Volume 18, No 4 | International Journal of Radiation Research, October 2020

Grading evaluation study of atlas based auto-segmentation of organs at risk in thorax

Y.C. Ying[1,2#], J.F. Cheng[2#], H. Wang[2], H.L. Gu[2], H. Chen[2], Y. Shao[2], Y.H. Duan[2], A.H. Feng[2], X.L. Fu[2], H. Quan[1*], Z.Y. Xu[2*]

[1]Key Laboratory of Artificial Micro- and Nano-structures of Ministry of Education and Center for Electronic Microscopy and Department of Physics, Wuhan University, Wuhan, China
[2]Department of Radiation Oncology, Shanghai Chest Hospital Shanghai Jiao Tong University, Shanghai, China

► ***Original article***

Corresponding authors:
Hong Quan, Ph.D.,
E-mail: *csp6606@sina.com*
Zhiyong Xu, Ph.D.,
E-mail: *xzyong12@vip.sina.com*

Revised: October 2019
Accepted: November 2019

Int. J. Radiat. Res., October 2020; 18(4): 647-656

DOI: 10.18869/acadpub.ijrr.18.4.647

[#]Y.C. Ying and J.F. Cheng contributed equally to this work and should be regarded as co-first authors.

ABSTRACT

Background: The grading evaluation of atlas based auto-segmentation (ABAS) of organs at risk (OARs) in thorax was studied. ***Materials and Methods:*** Forty patients with thoracic cancer were included in this study, and for each thirteen thoracic OARs were delineated by an experienced radiation oncologist. The patients were randomly grouped into the training and the test dataset (20 each). The investigated ABAS strategies included single-atlas (Single), majority voting with 5 atlas matches (MV5) and simultaneous truth and performance level estimation (STAPLE) with 5 atlas matches (ST5). The Dice similarity coefficient (DSC), the difference of the Euclidean distance between centers of mass (ΔCMD), the difference of volume (ΔV), maximum Hausdorff distance (MHD) and average Hausdorff distance (AHD) between auto-segmented and manual contours were calculated. ***Results:*** Most of the index values (33/65) of ST5 were optimal. There were differences in the grading results for the five indexes. With DSC, five, four and four OARs were graded into Level 3, Level 2 and Level 1, respectively. The mean DSC values ranged from 0.88 to 0.96, from 0.73 to 0.79, and from 0.53 to 0.62 for the Level 3, Level 2 and Level 1, respectively. ***Conclusion:*** Grading evaluation of ABAS of thoracic OARs based on the DSC proved to be feasible and relatively more reliable. The thoracic OARs auto-segmentation was divided into three levels based on the DSC. Level 3 OARs can be auto-segmented, Level 2 OARs delineations need to be manually modified after the auto-segmentation, and Level 1 OARs are not recommended for the auto-segmentation.

Keywords: *Atlas based auto-segmentation; grading evaluation; thorax; organs at risk.*

INTRODUCTION

The current radiotherapy strategy is to improve the local control rate of the tumor as much as possible to reduce the possibility of recurrence, while having a quantitative understanding of the radiation dose to organs at risk (OARs) [1], so as to avoid normal tissue complications induced by excess radiation dose, leading to a sharp decline in the patient quality of life. Therefore, accurately delineating the normal tissue contour is one of the important prerequisites for precise radiotherapy.

Although manual delineation is the gold standard for delineating normal tissue contours [2], this work can be time consuming and laborious. Even if manual delineation is conducted according to the guidelines, there are still intra and inter-observer variabilities. These variabilities may affect the evaluation of the quality of radiotherapy plans, which is one of the main sources of error in radiotherapy plans [3]. To overcome these shortcomings, auto-segmentation algorithm has been

developed and start to be widely used. At present, most commercial auto-segmentation software uses atlas-based auto-segmentation (ABAS) algorithm [4].

Auto-segmentation of thoracic OARs has been reported in literatures [5-8]. The OARs of study included the lungs, spinal cord, heart, esophagus and trachea. Great vessel was not taken into account. Eric *et al.* [9] outlined the substructure of the heart, but did not include other important thoracic OARs such as lungs and spinal cord. In this paper, we attempt to outline the necessary thoracic OARs as comprehensively as possible, making the results more universal.

Generally, there are four types of geometric evaluation indexes to verify the accuracy of auto -segmentation software. These indexes are Dice similarity coefficient (DSC), moment, maximum Hausdorff distance (MHD) and average Hausdorff distance (AHD) [4]. Some studies on the auto-segmentation of the thoracic OARs involved only two or three types of indexes, which was not sufficient to fully evaluate the accuracy of auto-segmentation [5-8]. For example, Eduard *et al.* [8] used DSC and MHD for the thoracic and abdomen OARs. In this study, we investigated the use of all four types of indexes for the purpose. Among them, the moment index includes the difference of the Euclidean distance between centers of mass (ΔCMD) and the difference of volume (ΔV), which could present more details for the auto-segmented contours of the unsatisfactory performance.

At present, there were few literatures on the grading evaluation of auto-segmentation software to delineate the thoracic OARs for clinical use. Most studies [5-9] simply calculated the geometrical indexes of each OAR. Delia *et al.* [10] combined the DSC, ΔCMD and AHD indexes in the auto-segmentation study of breast cancer to access the accuracy levels of each OAR. There was also a literature that added subjective scoring [11], but their research did not focus on whether each OAR could be generated using auto-segmentation software. In this study, we graded the accuracy of auto-segmentation of thoracic OARs as comprehensively as possible by the five indexes in three levels.

Accurate identification of OARs in thorax is difficult for all OARs with ABAS to date. However, some thoracic OARs can be fairly accurately segmented with ABAS, while the segmentation accuracy of some other OARs can be limited. Therefore, it is necessary to develop a grading method to fully evaluate the performance of auto-segmentation of thoracic OARs so that ABAS can be properly used on the segmentation of thoracic OARs. This is, to our knowledge, the first time that comprehensive geometric indexes were used to gradedly evaluate ABAS based on comprehensively grading the thoracic OARs. More importantly, we graded ABAS into three different levels under the consideration of clinical feasibility. In the present work, grading evaluation of ABAS of fourteen kinds of thoracic OARs which include the left lung, right lung, spinal cord, heart, esophagus, chest wall, aorta, pulmonary artery, pulmonary vein, superior vena cava, inferior vena cava, skin, trachea and brachial plexus on computed tomography (CT) images was studied. Three auto-segmentation methods were compared with manual delineation. Five geometric indexes were used to quantitatively evaluate the accuracy of ABAS. The time difference between auto-segmentation and manual delineation was also compared.

MATERIALS AND METHODS

Patient selection and contour methods

We retrospectively selected forty patients with thoracic malignant tumors treated in our center between November and December 2018. We included patients with lung, esophageal and thymic tumors to ensure the diversity of atlas library. CT scans of each patient were obtained by a Siemens Somatom Definition AS CT Scanner System (Siemens Healthcare, Erlangen, Germany). The slice thickness of the CT scans was 3mm. The images were transferred to Pinnacle^3 treatment planning system (TPS) v9.10 (Philips Healthy, Fitchburg, WI, USA). Following the Radiation Therapy Oncology Group (RTOG) guidelines [12], an experienced radiotherapist in our center manually delineated thirteen thoracic OARs, including the left lung (L

Lung), right lung (R Lung), spinal cord (SC), heart, esophagus (ESO), chest wall (CW), aorta (AOR), pulmonary artery (PA), pulmonary in (PV), superior vena cava (SVC), inferior vena cava (IVC), skin and trachea of forty patients on the Pinnacle TPS. The brachial plexus was not included because of the limitation of CT contrast and resolution.

Forty patients were randomly divided into two groups, the atlas training dataset and the test dataset. The atlas training dataset contained twenty patients, and the other twenty patients were included into the test dataset. According to the modeling requirements of the latest version of commercial software MIM 6.8.7 (MIMvista Corp., Cleveland, US-OH), one patient with average anatomy in training dataset was set to the model atlas, and the other nineteen were set to the object atlases. For the test dataset, we used the ABAS tool of MIM for OARs auto-segmentation.

Atlas based auto-segmentation

Description of ABAS tool

The ABAS is a method to segment new images based on previously segmented images. The primary factor to ensure the accuracy of ABAS is the accuracy of image registration. Differences in the anatomical structure will cause registration errors, and the determination of the average patient will alleviate this situation. So, the first step is to select an average patient as the model atlas, the rest as object atlases, and then registering the object atlases one by one to the model atlas to get the corresponding spatial correspondence. The above steps are the establishment process of the atlas library.

When a new image needs to be segmented, it will be registered to the model atlas and the corresponding spatial correspondence is compared to those of the object atlases. Then the most similar object atlases are selected from the atlas library, that is, the best match atlases, and their contours are propagated to the new image. Atlas selection includes single-atlas and multi-atlas. Single-atlas is to select one best match atlas from the atlas library. In order to improve the robustness of image segmentation, multi-atlas has more than one best match atlases, which is related to the fusion of multiple atlas tags. At present, common tag fusion algorithms are major voting algorithm [13] and simultaneous truth and performance level evaluation (STAPLE) algorithm [14]. The majority voting algorithm selects the tags of each voxel that most atlases appeared. The STAPLE algorithm calculates a probability model based on the similarity between each selected atlas and the new image, and weighted fuses the tags of each atlas spread on each voxel.

Implementation of ABAS tool

The MIM 6.8.7 was used to create atlas library by the twenty training patients for thoracic OARs. The CT images of the twenty test patients were transmitted to MIM, and their auto -segmented contours were obtained after setting the auto-segmentation region, OARs, atlas selection and fusion algorithms. This study used two atlas selection methods both the single-atlas and multi-atlas. And for the test patients using multi-atlas, 5 best match atlases were selected based on the research results of Pirozzi [15]. For the multi-atlas mode, the two major fusion algorithms, major voting and STAPLE, were selected. Therefore, we studied three auto-segmented contours based on different algorithms: single-atlas based contour (Single); multi-atlas based contour with majority voting algorithm (MV5); multi-atlas based contour with STAPLE algorithm (ST5). The manual contour (MC) was used as the gold standard to evaluate the OARs' performance by the above three auto-segmentation methods.

Geometric evaluation

In order to grade the accuracy of ABAS tools for the auto-segmentation of thoracic OARs, the Dice similarity coefficient (DSC), moment, maximum Hausdorff distance (MHD) and average Hausdorff distance (AHD) were used to access the geometric differences between auto-segmented and manual contours.

The DSC [16] was calculated by using equation 1.

$$DSC = \frac{2|V_{manual} \cap V_{atlas}|}{|V_{manual} + V_{atlas}|} \quad (1)$$

Where; V_{manual} is the volume of manual contour

and V_{atlas} is the volume of the auto-segmented contour. The range of DSC is 0-1. If DSC is 1, then the two contours are coincident perfectly, and if DSC is 0, then the two don't overlap at all.

Moment metrics include the difference of the Euclidean distance between centers of mass (ΔCMD) and the difference of volume (ΔV).

The MHD [17] refers to the maximum distance between two point sets of the two contours, and is sensitive to the region with the largest difference in segmentation. The metric is commonly used in auto-segmentation studies.

The AHD [17] describes the average distance between two contours. The smaller the AHD is, the smaller the difference is between them. When the DSC is close to 1, the AHD might be a good index for distinguishing contours' difference.

Statistical analysis and tests

Statistical analysis of these geometric indexes with different atlas selections and fusion algorithms were performed using the Wilcoxon signed ranks tests with $p<0.05$ considered statistically significant. All analyses were performed using SPSS version 17.0 (SPSS, Chicago, IL, USA).

Clinical Efficiency

The times taken by MIM's ABAS tool to auto-segment the thirteen thoracic OARs involved in this study were also recorded.

The establishment of standard of grading evaluation

At present, there was no universal standard of the grading evaluation of auto-segmentation of thoracic OARs. Most of the literatures used the DSC [5-10]. In this paper, DSC was used as the main index for grading evaluation, and the other four indexes were also studied and their grading results were compared with DSC's. After the best of three auto-segmentation methods of Single, MV5 and ST5 was selected, the performance of fourteen thoracic OARs with the best method was graded into several levels.

RESULTS

Atlas accuracy evaluation with three different auto-segmentation methods

Tables 1 and 2 shown the mean values and statistical analysis of the five geometric indexes of the thirteen thoracic OARs' auto-segmented contours generated by ST5, MV5 and Single methods.

ST5 was the best performing auto-segmentation method. Most of the indexes (33/65) were the best values, compared to those derived with other methods. The mean DSCs of eight OARs were higher than those of the MV5 and Single methods. The mean ΔCMDs of nine OARs were smaller than the other two. In the method, the mean values of ΔV and AHD of six OARs were the smallest, and the mean values of MHD of four OARs were the smallest.

The overall accuracy of MV5 and Single methods was lower than ST5, but some indexes were the best, which could be divided into the following three cases. The first case was that the mean ΔCMD (R Lung), ΔV (AOR, trachea and ESO), MHD (heart, AOR and IVC), AHD (AOR and ESO) of MV5 method and the mean ΔCMD (Chest Wall) of Single method were the best, but there were no statistically significant differences in these indexes compared with ST5. The second case was that the mean value of all the indexes of R Lung, L Lung and skin auto-segmented contours generated by MV5 were the best, except the mean ΔCMD of R Lung, which had statistically significant differences compared with ST5. But the mean DSCs of all the methods were high (DSC>0.93). The third case was that the mean values of all the indexes of chest wall and trachea auto-segmented contours generated by MV5 were the best, and there were significant differences in these indexes between MV5 and ST5, except for the mean ΔCMD of chest wall, ΔCMD and ΔV of trachea.

In order to investigate the performance of auto-segmentation of each thoracic OAR by MIM, the following analyses were conducted using ST5 method.

Table 1. Mean value of the five indexes with STAPLE 5 (ST5), majority voting 5 (MV5) and single atlas (Single) auto-segmentation methods.

Structure	DSC			ΔCMD (cm)			ΔV (%)			MHD (cm)			AHD (cm)		
	ST5	MV5	Single	ST5	MV5	Single	ST5	MV5	Single	ST5	MV5	Single	ST5	MV5	Single
R Lung	0.96	**0.97**	0.96	0.13	**0.08**	0.13	6.50	**2.48**	4.43	2.18	**1.71**	2.17	0.11	**0.07**	0.10
L Lung	0.94	**0.96**	0.95	0.28	**0.12**	0.15	9.97	**3.25**	5.88	3.45	**1.97**	3.05	0.17	**0.08**	0.13
Skin	0.93	**0.97**	0.96	2.06	**0.95**	1.43	11.24	**6.09**	8.29	8.85	**6.41**	7.14	0.58	**0.27**	0.37
Heart	**0.90**	0.89	0.87	**0.40**	0.42	0.49	**7.37**	8.55	8.12	1.85	**1.82**	2.26	**0.24**	0.25	0.31
Spinal Cord	**0.88**	0.86	0.82	**1.34**	1.79	2.26	**8.80**	15.04	17.63	**2.74**	3.46	4.44	**0.13**	0.18	0.31
AOR	**0.79**	0.78	0.75	**0.92**	0.97	1.25	24.49	**18.83**	18.98	2.72	**2.52**	3.32	0.28	**0.26**	0.34
Chest Wall	0.77	**0.83**	0.82	1.29	1.12	**1.05**	39.41	**9.93**	13.43	6.51	**3.47**	4.13	0.46	**0.24**	0.28
Trachea	0.75	**0.79**	0.73	**0.53**	0.57	1.01	34.06	**11.31**	23.67	4.27	**2.20**	3.34	0.25	**0.14**	0.25
PA	**0.73**	0.68	0.62	**0.71**	0.76	0.87	**15.95**	28.24	25.40	**2.12**	2.37	2.33	**0.28**	0.32	0.38
SVC	**0.62**	0.55	0.56	**1.17**	1.33	1.26	**28.53**	44.40	34.80	**1.87**	2.26	2.26	**0.33**	0.42	0.42
ESO	**0.57**	0.54	0.50	**0.89**	1.00	1.82	32.77	**31.73**	33.27	2.10	**2.05**	2.98	0.29	**0.28**	0.49
IVC	**0.56**	0.48	0.41	**0.91**	0.96	1.34	**30.16**	46.83	50.24	2.17	**2.06**	3.85	**0.43**	0.48	0.88
PV	**0.53**	0.43	0.49	**1.00**	1.10	1.02	**35.44**	44.92	38.60	**2.65**	2.72	2.99	**0.44**	0.54	0.52

Table 2. Results of the statistical analysis of STAPLE 5 (ST5), majority voting 5 (MV5) and single-atlas (Single) auto-segmentation methods and the time of manual delineation. a: a statistical difference between ST5 and MV5 ($p<0.05$), b: a statistical difference between ST5 and Single ($p<0.05$), c: a statistical difference between MV5 and Single ($p<0.05$), -: the differences between the three is not statistically significant.

Structure	*p*					T (min)
	DSC	ΔCMD (cm)	ΔV (%)	MHD (cm)	AHD (cm)	
R Lung	a, c	c	a, c	a, c	a, c	7.00
L Lung	a, c	a	a, c	a, c	a, c	7.10
Skin	a, b	a, b	a, b	a, b	a, b	/
Heart	b, c	-	-	-	b, c	9.50
Spinal Cord	a, b, c	a, b	a, b	a, b, c	a, b, c	3.20
AOR	b, c	-	-	c	c	6.20
Chest Wall	a, b	-	a, b, c	a, b	a, b	19.60
Trachea	a, c	b, c	c	a, c	a, b, c	17.30
PA	a, b, c	b	a, b	a, b	a, b, c	4.80
SVC	a	a	a, b, c	a, b	a, b	2.00
ESO	b	b, c	-	a, b, c	b, c	12.00
IVC	a, b, c	b, c	a, b	b, c	b, c	3.70
PV	a, c	-	-	-	a	5.00

Auto-segmentation results for ST5 method

Figure 1 showed the consistency between the thirteen OARs' auto-segmented contours generated by ST5 and manual contours. For the R Lung, L Lung, skin, heart and spinal cord, the mean value ranges of DSC, ΔCMD, ΔV, MHD and AHD were 0.88-0.96, 0.13-2.06 cm, 6.50%-11.24%, 1.85-8.85 cm, 0.11-0.58 cm, respectively.

For the AOR, chest wall, trachea and PA, the mean value ranges of DSC, ΔCMD, ΔV, MHD and AHD were 0.73-0.79, 0.53-1.29 cm, 15.95%-39.41%, 2.12-6.51 cm, 0.28-0.46 cm, respectively.

For the SVC, ESO, IVC and PV, the mean value ranges of DSC, ΔCMD, ΔV, MHD and AHD were 0.53-0.62, 0.89-1.17 cm, 28.53%-35.44%, 1.87-2.65 cm, 0.29-0.44 cm, respectively.

Clinical efficiency of auto-segmentation

The average times of these thirteen OARs manually delineated by the radiotherapist are shown in table 2. The total time for manual delineation was 97.4 minutes, and the total time for OARs auto-segmentation one by one was 31.0 minutes. If all the thirteen OARs were

auto-segmented in one time, it only took 3.2 minutes. Therefore, auto-segmentation of the thirteen OARs saved by 96.7% compared to manual delineation.

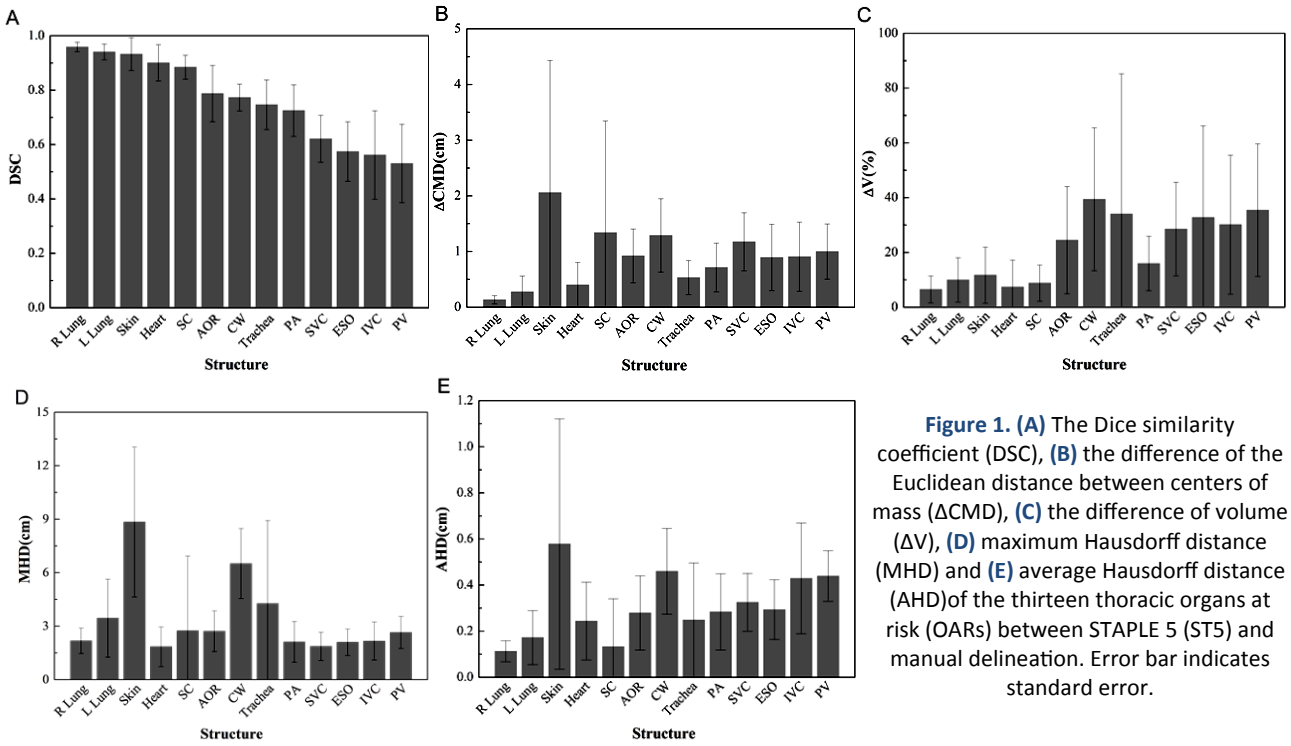

Figure 1. (A) The Dice similarity coefficient (DSC), **(B)** the difference of the Euclidean distance between centers of mass (ΔCMD), **(C)** the difference of volume (ΔV), **(D)** maximum Hausdorff distance (MHD) and **(E)** average Hausdorff distance (AHD)of the thirteen thoracic organs at risk (OARs) between STAPLE 5 (ST5) and manual delineation. Error bar indicates standard error.

Clinical efficiency of auto-segmentation

The average times of these thirteen OARs manually delineated by the radiotherapist are shown in table 2. The total time for manual delineation was 97.4 minutes, and the total time for OARs auto-segmentation one by one was 31.0 minutes. If all the thirteen OARs were auto-segmented in one time, it only took 3.2 minutes. Therefore, auto-segmentation of the thirteen OARs saved by 96.7% compared to manual delineation.

The standard of grading evaluation

To further investigate the accuracy and applicability of the auto-segmentation tools on the thoracic OARs, attempts were made to grade the results of the auto-segmentation for the thirteen thoracic OARs based on the five geometric indexes. As shown in table 3, OARs auto-segmentation was divided into three levels by each index. Namely, Level 1 OARs were not recommended to use auto-segmentation, and Level 2 OARs required manual modification after auto-segmentation, and Level 3 OARs could completely replace manual delineation.

Table 3. The standard of grading evaluation by five geometric indexes

Level	DSC	ΔCMD (cm)	ΔV (%)	MHD (cm)	AHD (cm)
1	(0-0.7]	(1.0-∞)	(20-∞)	(2.2-∞)	(0.4-∞)
2	(0.7-0.8]	(0.5-1.0]	(10-20]	(1.0-2.2]	(0.2-0.4]
3	(0.8-1.0]	[0-0.5]	[0-10]	[0-1.0]	[0-0.2]

Grading evaluation of thirteen thoracic OARs using ST5 method

According to the standard of grading evaluation, the results of the auto-segmentation for the thirteen thoracic OARs based on the five geometric indexes, which shown in table 4 and figure 2. According to the DSC, the R Lung, L Lung, skin, heart and spinal cord were Level 3 (mean DSC range: 0.88-0.96), which could completely replace manual delineation. The AOR, chest wall, trachea and PA were Level 2 (mean DSC range: 0.73-0.79), which required to

be manually modified after auto-segmentation. The SVC, ESO, IVC and PV were Level 1 (mean DSC range: 0.53-0.62), which could not use auto-segmentation.

There were differences between the grading results of the other four indexes and those of the DSC. For the spinal cord, it was rated as Level 3 by DSC, but according to ΔCMD, MHD that was rated as Level 1. Similarly, the chest wall and skin were rated higher by DSC. However, there were also cases where the grade according to DSC was lower than the other indexes. For small OARs such as SVC, IVC, it was rated as Level 1 by DSC, but according to ΔCMD, MHD and AHD that were rated as Level 2. Moreover, the grading results of most of the OARs evaluated by ΔV were the same as those of DSC, except the grades of the skin, AOR, chest wall and trachea were one level lower than those of DSC.

Table 4. Grading evaluation results of the thirteen thoracic organs at risk (OARs) by each of the five indexes.

	DSC	ΔCMD	ΔV	MHD	AHD
R Lung	3	3	3	2	3
L Lung	3	3	3	1	3
Skin	3	1	2	1	1
Heart	3	3	3	2	2
Spinal Cord	3	1	3	1	3
AOR	2	2	1	1	2
Chest Wall	2	1	1	1	1
Trachea	2	2	1	1	2
PA	2	2	2	2	2
SVC	1	1	1	2	2
ESO	1	2	1	2	2
IVC	1	2	1	2	1
PV	1	2	1	1	1

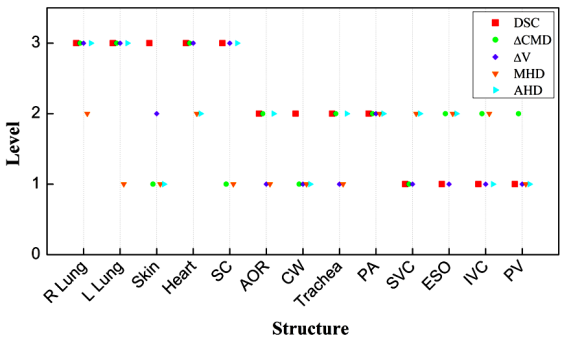

Figure 2. Grading evaluation results of the thirteen thoracic organs at risk (OARs) by the Dice similarity coefficient (DSC), the difference of the Euclidean distance between centers of mass (ΔCMD), the difference of volume (ΔV), maximum Hausdorff distance (MHD) and average Hausdorff distance (AHD)

DISCUSSION

In this paper, we graded the auto-segmentation of thirteen important thoracic OARs using ST5, MV5 and Single methods in the latest version of MIM using five geometric indices of DSC, ΔCMD, ΔV, MHD and AHD. The results showed that most of the indexes of ST5 method were better, while the MV5 method only had the advantage in automatically segmenting the chest wall and trachea (p<0.05). The Single method was inferior to the other two methods in general. Therefore, ST5 method was chosen as the primary one for the grading evaluation. We divided the performance of the thirteen OARs auto-segmentation based on ST5 method into three levels. This provided a clear scope of application of MIM's auto-segmentation of the thirteen thoracic OARs. The research methods used in this study are also applicable to the grading evaluation of other auto-segmentation software applications.

In this study, forty cases were randomly divided into the training dataset and the test dataset on average. The two datasets did not overlap, which means the patients whose OARs auto-segmentation were graded did not include those in the training dataset.

La Macchia *et al.* [18] evaluated the accuracy of three commercial auto-segmentation software for segmentation of sixteen OARs in head and neck. Since our center primarily treat thoracic tumor patients, the auto-segmentation study of thoracic OARs is the focus. We attempted to incorporate more thoracic OARs into our study such as the L Lung, R Lung, spinal cord, heart, ESO, thoracic wall, AOR, PA, PV, SVC, IVC, skin, trachea and brachial plexus, etc. However, due to the limitation of CT contrast and resolution, it was difficult to identify complex brachial plexus by human expert [19], and MIM could not currently automatically segment it. It might be necessary to rely on multimodal image fusion to achieve auto-segmentation of the brachial plexus.

Previous clinical evaluations based on auto-segmentation of thoracic OARs showed that both ABAS and deep learning techniques could save time, which is consistent with our

study [11]. Compared to manually delineating thirteen thoracic OARs, MIM achieved a significant time benefit by 96.7%. The sum time for automatically segmenting OARs one by one was 31.0 minutes, which was much longer than the time for segmenting all OARs together (3.2 minutes). The reason was that MIM took about 2 minutes to select atlas matches for each auto-segmentation. In addition, because MIM and the common TPS have the ability to auto-segment the skin, manual delineation time of skin was not recorded.

Our results showed that ABAS was not completely applicable for thoracic OARs, so grading auto-segmentation accuracy of OARs was an important task. We divided the segmentation accuracy of each OAR into three levels according to the DSC (table 4). For the R Lung, L Lung, skin, heart and spinal cord, the accuracy of auto-segmentation was good (DSC: 0.88-0.96), So they did not need manual modification. These Level 3 OARs could be auto-segmented. For the AOR, chest wall, trachea and PA, the accuracy of auto-segmentation was medium (DSC: 0.73-0.79). These auto-segmented contours needed to be modified in part, but there was still considerable time benefit compared with manual delineation slice-by-slice. These Level 2 OARs required manually modification after auto-segmentation. For the SVC, ESO, IVC and PV, the accuracy of auto-segmentation was poor (DSC: 0.53-0.62). These Level 1 OARs should be delineated manually rather than auto-segmentation. The grading results of the above OARs except SVC based on the study of breast cancer were consistent with ours [10]. The brachial plexus, which could not currently be automatically segmented by MIM, should also belonged to Level 1.

At the same time, we also graded auto-segmentation of each OAR according to the other four indexes: ΔCMD, ΔV, MHD and AHD, and found that the results were somewhat different from the DSC. For the SVC and IVC, although DSC were 0.62/0.56 (Level 1), the value of the distance indexes of ΔCMD, MHD and AHD were 1.17/0.91 cm, 1.87/2.17 cm and 0.33/0.43 cm (Level 2), respectively. The main reason was that these organs were small. Even if the segmentation of them was not good, the distance difference was not large. However, the situation of the spinal cord was opposite, in which the DSC was 0.88 (Level 3), and the ΔCMD and MHD were as high as 1.34 cm and 2.74 cm (Level 1), respectively. This was because the structure of the spinal cord was cylindrical, and it was difficult to judge the starting and ending slices when the auto-segmentation was performed, therefore the ΔCMD and MHD of the superior-inferior (SI) direction were large. And chest wall and skin had similar situations. For trachea, the grading result of ΔV (Level 3) was inconsistent with DSC (Level 2), which was caused by three outliers (ΔV: 136.27%, 168.98%, 146.62%). After removing these outliers, ΔV and DSC were 13.49% (Level 2) and 0.78 (Level 2), respectively. Although these four indexes could not be directly used for grading evaluation, they could be used to analyze the inconsistencies between auto-segmented and manual contours.

Because the size and shape of each OAR are different, the geometric indexes will be affected by these characteristics of OARs [20]. Future work will grade each OAR according to specific range of geometric indexes, combined with subjective scores, rather than using unified indexes range to grade all OARs.

The effect of contour differences on dosimetry has not been studied in this paper. Lo *et al.* [21] reported that differences in contours caused dose differences in the peer review of the lung cancer target and normal tissues delineation. Robert *et al.* [2] found in the diametric evaluation of auto-segmentation for breast cancer patients that although the geometric differences of the left anterior descending artery between manual delineation and auto-segmentation were large, no dosimetric differences were caused. Therefore, we did not conduct a study of dosimetric differences. Grading evaluation of OARs auto-segmentation by geometric indexes and dosimetric indexes can be studied in the future. The combination of the above two kinds of indexes may improve the accuracy of grading. At present, the performance of ABAS tools is

mainly limited by two factors: the contrast of the image and the volume of the object [11]. For soft tissues with low contrast, ABAS tools are not easy to identify the boundary, and even radiation oncologists need great efforts to determine the exact boundary. In recent years, machine learning technology, especially deep learning method, shows promising results in its use in medical imaging specialty [22-24]. Therefore, machine learning technology may become the main development direction of auto-segmentation software in the future.

CONCLUSION

Grading evaluation of ABAS of thoracic OARs according to the DSC is feasible. Thoracic OARs auto-segmentation was graded into three levels, namely Level 3, 2, 1. The grading results may be useful in providing guidance for of the future ABAS development is to improve the algorithm so that more OARs can be automatically segmented.

***Conflicts of interest*:** *Declared none.*

REFERENCES

1. D'Andrea M, Benassi M, Strigari L (2016) Modeling Radiotherapy Induced Normal Tissue Complications: An Overview beyond Phenomenological Models. *Comput Math Methods Med*, ***2016:*** *1-9*.
2. Kaderka R, Gillespie EF, Mundt RC, Bryant AK, Sanudo-Thomas CB, Harrison AL, Wouters EL, Moiseenko V, Moore KL, Atwood TF, Murphy JD (2019) Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. *Radiother Oncol*, ***131:*** *215-220*.
3. Herk MV (2004) Errors and margins in radiotherapy. *Semin Radiat Oncol*, ***14:*** *52-64*.
4. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, Yang J (2014) Vision 20/20: Perspectives on automated image segmentation for radiotherapy. *Med Phys*, ***41:*** *050902*.
5. Kim J, Han J, Ailawadi S, Baker J, Hsia A, Xu Z, Ryu S (2016) Multi-Atlas Based Automatic Organ Segmentation for Lung Radiotherapy Planning. *Med Phys*, ***43:*** *3433*.
6. Meng Q, Kitasaka T, Nimura Y, Oda M, Ueno J, Mori K (2017) Automatic segmentation of airway tree based on local intensity filter and machine learning technique in 3D chest CT volume. *Int J Comput Assist Radiol Surg*, ***12:*** *245-261*.
7. Rebouças Filho PP, Cortez PC, Da SBA, Vh CA, Jm RST (2017) Novel and powerful 3D adaptive crisp active contour method applied in the segmentation of CT lung images. *Med Image Anal*, ***35:*** *503-516*.
8. Eduard S, Marcus DM, Tim F (2014) Multiatlas segmentation of thoracic and abdominal anatomy with level set-based local search. *J Appl Clin Med Phys*, ***15:*** *22–38*.
9. Morris ED, Ghanem AI, Pantelic MV, Walker EM, Han X, Glide-Hurst CK (2019) Cardiac Substructure Segmentation and Dosimetry Using a Novel Hybrid Magnetic Resonance and Computed Tomography Cardiac Atlas. *Int J Radiat Oncol Biol Phys*, ***103:*** *985-993*.
10. Ciardo D, Gerardi MA, Vigorito S, Morra A, Dell'acqua V, Diaz FJ, Cattani F, Zaffino P, Ricotti R, Spadea MF, Riboldi M, Orecchia R, Baroni G, Leonardi MC, Jereczek-Fossa BA (2017) Atlas-based segmentation in breast cancer radiotherapy: Evaluation of specific and generic-purpose atlases. *Breast*, ***32:*** *44-52*.
11. Lustberg T, van Soest J, Gooding M, Peressutti D, Aljabar P, van der Stoep J, van Elmpt W, Dekker A (2018) Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiother Oncol*, ***126:*** *312-317*.
12. Kong FM, L Quint MM, Bradley J (2012) Atlas for organs at risk (OAR) in thoracic radiation therapy. Available from: https://www.rtog.org/CoreLab/ContouringAtlases/LungAtlas.aspx.
13. Heckemann RA, Hajnal JV, Paul A, Daniel R, Alexander H (2006) Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *Neuroimage*, ***33:*** *115-126*.
14. Warfield SK, Zou KH, Wells WM (2004) Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Trans Med Imaging*, ***23:*** *903-921*.
15. Pirozzi S, Horvat M, Piper J, Nelson A (2012) Atlas-Based Segmentation: Evaluation of a Multi-Atlas Approach for Lung Cancer. *Med Phys*, ***39:*** *3677*.
16. Dice LR (1945) Measures of the amount of ecologic association between species. *Ecology*, ***26:*** *297-302*.
17. Sim DG, Kwon OK, Park RH (1999) Object matching algorithms using robust Hausdorff distance measures. *IEEE Transactions on Image Processing*, ***8:*** *425-429*.
18. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, Lomax AJ, Widesott L (2012) Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol*, ***7:*** *160*.
19. Myo M, Daniel R, Elly K, Michael P, Scott C, Lydia Z, Karen W, John S, Siddhartha B (2014) External evaluation of the radiation therapy oncology group brachial plexus contour-

ing protocol: several issues identified. *J Med Imag Radiat Oncol*, ***58:*** *360-368*.
20. Isambert A, Dhermain F, Bidault F, Commowick O, Bondiau PY, Malandain G, Lefkopoulos D (2008) Evaluation of an atlas-based automatic segmentation software for the delineation of brain organs at risk in a radiation therapy clinical context. *Radiother Oncol*, ***87:*** *93-99*.
21. Lo AC, Liu M, Chan E, Lund C, Truong PT, Loewen S, Cao J, Schellenberg D, Carolan H, Berrang T (2014) The Impact of Peer Review of Volume Delineation in Stereotactic Body Radiation Therapy Planning for Primary Lung Cancer: A Multicenter Quality Assurance Study. *J Thorac Oncol*, ***9:*** *527-533*.
22. van Rooij W, Dahele M, Ribeiro Brandao H, Delaney AR, Slotman BJ, Verbakel WF (2019) Deep learning-based delineation of head and neck organs-at-risk: geometric and dosimetric evaluation. *Int J Radiat Oncol Biol Phys, 104.*
23. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, Kamal MJ, Meheissen MAM, Mohamed ASR, Rao A, Williams B, Wong A, Yang J, Aristophanous M (2018) Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. *Int J Radiat Oncol Biol Phys*, ***101:*** *468-478*.
24. Hu P, Wu F, Peng J, Liang P, Kong D (2016) Automatic 3D liver segmentation based on deep learning and globally optimized surface evolution. *Phys Med Biol*, ***61:*** *8676*.