View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HKU Scholars Hub

BJR

© 2015 The Authors. Published by the British Institute of Radiology

Received: 5 November 2014 | Revised: 20 August 2015 | Accepted: 4 November 2015

doi: 10.1259/bjr.20140732

Cite this article as:
Wong WKH, Leung LHT, Kwong DLW. Evaluation and optimization of the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy. *Br J Radiol* 2016; **89**: 20140732.

SHORT COMMUNICATION

Evaluation and optimization of the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy

[1]WICGER K H WONG, MSc, [1]LUCULLUS H T LEUNG, PhD and [2]DORA L W KWONG, MB BS, MD

[1]Department of Oncology, Princess Margaret Hospital, Kwai Chung, Hong Kong
[2]Department of Clinical Oncology, University of Hong Kong, Pok Fu Lam, Hong Kong

Address correspondence to: Mr Wicger K H Wong
E-mail: *wongkhw@ha.org.hk*

Objective: To evaluate and optimize the parameters used in multiple-atlas-based segmentation of prostate cancers in radiation therapy.
Methods: A retrospective study was conducted, and the accuracy of the multiple-atlas-based segmentation was tested on 30 patients. The effect of library size (LS), number of atlases used for contour averaging and the contour averaging strategy were also studied. The autogenerated contours were compared with the manually drawn contours. Dice similarity coefficient (DSC) and Hausdorff distance were used to evaluate the segmentation agreement.
Results: Mixed results were found between simultaneous truth and performance level estimation (STAPLE) and majority vote (MV) strategies. Multiple-atlas approaches were relatively insensitive to LS. A LS of ten was adequate, and further increase in the LS only showed insignificant gain. Multiple atlas performed better than single atlas for most of the time. Using more atlases did not guarantee better performance, with five atlases performing better than ten atlases. With our recommended setting, the median DSC for the bladder, rectum, prostate, seminal vesicle and femurs was 0.90, 0.77, 0.84, 0.56 and 0.95, respectively.
Conclusion: Our study shows that multiple-atlas-based strategies have better accuracy than single-atlas approach. STAPLE is preferred, and a LS of ten is adequate for prostate cases. Using five atlases for contour averaging is recommended. The contouring accuracy of seminal vesicle still needs improvement, and manual editing is still required for the other structures.
Advances in knowledge: This article provides a better understanding of the influence of the parameters used in multiple-atlas-based segmentation of prostate cancers.

INTRODUCTION

Much of the time in intensity-modulated radiotherapy planning is spent on the tedious contouring task. The contouring time for prostate cases can be up to 17.5 min.[1] Apart from being time consuming, manual contouring also suffers from interobserver variability.[2]

Single-atlas-based segmentation was introduced and several studies[3,4] showed that atlas-based automated contouring can reduce the contour time and the interobserver variability. However, the resultant accuracy was unsatisfactory at early stage development.[4] Multiple-atlas method was suggested to address the problem, and many different averaging strategies were proposed including simultaneous truth and performance level estimation (STAPLE) and majority vote (MV).

The commercial software MIMVista v. 6.0 (MIMVista Corp., Cleveland, OH) employed the multiple-atlas approach in the autosegmentation feature. Two averaging strategies, STAPLE and MV, can be used in the autosegmentation process. They differ from each other in the way that they assign the weighting to each of the segmentations.[5] Apart from choosing the averaging method, the program also allows the user to have the flexibility to set the number of atlases chosen from the library for contour averaging.

The aim of this study is to validate the accuracy of the contours generated by the software and to study the influence of the number of atlases chosen, the size of library and the averaging strategies to the autosegmentation process of prostate cancers.

METHODS AND MATERIALS

A retrospective study was performed and Supplementary Figure A outlines the procedures, and the details are described as follows.

Description of image data
80 patients with prostate cancer who had received intensity-modulated radiotherapy treatment were randomly selected. All of the cases had CT scan performed with the 2.5-mm slice scanning protocol and with bladder-filling protocol employed. The bladder, rectum, prostate gland, seminal vesicles and femurs were delineated by one of our oncologists and validated by two senior radiation oncologists. The organs at risk and target volume were delineated using departmental guideline based on the Radiation Therapy Oncology Group reference.[6]

Multiple-atlas-based segmentation
The effects on the segmentation accuracy of the following three factors were studied.

Multiple-atlas combination method
(1) STAPLE
(2) MV.

*Number of atlases chosen from the library (*n = *number of atlases chosen)*
Different number of atlases ($n = 1$, 3, 5 or 10) were chosen from the library to create the combined contour.

Library size
Three library sets were built with 10, 30 and 50 atlases, respectively.

30 sample cases were used to evaluate the accuracy and influence of the above factors. A complete set of structures was generated for each of the sample cases for each setting. Similarity measure was then calculated to assess the agreement between the generated contours and the contours drawn by oncologists.

Similarity measurements
Dice similarity coefficient (DSC) and Hausdorff distance (HD) were used to assess the degree of agreement. The DSC measures the volumetric overlap between two contour sets,[7] whereas HD is a distance measure of the mismatch of two contour sets[8] (for definition, see Supplementary material).

Running time
The autosegmentation was performed with a standalone workstation (Intel® Xeon® central processing unit E5620, 2.4 GHz, 24 GB of random-access memory, Dell Inc., Plano, TX; Microsoft® Windows® XP Professional ×64 edition, Microsoft, Redmond, WA). The computation time for each setting was recorded.

Statistical analyses
The DSC in segmentation studies usually violates the normal distribution assumption of parametric statistical test. Logit transform were performed for the DSC to allow appropriate statistical inferences,[1] where

$$\text{logit}(\text{DSC}) = \ln[\text{DSC}/(1 - \text{DSC})].$$

All the statistical tests were performed with logit(DSC). However, for better direct comparison with other literatures, the DSC values were presented in their general form in the following text unless stated explicitly.

Simultaneous truth and performance level estimation versus *majority vote*
Paired *t*-test was used to compare the mean logit(DSC) measurements for the contours generated by the two segmentation strategies. For the HD measurements, Wilcoxon signed-ranks tests were performed.

Number of atlases used and the library size
Analyses of variance were performed to compare the mean logit(DSC) between groups followed by pairwise *t*-tests with Bonferroni correction. For the HD measurements, Friedman test was used to test the difference between groups followed by Wilcoxon signed-rank test with Bonferroni correction.

RESULTS
Segmentation accuracy
The highest degree of agreement was found for femurs with the median DSC ranging from 0.93 to 0.95, and the HD ranging from 12.9 to 14.9 mm (see Supplementary Tables A and B). The degree of agreement showed no dependence on the choice of parameters for the femurs. No significant difference was found between any groups on the segmentation agreement.

The femurs were followed by the bladder, prostate and rectum with a median DSC range of 0.84–0.93, 0.73–0.86 and 0.68–0.78, respectively. The seminal vesicle showed the lowest agreement, with the median DSC ranging from 0.33 to 0.64.

Figure 1 summarizes the segmentation results for each structure.

Segmentation strategy
Mixed results were found when comparing the two segmentation strategies, and Table 1 shows the statistical results for the two segmentation methods.

The differences between the two segmentation methods were more likely to be found when large number of atlases ($n = 10$) were used to calculate the average contour.

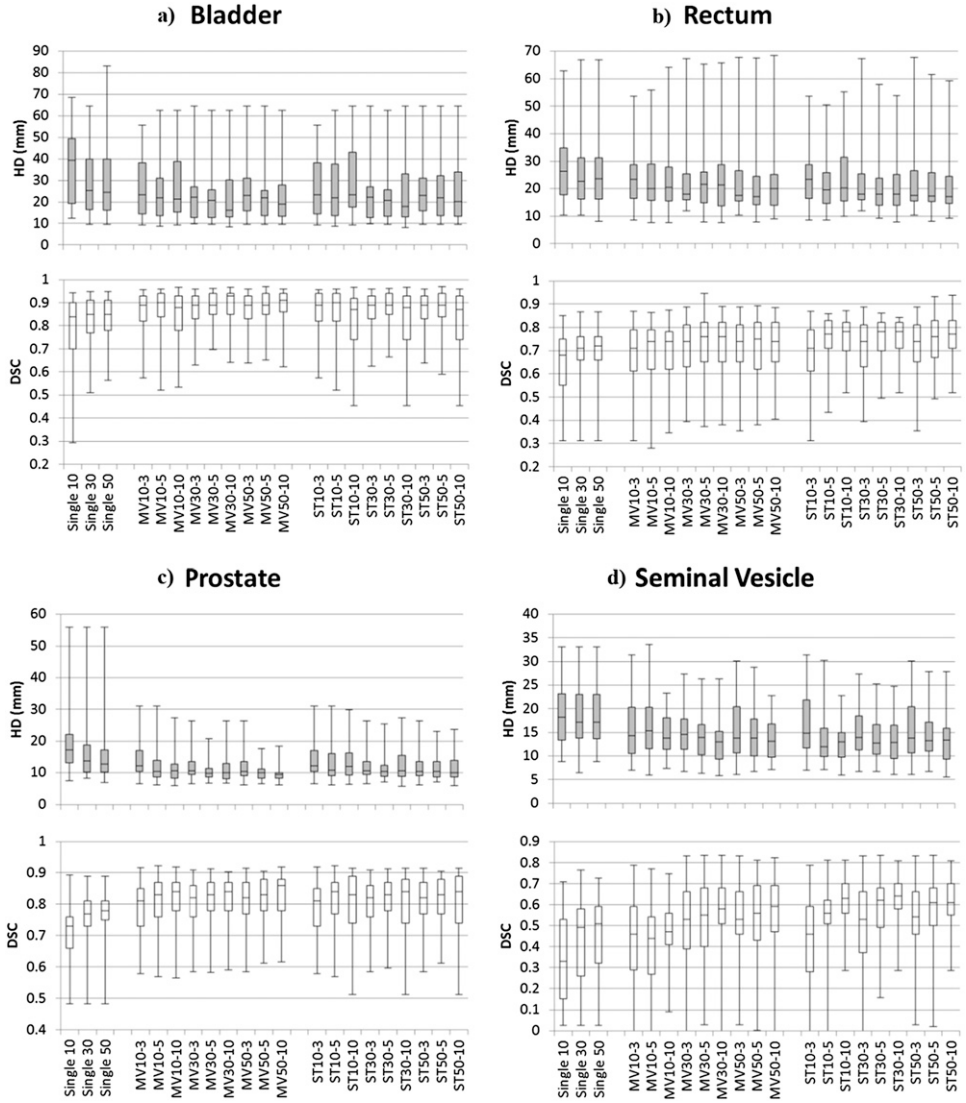
Both of the DSC and HD evaluation showed that MV performed better in the bladder and prostate, whereas STAPLE performed better in the rectum. The DSC evaluation indicated STAPLE performed better in the seminal vesicle.

Number of atlases used
The effect of the number of atlases used was tested with different library sizes (LSs) and structures. Among the 36 comparing groups, 31 of them showed significant difference between groups with different number of atlases used.

For those comparing groups with significant differences found in the analyses of variance, the single-atlas approach always performed inferiorly to the multiple-atlas approach.

For the multiple-atlas approach, the groups which used five atlases ($n = 5$) were the best amongst the three groups. Figure

Figure 1. Results of the dice similarity coefficient (DSC) and Hausdorff distance (HD) analyses. MV, the majority voting group with the library size and the number of atlases being used; Single, single atlas approach with the library size; ST, simultaneous truth and performance level estimation with the library size and number of atlases being used.



2a,b summarizes the Bonferroni *post hoc* comparison for the different number of atlases used.

Library size

Table 2 summarizes the results of the statistical tests for groups with different LS.

For the single-atlas approach, statistical increases in logit(DSC) value were found in the rectum and prostate when the LS was increased from 10 to 50. The HD evaluation also showed a significant difference for the prostate when the LS was increased from 10 to 50.

For the multiple-atlas approach, no trend was observed in logit(DSC) value with the increase of LS, except for the seminal vesicle and the majority voting group with ten atlases being used (MV10). In addition, except the groups with ten atlases being used, the HD evaluation showed that there was no significant difference in the HD value for different LSs for the rectum and bladder. However, for the prostate groups which used three atlases for averaging, it showed that the HD value was significantly smaller when a larger LS was used.

Running time

A linear relationship was found between the number of atlases used and the computation time (Supplementary Figure B). On the other hand, increasing the LS from 10 to 50 increased the computation time only by a few seconds.

DISCUSSION

In this study, we presented an investigation on the factors that will affect the segmentation accuracy in the multiple-atlas approach. Our results indicated that the choice of the

Table 1. Comparison of results between majority vote (MV) and simultaneous truth and performance level estimation (STAPLE)

Library size		10			30			50		
Number of atlases used		3	5	10	3	5	10	3	5	10
Bladder										
Paired *t*-test [logit(DSC)]	*p*-value	>0.1	>0.05	*<0.001*	>0.1	>0.05	*<0.001*	>0.1	>0.1	*<0.001*
Mean diff.[a]		0.00	0.00	0.02	0.00	0.00	0.05	0.00	0.01	0.05
Wilcoxon test (HD)	*p*-value	>0.1	>0.1	*<0.001*	>0.1	>0.1	*0.001*	>0.1	>0.1	*0.023*
Mean diff. (mm)[b]		0.0	−1.0	−3.6	0.0	−0.5	−2.0	0.0	−0.6	−1.8
Rectum										
Paired *t*-test [logit(DSC)]	*p*-value	>0.1	*0.002*	*0.009*	>0.1	>0.1	>0.1	>0.1	*0.009*	*0.037*
Mean diff.[a]		0.00	−0.06	−0.07	0.00	−0.03	−0.05	0.00	−0.04	−0.05
Wilcoxon test (HD)	*p*-value	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Mean diff. (mm)[b]		0.0	1.3	−1.5	0.0	1.5	1.4	0.0	2.5	1.5
Prostate										
Paired *t*-test [logit(DSC)]	*p*-value	>0.1	>0.1	*0.036*	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Mean diff.[a]		0.00	0.00	0.02	0.00	0.00	0.02	0.00	0.01	0.02
Wilcoxon test (HD)	*p*-value	>0.1	>0.1	>0.1	>0.1	>0.1	*0.005*	>0.1	*0.015*	*0.001*
Mean diff. (mm)[b]		0.0	−1.0	−1.8	0.0	−1.1	−1.1	0.0	−1.2	−1.9
Seminal vesicle										
Paired *t*-test [logit(DSC)]	*p*-value	*0.018*	*<0.001*	*<0.001*	>0.1	*0.003*	>0.05	>0.1	>0.05	>0.05
Mean diff.[a]		0.01	−0.12	−0.16	0.00	−0.06	−0.06	0.00	−0.04	−0.06
Wilcoxon test (HD)	*p*-value	*0.013*	*0.006*	*0.047*	>0.1	>0.1	>0.1	>0.1	>0.1	>0.1
Mean diff. (mm)[b]		−0.4	2.3	1.6	−0.1	0.2	−0.3	−0.1	0.0	−0.6

diff., difference; DSC, dice similarity coefficient; HD, Hausdorff distance.
[a]The mean difference in this table is presented in DSC. Positive mean difference indicates better performance for the MV group and negative mean difference indicates better performance for the STAPLE group.
[b]Positive mean difference indicates better performance for the STAPLE group and negative mean difference indicates better performance for the MV group.

parameters used in the segmentation process affected the resultant accuracy.

Theoretically, single-atlas-based segmentation is prone to various sources of error, including registration error and the existence of segmentation error of the atlas.[9] The multiple-atlas approach can minimize the above errors by averaging the structure contour of each atlas.

Our study showed that the segmentation accuracy of both STAPLE and MV were significantly better than the single-atlas approach. With the DSC evaluation, there was no multiple-atlas group that performed worse than the single-atlas approach.

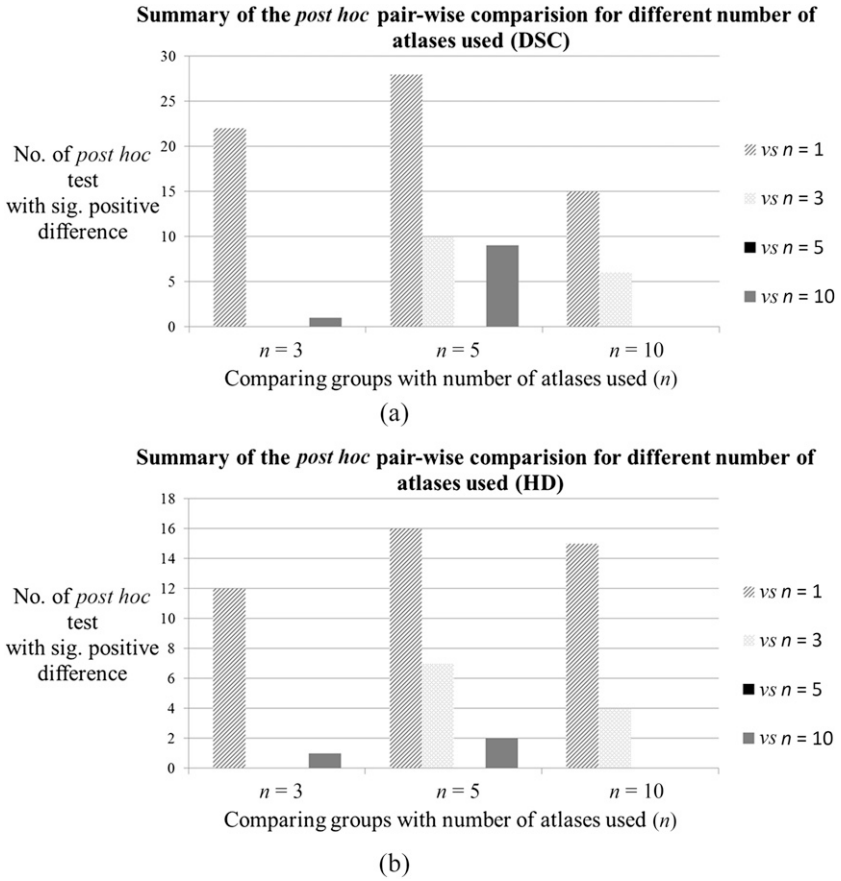
There are studies comparing different atlas combination methods, but there is no consensus among researchers regarding the performance of them.[10,11] Mixed results were found between the two methods in our study. The differences between the two methods were more pronounced when more atlases were used for the averaging process. The choice of the averaging method is more critical when only ten atlases were used. Although differences were found between the two methods, none of them consistently performs better than the other.

Apart from the atlas combination method, our study also showed that the number of atlases used in the combination process has a significant effect on the segmentation accuracy. Our results showed that using five atlases is the optimal number for prostate cases. Using ten atlases did not improve the segmentation accuracy. Moreover, taking into account that there was a linear relationship between the number of atlases used and the computation time, using five atlases could almost save half of the computation time. Therefore, we recommend using five atlases in the averaging process for prostate cancers.

Multiple-atlases approach did not only result in better agreement but was also superior to the single-atlas approach in the dependence of LS. The single-atlas approach is more sensitive to the LS. The segmentation accuracy can be increased by building a larger library.

In contrast to the single-atlas approach, both of the multiple-atlas segmentation methods are relatively insensitive to the LS if

Figure 2. (a) Summary of the dice similarity coefficient (DSC) Bonferroni *post hoc* comparison for the different number of atlases used. (b) Summary of the Hausdorff distance (HD) Bonferroni *post hoc* comparison for the different number of atlases used. No., number; sig., significant.



compared with the single-atlas method. No trend was observed in DSC value with the increase of LS except for the seminal vesicle and the MV10.

For the seminal vesicle, *post hoc* test showed no significant difference for different LSs for the STAPLE group, whereas the accuracy of the MV group was dependent on the LS.

Significant differences were also found in the bladder and rectum for the MV10. Bonferroni *post hoc* tests showed that the groups with a LS of 10 were worse than those of 30 and 50. The MV10 used ten atlases for contour combination. Choosing a LS of ten means that the software has to use all the atlases available and leaving it no degree of freedom to choose the appropriate atlases from the library. Apart from this special situation where all the atlases were used, no difference was found for different LSs for other comparing groups. Being relatively insensitive to the LS allows users to use smaller LSs and be able to create their library in a short period of time. Regarding the dependence of LS, STAPLE is more independent of it and is preferred especially when only small-sized libraries are available.

The results of our study showed that the multiple-atlas approaches have significant improvement on segmentation accuracy in prostate cases. However, structures such as the seminal vesicle still require a high degree of modification. With our recommended settings (STAPLE, $n = 5$, LS of ten), the median DSC value was 0.56. The failure may be due to the great variability of the size and shape of the seminal vesicle. In addition, the seminal vesicle is a small-sized organ; the program finds the best-matched atlas from the library according to the similarity of the whole pelvis instead of the seminal vesicle on its own. Another reason for the failure on segmenting the seminal vesicle is that the interobserver agreement is low. A recent study indicated that the use of atlas-based auto-segmentation (ABAS) is not beneficial for structures with low interobserver agreement.[12]

Apart from the failure for the seminal vesicle, rectum had a median DSC value of 0.77 with our recommended settings while the bladder and prostate had a median DSC value of 0.90 and 0.84, respectively. The left and right femurs had excellent agreement with median DSC value of 0.95 and 0.94. With our recommended settings, the resultant agreements were comparable with interrater variability reported in other studies.[13] One of the limitations of this study is that the contouring time was not registered and reference of other studies was adopted. The time-saving ability of ABAS had been demonstrated by another study[3] with comparable DSC

Table 2. Multiple-group comparison and Bonferroni *post hoc* test to test the difference in logit transform performed for the dice similarity coefficient [logit(DSC)] and Hausdorff distance (HD) between groups with different library sizes (LSs)

Number of atlases used		1	3	5	10
Bladder					
MV	*p*-value [logit(DSC)]	*0.047*	>0.1	>0.1	*<0.001*
	p-value (HD)	>0.1	>0.1	>0.1	*<0.001*
	Post hoc test	No difference			*a,b,c,d*
STAPLE	*p*-value [logit(DSC)]		>0.1	>0.1	>0.1
	p-value (HD)		>0.1	>0.1	*<0.001*
	Post hoc test				*c,d*
Rectum					
MV	*p*-value [logit(DSC)]	*0.030*	>0.05	>0.05	*0.020*
	p-value (HD)	>0.1	>0.1	>0.1	>0.1
	Post hoc test	*a*			*a,b*
STAPLE	*p*-value [logit(DSC)]		>0.05	>0.1	>0.1
	p-value (HD)	>0.1	>0.1	>0.1	*0.006*
	Post hoc test				c,d
Prostate					
MV	*p*-value [logit(DSC)]	*0.020*	>0.1	>0.1	>0.1
	p-value (HD)	*0.022*	*0.001*	>0.1	*0.017*
	Post hoc test	*a,b,c,d*	*c,d*		c,e
STAPLE	*p*-value [logit(DSC)]		>0.1	>0.1	>0.1
	p-value (HD)		*0.001*	>0.1	*0.015*
	Post hoc test		*c,d*		c
Seminal vesicle					
MV	*p*-value [logit(DSC)]	*0.024*	>0.05	*0.012*	>0.1
	p-value (HD)	>0.1	>0.1	>0.1	*0.026*
	Post hoc test	*a*		*a*	*c,d*
STAPLE	*p*-value [logit(DSC)]		*0.048*	>0.1	>0.1
	p-value (HD)		>0.1	>0.1	>0.1
	Post hoc test		No difference		

MV, majority vote; STAPLE, simultaneous truth and performance level estimation.
[a]The logit(DSC) value of the group with LS of 50 > the group with LS of 10; $p < 0.05$.
[b]The logit(DSC) value of the group with LS of 30 > the group with LS of 10; $p < 0.05$.
[c]The HD value of the group with LS equal to 50 < the group with LS of 10; $p < 0.05$.
[d]The HD value of the group with LS equal to 30 < the group with LS of 10; $p < 0.05$.
[e]The HD value of the group with LS equal to 50 < the group with LS of 30; $p < 0.05$.

results; they found that the use of ABAS can reduce the contouring time by >70% in prostate cases. Another study with inferior DSC result also demonstrated a reduction of delineation time by 35%.[14]

CONCLUSIONS

Both of the STAPLE and MV strategies performed better than the single-atlas approach. Although mixed results were found between the two average methods, STAPLE is preferred as its performance is less sensitive to the LS. The multiple-atlas approach has the advantage of being able to use smaller LSs. A LS of ten is adequate for prostate cases and further increases of the LS only shows insignificant gain. Five atlases are recommended to be used to generate the averaged contours. Further increase in the number of atlases shows no gain in the accuracy but significantly increases the computation time.

REFERENCES

1. Hwee J, Louie AV, Gaede S, Bauman G, D'Souza D, Sexton T, et al. Technology assessment of automated atlas based segmentation in prostate bed contouring. *Radiat Oncol* 2011; **6**: 110. doi: 10.1186/1748-717X-6-110
2. Valicenti RK, Sweet JW, Hauck WW, Hudes RS, Lee T, Dicker AP, et al. Variation of clinical target volume definition in three dimensional conformal radiotherapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 1999; **44**: 931–5. doi: 10.1016/S0360-3016(99)00090-5
3. La Macchia M, Fellin F, Amichetti M, Cianchetti M, Gianolini S, Paola V, et al. Systematic evaluation of three different commercial software solutions for automatic segmentation for adaptive therapy in head-and-neck, prostate and pleural cancer. *Radiat Oncol* 2012; **7**: 160. doi: 10.1186/1748-717X-7-160
4. Pejavar S, Yom SS, Hwang A, Speight J, Gottschalk A, Hsu IC, et al. Computer-assisted, atlas-based segmentation for target volume delineation in whole pelvic IMRT for prostate cancer. *Int J Radiat Oncol Biol Phys* 2008; **72**: S148. doi: 10.1016/j.ijrobp.2008.06.476
5. 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; **23**: 903–21. doi: 10.1109/TMI.2004.828354
6. Michalski JM, Lawton C, El Naqa I, Ritter M, O'Meara E, Seider MJ, et al. Development of RTOG consensus guidelines for the definition of the clinical target volume for post-operative conformal radiation therapy for prostate cancer. *Int J Radiat Oncol Biol Phys* 2010; **76**: 361–8. doi: 10.1016/j.ijrobp.2009.02.006
7. Al-Mayah A, Moseley J, Hunter S, Velec M, Chau L, Breen S, et al. Biomechanical-based image registration for head and neck radiation treatment. *Phys Med Biol* 2010; **55**: 6491–500. doi: 10.1088/0031-9155/55/21/010
8. Huttenlocher DP, Klanderman GA, Rucklidge WJ. Comparing images using the Hausdorff distance. *IEEE Trans Pattern Anal Mach Intell* 1993; **9**: 850–63. doi: 10.1109/34.232073
9. Aljabar P, Heckemann RA, Hammers A, Hajnal JV, Rueckert D. Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. *Neuroimage* 2009; **46**: 726–38. doi: 10.1016/j.neuroimage.2009.02.018
10. Sabuncu MR, Yeo BTT, van Leemput K, Fischl B, Golland P. A generative model for image segmentation based on label fusion. *IEEE Trans Med Imaging* 2010; **29**: 1714–29. doi: 10.1109/TMI.2010.2050897
11. Klein S, van der Heide UA, Lips IM, van Vulpen M, Staring M, Pluim JP. Automatic segmentation of the prostate in 3D MR images by atlas matching using localized mutual information. *Med Phys* 2008; **35**: 1407–17. doi: 10.1118/1.2842076
12. Langmack KA, Perry C, Sinstead C, Mills J, Saunders D. The utility of atlas-assisted segmentation in the male pelvis is dependent on the interobserver agreement of the structures segmented. *Br J Radiol* 2014, **87**: 20140299. doi: 10.1259/bjr.20140299
13. Sharp G, Fritscher KD, Pekar V, Peroni M, Shusharina N, Veeraraghavan H, et al. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Med Phys* 2014; **41**: 050902. doi: 10.1118/1.4871620
14. Gambacorta MA, Valentini C, Dinapoli N, Boldrini L, Caria N, Barba MC, et al. Clinical validation of atlas-based auto-segmentation of pelvic volumes and normal tissue in rectal tumors using auto-segmentation computed system. *Acta Oncol* 2013; **52**: 1676–81. doi: 10.3109/0284186X.2012.754989