Assessment of Multiple Atlas-Based Segmentation in Prostate Bed Contouring

Eddie Liu [1] , Arvand Barghi [2] , George Rodrigues [3]

1. London Health Sciences Centre and University of Western Ontario, London Ontario Canada 2. Department of Physics and Astronomy, London Health Sciences Centre and University of Western Ontario, London Ontario Canada 3. Department of Radiation Oncology, London Regional Cancer Program, London, Ontario, CA; Schulich School of Medicine & Dentistry, Western University, London, Ontario, CA

✉ **Corresponding author:** George Rodrigues, george.rodrigues@lhsc.on.ca
Disclosures can be found in Additional Information at the end of the article

Abstract

Dice similarity coefficients (DSC) of single best matching (SBM) and multiple best matching (MBM) prostate bed automated atlas-based segmentation (AABS) contours were compared to an expert panel gold standard. DSC scores improved with MBM in bladder (0.73-0.82) and penile bulb (0.40-0.54), with no improvement in other organs.

Introduction

Advancements in radiotherapy, such as intensity-modulated radiotherapy (IMRT) and image-guided radiotherapy (IGRT), improve therapeutic outcome by facilitating dose escalation in target tissue, while sparing adjacent normal tissues [1]. These advancements also demand high-contouring accuracy from radiation oncologists in order to optimize patient outcomes. Yet, manual contouring can be tedious and time-consuming. Furthermore, inter- and intra-observer variability is a major source of uncertainty [2]. To reduce time spent contouring as well as to decrease uncertainty, automated contouring techniques are being increasingly explored in the medical literature.

The use of computer-assisted auto-contouring algorithms, such as automated atlas-based segmentation (AABS), are a promising new approach to overcome the limitations of manual contouring. AABS begins by automatically selecting from a database of pre-contoured CT's, a best match to the patient simulation CT. It then performs a deformable registration of the selected contour to better match the patient anatomy between the two CTs. AABS algorithms can function using a single best match (SBM), where only one pre-contoured CT is used from a database, or a multiple best match (MBM), where a number of best matching contours are retrieved and combined to generate the contour using an algorithm.

The performance of AABS is a focus of research in the field of radiation oncology. Previous studies have demonstrated that AABS decreases inter- and intra-observer variability as well as contouring time in multiple cancer types [1, 3-5]. Yet, some research suggests the need for further improvement of AABS approaches, as illustrated by Hwee, et al. where they found that only 12% of their auto-contoured images were considered clinically acceptable by blinded

human observers [1]. Previous studies have not characterized, in particular, any differences between single best matching (SBM) and multiple best matching (MBM) approaches, or differences between contours when the number of best matches is varied when using MBM. Thus, the objective of this study is to investigate the potential for improvement of automated contours using commercially available contouring software that features AABS with MBM capabilities.

Materials And Methods

Five pelvic CT simulation datasets (512 x 512 pixels, 3mm slice thickness, 120 kVp) of five different prostate bed patients were each contoured by an expert panel of five radiation oncologists [1]. The six structures specifically delineated were prostate bed, rectum, bladder, penile bulb, and left and right femoral heads. A consensus contour for each structure was generated using the simultaneous truth and performance level estimate (STAPLE) algorithm [6]. The STAPLE algorithm estimates the true volume of a structure from a collection of observer contours as inputs. The STAPLE consensus contours were taken as the gold standard for investigational (automated) contours to be compared against.

A previously developed atlas database [1] was used for AABS auto-contouring. Commercially available software (MIM Software Inc, Cleveland OH, USA) was used to perform AABS, since it features not only SBM but also a 'multi-atlas' tool that allows MBM. In the case of MBM, the software generates the final segmentation from the volume of overlap between at least half of the indexed contours (for example two of three, two of four, three of five, etc.) (Figure 1). In this study, MBM of up to 10 best matches was explored. For each of the five patients and for each of the six structures, 10 AABS contours were generated by ranging from one to 10 best matches. Thus, a total of 300 AABS contours (10 AABS x six structures x five patient datasets) were generated and compared against the six STAPLE consensus contours to generate study datapoints.

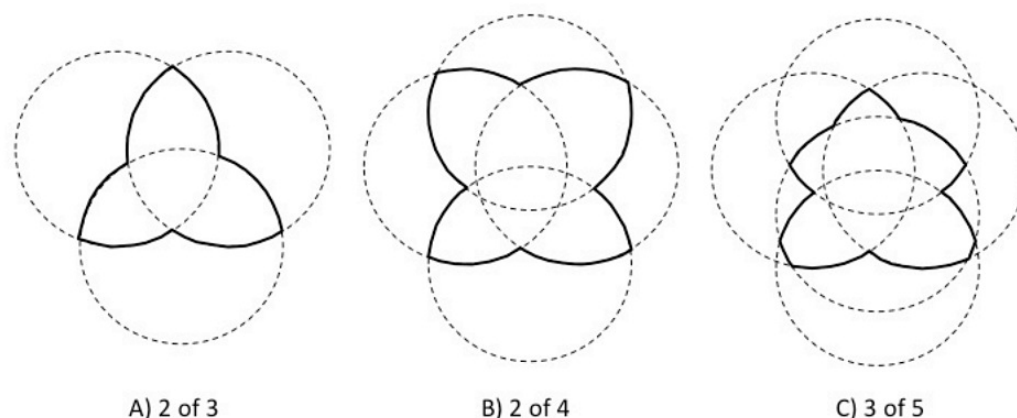A) 2 of 3   B) 2 of 4   C) 3 of 5

FIGURE 1: Illustration of MBM overlap algorithm

Illustration of MBM overlap algorithm, where the final contour is defined by the volume of overlap of at least half of the best matching contours in the case of a) 2 of 3, b) 2 of 4, and c) 3 of 5.

StructSure software (StructSure TM, Standard Imaging Inc., Middletown, WI, USA) was used to calculate Dice similarity coefficient (DSC) [6]. The DSC is defined as:

$$DSC = \frac{2|V \cap V_c|}{|V| + |V_c|}$$

where V is the volume within a contour given by a single observer, Vc is volume within the consensus contour, and $\cap$ denotes the volume of overlap between the two contours [7]. Since DSC is a coefficient, the results are logit-transformed prior to statistical analysis to ensure normality [8]. ANOVA testing was used to estimate statistical significance of any correlations between the number of best-matches and logit (DSC).

Results

Table *1* summarizes the results of mean and standard deviation DSC scores, averaged over the five patients, for each structure and number of best matches from one to 10. Bladder and penile bulb show a statistically significant improvement in DSC score which gradually improves as the number of best matches are increased, from DSC of 0.73 with one best match to 0.82 with 10 best matches for bladder ($p < 0.001$), and from a DSC of 0.40 with one best match to 0.54 with 10 best matches for penile bulb ($p = 0.047$). The rectum also showed an observed improvement from 0.56 to 0.67, but this finding was not found to be statistically significant ($p = 0.509$). The remaining structures did not show improvement as the number of best matches was increased. The left and right femoral heads had high, unchanging DSC scores of 0.90-0.93. The prostate bed had relatively low, unchanging DSC scores of 0.62-0.67.

	Mean DC Multiatlas Contour versus STAPLE (SD)						Mean Time Elapsed (s) for Prostate Bed Contour
Atlas #	Bladder	Left Femur	Right Femur	Penile Bulb	Prostate Bed	Rectum	
1	0.73 (0.18)	0.92 (0.01)	0.92 (0.00)	0.40 (0.11)	0.62 (0.06)	0.56 (0.09)	32.0 (4.1)
2	0.71 (0.20)	0.90 (0.04)	0.92 (0.02)	0.46 (0.09)	0.68 (0.06)	0.61 (0.05)	54.4 (6.1)
3	0.78 (0.15)	0.92 (0.01)	0.93 (0.01)	0.43 (0.10)	0.64 (0.07)	0.60 (0.21)	73.6 (3.7)
4	0.78 (0.17)	0.90 (0.03)	0.92 (0.03)	0.48 (0.17)	0.63 (0.09)	0.61 (0.14)	97.0 (10.7)
5	0.79 (0.16)	0.92 (0.01)	0.92 (0.02)	0.50 (0.15)	0.65 (0.07)	0.62 (0.16)	122.2 (13.1)
6	0.79 (0.15)	0.92 (0.01)	0.92 (0.02)	0.54 (0.20)	0.67 (0.05)	0.63 (0.14)	144.8 (16.7)
7	0.80 (0.16)	0.92 (0.02)	0.92 (0.02)	0.54 (0.20)	0.62 (0.09)	0.62 (0.17)	171.4 (18.3)
8	0.82 (0.13)	0.92 (0.01)	0.92 (0.02)	0.52 (0.19)	0.64 (0.08)	0.63 (0.15)	202.4 (21.7)
9	0.84 (0.12)	0.92 (0.02)	0.91 (0.03)	0.55 (0.16)	0.63 (0.07)	0.63 (0.16)	233.8 (23.7)
10	0.82 (0.15)	0.92 (0.02)	0.91 (0.03)	0.54 (0.17)	0.63 (0.08)	0.67 (0.14)	257.4 (25.8)
P Value (1 vs. 10) of Logit(DC)	<0.001	--	0.740	0.047	0.315	0.509	

TABLE 1: Mean DSC results for bladder, femoral heads, penile bulb, prostate bed and rectum with 1-10 best matches.

Also showing p values for 1 vs 10 best matches of logit (DSC) from ANOVA testing. For the left femoral head, there is no p value reported since both 1 and 10 were 0.92. Right-most column shows the mean elapsed time for completion of a prostate bed auto-contour for 1-10 best matches.

Two factors, which may contribute to failure of AABS, are variability of anatomy between patients and poor contrast of structures and their background on CT. AABS relies on similarity between patient anatomies, so that a pre-contoured CT can be used to closely approximate the current structure for automated contouring. Large databases of atlases are used to increase the chances of a best match being similar to the current patient's anatomy. Deformable registration is also performed to improve the match between CT scans. Yet, structures that have poor contrast pose problems for deformable registration algorithms, which rely on CT contrast differences. Both of these factors can be structure-specific. Thus, it is expected that AABS will have variable success depending on the features of the structure involved. High-contrast, consistently shaped structures are likely to be well-suited to AABS techniques, whereas low-contrast structures with variable anatomy are more likely to be poorly suited.

Discussion

The results of this study show a benefit of using MBM in relation to contouring of certain structures. In particular, the bladder and penile bulb contours demonstrated a marked improvement with increased best match number. The MBM approach was able to improve the DSC of penile bulb but not of prostate bed. The femoral heads had the highest DSC scores, which were achieved even with SBM. The high DSC score of femoral heads can be attributed to their relatively consistent shape between different patients, as well as their very high degree of

contrast. The bladder also has a high contrast border, which may explain the relatively high DSC scores of this organ. Yet, the bladder can have a somewhat variable shape, which may explain why the bladder DSC improves gradually as best match number is increased. In contrast, the prostate bed is highly variable from patient to patient, and is also a relatively low-contrast target with ill-defined borders. This may have contributed to the poor DSC scores of prostate bed in this study as well as in previous AABS studies.

There are several limitations to the study. First, only one contouring algorithm was investigated. The overlapping algorithm used to combine the MBM contours into a single contour was one of many possible MBM approaches. The results described in this manuscript are not necessarily generalizable to other AABS software solutions that use different contouring algorithms. Other new algorithms for multi-atlas contouring are emerging and show promising improvements in accuracy [9-11]. Furthermore, the number of atlases in the database was fixed but large. It is not clear whether the contours could be further improved by increasing the number of pre-contoured CT datasets in the AABS library. Another limitation that is present in this study is the uncertainty of the gold standard. As with all studies investigating contouring variability, the definition of a true gold standard is challenging. In this study, the gold standard was generated using the STAPLE algorithm. This approach may reduce the subjectivity of gold standard contour definition. Yet, it is still unclear exactly how gold standard contours are to be best generated in these studies.

Conclusions

Future work includes the identification of other structures that could benefit from a MBM approach. Pirozzi, et al. recently found that a multi-atlas approach for lung cancer resulted in significantly more accurate contours than compared to a single best matched index [10]. The organs in that study included esophagus, spinal cord, heart, left lung, right lung, and trachea. However, it did not specify whether the improvement was seen in all of these structures or just a select few. Future work will also focus on investigation into the effect of increasing the size and/or contents of the AABS library on observed Dice coefficients between automated contours and clinical gold standards.

Additional Information

Disclosures

Human subjects: Consent was obtained by all participants in this study. The University of Western Ontario ethics committee issued approval N/A. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the submitted work.

References

1. Hwee J, Louie AV, Gaede S, et al.: Technology assessment of automated atlas-based segmentation in prostate bed contouring. Radiat Oncol. 2011, 6:110.
2. Mitchell DM, Perry L, Smith S, et al.: Assessing the effect of a contouring protocol on postprostatectomy radiotherapy clinical target volumes and interphysician variation. Int J Radiat Oncol Biol Phys. 2009, 75:990-3.
3. Stapleford LJ, Lawson JD, Perkins C, et al.: Evaluation of automatic atlas-based lymph node

segmentation for head-and-neck cancer. Int J Radiat Oncol Biol Phys. 2010, 77:959-66.

4. van Baardwijk A, Bosmans G, Boersma L, et al.: PET-CT-based auto-contouring in non-small-cell lung cancer correlates with pathology and reduces interobserver variability in the delineation of the primary tumor and involved nodal volumes. Int J Radiat Oncol Biol Phys . 2007, 68:771-8.

5. Young AV, Wortham A, Wernick I, Evans A, Ennis RD: Atlas-based segmentation improves consistency and decreases time required for contouring postoperative endometrial cancer nodal volumes. Int J Radiat Oncol Biol Phys . 2011, 79:943-7.

6. Warfield SK, Zou KH, Wells WM: Simultaneous Truth and Performance Level Estimation (STAPLE): An algorithm for the validation of image segmentation. IEEE Transactions on Medical Imaging . 2004, 23:903-21.

7. Dice LR: Measures of the amount of ecologic association between species . Ecology. 1945, 26:297–302.

8. Berkson J: Application of the logistic function to bio-assay . Jour Amer Statist Assoc . 1944, 39:357-65.

9. Arbisser A, Sharp G, Golland P, Shusharina N: SU-E-J-101: Weighted Voting Method for Multi-Atlas Segmentation in CT Scans. Med Phys. 2012, 39:3675-6.

10. Pirozzi S, Horvat M, Piper J, Nelson A: SU-E-J-106: Atlas-Based Segmentation: Evaluation of a Multi-Atlas Approach for Lung Cancer. Med Phys . 2012, 39:3677.

11. Yang J, Garden A, Zhang Y, Zhang L, Court L, Dong L: WE-E-213CD-09: Multi-Atlas Fusion Using a Tissue Appearance Model. Med Phys. 2012, 39:3961.