual or potential conflict of interest.

Acknowledgment

This work was partially supported by the research grant from the Associazione Italiana per la Ricerca sul Cancro IG-13218 and by a research grant from Accuray Inc. entitled "Data collection and analysis of Tomotherapy and CyberKnife breast clinical studies, breast physics studies and prostate study". The Sponsors did not play any role in the study design, collection, analysis and interpretation of data, nor in the writing of the manuscript, nor in the decision to submit the manuscript for publication.

References

[1] Poortmans P, Marsiglia H, De las Heras M, Algara M. Clinical and technological transition in breast cancer. Rep Pract Oncol Radiother 2013;18(6):345–52.
[2] Iglesias JE, Sabuncu MR. Multi-atlas segmentation of biomedical images: a survey. Med Image Anal 2015 Aug;24(1):205–19. http://dx.doi.org/10.1016/j.media.2015.06.012.
[3] Miles EA, Clark CH, Urbano MT, Bidmead M, Dearnaley DP, Harrington KJ, et al. The impact of introducing intensity modulated radiotherapy into routine clinical practice. Radiother Oncol 2005;77:241–6.
[4] Hong TS, Tome WA, Harari PM. Heterogeneity in head and. neck IMRT target design and clinical practice. Radiother Oncol 2012;103:92–8.
[5] O'Donnell H, Cooke K, Walsh N, Plowman PN. Early experience of tomotherapy-based intensity-modulated radiotherapy for breast cancer treatment. Clin Oncol 2009;21(4):294–301.
[6] Abo-Madyan Y, Aziz MH, Aly MM, Schneider F, Sperk E, Clausen S, et al. Second cancer risk after 3D-CRT, IMRT and VMAT for breast cancer. Radiother Oncol 2014;110(3):471–6.
[7] Petersen RP, Truong PT, Kader HA, Berthelet E, Lee JC, Hilts ML, et al. Target volume delineation for partial breast radiotherapy planning: clinical characteristics associated with low interobserver concordance. Int J Radiat Oncol Biol Phys 2007;69:41–8.
[8] 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;50:1366–72.
[9] Landis DM, Luo W, Song J, Bellon JR, Punglia RS, Wong JS, et al. Variability among breast radiation oncologists in delineation of the postsurgical lumpectomy cavity. Int J Radiat Oncol Biol Phys 2007;67:1299–308.
[10] Pitkanen MA, Holli KA, Ojala AT, Laippala P. Quality assurance in radiotherapy of breast cancer – variability in planning target volume delineation. Acta Oncol 2001;40:50–5.
[11] Struikmans H, Warlam-Rodenhuis C, Stam T, Stapper G, Tersteeg RJ, Bol GH, et al. Interobserver variability of clinical target volume delineation of glandular breast tissue and of boost volume in tangential breast irradiation. Radiother Oncol 2005;76:293–9.
[12] Li XA, Tai A, Arthur DW, Buchholz TA, Macdonald S, Marks LB, Moran JM, et al., Radiation Therapy Oncology Group Multi-Institutional and Multiobserver Study. Variability of target and normal structure delineation for breast cancer radiotherapy: an RTOG Multi-Institutional and Multiobserver Study. Int J Radiat Oncol Biol Phys 2009;73:944.
[13] Van Mourik AM, Elkhuizen PHM, Minkema D, Duppen JC, van Vliet-Vroegindeweij C. Multiinstitutional study on target volume delineation variation in breast radiotherapy in the presence of guidelines. Radiother Oncol March 2010;2010(94):286–91.
[14] Ramus L, Malandain G. Assessing selection methods in the context of multi-atlas based segmentation. In: Biomedical imaging: from nano to macro, 2010 IEEE international symposium on (pp. 1321–1324). IEEE; April 2010.
[15] Warfield SK, Zou KH, Wells WM. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. IEEE Trans Med Imaging 2004 Jul;23(7):903–21.
[16] Peroni M, Spadea MF, Riboldi M, Falcone S, Vaccaro C, Sharp GC, et al. Validation of automatic contour propagation for 4D treatment planning using multiple metrics. Technol Cancer Res Treat 2013 Dec;12(6):501–10.
[17] Dong-Gyu S, Oh-Kyu K, Rae-Hong P. Object matching algorithms using robust Hausdorff distance measures. IEEE Trans Image Process 1999;8:425–9.
[18] Yoo TS, Ackerman MJ, Lorensen WE, Schroeder W, Chalana V, Aylward S, et al. Design for an image processing API: a technical report on ITK - the Insight Toolkit. Stud Health Technol Inf 2002;85:586–92.

[19] Dice LR. Measures of the amount of ecologic association between species. Ecology 1945;26(3):297–302.
[20] Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau PY, Malandain G, et al. Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. Radiat Oncol 2008;87(1):93–9.
[21] Eldesoky AR, Yates ES, Nyeng TB, Thomsen MS, Nielsen HM, Poortmans P, 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 Sep 30. pii: S0167-8140(16)34303–1.
[22] Reed VK, Woodward WA, Zhang L, Strom EA, Perkins GH, Tereffe W, et al. Automatic segmentation of whole breast using atlas approach and deformable image registration. Int J Radiat Oncol Biol Phys 2009;73(5):1493–500.
[23] Daisne JF, Blumhofer A. Atlas-based automatic segmentation of head and neck organs at risk and nodal target volumes: a clinical validation. Radiat Oncol 2013;8(1):1.
[24] Stapleford LJ, Lawson JD, Perkins C, Edelman S, Davis L, McDonald MW, et al. Evaluation of automatic atlas-based lymph node segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys 2010;77(3):959–66.
[25] 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 Jan;102(1):68–73.
[26] Velker VM, Rodrigues GB, Dinniwell R, Hwee J, Louie AV. Creation of RTOG compliant patient CT-atlases for automated atlas based contouring of local regional breast and high-risk prostate cancers. Radiat Oncol 2013 Jul 25;8:188.
[27] Van de Velde J, Wouters J, Vercauteren T, De Gersem W, Achten E, De Neve W, et al. Optimal number of atlases and label fusion for automatic multi-atlas-based brachial plexus contouring in radiotherapy treatment planning. Radiat Oncol 2016;11(1):1–9.
[28] Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Radiat Oncol 2014;112(3):317–20.
[29] Zaffino P, Ciardo D, Piperno G, Travaini LL, Comi S, Ferrari A, et al. Radiotherapy of hodgkin and non-hodgkin lymphoma a nonrigid image-based registration method for automatic localization of prechemotherapy gross tumor volume. Technol cancer Res Treat 2016;15(2):355–64.
[30] Voet PW, Dirkx ML, Teguh DN, Hoogeman MS, Levendag PC, Heijmen BJ. 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.
[31] Zaffino P, Fritscher K, Peroni M, Spadea MF, Schubert R, Sharp G. OC-0180: atlas selection strategies for multi atlas based segmentation algorithm for head and neck radiotherapy. Radiother Oncol 2014;111:S70–1.
[32] Schreibmann E, Marcus DM, Fox T. Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. J Appl Clin Med Phys 2014 Jul 8;15(4):4468.