Advances in Radiation Oncology (2024) 9, 101417

advances
in radiation oncology

www.advancesradonc.org

Scientific Article

Prospective Evaluation of Automated Contouring for CT-Based Brachytherapy for Gynecologic Malignancies

Abigayle C. Kraus, MD,[a] Zohaib Iqbal, PhD,[b] Rex A. Cardan, PhD,[b] Richard A. Popple, PhD,[b] Dennis N. Stanley, PhD,[b] Sui Shen, PhD,[b] Joel A. Pogue, PhD,[b] Xingen Wu, PhD,[b] Kevin Lee, MD, PhD,[b] Samuel Marcrom, MD,[b,1] and Carlos E. Cardenas, PhD[b,1,*]

[a]*Heersink School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama; and* [b]*Department of Radiation Oncology, University of Alabama at Birmingham, Birmingham, Alabama*

Received 15 May 2023; accepted 30 November 2023

Purpose: The use of deep learning to auto-contour organs at risk (OARs) in gynecologic radiation treatment is well established. Yet, there is limited data investigating the prospective use of auto-contouring in clinical practice. In this study, we assess the accuracy and efficiency of auto-contouring OARs for computed tomography−based brachytherapy treatment planning of gynecologic malignancies.
Methods and Materials: An inhouse contouring tool automatically delineated 5 OARs in gynecologic radiation treatment planning: the bladder, small bowel, sigmoid, rectum, and urethra. Accuracy of each auto-contour was evaluated using a 5-point Likert scale: a score of 5 indicated the contour could be used without edits, while a score of 1 indicated the contour was unusable. During scoring, automated contours were edited and subsequently used for treatment planning. Dice similarity coefficient, mean surface distance, 95% Hausdorff distance, Hausdorff distance, and dosimetric changes between original and edited contours were calculated. Contour approval time and total planning time of a prospective auto-contoured (AC) cohort were compared with times from a retrospective manually contoured (MC) cohort.
Results: Thirty AC cases from January 2022 to July 2022 and 31 MC cases from July 2021 to January 2022 were included. The mean (±SD) Likert score for each OAR was the following: bladder 4.77 (±0.58), small bowel 3.96 (±0.91), sigmoid colon 3.92 (±0.81), rectum 4.6 (±0.71), and urethra 4.27 (±0.78). No ACs required major edits. All OARs had a mean Dice similarity coefficient > 0.86, mean surface distance < 0.48 mm, 95% Hausdorff distance < 3.2 mm, and Hausdorff distance < 10.32 mm between original and edited contours. There was no significant difference in dose-volume histogram metrics (D2.0 cc/D0.1 cc) between original and edited contours (P values > .05). The average time to plan approval in the AC cohort was 19% less than the MC cohort. (AC vs MC, 117.0 + 18.0 minutes vs 144.9 ± 64.5 minutes, $P = .045$).
Conclusions: Automated contouring is useful and accurate in clinical practice. Auto-contouring OARs streamlines radiation treatment workflows and decreases time required to design and approve gynecologic brachytherapy plans.

This work was previously presented at the American Society for Radiation Oncology 2022 annual meeting.

Sources of support: Dr Cardenas received funding from the University of Alabama at Birmingham and the National Institutions of Health/National Cancer Institute award (LRP0000018407) that is related to the work presented in this study.

Research data are stored in an institutional repository and will be shared upon request to the corresponding author.

[1] S.M. and C.E.C. equally contributed to this work as senior authors.

[*] Corresponding author: Carlos E. Cardenas, PhD; Email: cecardenas@uabmc.edu

https://doi.org/10.1016/j.adro.2023.101417

Introduction

Brachytherapy is an integral component for the definitive treatment of locally advanced cervical cancer and other gynecologic malignancies. This modality has matured for over a century, progressing from conventional “Point A” planning with tandem and ovoid applicators to 3-dimensional image-guided brachytherapy leveraging a variety of intracavitary and interstitial applicators.[1] Despite advancements in applicator development, dose prescription methods, and standardization of clinical target delineation, brachytherapy remains a costly, time-consuming task for physicians. Time-driven, activity-based cost evaluations comparing breast and cervical brachytherapy to traditional external beam radiation have shown that brachytherapy is more costly, requiring substantially more time from physicians and medical physicists.[2-4] With the advent of artificial intelligence, brachytherapy may be at another precipice of evolution if automatic contouring, or auto-contouring (AC), is introduced into the treatment planning process.

Delineation of clinical target volumes (CTVs) and organs at risk (OARs) is a time-intensive process for physicians and is known to have large inter- and intraobserver variability. In the past decade, a fourth-generation of deep learning via convolutional neural networks has been introduced to AC CTVs and OARs in a variety of disease sites.[5] Most gynecologic studies have shown clinical acceptability of AC in computed tomography (CT)−based external beam radiation treatment.[6-12] In gynecologic brachytherapy, needles and applicators are implanted into the patient, making it more desirable to shorten the time for contouring. Studies have demonstrated successful auto-digitization of applicators and delineation of CTVs and OARs in both CT- and magnetic resonance imaging (MRI)−based planning.[13-19] However, these retrospective analyses contain limited data on the effect of integration of AC into a clinical workflow where high-dose brachytherapy can be administered shortly after applicator placement.

The objective of our study was to validate the efficiency and accuracy of AC in CT-based gynecologic brachytherapy. First, we evaluated the integration of AC into clinical practice with the hypothesis that AC would decrease physician contour approval time and total planning time. Second, we evaluated the clinical acceptability of AC by physician review of AC and by comparing the geometric and dosimetric differences between original and edited contours. To the best of our knowledge, this is the first study that prospectively evaluates the accuracy and efficiency of AC in gynecologic brachytherapy treatment planning. Here we demonstrate successful implementation of a deep learning contouring tool at a large academic medical center. This addition is important as it validates previously published retrospective studies and provides a framework on how to effectively implement AC in clinical practice.

Methods and Materials

AC model

Institutional review board approval was requested and obtained. Brachytherapy simulation CT scans from 50 patients with gynecologic cancer previously treated at our institution were used to train the AC model. Each patient was treated using an interstitial applicator and received intravenous contrast during CT simulation to improve urethral and bladder delineation. Bladder, small bowel, sigmoid colon, rectum, and urethra clinical contours were retrospectively inspected and edited, if needed, to standardize contouring practices before training the model. A 3-dimensional (3D) U-Net architecture was used to automatically contour bladder, small bowel, sigmoid colon, rectum, and urethra volumes. The deep learning network was trained to 1000 epochs using a 5-fold cross-validation technique and employed early stopping to avoid overfitting. The model was trained using 2 RTX 3090 Nvidia GPUs using a dedicated training workstation (Lambda Inc, San Francisco, CA). During testing, test-time augmentations were used and predictions from the 5 trained model weights are combined for improved contouring accuracy.[20] The training cohort included cases previously treated using Syed (Alpha-Omega Services, Bellflower, CA) or tandem and ring (Varian Medical Systems, Palo Alto, CA) applicators; patients were simulated following institutional guidelines: with a full bladder technique and a Foley catheter inflated with a small amount of contrast material (Omnipaque) for improved urethral identification, and using 1-mm axial slices acquired on a Philips Brilliance 64 CT scanner (Philips Health care, Cleveland, OH).

Architecture and training parameters

A hyperparameter search was performed to identify the optimal parameters (eg, kernel size, resolution steps) on a modified 3D U-Net architecture.[21] Our U-Net model used a residual function (short-connections) similar to that described by Milletari et al[22] and used batch normalization[23] after each 3D convolutional layer. The same architecture was used to train 5 separate model weights with random initialization of weights. These models were used in an ensemble approach to further improve the confidence in the resulting segmentation.[20] The model was trained using the Adam optimizer with a learning rate of

0.001 and early stopping regularization to avoid overfitting of the models. Commonly used data augmentations (eg, translation, rotation) were used during training. The loss was set to the sum of the Dice loss and cross-entropy loss ($L = L_{Dice} + L_{CE}$). Before training, CT scans were preprocessed using linear HU transformation with predefined window/level settings (-600, 1200 Housfield Units) to have values from 0 to 1 (ie, -600 HU → 0 and 1200 HU → 1) and to have isotropic (1 mm × 1 mm × 1 mm) voxel spacing. A patch-based approach was used for training and inference with an input size of 128 × 128 × 48 voxels in the x, y, and z coordinates. During inference, predictions from the 5 models were combined using majority voting.

Clinical implementation of AC

Before clinical implementation, AC quality was evaluated retrospectively on 10 cases not included in our model training cohort. For each patient, ACs were reviewed on a slice-by-slice basis and scored using a 5-point Likert scale, shown in Table 1. Scores of 1 and 2 indicated that the AC was so deficient that it was not helpful to the treating physician. A score of 3 indicated the AC required minor edits for clinical acceptability but was still more useful than starting from scratch. ACs scored a 4 or 5 were considered clinically safe and accurate, as minor edits were due to stylistic difference and not clinically important (score of 4), or no edits were required, and the AC could be used as is (score of 5). If a majority (>75%) of ACs retrospectively evaluated received favorable scores (either 4 or higher), then this would satisfy our quality standard and suggest the model was useful for prospective clinical use.

Workflow efficiency

To assess the efficiency of AC in clinical practice, we compared the planning time of a prospective, AC cohort to the planning time of a retrospectively selected, manually contoured (MC) cohort. The MC cohort cases selected included all intracavitary and interstitial gynecologic cancer cases treated between August 2021 through January 2022. Demographics of these 2 cohorts are given in Table 2. While there are some slight differences in primary disease staging between AC and MC cohorts, these differences were not considered to inherently affect normal tissue contours and contouring time. In both the AC and MC cohorts, planning and contouring were synchronized (ie, performed in parallel) by duplicating the CT scan immediately after import into the treatment planning system (TPS); 1 structure set was designated for planning, and the second structure set was designated for contouring. This enabled a comparison between MC and AC cohorts which was minimally biased by treatment planning differences. This approach of synchronizing planning and contouring for both the AC and MC cohorts, by duplicating the CT image immediately after import into the TPS, was a critical step in our methodology. By assigning 1 structure set for planning and a separate, identical set for contouring, we effectively isolated the contouring process from the planning process. This isolation meant that any time variations observed could be more confidently attributed to the differences in contouring methods (AC vs MC) rather than variations in the planning process itself. For instance, if 1 cohort had received more complex treatment plans by default, it could have inadvertently introduced additional time requirements for contouring due to the complexity of the plans. By standardizing the planning aspect across both cohorts, we minimized such confounding variables. Furthermore, this standardization ensured that each cohort's planning time was not influenced by potentially variable factors such as planner experience, planning complexity, or software performance during the planning phase. As a result, any significant differences in the time required for contouring could be more reliably linked to the efficiency of the AC software as opposed to other aspects of the treatment planning process. This careful control of

Table 1 Scoring criteria of 5-point Likert scale for model accuracy

Score	Description
5: Use as-is	Clinically acceptable, could be used for treatment without change.
4: Minor edits are not necessary	Stylistic differences, but not clinically important. The current contours are acceptable.
3: Minors edits are necessary	Edits are clinically important, but it is more efficient to edit the automatically generated contours than start from scratch.
2: Major edits	Edits are required to ensure appropriate treatment and sufficiently significant that the user would prefer to start from scratch.
1: Unusable	The automatically generated contours are so poor that they are unusable.

The treating physician qualitatively scored each automatically generated contour on its accuracy using a 5-point Likert scale.

Table 2 Descriptive characteristics of manually and automatically contoured cohorts

Characteristic	No. (%)	
	MC cohort (N, %)	AC cohort (N, %)
Age, mean (range), y	52 (30-78)	54 (28-81)
High-risk CTV, mean (range), cc	50 (28-176)	49 (16-189)
Cervical cancer (FIGO stage)		
IA1	2 (6)	2 (7)
IB2	4 (13)	0 (0)
IB3	0 (0)	3 (10)
IIA1	1 (3)	2 (7)
IIA2	5 (16)	4 (13)
IIB	1 (3)	3 (10)
IIIA	0 (0)	1 (3)
IIIB	3 (10)	0 (0)
IIIC1	5 (16)	9 (30)
IIIC2	0 (0)	0 (0)
IV	1 (3)	4 (13)
Endometrial cancer (FIGO stage)		
IA	2 (6)	0 (0)
IIIA	1 (3)	0 (0)
IIIB	1 (3)	0 (0)
Vaginal cancer (FIGO stage)		
IB	1 (3)	0 (0)
IIB	1 (3)	1 (3)
IVA	2 (6)	1 (3)
Ovarian cancer - Stage II	1 (3)	0 (0)
External beam radiation therapy		
Yes	30 (97)	30 (100)
No	1 (3)	0 (0)
Brachytherapy total dose/fx		
24 Gy/3 fx	8 (26)	7 (23)
24 Gy/4 fx	9 (29)	6 (19)
28 Gy/4 fx	12 (39)	15 (48)
25 Gy/5 fx	0 (0)	2 (6)
30 Gy/5 fx	1 (3)	0 (0)
35 Gy/5 fx	1 (3)	0 (0)

Abbreviations: AC= automatically contoured; CTV = clinical target volume; FIGO = International Federation of Gynecology and Obstetrics; fx = fractions; MC = manually contoured,

variables thus allowed for a more direct and unbiased comparison between the MC and AC cohorts, providing clearer insights into the effect of AC implementation on clinical workflow efficiency.

In this study, we define planning time as the time difference between import of the CT image into the TPS (Eclipse 16.1, Varian Medical Systems, Palo Alto, CA), and final physics plan approval, which are recorded in the TPS time stamps. Similarly, we quantified time to contour approval by measuring the time difference between import of the CT image into the TPS and the last modification made to plan’s structure set, as recorded by time stamps within the TPS. The MC group was composed of 31 patients who were treated immediately before the implementation of the AC tool. Mean, SD, and SE of each cohort's planning times were analyzed via the unpaired *t* test. In addition, physician slice-by-slice review of AC OARs and overall contouring time, including additions of clinical targets, were measured for each case in the AC cohort. Here, the overall contouring time captures the amount of time spent by the treating physician editing AC (if needed) and MC of additional structures such as targets.

Prospective use quantitative and qualitative evaluation

Our inhouse developed AC system maintains a database of previously AC structures from clinical cases. Using the unedited, original version of the ACs, we quantitatively evaluated edits made to ACs before clinical use using standard contour overlap and distance metrics. The Dice similarity coefficient (DSC) is a metric that measures twice the volume of overlap between 2 volumes divided by the sum of the volumes $\left(DSC = \frac{2 \times A \cap B}{A + B}\right)$.[24] A DSC equaling 1 describes 2 structures with perfect overlap, whereas a DSC of 0 indicates no overlap between 2 structures. The mean surface distance (MSD), 95% Hausdorff distance (95HD), and max HD were calculated to quantify the overall extent of the edits made to deem the edited ACs clinically usable.[25] A larger distance metric represents greater difference between the original and edited ACs. In addition, we compare the passing rate for dosimetric goals measured on edited contours and original contours. Clinical goals evaluated in standard practice at our institution are included in Table E1.

Furthermore, a single board-certified radiation oncologist with extensive brachytherapy expertise, who was also the treating physician, qualitatively evaluated each OARs' AC. The 5-point Likert scale shown in Table 1 was used in this evaluation. The distribution of Likert scores was analyzed and further evaluated by calculating the mean, SD, and SE for each OAR.

Statistical methods

The unpaired Welch's *t* test was used to test our hypothesis that AC resulted in less time between CT simulation and plan approval. A *P* value < .05 was considered significant. All statistical tests, where applicable, are 2-sided. To elucidate the relationship between the qualitative assessments of the OAR contours and the quantitative changes made during editing, we employed Spearman's rank correlation coefficient. This nonparametric measure was chosen to determine the strength and direction of the association between the ordinal data from the Likert scale evaluations—reflecting the radiation oncologist's subjective judgment of contour quality—and the continuous variables derived from geometric similarity and dose-volume histogram (DVH) metrics. By applying Spearman's rank correlation, we aim to identify whether a higher qualitative score (indicating a better-perceived quality of the AC by the expert) correlates with a closer geometric similarity between the original, ACs and the manually edited, clinically used contours. Furthermore, to compare the DVH metrics derived from the automatically generated contours with those from the clinically used contours, we used the Wilcoxon signed-rank test. All statistics were calculated using SAS version 9.4 (SAS Institute).

Results

In January 2022, our institution designed and implemented an inhouse contouring tool to automatically delineate 5 OARs in CT-based gynecology brachytherapy: the bladder, small bowel, sigmoid colon, rectum, and urethra. Retrospective evaluation resulted in a majority of ACs receiving scores of 4 or 5 (47/60), meeting our prerequisite goal (>75%) for clinical release of the AC tool. From implementation to July 2022, 30 patients were treated using the AC tool and were identified as the prospective AC cohort.

Workflow efficiency

Time to contour approval of the AC and MC cohorts were compared, with results shown in Figure 1A and B. Average contour approval time was 63 minutes (SD, 18 minutes; range, 40-111 minutes) for the AC group and 99 minutes (SD, 68 minutes; range, 29-304 minutes) for the MC group (*P* < .0071). Clinical implementation of AC significantly decreased total planning time by 19% (*P* < .045). Furthermore, the variance in total planning time among AC cases was 33 minutes compared with 65 minutes for the MC group. Lastly, physician slice-by-slice review and editing of ACs took an average 4.6 minutes (SD, 3.1 minutes; range, 1.7-13.4 minutes). Total contouring time, including delineation of clinical targets with fused MRI, took an average 15.8 minutes (SD, 5.3 minutes; range, 7.3-28.2 minutes). All AC OARs were generated and automatically imported into the TPS < 5 minutes.

AC accuracy

The mean DSC, MSD, 95HD, and HD between the original and edited ACs of each OAR are given in Table 3.

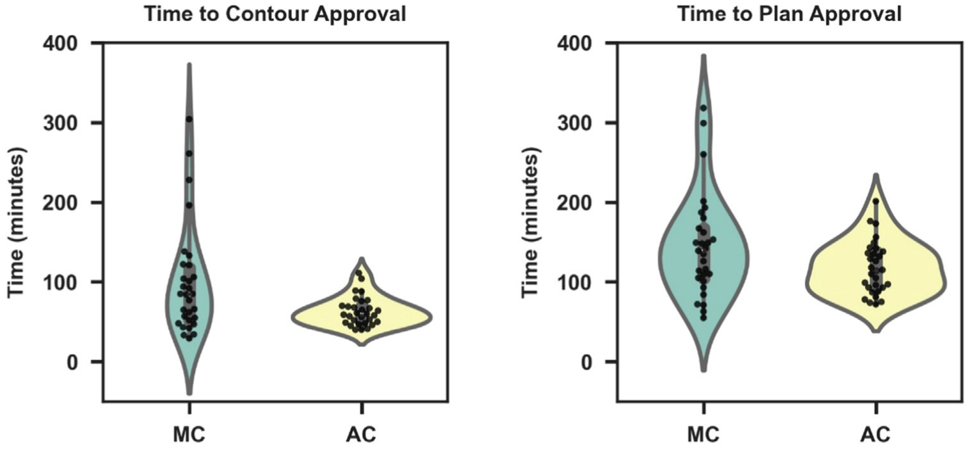

Figure 1 Comparison of contouring and total planning times between manually contoured (MC) and automatically contoured (AC) cases. (A) Average contouring time, defined as the time between computed tomography image importation to treatment planning systems and contour completion, was reduced by 36 minutes in the AC cohort when compared to a MC group (p = .0071). (B) Total planning time was defined as the time between computed tomography image importation to treatment planning systems and plan approval. AC reduced total planning time by 19% (*p* = .045).

Among all OAR contours, 76% had a DSC > 0.95. Urethral contours accounted for 68% of contours with a DSC < 0.95. All structures had a mean MSD < 0.48 mm, a mean 95HD < 3.2 mm, and a mean HD < 10.3 mm. Across all geometric measures, the sigmoid colon and urethra showed the largest difference between the original and edited ACs. D2 cc or D0.1 cc difference between the original and edited ACs of each OAR were evaluated using D2 cc (except for the urethra where D0.1 cc was used). Overall, there was no statistically significant difference between the DVH metrics from the original and edited ACs (*P* values > .05 for all OARs; Table 3).

The distribution of Likert scores by OARs are given in Figure 2. The mean Likert score for each OAR was the following: bladder 4.77 (SD, 0.58), small bowel 3.96 (SD, 0.91), sigmoid colon 3.92 (SD, 0.81), rectum 4.6 (SD, 0.71), and urethra 4.27 (SD, 0.78). No automatically generated contour received a Likert score of 1 or 2. Small bowel ACs were considered the least accurate with 42% of contours receiving a score of 3. However, all OARs, including the small bowel, received a majority of AC scores between 4 and 5.

The Spearman's rank correlation analysis revealed the degree of association between the Likert scores, assigned based on qualitative expert evaluation, and the geometric similarity and DVH metrics that characterize the differences between the original ACs and their edited counterparts for all OARs. There was a weakly positive relationship between Likert score and DSC (r(126) = 0.28; *P* = .002), while there was a negative relationship between Likert score and MSD (r(126) = -0.36; *P* < .001), 95HD (r(126) = -0.36; *P* < .001.), and HD (r(126) = -0.49;

Table 3 Summary of 5-point Likert scores for the original auto-contours, geometric similarity metrics and D2cc difference between the original and edited auto-contours

Structure	Bladder	Small bowel	Sigmoid colon	Rectum	Urethra
5-point Likert score	4.8 ± 0.6	4.0 ± 0.9	3.9 ± 0.8	4.6 ± 0.7	4.3 ± 0.8
DSC (%)	0.99 ± 0.00	0.99 ± 0.01	0.96 ± 0.04	0.98 ± 0.10	0.87 ± 0.10
MSD (mm)	0.05 ± 0.03	0.27 ± 0.44	0.45 ± 0.87	0.16 ± 0.24	0.47 ± 0.88
95HD (mm)	0.4 ± 0.2	1. 5 ± 2.5	3.2 ± 7.1	1.0 ± 2.1	2.4 ± 5.6
HD (mm)	1. 8 ± 0.6	10.2 ± 13.1	10.3 ± 14.3	4.0 ± 5.3	3.6 ± 7.5
Original D2cc (Gy)*	17.8 ± 6.7	7.5 ± 4.4	10.7 ± 4.9	13.0 ± 5.5	8.7 ± 5.8*
Edited D2cc (Gy)*	17.8 ± 6.7	7.9 ± 4.8	11.1 ± 5.4	12.9 ± 5.7	8.7 ± 6.0*
D2cc difference (Gy)*	0.03 ± 0.26	−0.45 ± 1.11	−0.41 ± 1.35	0.07 ± 0.60	−0.08 ± 0.40*
P value (original vs edited D2cc)	.358	.795	.787	.704	.542

Abbreviations: 95HD = 95% Hausdorff distance; DSC = Dice similarity coefficient; HD= Hausdorff distance; MSD = mean surface distance.

*All dose volume histogram metrics were evaluated with D2cc, except for the urethra, for which D0.1cc was used. *P* values comparing original D2cc (Gy) versus edited D2cc (Gy) values were calculated using the 2-tailed Wilcoxon signed-rank test. All measures reported in the format mean ± SD.

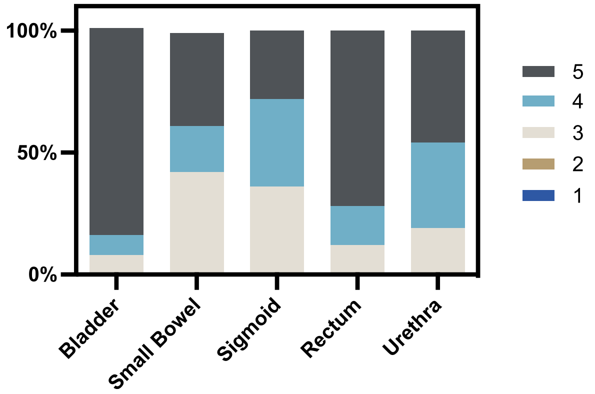

Figure 2 Distribution of 5-point Likert scores by organ at risk. All auto-contours of each organ at risk were evaluated by the treating physician. No auto-contours received lower than a score of 3, meaning no auto-contour required major edits, and all auto-contours were clinically helpful.

$P < .0010$). Lastly, there was a weakly negative relationship between the Likert score and percent change in D2 cc (r(126) = -0.22; $P = .011$).

Discussion

In this study, we constructed an inhouse contouring tool to automatically delineate OARs in CT-based gynecologic brachytherapy, then successfully implemented this tool into clinical practice at a large academic medical center. The contouring tool reduced time to contour approval, total planning time, and time variation among AC cases compared with an MC cohort. When reviewed by the treating physician, all automatically generated contours were scored as clinically helpful, with no contours requiring major edits. Furthermore, when comparing dosimetric parameters, edited ACs had no significant difference in D2 cc compared with their original, unedited counterparts. Overall, this is the first study to discuss the integration of AC into clinical practice and to prospectively demonstrate that contouring via deep learning is an efficient, accurate modality to automatically contour OARs in CT-based gynecologic brachytherapy.

To increase objectivity in the current analysis, we chose to use TPS time stamps to measure contour approval and overall planning times in our retrospective and prospective cohorts. A key factor in this decision was that there were no changes in clinical workflow during AC implementation, meaning that the contouring/planning workflows were not optimized around the AC application. In our clinical practice, the brachytherapy medical physicist, who is present during CT simulation, is in charge of importing the planning CT to the TPS and preparing image data (MRIs, etc) for planning. It is not uncommon for there to be some waiting time during this hand-off process, as the clinical team member may be attending to multiple clinical needs. In the thoughtful design of this study, our team concluded that using the TPS time stamps provides a more “real-life” quantification of time which includes this inherent hand-off time observed in the overall planning process.

The time saving benefit of automated contouring can largely be attributed to the standardization of time to contour approval among cases. Time to contour approval of the MC cohort ranged from 29.0 to 304.0 minutes, while the AC cohort ranged from 71.0 to 111.0 minutes. During MC, radiation oncologists may have additional clinical responsibilities (such as attending clinical needs of the brachytherapy patient, providing simulation CT/stereotactic body radiation therapy coverage, and seeing patients under treatment), that could lead to additional interruptions in the contouring process. The implementation of AC may decrease these interruptions and eliminate large time variation among cases and could explain the difference in “contour approval” times observed in our data. In addition, we found our contouring tool facilitated simultaneous completion of treatment planning requirements; while ACs for OARs were generated outside the TPS, the treating physician could begin delineating clinical targets, and the medical physicist could start digitizing applicators. Synchronizing these tasks promoted a streamlined workflow with quicker planning times. This, in particular, could improve the comfort of patients receiving inpatient or outpatient brachytherapy while also creating the potential for more patients to be treated in the same day.

Our AC tool was highly accurate based on physician scoring and geometric concordance between the original and edited contours, which quantitatively measured the extent of contour editing completed by the treating physician. Bladder ACs consistently scored the highest among all criteria (mean Likert, 4.77; DSC, 0.99; MSD, 0.052 mm; 95HD, 0.39 mm; HD, 1.78 mm), which can largely be attributed to bladder filling protocols at our institution and the consistent shape and size of the bladder. In contrast to the bladder, the architecture of the small bowel and sigmoid colon can vary greatly among patients; thus, we observed less accurate Likert scoring and geometric similarity metrics for these 2 OARs. Similar results have been highlighted in previous studies, given in Table 4. Interestingly, the urethra was the only OAR with a mean DSC < 0.95 and had relatively poor geometric concordance despite a high Likert score versus other structures (4.27). This is most likely due to the urethra’s small size, causing any contour edits to have a disproportionately large effects on the spatial overlay between original and edited contours.

We observed no significant difference in DVH metrics between the original and edited contours among all OARs. Not surprisingly, OARs with relatively poor geometric concordance, specifically the small bowel and

Table 4 Results of geometric evaluation from previous studies

Study	Our method	Mohammadi et al.[18]	Jiang et al.[16]	Zhang et al.[17]	Yoganathan et al.[19]
Radiation modality	CT-based brachytherapy, original vs edited auto-contour	CT-based brachytherapy, manual vs auto-contour	CT-based brachytherapy, manual vs auto-contour	CT-based brachytherapy, manual vs auto-contour	MRI-based brachytherapy, manual vs auto-contour
DSC (%)	Bladder, 0.995 ± 0.003	Bladder, 0.974	Bladder, 0.860 ± 0.086	Bladder, 0.869 ± 0.032	Bladder, 0.90 ± 0.05
	Rectum, 0.984 ± 0.097	Rectum, 0.972	Rectum, 0.858 ± 0.089	Rectum, 0.821 ± 0.05	Rectum, 0.76 ± 0.07
	Sigmoid, 0.964 ± 0.043	Sigmoid, 94.1	Sigmoid, 0.664 ± 0.123	Sigmoid, 0.645 ± 0.079	Sigmoid, 0.65 ± 0.12
	Small bowel, 0.987 ± 0.014		Small bowel, 0.563 ± 0.129	Small bowel, 0.803 ± 0.058	Small bowel, 0.54 ± 0.12
	Urethra, 0.868 ± 0.097				
MSD (mm)	Bladder, 0.052 ± 0.032	-	-	-	-
	Rectum, 0.156 ± 0.238				
	Sigmoid, 0.452 ± 0.870				
	Small bowel, 0.269 ± 0.443				
	Urethra, 0.474 ± 0.876				
95HD (mm)	Bladder, 0.399 ± 0.211	Bladder, 1.64	-	-	Bladder, 6.28 ± 3.42
	Rectum, 1.005 ± 2.112	Rectum, 1.37			Rectum, 8.20 ± 4.07
	Sigmoid, 3.215 ± 7.055	Sigmoid, 1.84			Sigmoid, 20.44 ± 11.70
	Small bowel, 01.446 ± 2.480				Small bowel, 22.3 ± 13.66
	Urethra, 2.447 ± 5.612				
HD (mm)	Bladder, 1.778 ± 0.610	Bladder, 3.51	Bladder, 19.981 ± 11.418	Bladder, 12.1 ± 4.0	-
	Rectum, 3.968 ± 5.322	Rectum, 1.89	Rectum, 12.273 ± 8.080	Rectum, 9.2 ± 4.6	
	Sigmoid, 10.320 ± 14.330	Sigmoid, 2.82	Sigmoid, 98.409 ± 50.984	Sigmoid, 19.6 ± 8.7	
	Small bowel, 10.177 ± 13.123		Small bowel, 68.123 ± 33.781	Small bowel, 27.8 ± 10.8	
	Urethra, 3.629 ± 7.499				

Abbreviations: 95HD = 95% Hausdorff distance; CT = computed tomography; DSC = Dice similarity coefficient; MRI = magnetic resonance imaging; MSD = mean surface distance; HD= Hausdorff distance. There have been limited studies evaluating the accuracy of auto-contouring in gynecologic brachytherapy. If a previous study compared multiple models, the most accurate model is reported. Measures reported in mean ± SD.

sigmoid colon, had the greatest DVH parameter difference between the original and edited contours (mean Δ%D2 cc, 6.2% and 4.7% respectively). Even with this difference, original DVH metrics of the small bowel and sigmoid colon met clinical goals in 100% and 93% of cases, respectively (D2 cc < 18.00 Gy). The rectum had the largest number of cases violating DVH metric planning goals (36.7%); however, this was true for both the original and edited contours, and the percent change in D2 cc among all rectal contours was <2%. This suggests changes to these rectal contours were overall not clinically significant, which supports their mean Likert score of 4.60. While DVH comparisons showed that physician edits to ACs resulted in no statistical difference in DVH metrics for these OARs (Table 3), it remains to be determined whether planning on unedited ACs results in clinically acceptable plans on edited structures.

Few studies have evaluated automated contouring of female pelvic structures for CT-based brachytherapy treatment. Jiang and colleagues estimated AC decreased OAR delineation time by 60 to 75%.[15] Our study strengthens their conclusion as we compared times between manually and automatically contoured cohorts, while they compared AC duration versus the estimated duration of MC at their institution. To the best of our knowledge, this is the only previous study regarding CT-based brachytherapy that has discussed time-saving benefits of AC. However, there are multiple studies covering nongynecologic disease sites that have demonstrated reduction of planning time up to 40%.[26-28]

All previous studies evaluating the accuracy of AC in gynecologic brachytherapy have been retrospective. These studies, given in Table 4, have compared geometric concordance between manually delineated contours used in the patient care and automatic generated contours. Mohammed et al compared contours of 2 deep learning models (ResU-Net and UNet) with previously manually delineated contours of the bladder, rectum, and sigmoid. Jiang et al and Zhang et al evaluated multiple AC models for the bladder, rectum, sigmoid, and small bowel. Among the overlap/distance metrics evaluated in the present study (DSC, 95HD, and HD), our model resulted in greater concordance between compared structures. It is important to note that while we compared original ACs with subsequently edited versions, previous studies evaluated similarity between MCs and ACs. Thus, it is expected our model would produce higher geometric concordance, and we cannot suggest our model demonstrated improved performance. However, our study is the only prospective evaluation of AC's integration into brachytherapy practice, and the concordance between original and edited contours could be a more clinically relevant metric because it quantifies the number of edits required to make a contour ready for patient care. Rather than assessing whether our model worked better than previous models, the results of our study validate previous retrospective data and demonstrate that AC can be successfully used in patient care.

There are several limitations to our study. When comparing time metrics between manual and AC, we used a retrospective selected MC cohort and a prospectively selected AC cohort; thus, there could be confounding bias not accounted for due to nonrandomized sampling. Furthermore, as mentioned above, all previous studies have compared geometric concordance between MCs and ACs, while we have evaluated geometric similarities between an original AC and its physician-edited counterpart. Thus, we are unable to make reliable conclusions on whether our contouring tool outperforms previously described models. Furthermore, the current study does not consider interobserver variation or variation in clinical practices when evaluating edits of the clinically used ACs, and it remains to be determined how other (ie, nontreating) physicians would have edited the ACs before planning.

Conclusion

Brachytherapy is a critical component in the definitive treatment of multiple gynecologic cancers and has evolved over time to improve patient outcomes via improvements in technique, applicators, and treatment planning software. We present the first clinical implementation of automated contouring into brachytherapy practice at a large academic medical center and demonstrate that AC is an accurate and reliable tool to delineate female pelvic structures in CT-based brachytherapy. Looking forward, these results further push the evolution of brachytherapy as AC increases clinical efficiency of gynecologic cancer care and improves the patient experience in the treatment process.

Disclosures

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.adro.2023.101417.

References

1. Srivastava A, Datta NR. Brachytherapy in cancer cervix: Time to move ahead from point A? *World J Clin Oncol*. 2014;5:764-774.

2. Bauer-Nilsen K, Hill C, Trifiletti DM, et al. Evaluation of delivery costs for external beam radiation therapy and brachytherapy for locally advanced cervical cancer using time-driven activity-based costing. *Int J Radiat Oncol Biol Phys*. 2018;100:88-94.
3. Ning MS, Klopp AH, Jhingran A, et al. Quantifying institutional resource utilization of adjuvant brachytherapy and intensity-modulated radiation therapy for endometrial cancer via time-driven activity-based costing. *Brachytherapy*. 2019;18:445-452.
4. Mulherkar R, Keller A, Showalter TN, Thaker N, Beriwal S. A primer on time-driven activity-based costing in brachytherapy. *Brachytherapy*. 2022;21:43-48.
5. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol*. 2019;29:185-197.
6. Rhee DJ, Jhingran A, Rigaud B, et al. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*. 2020;47:5648-5658.
7. Rigaud B, Anderson BM, Yu ZH, et al. Automatic segmentation using deep learning to enable online dose optimization during adaptive radiation therapy of cervical cancer. *Int J Radiat Oncol Biol Phys*. 2021;109:1096-1110.
8. Ma CY, Zhou JY, Xu XT, et al. Deep learning-based auto-segmentation of clinical target volumes for radiotherapy treatment of cervical cancer. *J Appl Clin Med Phys*. 2022;23:e13470.
9. Liu Z, Liu X, Guan H, et al. Development and validation of a deep learning algorithm for auto-delineation of clinical target volume and organs at risk in cervical cancer radiotherapy. *Radiother Oncol*. 2020;153:172-179.
10. Liu Z, Liu X, Xiao B, et al. Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network. *Phys Medica*. 2020;69:184-191.
11. Wang Z, Chang Y, Peng Z, et al. Evaluation of deep learning-based auto-segmentation algorithms for delineating clinical target volume and organs at risk involving data for 125 cervical cancer patients. *J Appl Clin Med Phys*. 2020;21:272-279.
12. Ding Y, Chen Z, Wang Z, et al. Three-dimensional deep neural network for automatic delineation of cervical cancer in planning computed tomography images. *J Appl Clin Med Phys*. 2022;23:e13566.
13. Hu H, Yang Q, Li J, et al. Deep learning applications in automatic segmentation and reconstruction in CT-based cervix brachytherapy. *J Contemp Brachytherapy*. 2021;13:325-330.
14. Ecker S, Zimmermann L, Heilemann G, et al. Neural network-assisted automated image registration for MRI-guided adaptive brachytherapy in cervical cancer. *Z Med Phys*. 2022;32:488-499.
15. Wong J, Kolbeck C, Giambattista J, Giambattista JA, Huang V, Jaswal JK. Deep learning-based auto-segmentation for pelvic organs at risk and clinical target volumes in intracavitary high dose rate brachytherapy. *Int J Radiat Oncol*. 2020;108:e284.
16. Jiang X, Wang F, Chen Y, Yan S. RefineNet-based automatic delineation of the clinical target volume and organs at risk for three-dimensional brachytherapy for cervical cancer. *Ann Transl Med*. 2021;9:1721.
17. Zhang D, Yang Z, Jiang S, Zhou Z, Meng M, Wang W. Automatic segmentation and applicator reconstruction for CT-based brachytherapy of cervical cancer using 3D convolutional neural networks. *J Appl Clin Med Phys*. 2020;21:158-169.
18. Mohammadi R, Shokatian I, Salehi M, Arabi H, Shiri I, Zaidi H. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. *Radiother Oncol*. 2021;159:231-240.
19. Yoganathan SA, Paul SN, Paloor S, et al. Automatic segmentation of magnetic resonance images for high-dose-rate cervical cancer brachytherapy using deep learning. *Med Phys*. 2022;49:1571-1584.
20. Cardenas CE, Beadle BM, Garden AS, et al. Generating high-quality lymph node clinical target volumes for head and neck cancer radiation therapy using a fully automated deep learning-based approach. *Int J Radiat Oncol Biol Phys*. 2021;109(3):801-812.
21. Cardenas CE, Anderson BM, Aristophanous M, et al. Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks. *Phys Med Biol*. 2018;63:215026.
22. Milletari F, Navab N, Ahmadi SA. V-Net: Fully convolutional neural networks for volumetric medical image segmentation. *Paper presented at: 2016 Fourth International Conference on 3D Vision (3DV)*. Stanford, CA; 2016. October 25-28.
23. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Paper presented at: 32nd International Conference on Machine Learning*. Lille, France; 2015. July 6-11.
24. Dice LR. Measures of the amount of ecologic association between species. *Ecology*. 1945;26:297-302.
25. Cardenas CE, Mohamed ASR, Tao R, et al. Prospective qualitative and quantitative analysis of real-time peer review quality assurance rounds incorporating direct physical examination for head and neck cancer radiation therapy. *Int J Radiat Oncol Biol Phys*. 2017;98:532-540.
26. Zhu J, Liu Y, Zhang J, Wang Y, Chen L. Preliminary clinical study of the differences between interobserver evaluation and deep convolutional neural network-based segmentation of multiple organs at risk in CT images of lung cancer. *Front Oncol*. 2019;9:627.
27. Byun HK, Chang JS, Choi MS, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. *Radiat Oncol*. 2021;16:203.
28. Lin L, Dou Q, Jin YM, et al. Deep learning for automated contouring of primary tumor volumes by MRI for nasopharyngeal carcinoma. *Radiology*. 2019;291:677-686.