

8-2021

## AUTOMATION OF RADIATION TREATMENT PLANNING FOR CERVICAL CANCER

Dong Joo Rhee

Follow this and additional works at: [https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations](https://digitalcommons.library.tmc.edu/utgsbs_dissertations)



Part of the [Artificial Intelligence and Robotics Commons](#), and the [Radiation Medicine Commons](#)

---

### Recommended Citation

Rhee, Dong Joo, "AUTOMATION OF RADIATION TREATMENT PLANNING FOR CERVICAL CANCER" (2021).  
*The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences  
Dissertations and Theses (Open Access)*. 1112.  
[https://digitalcommons.library.tmc.edu/utgsbs\\_dissertations/1112](https://digitalcommons.library.tmc.edu/utgsbs_dissertations/1112)

This Dissertation (PhD) is brought to you for free and open access by the The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences at DigitalCommons@TMC. It has been accepted for inclusion in The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences Dissertations and Theses (Open Access) by an authorized administrator of DigitalCommons@TMC. For more information, please contact [digitalcommons@library.tmc.edu](mailto:digitalcommons@library.tmc.edu).

# AUTOMATION OF RADIATION TREATMENT PLANNING FOR CERVICAL CANCER

by

*Dong Joo Rhee, M.S.*

APPROVED:

\_\_\_\_\_  
Laurence E. Court, Ph.D.  
Advisory Professor

\_\_\_\_\_  
Carlos E. Cardenas, Ph.D.

\_\_\_\_\_  
Anuja Jhingran, Ph.D.

\_\_\_\_\_  
Stephen F. Kry, Ph.D.

\_\_\_\_\_  
Surendra Prajapati, Ph.D.

APPROVED:

\_\_\_\_\_  
Dean, The University of Texas  
MD Anderson Cancer Center UTHHealth Graduate School of Biomedical Sciences

AUTOMATION OF RADIATION TREATMENT PLANNING FOR CERVICAL CANCER

A

DISSERTATION

Presented to the Faculty of

The University of Texas

MD Anderson Cancer Center UTHealth

Graduate School of Biomedical Sciences

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

by

Dong Joo Rhee, M.S.  
Houston, Texas

AUGUST, 2021

## Acknowledgments

My Ph.D. was started with Hurricane Harvey. I had to get through Winter Storm Uri and spent about a year and a half of my Ph.D. working from home during the Covid-19 pandemic. Nonetheless, I could safely go through my journey to become a Ph.D. and enjoyed every single aspect of it. It was not possible without the help and support of my advisor, colleagues, friends, and family.

I would like to start by thanking my advisor, Dr. Laurence Court for his unlimited support for me. I cannot think of a way to become a better advisor than you. My career goal is to become a PI and advisor like you.

Thanks to the advisory committee members, Drs. Carlos Cardenas, Anuja Jhingran, Stephen Kry, Surendra Prajapati, and Sastry Vedam. Every piece of advice you gave me made me and my project much better.

Thanks to every member in Court Lab for being great colleagues and friends. I would not be able to go this far without your support. Special thanks to Drs. Tucker Netherton, Callistus Nguyen, Joy Zhang, Jinzhong Yang for your help and advice, Kai Huang for field-in-field code development, Raymond Mumme and Raphael Douglas for data generation and collection, Niki Olanrewaju and Don Hancock for plan review and RapidPlan model development.

Thanks to our internal and external collaborators, Drs. Hannah Simonds, Beth Beadle, Ingrid White, Nazia Fakie, Alicia Sherriff, William Shaw, Chris Trauernicht,

Kristy Brock, Bastien Rigaud, Chidinma Anakwenze, Nicolette Taku, and Adam Garden for providing data and codes, and reviewing the contours and plans.

Thanks to my classmates, Yulun He, Sharbacha Edward, Yasaman Barekatin, Brandon Luckett, Shannon Hartzell. I can still vividly remember the days and nights we spent together at the first-year office. I could successfully complete all the coursework with your help and support.

Thanks to my Korean friends in Houston. Dr. Eun Young Han, I will never forget our Pho tour and your mental support. Thanks Dr. Rumi Lee, Jiah Yang, Kunhee Kim, Dong Ho Shin, Wonsuh Sung for becoming my soju buddies and having fun with me whenever I accomplished something. I was able to overcome homesickness with your support.

Finally, thanks to my family, Tack Gyoo Rhee, Sun Hee Kim, and Yoo Joo Rhee for both mental and material supports. I was not even able to start this journey without your support. I always feel special about having such a great family.

# AUTOMATION OF RADIATION TREATMENT PLANNING FOR CERVICAL CANCER

Dong Joo Rhee, M.S.\*

Advisory Professor: Laurence Court, Ph.D.

Cervical cancer is one of the most common cancer in low- and middle-income countries (LMICs). The mortality rate can be reduced if radiation treatment becomes widely available. However, due to the lack of radiation treatment facilities and human resources, many cervical cancer patients in Africa are not able to receive timely treatments or advanced therapies. To increase the availability of radiation treatment in low-and middle-income countries (LMICs) including African countries, many attempts have been made to reduce the cost of medical linear accelerators. However, increasing the number of treatment machines would not instantly resolve the issues, as there would be insufficient trained and experienced medical staff to create high-quality radiation treatment plans. To fill the gap, we automated the entire radiation treatment planning process by automating the contouring, planning, and quality assurance (QA) processes in cervical cancer radiation treatment.

To create a high-quality radiation treatment plan, accurate contours must be generated first. We used convolutional neural networks (CNN), one of the most effective deep learning techniques for image processing, to create an auto-contouring model for 3 clinical target volumes (CTVs) and 12 normal structures

for cervical cancer radiation treatment and showed that 93% of the automatically generated contours were clinically acceptable.

For planning, we automated 3 treatment delivery techniques including 2D 4-field-box, 3D conformal radiation therapy (3D-CRT), and volumetric-modulated arc therapy (VMAT). We also automated the field-in-field (FIF) technique to reduce hotspots in the automatically generated 4-field-box and 3D-CRT plans. Each beam delivery technique was evaluated on 35 retrospective patient datasets from South Africa, and on average, 95% of the automatically generated plans were clinically acceptable.

As clinically unacceptable plans were mostly caused by inaccurately generated contours, the quality of the contours should be verified to ensure the quality of the plans. To automatically detect clinically unacceptable contours, we developed an automated contour QA method using two independently developed auto-contouring systems. We hypothesized that if one of the two independently developed auto-contouring systems failed, the discrepancy between the two contours would be substantial enough to be identified by measuring the similarity between the two contours. We found that more than 90% of the contouring errors can be detected with an appropriate choice of similarity metrics.

In conclusion, the majority of the automatically generated contours and plans for cervical cancer radiation treatment were clinically acceptable. Furthermore, errors in the contours can be flagged by the contour QA method. The entire system has been implemented to the Radiation Planning Assistant (RPA), a web-based toolbox for automated planning, to help cervical cancer patients in LMICs.

## Table of Contents

Approvals .....	i
Title Page .....	ii
Acknowledgments.....	iii
Table of Contents .....	vii
List of Illustrations .....	xi
List of Tables .....	xv
Chapter 1 : Introduction .....	1
Chapter 2 : Purpose and Central Hypothesis.....	3
Chapter 3 : Automatic contouring system for cervical cancer using convolutional neural networks .....	5
3.1 Introduction .....	5
3.2 Methods .....	6
3.2.1 Training parameters .....	8
3.2.2. Bony structures .....	10
3.2.3. CTVs.....	10
3.2.3.1 Primary CTV.....	10
3.2.3.2 Pelvic lymph node CTV .....	12
3.2.3.3 Para-aortic lymph node (PAN) CTV .....	13
3.2.4 Organs at risk .....	14



3.2.5 Test dataset .....	17
3.3 Results .....	19
3.3.1 Model accuracy .....	19
3.3.2 Physician review .....	20
3.4 Discussion .....	24
3.4.1. Quantitative results.....	25
3.4.2. Failure cases from physician's review .....	28
3.5 Conclusion.....	31
Chapter 4 : Development of a quality assurance system to detect errors in automatically generated contours for cervical cancer. ....	32
4.1 Introduction .....	32
4.2 Methods.....	33
4.2.1 Two deep-learning-based auto-contouring systems .....	33
4.2.2 Quantitative metrics.....	34
4.2.3 Error detection model with support vector machine for the soft-tissue structures .....	35
4.2.4 Data acquisition.....	38
4.2.5 Error detection model for the bony structures .....	42
4.3 Results .....	42
4.3.1 Soft-tissue structures.....	42
4.3.1.1 Single-metric analysis.....	42
4.3.1.2 Multi-metric analysis.....	51

4.3.2 Bony structures .....	54
4.4 Discussion .....	55
4.5 Conclusion.....	61
Chapter 5 : Automated radiation treatment planning for cervical cancer radiation treatment.....	62
5.1 Introduction .....	62
5.2 Methods .....	64
5.2.1 4-field-box plans with bony landmarks .....	64
5.2.2 3D-CRT plans with the CTV contours .....	65
5.2.3 Field-in-field technique .....	66
5.2.3.1 MU optimization algorithm.....	69
5.2.4 VMAT .....	70
5.2.5 Plan review .....	71
5.3 Results .....	73
5.3.1 4-field-box .....	75
5.3.2 3D-CRT .....	75
5.3.3 VMAT .....	76
5.4 Discussion .....	77
5.4.1 4-field-box .....	78
5.4.2 3D-CRT .....	78
5.4.3 VMAT .....	79

5.5 Conclusion.....	80
Chapter 6 : Discussion and Conclusion .....	81
6.1 Project Summary.....	81
6.2 Study Limitations and Future Directions.....	82
6.3 Conclusion.....	85
Appendix A.....	86
Bibliography .....	87
Vita .....	101

## List of Illustrations

Figure 1. Application of the CNN-based classification and segmentation models to a CT scan. (a) The presence or absence of the organ of interest (in this case, femurs) was evaluated on each CT slice, (b) the cranial-caudal extent of the organ of interest was determined with post-processing, and (c) the slices that were classified to contain the organ of interest were used in the segmentation model to generate contours.....	8
Figure 2. Segmentation using cropped 3D images for better accuracy. (a) Resize the CT from 512x512 to 256x256 pixels and then segment the organ of interest and find the center of mass, (b) crop the region around the segmented organ on the original 512x512 CT scan, and (c) re-segment the organ of interest on the cropped image. ....	12
Figure 3. Flowchart of the semi-automated data curation method to identify incorrect clinical contours. Data were randomly split into 2 groups, and 2 auto-segmentation models were trained with each dataset. Then, each segmentation model was applied to the other group of data to create contours. If the Sørensen-Dice similarity coefficient (DSC) was lower than the threshold value, the original contour was manually reviewed and deleted if incorrect. ....	15
Figure 4. Overall flowchart of the developed auto-contouring system for cervical cancer. (a) The slice-by-slice classification was conducted to identify CT slices that contain a target structure, and the process is visually demonstrated in Figure 1. (b) Bony structures were contoured as described in 2.B. (c) Spinal cord, vagina, and PAN CTV were contoured with the 2D FCN-8s segmentation architecture. (d) Other structures (the organs-at-risk and the primary and the nodal CTVs) were contoured as demonstrated in Figure 2. (e) Extra steps were required for the nodal CTV contours as described in 2.C.2. ....	16
Figure 5. The distributions of Sørensen-Dice similarity coefficients (DSC) between the ground truth and the automatically generated contours of 14 structures. ....	22

Figure 6. The outlier contours from the internal test dataset. The ground truth contours (green) and outliers (red) are given for (a) primary CTV (Sørensen-Dice similarity coefficient [DSC] = 0.43), (b) bladder (DSC = 0.21), (c) L4 and L5 vertebral bodies (DSC = 0.0 each), and (d) PAN CTV (DSC = 0.43). .....	22
Figure 7. Examples of automatically generated contours (red) versus ground truth (green) from physician's manual review of contours for the primary CTV, bladder, and rectum. ....	24
Figure 8. Example of missing the vagina contour due to the gas-filled rectum .....	29
Figure 9. The aortic bifurcation is clearly defined in the red box in (a), whereas the aortic bifurcation is barely identifiable in the red box in (b). The adjustment of the window level did not improve the visual inspection.....	30
Figure 10. Demonstration of the how the SVM-based contour QA model was developed. The red and green data points represent the clinically unacceptable and acceptable contours, respectively. (a) Using a single metric and determine the threshold between the clinically acceptable and unacceptable contours. Since the input is one dimensional, SVM with the linear kernel can only be used. (b) Using a combination of the metrics to determine the threshold either with a linear or a non-linear kernel.....	37
Figure 11. Examples of manually generated, clinically acceptable (green) and unacceptable (red) contours for the (a) UteroCervix, (b) Bladder, (c) Right Kidney, and (d) Rectum. (e) The reference auto-contour (yellow) was clinically unacceptable when the verification auto-contour (blue) was clinically acceptable. (f) Both the reference and the verification auto-contours were clinically unacceptable. ....	40
Figure 12. (a) Diagram demonstrating the data acquisition process for automatic contour QA model development, and (b) demonstrating that each set was split equally into 3 for 3-fold cross-validation. ....	41

Figure 13. Average accuracies of the contour QA model with an individual metric for each structure with various penalty parameters, C. The error bar represents $\pm 1$ standard deviation from three-fold cross-validation. ....	44
Figure 14. The ROC curves with a surface DSC with a tolerance of 2 mm, the best metric to predict the clinical acceptability of the automatically generated contours. ....	50
Figure 15. Average accuracies of SVM model with multiple metrics for each structure. The error bar represents $\pm 1$ standard deviation. Four different kernels (linear, polynomial, radial basis function (rbf), and sigmoid) were tested.....	52
Figure 16. Distributions of DSC and HD <sub>100</sub> for the clinically acceptable (green markers) and unacceptable (red markers) cases for the 6 bony structures. The thresholds between the clinically acceptable and unacceptable contours were manually determined based on the distributions.....	55
Figure 17. The surface DSC distributions of the clinically acceptable and unacceptable kidney contours with (left) and without (right) the manually generated contours. The thresholds can be confidently determined with the manual contours, whereas the threshold can be anywhere between the blue and red dashed lines without the manual contours due to an insufficient amount of data. ....	59
Figure 18. Synthetic PTV structure was defined based on the beam apertures for the 4-field-box plans. Firstly, the beam path from each beam angle was converted into a 3D binary mask, and then the volume overlapped by each mask was defined to be the region of hot spot detection (RHD). Finally, the synthetic PTV was created from 7 mm shrinkage from the RHD, the overlapped volume. ....	65
Figure 19. Workflow of the automated 3D-CRT system for cervical cancer. The PTV was derived from the automatically generated CTVs. The beam apertures were determined with a	

user-defined uniform margin (7 mm in this study) around the projected PTV. The dose was calculated with a pre-defined MU.....	66
Figure 20. Overall flowchart of the FIF automation algorithm. The processes in the white, blue, and yellow boxes were conducted in the RPA system, treatment planning system (Eclipse), and our FIF automation algorithm, respectively. ....	68
Figure 21. Demonstration of the beam shape determination for a sub-field. The pink region represents the hot spot, and the beam shape of the sub-field was defined not to include the pink region from the projected gantry angle. ....	68
Figure 22. PTV and RHD definitions for the FIF optimization algorithm. ....	69
Figure 23. Demonstration of the customized plans for Reviewer #1 (left) and Reviewer #2 (right) for (a) 4-field-box, (b) 3D-CRT, and (c) VMAT techniques. The thick lines represent the PTV (red), 105% isodose line (pink), 100% isodose line (yellow), and 95% isodose line (green). ....	74
Figure 24. Examples of the plans scored 3 or lower. The thick yellow lines represent 100% isodose line and the thick red lines represent the PTV. (a) 4-field-box plan scored 3 because of the excessive dose to the bowel. (b) 3D-CRT plan scored 3 as the PTV (red) was not fully covered by 100% isodose line near the bowel. (c) VMAT plan scored 2 as the PTV (red) was not fully covered by 100% isodose line near the rectum. ....	77

## List of Tables

Table 1. The number of CT scans used for training and validation for each structure .....	17
Table 2. Sørensen-Dice similarity coefficients (in percentage), mean surface distance (in cm), and Hausdorff distance (in cm) between our CNN-based model and clinical contours from 140 internal test CT scans .....	21
Table 3. Qualitative scores of the automatically generated contours on 30 external CT scans.	23
Table 4. Summary of CNN-based auto-contouring results for normal structures in pelvic CTs from other groups .....	27
Table 5. List of the combined metrics used in the multi-metric analysis .....	38
Table 6. Changes in accuracy when applying the average threshold of various structures instead of optimal thresholds for each structure. ....	46
Table 7. Overall accuracies, sensitivities, and specificities with maximized accuracy through the SVM, fixed sensitivity of 0.90, and fixed sensitivity of 0.95 when surface DSC with a thickness of 2 mm was used. ....	47
Table 8. AUCs of each structure and each metric. 95% confidence interval (CI) for AUCs were derived with the bootstrapping method with n=2000.....	49
Table 9. Overall accuracies, sensitivities, and specificities from the single-metric and multi-metric analyses, when SVM was used with the linear kernel. ....	53
Table 10. Likert scale to score automatically generated radiotherapy plans .....	72
Table 11. Physician scoring results for each technique with each review session. The plan criteria for the coverage and the maximum dose were presented. The reviewer numbers are arbitrarily assigned. ....	73



## Chapter 1 : Introduction

Cervical cancer is the second most common cancer for women in Africa and the 5-year survival rate is only 21% versus 70% in the United States [1], [2]. The most cost-effective treatment to increase the survival rate of cervical cancer patients in low- and middle-income countries (LMICs) is radiation treatment [3]. However, due to the lack of radiation treatment facilities and human resources, many cervical cancer patients in Africa are not able to receive timely treatments or advanced therapies. To increase the availability of radiation treatment in low- and middle-income countries (LMICs) including African countries, many attempts have been made to reduce the cost of medical linear accelerators [3]–[6]. However, even if medical accelerators are provided to hospitals in LMICs immediately, LMICs will still lack radiation oncologists, medical physicists, and dosimetrists who are needed to create high-quality radiation treatment plans [6].

A potential solution for the lack of experts in radiation treatment planning is to automate the radiation treatment planning process. An automatic radiation treatment planning system for cervical cancer patients will address the shortage of treatment planning staff, and subsequently increase the survival rate for cervical cancer patients in LMICs. Some work on automated radiation treatment were conducted, but they are mostly restricted to a 2D or a simple 3D radiation treatment in limited conditions [7], [8], mostly due to inadequate quality of currently available auto-contouring systems. We used convolutional neural networks (CNN), one of the most effective deep learning techniques for image processing [9]–[13], to create an auto-contouring model for 3 CTVs and 12 normal structures in the female pelvis.

In deep learning, the number and the accuracy of training data determine the accuracy of the final model. We have collected the largest number of clinical data to train the CT-based auto-contouring model to date [16]. We have applied a unique deep learning technique [18] to curate a large number of training data to improve the accuracy of the model with minimal human efforts. Furthermore, we have developed a method to verify the clinical acceptability of the automatically generated contours through comparison between two independently generated auto-contours. The similarity between the two automatically generated contours were quantified to determine whether the reference contour was correctly generated. This method will substantially reduce the risk of delivering an incorrect plan to a patient.

Considering the different clinical conditions in various countries and hospitals, we developed a CT-based auto-planning system for cervical cancer patients with 3 treatment techniques (Bony-structure-based 4-field-box therapy, soft-tissue-based 3D conformal radiation therapy (3D-CRT), and volumetric modulated arc therapy (VMAT)). Furthermore, the field-in-field (FIF) technique has also been automated to reduce excessively high doses in 4-field-box and 3D-CRT plans.

The goal of this study was to fully automate radiotherapy planning process for cervical cancer with the three different techniques. To determine if the automatically generated plans were clinically acceptable for actual patient treatment, we generated plans on retrospective patient CT scans from our partner hospitals in Africa. The automatically generated plans were evaluated by experienced radiation oncologists in MD Anderson Cancer Center and the partner hospitals in Africa.

## Chapter 2 : Purpose and Central Hypothesis

### **Central Hypothesis:**

We hypothesized that 90% of the automatically generated treatment plans with 3 different techniques for cervical cancer will be clinically acceptable and the automatic quality assurance tool can identify at least 90% of clinically unacceptable plans while specificity is higher than 80%.

### **Specific Aim 1:**

**Aim:** Automate contouring for cervical cancer radiation treatment.

**Hypothesis:** 90% of the contours generated by the auto-contouring system are clinically acceptable.

The work towards Aim 1 is presented in Chapter 3: Automatic contouring system for cervical cancer using convolutional neural networks.

### **Specific Aim 2:**

**Aim:** Automate radiation treatment planning for cervical cancer treatment.

**Hypothesis:** 90% of the automatically generated plans are clinically acceptable.

The work towards Aim 2 is presented is presented in Chapter 5: Automated radiation treatment planning for cervical cancer radiation treatment.

**Specific Aim 3:**

**Aim:** Develop a quality assurance tool to test the validity of cervical cancer radiation treatment plans.

**Hypothesis:** The overall QA system has the sensitivity higher than 90% with the specificity higher than 80%.

Aim 3.1: Develop a quality assurance system for the automatic contouring system.

The work towards Aim 3.1 is presented in Chapter 4: Development of a quality assurance system to detect errors in automatically generated contours for cervical cancer.

## Chapter 3 : Automatic contouring system for cervical cancer using convolutional neural networks

This chapter is based upon the following article:

Rhee, D.J., Jhingran, A., Rigaud, B., Netherton, T., Cardenas, C.E., Zhang, L., Vedam, S., Kry, S., Brock, K.K., Shaw, W., O'Reilly, F., Parkes, J., Burger, H., Fakie, N., Trauernicht, C., Simonds, H. and Court, L.E. (2020), Automatic contouring system for cervical cancer using convolutional neural networks. *Med. Phys.* doi:10.1002/mp.14467

### 3.1 Introduction

Manual contouring of tumors and normal structures is a very labor-intensive and time-consuming part of the radiation treatment planning process [14], [15]. "Wrong or inaccurate" contours drawn by physicians and dosimetrists constitute the highest and seventh-highest risk factors for failure of photon/electron external beam radiation treatment, respectively [16]. Most of these errors could be avoided if an accurate and reliable auto-contouring tool were available. In the past, various algorithms have been evaluated for the development of auto-contouring tools, with mixed success [17]–[19]. With the advent of deep learning, more specifically, convolutional neural networks (CNNs), this movement has been accelerated as CNNs outperformed most of the other algorithms in various segmentation tasks [13]. As a result, CNN-based auto-contouring systems for computed tomography (CT) images have been developed for various body sites, such as the head and neck [12], [20]–[23], thoracic region [24]–[27], abdomen [28]–[30], and pelvis [9], [31], [40]–[42], [32]–[39].

Although these approaches have generally been very successful, they are not yet accessible to cancer treatment centers where they would be most useful – those with limited resources that see a large number of cervical cancer patients, such as in Africa and other low- and middle-income countries (LMICs). In fact, cervical cancer is the second most common cancer in women in Africa [1], [2], and the most cost-effective treatment that increases the survival rate of cervical cancer patients in LMICs is radiation treatment [43]. To fill this gap, the Radiation Planning Assistant (RPA; [rpa.mdanderson.org](http://rpa.mdanderson.org)) [44], a web-based, fully automated radiotherapy contouring and planning generation system, is being developed to address the shortage of treatment planning staff and subsequently increase the survival rate for cancer patients in LMICs.

Although the potential of deep learning-based auto-contouring systems for pelvic structures has been explored in several previous studies, most of them were focused on prostate cancer [9], [32]–[34], [41], [42], and only a few papers have published results for the female pelvis [35], [36]. In this study, we developed an auto-contouring system that can contour the clinical treatment volumes (CTVs) and normal structures that are necessary for various cervical cancer radiation treatment planning techniques. The auto-contouring system in this work will be implemented with RPA to automatically generate high-quality radiation treatment for cervical cancer patients in LMICs.

### 3.2 Methods

Our CNN-based auto-contouring tool was developed to generate contours for 3 CTVs and 12 normal structures in the female pelvis: primary CTV, nodal CTV, PAN

CTV, bladder, rectum, spinal cord, left and right femurs, left and right kidneys, bowel space, sacrum, pelvic bone, L4 vertebral body, and L5 vertebral body. These structures were categorized into three groups: bony structures, organs at risk (OARs), and CTVs. These are the structures required to automate 4-field box, 3D-CRT, IMRT, and VMAT plans for cervical cancer [7], [45], [46].

First, the Inception-ResNet-V2 [47] classification architecture was trained to identify the extent of the structure in the cranial-caudal direction, as shown in Figure 1(a) and (b). This approach was taken to address the GPU memory limitation issue as well as to improve the accuracy of the automatically generated contours by allowing the subsequent segmentation model to process a restricted field of view [21]. Second, the segmentation models were applied to the CT slices that were classified to contain the organ of interest, as shown in Figure 1(c). Both the classification and the segmentation models were trained independently for each structure.

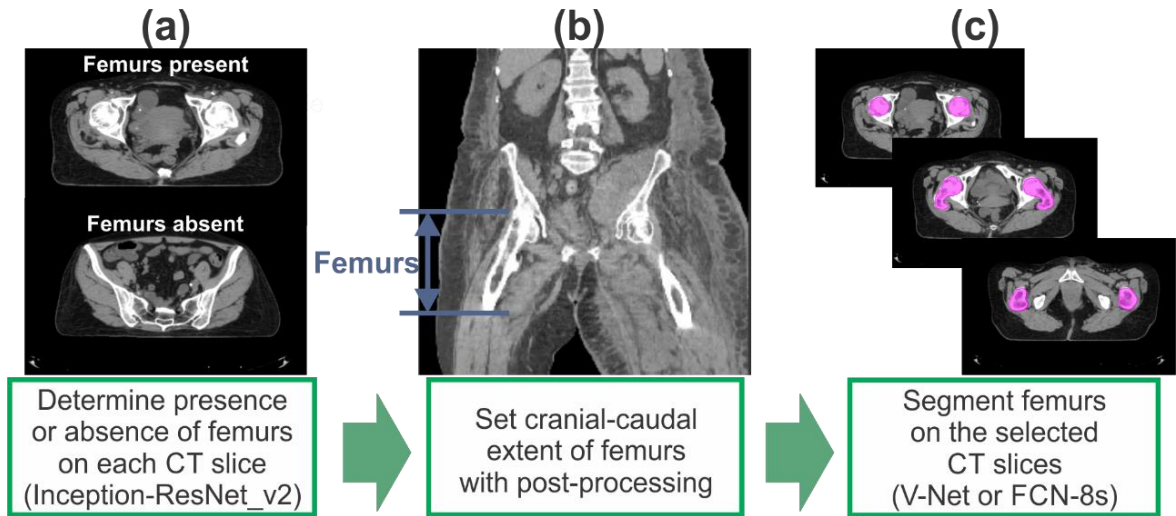


Figure 1. Application of the CNN-based classification and segmentation models to a CT scan. (a) The presence or absence of the organ of interest (in this case, femurs) was evaluated on each CT slice, (b) the cranial-caudal extent of the organ of interest was determined with post-processing, and (c) the slices that were classified to contain the organ of interest were used in the segmentation model to generate contours.

### 3.2.1 Training parameters

For the training and validation data, 2254 female pelvic CT scans from cancer patients who received radiation treatment from September 2004 to June 2018 at The University of Texas MD Anderson Cancer Center were used. Furthermore, 210 CT scans with kidney contours from the 2019 Kidney Tumor Segmentation Challenge (KiTS19) were added to our training data. The CT scans had pixel sizes in the transverse plane that ranged from 0.754 mm to 1.367 mm and slice thicknesses from 2.0 mm to 3.0 mm, except for 8 CT scans (3 were 5 mm, 3 were 4 mm, 1 was 1.5 mm, and 1 was 1.0 mm thick). All data were resampled to have the same voxel size of 1.17



mm x 1.17 mm x 2.5 mm. The CT numbers lower than -1000 HU and higher than 3000 HU were clipped and then linearly shifted to have a 0 to 4000 pixel intensity range.

An NVIDIA DGX Station with four V100 GPUs (16 GB RAM) was used to train our models. The loss function for the segmentation models was the Sørensen-Dice similarity coefficient (DSC) [48], [49], as this was our metric to determine the accuracy of the segmentation model. A weighted cross-entropy was used as a loss function for the classification model to compensate for the data imbalance between the number of slices with and without the organ of interest. The weight was determined to be the ratio of the number of absences to the number of presences. The Adam optimizer [50] was used as an optimization algorithm. The Adam optimizer's parameters, beta1, beta2, and epsilon, were set to 0.9, 0.999, and  $10^{-8}$ , respectively.

To select the 2D and 3D CNN segmentation architectures, we did a preliminary study on the spinal cord for 2D and the left kidney for 3D. The vanilla DeepLabv3+ [51] and the FCN-8s [10] with additional batch normalization layers at the end of every convolutional layer were trained to segment the spinal cord in 2D. The mean  $\pm$  standard deviation DSC were  $0.87 \pm 0.03$  and  $0.90 \pm 0.02$ , for the vanilla DeepLabv3+ and the modified FCN8-s, respectively, so the modified FCN-8s was chosen for our model. Similarly, the 3D U-Net [52] and the 3D V-Net [9] segmentation architectures were trained to segment the left kidney on CT images resized to have a 256x256x60 dimension. We added batch normalization layers at the end of every convolutional layer for both architectures. The mean  $\pm$  standard deviation DSC were  $0.93 \pm 0.04$  and  $0.93 \pm 0.04$ , for the U-Net and the V-Net, respectively. As there was no significant

difference between the two architectures, we chose the V-Net, which has residual connections in each stage.

### 3.2.2. Bony structures

The contours of the four bony structures (pelvic bone, sacrum, L4 vertebral body, and L5 vertebral body) were generated on 370 CT scans to train and validate the auto-contouring model. The pelvic bone was defined to be the traditional pelvic bone without the sacrum, as the sacrum was contoured as a separate structure. All the bony structure contours were automatically generated with a multi-atlas-based auto-contouring system (MACS) [17], [18], [53] first, and the automatically generated contours were manually reviewed and revised if necessary.

V-Net [9], a CNN-based 3D segmentation architecture, was used to segment the 4 bony structures. The input image for the segmentation architecture was resized to Nslicex256x256. A single segmentation model was used to contour the adjacent L4 and L5 vertebral bodies simultaneously. For data augmentation purposes, horizontal flip and rotation with random angles between  $-30^{\circ}$  and  $30^{\circ}$  along the axial axis were applied for these structures.

### 3.2.3. CTVs

#### 3.2.3.1 *Primary CTV*

We used the primary CTV described in the GEC-ESTRO II guideline [54] as the reference of the primary CTV for our cervical cancer patients. The primary CTV is defined to include the entire uterus and cervix (UteroCervix), the 20mm axially long vagina from the most inferior position of the UteroCervix, and the lateral parametria. To

train the model, 406, 490, and 487 UteroCervix, vagina and parametria contours, respectively, were either curated from clinical contours or manually generated from scratch by 4 physicians at MD Anderson Cancer Center.

V-Net was used to segment the UteroCervix and the parametria. Although the classification model restricted the field of view of the input images, the GPU memory was not sufficient to train the full-resolution CT images. To overcome this problem, we resized the input image to 256x256 pixels in the transverse plane, segmented the UteroCervix or the parametria, and estimated the center of mass of the structure. Then, we cropped the box that fully enclosed the structure and centered it on the center of mass of the prediction on the original CT scan with a 512x512 pixel image size. Finally, we applied the V-Net segmentation model to the cropped 3D image, as shown in Figure 2. This way, the final contour is predicted on the limited CT field of view with the original spatial resolution. This approach was inspired by the method proposed by Feng et al. [24] and applied to the rest of the CTVs and OARs that were segmented with the 3D segmentation model.

Although the cropped images were supposed to be centered at the center of mass of the organ in the prediction, the center was randomly chosen while training the model in each epoch for the data augmentation purpose. Furthermore, the random rotation between  $-30^{\circ}$  and  $30^{\circ}$  along the axial axis and the horizontal flip were also used for data augmentation. The same data augmentation techniques were applied to train the segmentation models for other CTVs and OARs.

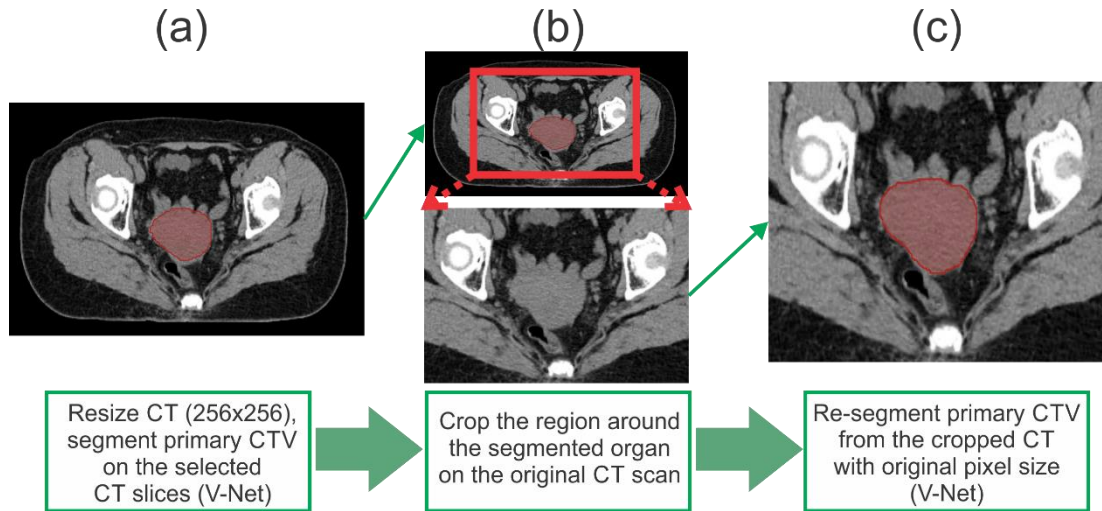


Figure 2. Segmentation using cropped 3D images for better accuracy. (a) Resize the CT from 512x512 to 256x256 pixels and then segment the organ of interest and find the center of mass, (b) crop the region around the segmented organ on the original 512x512 CT scan, and (c) re-segment the organ of interest on the cropped image.

The vagina auto-contouring model was developed using the 2D FCN-8s [10] model. To match with the GEC-ESTRO protocol, we applied the vagina auto-contouring model to the CT slices that were located axially within 20 mm from the inferior end of the UteroCervix contour. Then, the UteroCervix, parametria, and vagina contours were merged to generate the primary CTV contour. Both quantitative and qualitative analyses were performed on the final primary CTV contour, instead of on the individual structures.

### 3.2.3.2 Pelvic lymph node CTV

The nodal CTV covers the common iliac, external iliac, internal iliac, obturator, and presacral nodal regions as described in the GEC-ESTRO II guideline [54] for

intermediate-risk nodal CTV. To provide data for the training process, 250 nodal CTV contours were contoured by the same 4 physicians who contoured the primary CTV and later peer-reviewed to ensure high accuracy and consistency. As the lymph nodes and vessels are small and have CT numbers similar to those of muscles, a 3D segmentation model can sometimes miss a small part of the lymph nodes. To prevent this, FCN-8s [10], a 2D segmentation architecture, was also trained to auto-contour the nodal CTVs. The CT slices that were predicted to contain the nodal CTV contours by the 3D segmentation model were given to the 2D segmentation model for slice-by-slice prediction. In prediction, the sum of the nodal CTV contours from the 2D and 3D models was used as a final contour.

The superior border of the intermediate nodal CTV was determined at one slice below the bifurcation of the common iliac artery. To locate the superior border more accurately, a 3D segmentation model for the aorta near the bifurcation region was trained with 296 CT scans. The segmentation model was applied to a cropped region around the automatically generated L4 vertebral body contour to limit the field of view.

#### *3.2.3.3 Para-aortic lymph node (PAN) CTV*

The PAN CTV covers the para-aortic lymph nodes from the level of the renal veins to the aorta above the aortic bifurcation (i.e., one slice above the superior slice of the nodal CTV). In order to gather data sufficient for the PAN CTV segmentation model, we used 146 clinical contours, and all the contours were manually curated and revised if necessary. FCN-8s was used to auto-contour the PAN CTVs.

### 3.2.4 Organs at risk

OARs for cervical cancer radiation treatment include the bladder, rectum, spinal cord, left and right femurs, and left and right kidneys. The training and validation data for the OARs were acquired from clinical contours of the 2254 CT scans. Contours for each structure were considered to maximize the amount of data, and thus, the number of available structures in a single patient's data varied from 1 to 7. The total number of CT scans used for training and validation for each structure is shown in Table 1. Of these scans, 80% were used for training, and 20% were used for validation. Since the classification and the segmentation models were trained independently for each structure to avoid the class imbalance problem [55], the imbalance in the number of training data for each structure did not influence the model accuracy. As the contours were collected solely on the basis of their labels, review of these contours was required to confirm their accuracy. Owing to the substantial number of contours, we proposed a semi-automatic data curation method instead of manual review, as described in Figure 3. First, "unreviewed" contours and the corresponding scans were divided in half. Two CNN-based segmentation models, one for each half, were trained, and the contours were predicted on the other half of the dataset. If the DSC between the clinical contours and the predicted contours was lower than an arbitrarily determined threshold value (DSC=0.7 for the rectum, 0.8 for the remaining OARs), the original contour was manually reviewed, and any incorrect clinical contours were removed from the training dataset. Once the entire set of training data was reviewed, we repeated the process with the "refined" dataset from the beginning three times.

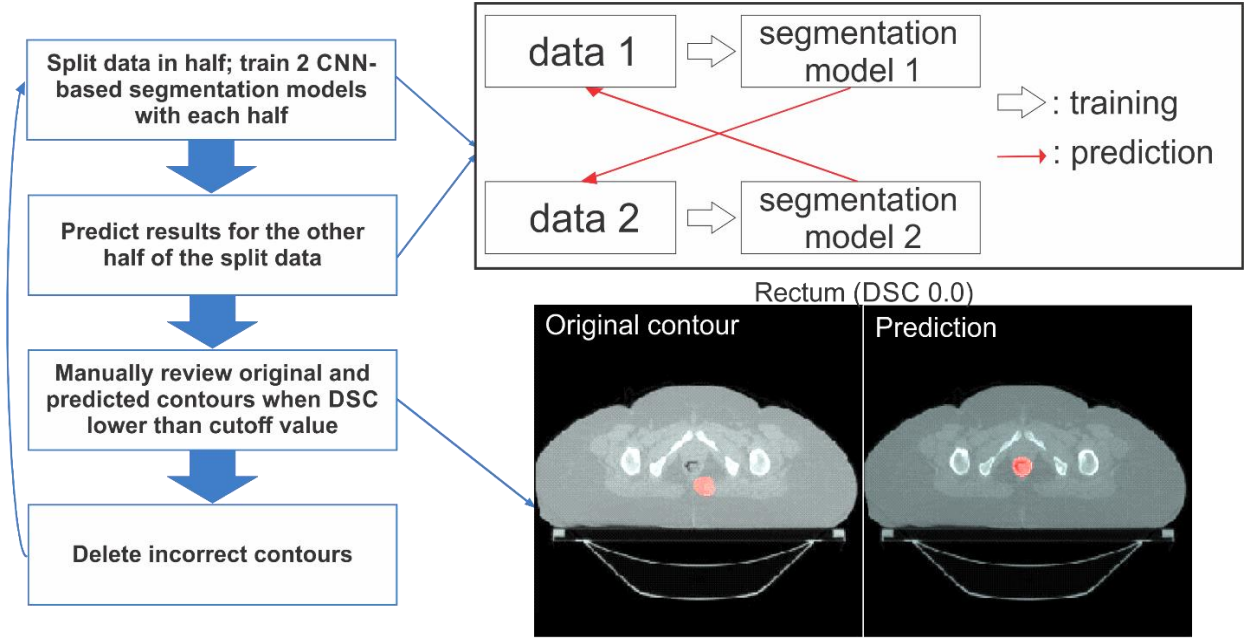


Figure 3. Flowchart of the semi-automated data curation method to identify incorrect clinical contours. Data were randomly split into 2 groups, and 2 auto-segmentation models were trained with each dataset. Then, each segmentation model was applied to the other group of data to create contours. If the Sørensen-Dice similarity coefficient (DSC) was lower than the threshold value, the original contour was manually reviewed and deleted if incorrect.

The left and right kidney contours from the KiTS19 dataset [56] were added to the training dataset. Abnormal kidneys with large tumors were excluded from the dataset, so 172 contours and 186 contours, respectively, for left and right kidneys were added from the total of 210 CT scans.

The definition of the bowel space contour varied, so we decided the bowel space to be the peritoneal cavity from the top of the left kidney to the middle of the rectum, then

subtract the OAR contours we automatically generated. The training dataset for the bowel space auto-contouring model was manually generated on 220 CT scans.

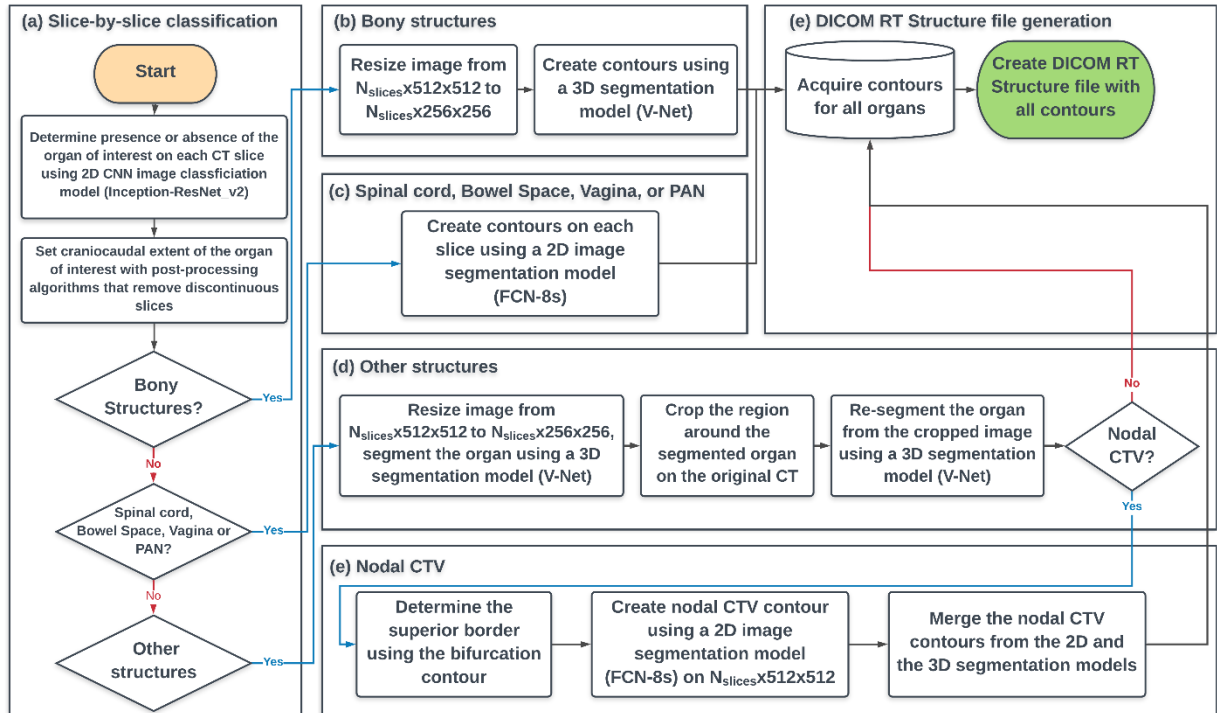


Figure 4. Overall flowchart of the developed auto-contouring system for cervical cancer.

(a) The slice-by-slice classification was conducted to identify CT slices that contain a target structure, and the process is visually demonstrated in Figure 1. (b) Bony structures were contoured as described in 2.B. (c) Spinal cord, vagina, and PAN CTV were contoured with the 2D FCN-8s segmentation architecture. (d) Other structures (the organs-at-risk and the primary and the nodal CTVs) were contoured as demonstrated in Figure 2. (e) Extra steps were required for the nodal CTV contours as described in 2.C.2.



All the OARs, except for the spinal cord and the bowel space, used 3D V-Net segmentation models and followed the steps described in Figure 2. For the spinal cord and the bowel space segmentation, a 2D FCN-8s model was used to generate the contour on each slice. The overall flowchart of the developed auto-contouring system is demonstrated in Figure 4.

Table 1. The number of CT scans used for training and validation for each structure

Structure	Number of training and validation datasets
UteroCervix	406
Vagina	490
Parametria	487
Nodal CTV	250
PAN CTV	146
Bladder	1678
Rectum	1514
Spinal cord	655
Femurs (left, right)	962, 983
Kidneys (left, right)	907, 943
Bowel space	220
Pelvic bone	370
Sacrum	370
L4/L5 vertebral bodies	370
CTV: clinical treatment volume; PAN: para-aortic lymph node	

### 3.2.5 Test dataset

For quantitative analysis of the auto-contouring system, CT scans and corresponding clinical contouring data from 140 female pelvic cancer patients who received radiation treatment at MD Anderson were used as the test dataset. All of the test CT scans were independent from the training and validation CT scans.

The contours of the CTVs were manually generated by physicians, and the contours of the bony structures and OARs were manually generated by medical physics researchers and reviewed by physicians. Some of the CT scans did not show all of the OARs, owing to the limited cranial-caudal extent. As the superior border of the PAN CTV can be slightly different, depending on the location of pathological nodes and physician judgment, we modified the superior borders of the automatically generated PAN CTV on the basis of the manually generated ground truth contour. We did the same for the inferior borders of the rectum and the spinal cord for similar reasons. The accuracy of the model was measured by the DSC, mean surface distance (MSD), and Hausdorff distance (HD) [53] between the automatically generated contours and the ground truth contours. Here, the definition of DSC is

$$DSC = \frac{2|A \cap B|}{|A| + |B|} \quad (1)$$

where A and B are the volumes defined by the reference and the verification contours, and the absolute brackets represent the number of voxels in each volume.

The definition of Hausdorff distance is

$$HD(X, Y) = \max \left\{ \max_{x \in X} \min_{y \in Y} d(x, y), \max_{y \in Y} \min_{x \in X} d(x, y) \right\} \quad (2)$$

where X, Y are the surfaces of the volumes A and B, respectively. x and y are the points on X and Y, and d(x,y) is the distance between the points x and y.

The definition of MSD is

$$MSD(X, Y) = \frac{1}{2} \left\{ \frac{1}{|X|} \sum_{x \in X} \min_{y \in Y} d(x, y) + \frac{1}{|Y|} \sum_{y \in Y} \min_{x \in X} d(y, x) \right\} \quad (3)$$

For qualitative analysis, contours were automatically generated using the auto-contouring system (Figure 4) for CT scans from 30 cervical cancer patients from 3 South African hospitals. This dataset was completely independent from the training dataset and the potential population target for the RPA system. The automatically generated contours were evaluated by an experienced radiation oncologist at MD Anderson and scored as needing no edits, minor edits, or major edits. For the contours scored as needing minor edits, revisions were preferred but not mandatory for the contours to be considered clinically acceptable.

### 3.3 Results

#### 3.3.1 Model accuracy

The DSC, MSD, and HD between the automatically generated contours and the internal test dataset for 140 CT scans are given in Table 2. Owing to the limited cranial-caudal extent, only 132, 129, and 127 contours were evaluated for the PAN, L4/L5 vertebral bodies, and kidneys, respectively; 2 patients did not have nodal CTV and 1 patient did not have spinal cord contours. All the CTVs had mean DSC > 0.76, mean MSD < 0.28 cm, and mean HD < 2.76 cm. All the normal structures had mean DSC > 0.81, mean MSD < 0.18 cm, and mean HD < 1.66 cm. All the bony structures had mean DSC > 0.90, mean MSD < 0.08 cm, and mean HD < 1.25 cm.

The overall boxplots of DSC for each structure are given in Figure 5. Although most of the automatically generated contours had DSC distribution within a certain range, low DSC outliers existed in the box plots, and some of these contours are shown in Figure 6. The failures in generating accurate contours often occurred when the CTVs and OARs were located near high-density material in the bowel, as shown in Figure 6(a). Contouring of the bladder occasionally failed when the border between the bladder and the uterus was vague, as shown in Figure 6(b). Contouring of L4 and L5 vertebral bodies sometimes failed when the segmentation model predicted L3 to be L4 and L4 to be L5, as shown in Figure 6(c). The automatically generated PAN CTV contours had low DSC values when the interface between the nodal CTV and the PAN CTV was incorrectly determined, as shown in Figure 6(d).

### 3.3.2 Physician review

Physician scoring of the automatically generated contours on the 30 external CT scans is shown in Table 3. Owing to the limited cranial-caudal extent, 28 contours were evaluated for the left and right kidneys. For the primary, nodal, and PAN CTVs, 73%, 70%, and 87% of the contours were clinically acceptable, respectively. For the bladder, rectum, bowel space, and right kidney, 90%, 93%, 93%, and 96% were clinically acceptable, respectively, and the other OARs were 100% clinically acceptable. For the bony structures, 93% and 97% of the L4 and L5 vertebral bodies were clinically acceptable, respectively, and the pelvic bone and sacrum were 100% clinically acceptable. Some of the minor edits and major edits are demonstrated in Figure 7.

Table 2. Sørensen-Dice similarity coefficients (in percentage), mean surface distance (in cm), and Hausdorff distance (in cm) between our CNN-based model and clinical contours from 140 internal test CT scans

<b>Structure</b>	<b>DSC (mean±SD)</b>	<b>MSD (mean±SD)</b>	<b>HD (mean±SD)</b>
Primary CTV	0.83±0.06	0.28±0.09	2.76±1.02
Nodal CTV	0.81±0.03	0.21±0.05	2.09±0.56
PAN CTV	0.76±0.09	0.27±0.16	2.00±1.00
Bladder	0.89±0.09	0.11±0.13	1.07±0.89
Rectum	0.81±0.09	0.18±0.14	1.66±1.17
Spinal cord	0.90±0.02	0.06±0.01	0.65±0.18
Femur, left	0.94±0.03	0.06±0.03	0.60±0.41
Femur, right	0.93±0.04	0.07±0.04	0.66±0.43
Kidney, left	0.94±0.02	0.08±0.03	0.76±0.28
Kidney, right	0.95±0.02	0.07±0.03	0.84±0.37
Pelvic bone	0.93±0.02	0.05±0.02	1.06±0.53
Sacrum	0.91±0.02	0.07±0.05	1.25±1.12
L4 vertebral body	0.91±0.15	0.07±0.15	0.53±0.36
L5 vertebral body	0.90±0.15	0.08±0.23	0.68±0.81

CNN: convolutional neural network; CT: computed tomography; DSC: Sørensen-Dice similarity coefficient; MSD: mean surface distance; SD: standard deviation

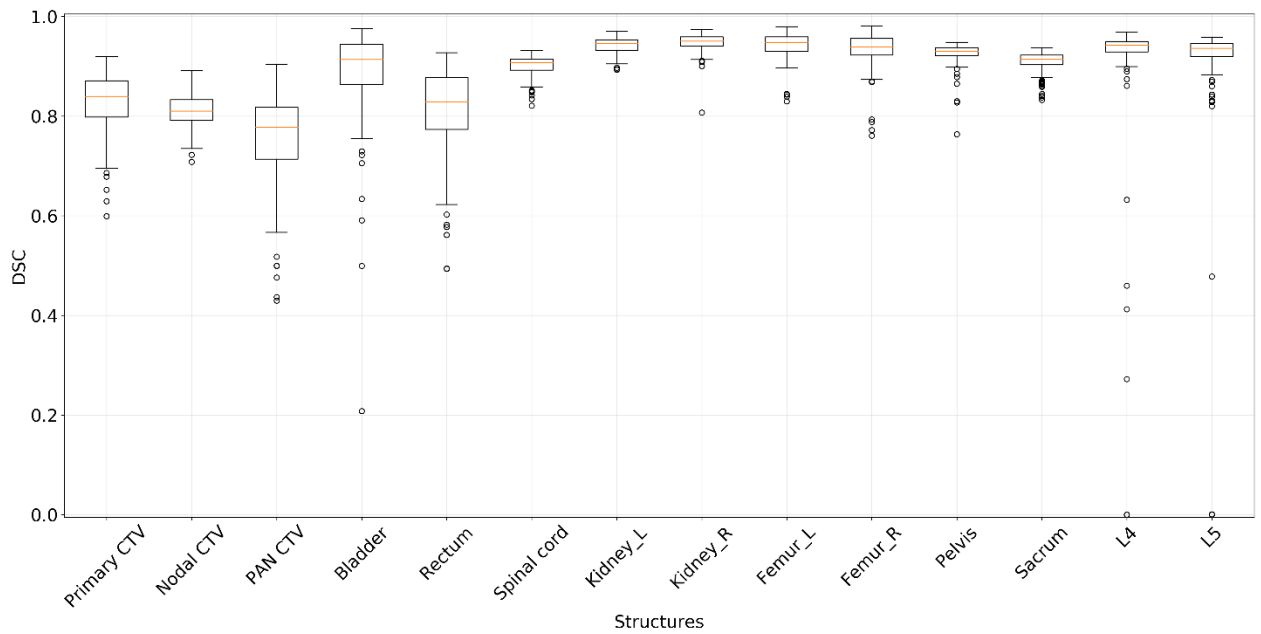


Figure 5. The distributions of Sørensen-Dice similarity coefficients (DSC) between the ground truth and the automatically generated contours of 14 structures.

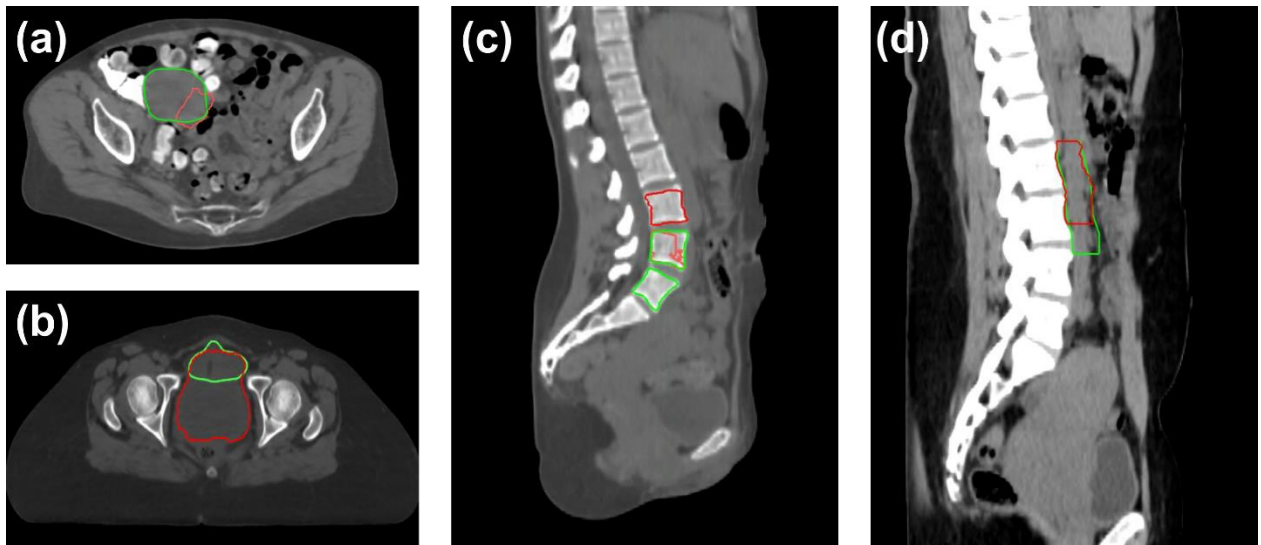


Figure 6. The outlier contours from the internal test dataset. The ground truth contours (green) and outliers (red) are given for (a) primary CTV (Sørensen-Dice similarity coefficient [DSC] = 0.43), (b) bladder (DSC = 0.21), (c) L4 and L5 vertebral bodies (DSC = 0.0 each), and (d) PAN CTV (DSC = 0.43).

Table 3. Qualitative scores of the automatically generated contours on 30 external CT scans

<b>Structure</b>	<b>No edits (%)</b>	<b>Minor edits (%)</b>	<b>Major edits (%)</b>
Primary CTV	6 (20%)	16 (53%)	8 (27%)
Nodal CTV	9 (30%)	12 (40%)	9 (30%)
PAN CTV	18 (60%)	8 (27%)	4 (13%)
Bladder	22 (73%)	5 (17%)	3 (10%)
Rectum	20 (67%)	8 (27%)	2 (7%)
Bowel Space	0 (0%)	28 (93%)	2 (7%)
Spinal cord	30 (100%)	0 (0%)	0 (0%)
Femur, left	27 (90%)	3 (10%)	0 (0%)
Femur, right	27 (90%)	3 (10%)	0 (0%)
Kidney, left	23 (82%)	5 (18%)	0 (0%)
Kidney, right	23 (82%)	4 (14%)	1 (4%)
Pelvic bone	24 (80%)	6 (20%)	0 (0%)
Sacrum	23 (77%)	7 (23%)	0 (0%)
L4 vertebral body	27 (90%)	1 (3%)	2 (7%)
L5 vertebral body	26 (87%)	3 (10%)	1 (3%)

CT: computed tomography; CTV: clinical treatment volume

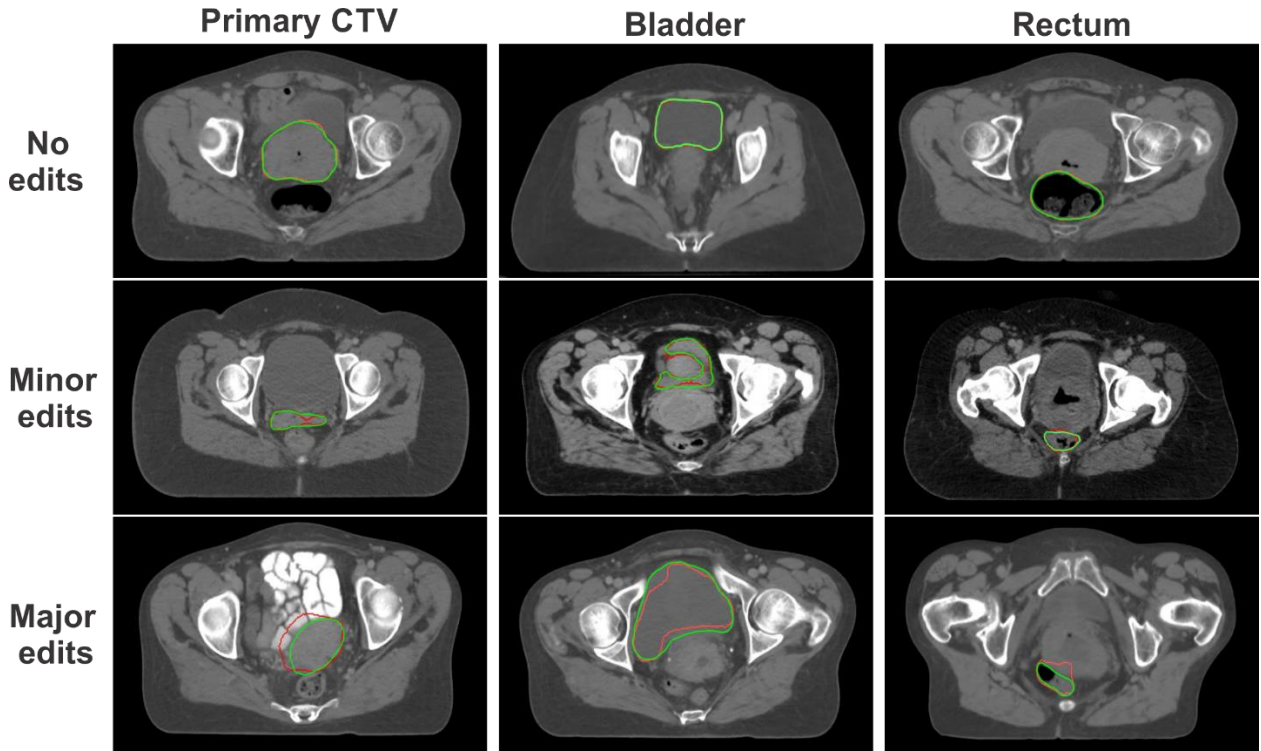


Figure 7. Examples of automatically generated contours (red) versus ground truth (green) from physician’s manual review of contours for the primary CTV, bladder, and rectum.

### 3.4 Discussion

We have developed a CNN-based auto-contouring tool for 3 CTVs and 12 normal structures in cervical cancer CTs that can be used for fully automated radiation treatment planning. The number of training, validation, and test CT scans we used to train and evaluate this model is the largest to date among deep learning-based female pelvis auto-contouring studies [35], [36]. We successfully acquired this high volume of data by using a semi-automatic data curation method. Also, to the best of our knowledge, we are the first to auto-contour nodal and PAN CTVs in the female pelvic



region using deep learning. We have demonstrated that our CNN-based auto-contouring system can accurately generate clinically acceptable contours for both CTVs and normal structures in multiple patient cohorts.

#### 3.4.1. Quantitative results

For the bony structures, 3.5% (5/140 from the quantitative analysis) of the L4 and L5 vertebral body contours were not clinically acceptable (i.e. outliers in the boxplot in Figure 5). Similarly, 6.7% (2/30 from the qualitative analysis) of the L4 and L5 vertebral body contours were not clinically acceptable (Table 3). Therefore, the overall failure rate for the bony structures was about 4%. This is a noticeable improvement compared to a previous study where the failure rate for the automatically generated contours in a multi-atlas-based auto-contouring system was about 10% [7].

The performances of deep learning-based auto-contouring systems for OARs in the female pelvis from other published literature are presented in

Table 4. As there is only 1 published paper on a deep learning-based auto-contouring system for cervical cancer, we have also included the state-of-the-art auto-contouring models for rectal and prostate cancers. Overall, the performance of our system is equivalent to or better than the auto-contouring system for cervical cancer developed by Liu et al. [36] for most of the structures.

Our quantitative test CT scans were randomly chosen from CTs of any female patient with an intact uterus, so the shape and volume of the bladder in the CT scans varied significantly. When we retrospectively tested our bladder segmentation model on 510 prostate patients with full bladders, the mean DSC was much improved at  $0.95 \pm 0.04$ . Compared with the state-of-the-art rectal and prostate models, our model performed at least as well in all structures except for the rectum. However, the direct comparison of auto-contouring models for different sites is not straightforward because the homogeneity of the structures in the test CT scans substantially changes the DSC, as shown in the accuracy of our 2 bladder models.

Table 4. Summary of CNN-based auto-contouring results for normal structures in pelvic CTs from other groups

Author	Sites	# test CTs	Structures	DSC results
Men et al. (2017)[39]	Rectal	60	Bladder	0.93
			Colon	0.62
			Intestine	0.65
			Femur_L	0.92
			Femur_R	0.92
			Rectal CTV	0.88
Kazemifar et al. (2018)[33]	Prostate	~26 (30% of 85)	Prostate	0.88
			Bladder	0.95
			Rectum	0.92
Balagopal et al. (2018)[34]	Prostate	~27 (Leave-one-out cross-validation, 20% of 135)	Prostate	0.90
			Bladder	0.95
			Rectum	0.84
			Femur_L	0.96
			Femur_R	0.95
Liu et al. (2020)[36]	Cervix	14	Bladder	0.92
			Bone marrow	0.85
			Rectum	0.79
			Small intestine	0.83
			Spinal cord	0.83
			Femur_L	0.91
			Femur_R	0.90
Our method	Cervix	140	Bladder (cervical cancer)	0.89
			Bladder (prostate cancer)	0.95
			Rectum	0.80
			Spinal cord	0.90
			Pelvic bone	0.93
			Sacrum	0.91
			Femur_L	0.94
			Femur_R	0.93

CNN: convolutional neural network; CT: computed tomography; DSC: Sørensen-Dice similarity coefficient; CTV: clinical treatment volume

#### 3.4.2. Failure cases from physician's review

The overall clinical acceptance rates were 79% for the CTVs and 97% for the OARs and bony structures. When high-density materials were located in the bowel, the auto-contouring system had a higher chance of creating inaccurate contours of the CTVs or OARs near the region, as shown in Figure 5 and Figure 6. These high-density materials were fecal matter resulting from a high-carb diet with minimal protein, fat, and fibers, which likely causes compacted slow-moving feces. This diet is more common in South Africa, the patient population for the external test dataset, than in the U.S, the patient population for the training and internal test datasets. As we acquire more CT data from such patients through the RPA system, we will be able to upgrade the auto-contouring system to achieve more robust results in these patients.

For the primary CTVs, 8/30 were scored as needing major edits. However, 3/8 failure cases were due to the underestimation of the vagina contour when significant gas or filling was seen in the rectum or sigmoid (diameter > 4 cm). Some parts of the vagina can be squeezed by the inflated rectum or sigmoid and the bladder, and the auto-contouring model often failed to contour the vagina properly under this circumstance, as shown in Figure 8. The clinicians should empty the rectums of such patients and rescan their patients according to the GEC-ESTRO protocol. Therefore, we should exclude those 3 cases and another 2 cases with the inflated rectums with the clinically acceptable primary CTVs from the qualitative evaluation to reflect a realistic clinical environment. With these, the acceptance rates for the primary CTV increase from 73% to 80%.

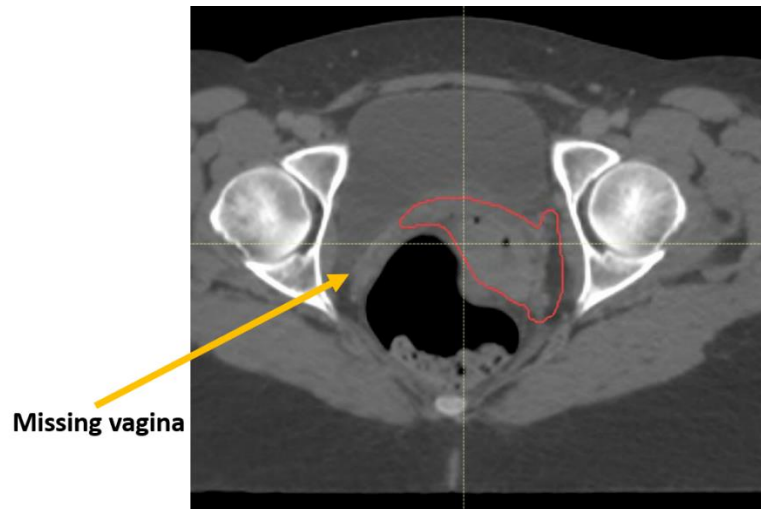


Figure 8. Example of missing the vagina contour due to the gas-filled rectum

For the nodal CTVs, 9/30 were scored as needing major edits; 1 was due to high-density fecal matter in the bowel, and 3 were due to failure to detect the superior border. The 3 cases did not have clear borders for vessels, as shown in Figure 9(b), and therefore, the bifurcation segmentation model did not perform appropriately. All 3 patients seemed to be underweight, based on their CT scans, so we believe that the poor contrast resolution was due to incorrect use of image acquisition parameters or the lack of fat in between the vessels. We need to further investigate our auto-contouring system in underweight patients and may need to adjust the CT acquisition parameters for these patients in the future.

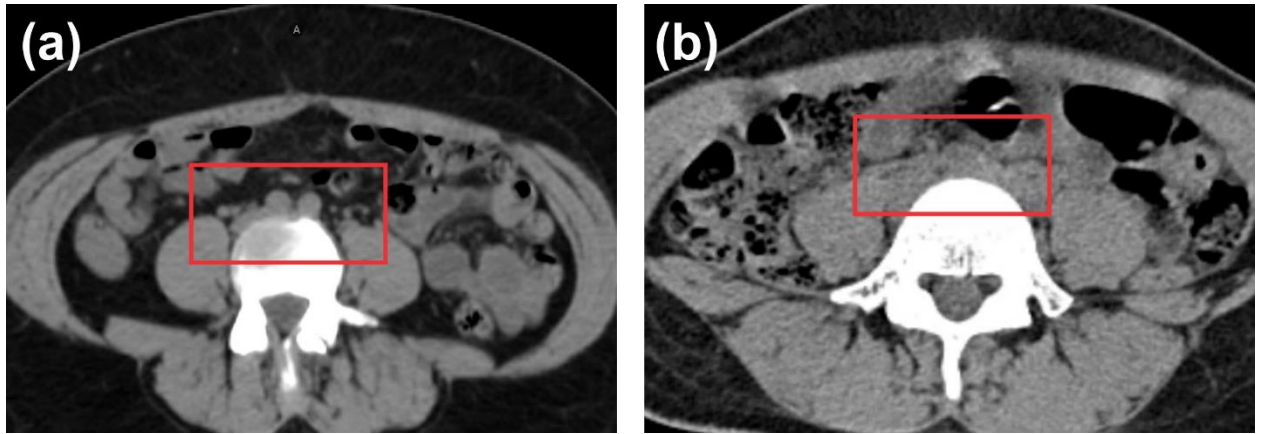


Figure 9. The aortic bifurcation is clearly defined in the red box in (a), whereas the aortic bifurcation is barely identifiable in the red box in (b). The adjustment of the window level did not improve the visual inspection.

We trained our model using a consistent, well-curated dataset from a single hospital and the publicly available kidney contours. Final physician review used images from 3 other hospitals, with different patient populations from the training dataset. Thus, the review results gave us some confidence in the ability of our model to successfully contour patients from a different patient population, as well as various CT scanners and imaging protocols. In this study, we did not examine the impact of inter-user variations on the physician assessment of these contours. Based on our experience with other sites [57], it is likely that an increased fraction of patients will be considered ‘minor edits’ instead of ‘no edits’ as we deploy the auto-contouring system to more hospitals. We will further assess and quantify the inter-user variability as we begin to deploy this system clinically.

### 3.5 Conclusion

We have demonstrated through both quantitative and qualitative studies that a CNN-based, auto-contouring tool can achieve clinically acceptable contours for most of the CTVs and normal structures in cervical cancer patients. We will implement our auto-contouring system to the Radiation Planning Assistant, accelerating the radiation treatment planning process in hospitals in low- and middle-income countries.

Chapter 4 : Development of a quality assurance system to detect errors in automatically generated contours for cervical cancer.

#### 4.1 Introduction

Although the auto-contouring system described in Chapter 3 performed rigorously, some of the auto-contours were still clinically unacceptable. AAPM Task Group 275 recently reported that failures in detecting contouring errors in treatment targets and normal structures are the largest and seventh largest risk factors in radiotherapy planning, respectively [16]. Because of this significant risk, automatic contouring error detection methods have been studied by several research groups. For example, Chen et al. [58] extracted the geometric features of organs and developed machine learning-based contouring error detection models for normal structures in the head and neck. McIntosh et al. [26] used image features to train a conditional random forest algorithm to detect contouring errors in thoracic structures. Hui et al. [59] used contour shapes in principal component and Procrustes analyses to detect contouring errors in pelvic structures. We previously demonstrated [21] that calculating the Dice similarity coefficient (DSC) between two automatically generated contours can be used to detect errors in one of the contours.

In this work, we have integrated multiple quantitative metrics to measure the similarity between two contours and used these metrics to provide quality assurance (QA) for the target and normal structure auto-contours generated from the auto-contouring system described in Chapter 3. We hypothesized that if one of the two independently developed auto-contouring systems failed, the discrepancy between the



two contours would be substantial. This discrepancy can be quantified using the similarity metrics of the two contours, such as the DSC, and errors in contours can be detected by analyzing these metrics. Even if both auto-contours from two independently developed systems fail simultaneously, it is very unlikely that they will fail similarly based on our previous study with head-and-neck normal structures [21]; thus, the discrepancy between the two contours will still be substantial. This study examined how to optimally flag incorrect contours by evaluating 11 different comparison metrics and evaluating different approaches to combining these metrics. To our knowledge, this is the first study to specifically cross-compare the results of two different deep learning-based auto-contouring approaches to identify contouring errors.

## 4.2 Methods

To develop the automatic contour QA system, we tested 11 quantitative metrics on six structures in the female pelvis: UteroCervix (uterus + cervix), nodal CTV, PAN (para-aortic nodal), bladder, rectum, and kidneys. Furthermore, the automatic contour QA system was developed for six bony structures (femurs, spinal cord, pelvis, sacrum, L4, and L5) using the method suggested by Rhee et al. [21] for the head-and-neck normal structures.

### 4.2.1 Two deep-learning-based auto-contouring systems

The reference auto-contouring system, which was used to generate the contours for clinical use, was the auto-contouring system described in Chapter 3. The verification auto-contouring system, which was used to test the clinical acceptability of the contours

from the reference system, was developed by Rigaud et al. [60]. The two auto-contouring systems were developed using completely separate training datasets.

As the Nodal CTV, PAN, and the four bony structures (Pelvis, Sacrum, L4, and L5) were not available in the original verification system, we trained the auto-contouring models for these structures using 140 CT scans to match all the structures. These 140 CT scans were from the training dataset for the verification auto-contouring system, and therefore, they were independent of the training dataset of the reference auto-contouring system. We used V-Net [9] for the 4 bony structures, FCN-8s [10] for the PAN, and the combination of the two architectures for the Nodal CTV to train the verification models, as described in Chapter 3.

#### 4.2.2 Quantitative metrics

To quantify the similarities between the two contours generated by the two auto-contouring systems, we used four widely used conventional metrics for contour comparison studies: Dice similarity coefficient (DSC), Hausdorff distance (HD<sub>100</sub>), 95% Hausdorff distance (HD<sub>95</sub>), and mean surface distance (MSD). We also tested the surface DSC, as suggested by Nikolov et al. [20] with 1, 2, 3, 4, 5, 7, and 10 mm tolerances. All of the metrics were calculated for the 3D contours; therefore, we obtained one metric per patient per structure.

The definitions of DSC, HD, and MSD are defined in equation (1), (2), and (3), respectively. The definition of 95% HD is similar to that of 100% HD, but it is based on the 95<sup>th</sup> percentile of the distance between the two contours, instead of the 100<sup>th</sup>

percentile for the regular HD. The acronym for the regular HD and 95% Hausdorff distance are HD\_100 and HD\_95, respectively, in this study.

The definition of surface DSC with the tolerance  $\tau$  is

$$Surface_{DSC(X,Y,\tau)} = \frac{|X \cap Y_B^\tau| + |Y \cap X_B^\tau|}{|X| + |Y|} \quad (4)$$

where  $X$ ,  $Y$  are the surfaces of the two contours,  $X_B^\tau$  is the border region of the surface  $X$ , and the border region consists of all the points that are within the tolerance distance  $\tau$  from the surface  $X$ . Therefore, the volume  $X^\tau$  has the shell thickness of  $2\tau$ . The acronym for the surface DSC with  $n$  mm tolerance is SDSC\_ $n$  in this study.

#### 4.2.3 Error detection model with support vector machine for the soft-tissue structures

We trained a support vector machine (SVM) [61], [62], a machine learning classification algorithm, to determine the most accurate metric for contouring error detection. We chose the SVM classification algorithm because it is a powerful and computationally fast machine learning algorithm. Furthermore, SVM is one of the most intuitive classification algorithms, [63] making it easy to interpret the final model.

We tested various combinations of metrics to develop the most accurate contouring error detection model. We first tested the models developed with the 11 quantitative metrics individually (single-metric analysis) using an SVM with the linear kernel, as this is the only kernel possible for a 1-D input, as shown in Figure 10(a). We tested the values of the penalty parameter  $C$  from 1 to 50 and applied the best value to calculate the final accuracies. To provide a more comprehensive evaluation of the performance,

we also performed ROC analysis and calculated the area under the ROC curve (AUC) on each metric and each structure.

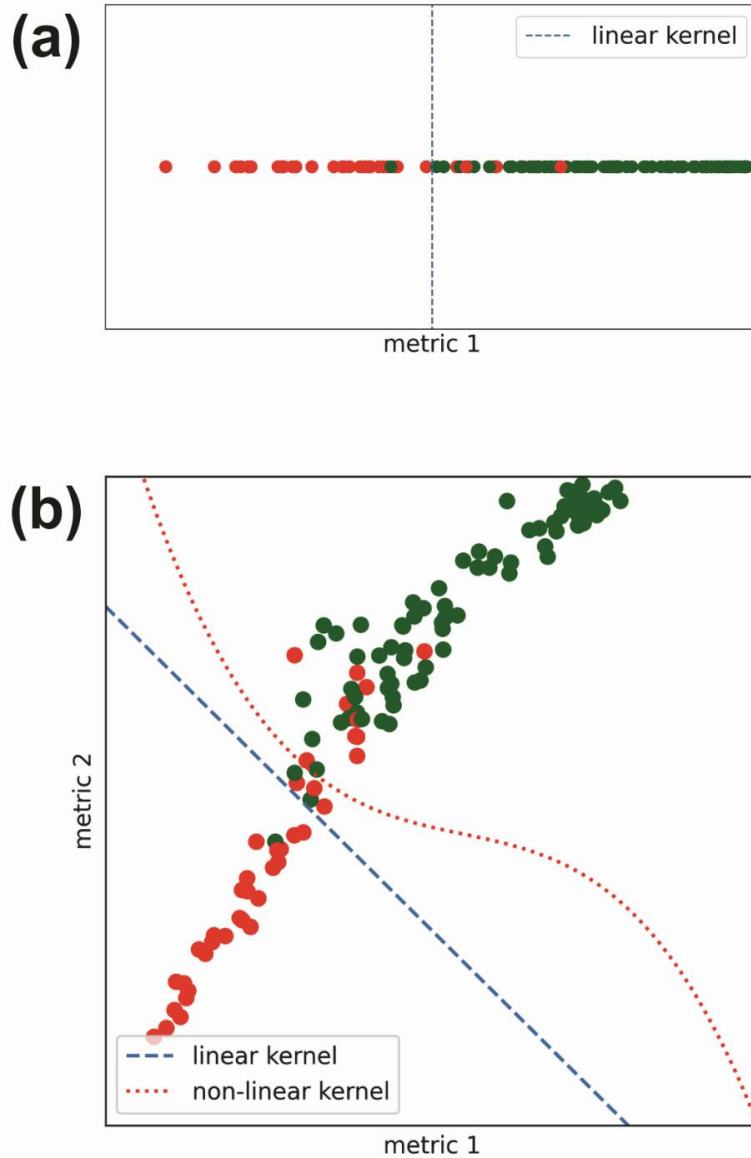


Figure 10. Demonstration of the how the SVM-based contour QA model was developed. The red and green data points represent the clinically unacceptable and acceptable contours, respectively. (a) Using a single metric and determine the threshold between the clinically acceptable and unacceptable contours. Since the input is one dimensional, SVM with the linear kernel can only be used. (b) Using a combination of the metrics to determine the threshold either with a linear or a non-linear kernel.

We also tested combinations of the 11 quantitative metrics on the basis of the results of the single-metric analysis (multi-metric analysis). The combinations tested are presented in Table 5. We tested linear, polynomial (with degree = 3), radial basis function, and sigmoid kernels in a multi-metric analysis. We tested the values of the penalty parameter C from 1 to 50 and found the best results for C larger than 5, with minimal variation between 5 and 20. Therefore, a penalty parameter of 10 was used for all metrics in the single-metric and multi-metric analyses.

Table 5. List of the combined metrics used in the multi-metric analysis

<b>Name</b>	<b>Metrics used</b>	<b>Description</b>
DSC_HD	DSC, HD_100	Most used quantitative metrics
Three_SDSC	SDSC 1 mm, 2 mm, 3 mm	Top 3 SDSC from single-metric analysis
Five_SDSC	SDSC 1 mm, 2 mm, 3 mm, 4 mm, 5 mm	Top 5 SDSC from single-metric analysis
Four_metrics	DSC, HD_100, HD_95, MSD	Four conventional quantitative metrics
Five_metrics	DSC, MSD, SDSC 1 mm, 2 mm, 3 mm	Two most effective conventional metrics + 3 most effective SDSCs
Seven_metrics	DSC, MSD, SDSC 1 mm, 2 mm, 3 mm, 4 mm, 5 mm	Two most effective conventional metrics + five most effective SDSCs
Nine_metrics	DSC, MSD, SDSC 1 mm, 2 mm, 3 mm, 4 mm, 5 mm, 7 mm, 10 mm	Two most effective conventional metrics + all SDSCs
All_metrics	DSC, HD_100, HD_95, MSD, SDSC 1 mm, 2 mm, 3 mm, 4 mm, 5 mm, 7 mm, 10 mm	All available metrics

\* DSC = Dice Similarity Coefficient, HD = Hausdorff Distance, MSD = Mean Surface Distance, SDSC = Surface Dice Similarity Coefficient

#### 4.2.4 Data acquisition

To train the SVM algorithm to determine the threshold between clinically acceptable and unacceptable contours, we needed a set of two automatic contours in

the same patients and organs. We created a set of auto-contours on 49 CT scans from MD Anderson and 38 CT scans from three hospitals in South Africa. The reference auto-contours were scored by one experienced radiation oncologist and one radiation oncology resident at MD Anderson. They each reviewed a subset of the contours and scored the contours as either needing no edits, minor edits, and major edits. For the contours scored as needing minor edits, revisions were preferred but not mandatory for the contours to be clinically acceptable, so the contours scored as needing major edits were considered as clinically unacceptable contours.

Furthermore, clinically acceptable and unacceptable contours for the 49 internal CT scans were manually created by radiation oncology residents at MD Anderson. The clinically unacceptable contours were manually introduced to mimic a potential error that can be made by a human or a deep learning algorithm as a result of a lack of experience or an unclear soft tissue border, as illustrated in Figure 11. Since most of the auto-contours were clinically acceptable, the number of clinically unacceptable contours were not sufficient to determine the robust thresholds. These manual contours were added to the dataset to fill this gap, and therefore enable the model to distinguish clinically acceptable and unacceptable contours more robustly.

Then, the quantitative metrics were calculated between the verification and the reference auto-contours for the internal and external dataset, between the verification auto-contours and the clinically acceptable manual contours for the internal dataset, and between the verification auto-contours and the clinically unacceptable manual contours for the internal dataset, as shown in Figure 12 (a). In total, this resulted in 185 calculated data points per metric per structure from the four sets of data. Each set of

data was split equally into three for three-fold cross-validation, as shown in Figure 12 (b).

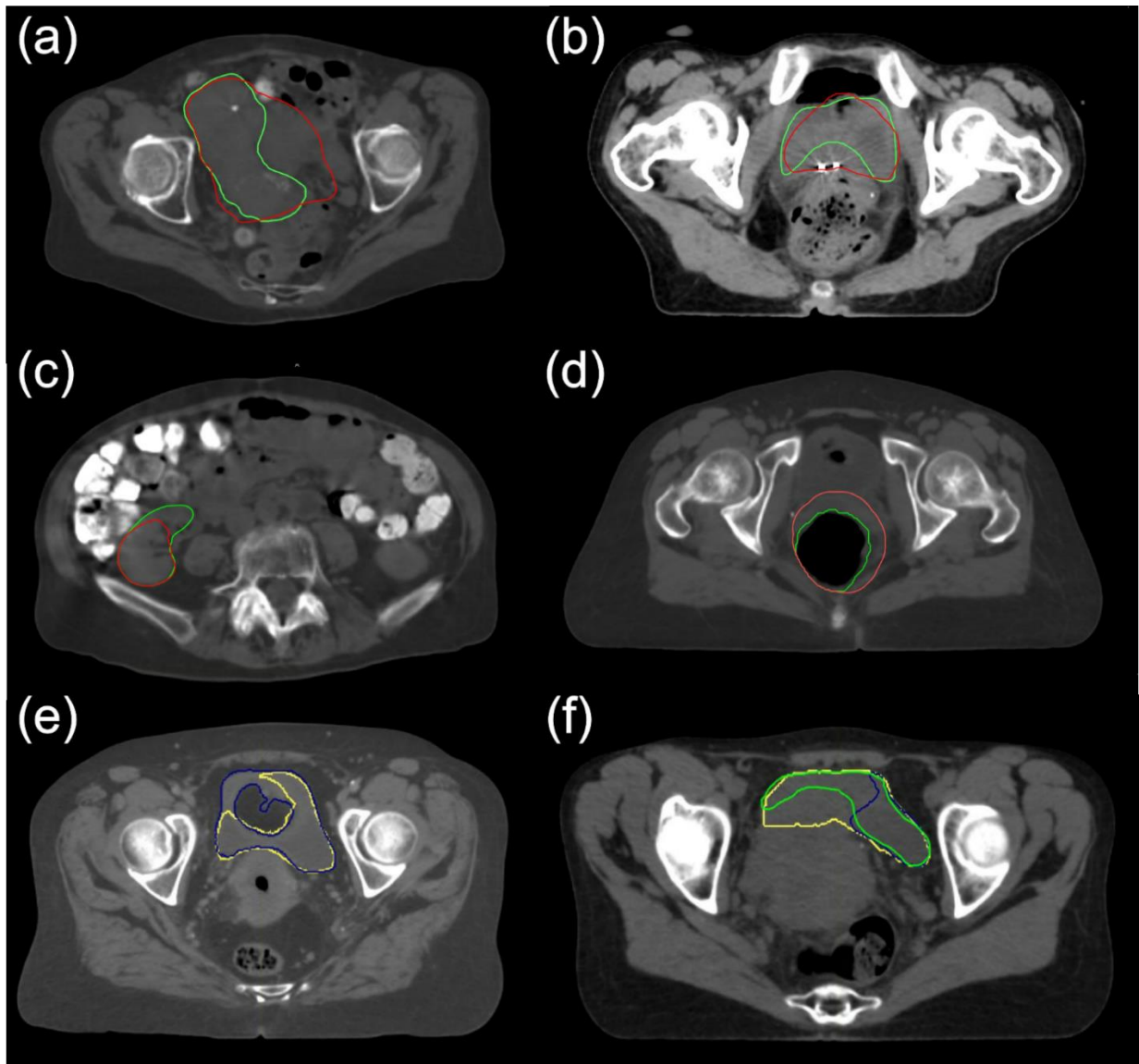


Figure 11. Examples of manually generated, clinically acceptable (green) and unacceptable (red) contours for the (a) UteroCervix, (b) Bladder, (c) Right Kidney, and (d) Rectum. (e) The reference auto-contour (yellow) was clinically unacceptable when the verification auto-contour (blue) was clinically acceptable. (f) Both the reference and the verification auto-contours were clinically unacceptable.



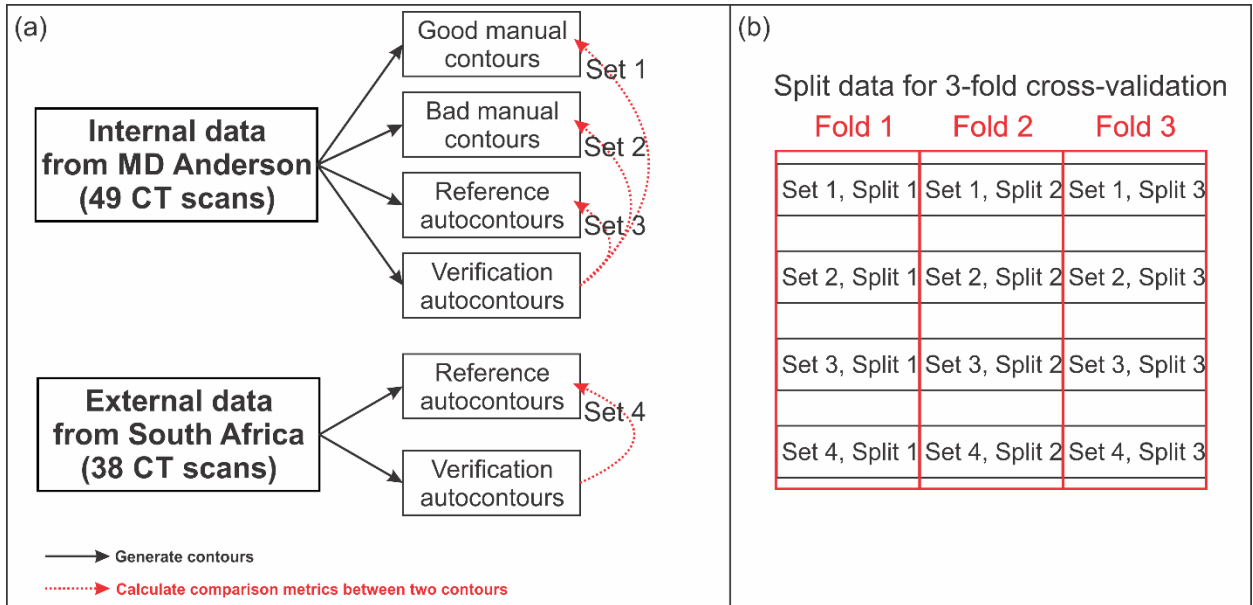


Figure 12. (a) Diagram demonstrating the data acquisition process for automatic contour QA model development, and (b) demonstrating that each set was split equally into 3 for 3-fold cross-validation.

We chose the 49 MD Anderson CT scans from the training dataset of the verification auto-contouring system. As a result, the verification auto-contours almost always accurately predicted the contours, and therefore, there were almost no false positives (i.e. the reference contour was clinically acceptable when the verification contour was not) on the 49 CT scans. False positives were attributed to errors in building an accurate SVM classification model, although they still supported the central hypothesis of the study because one of the contours was clinically unacceptable. Therefore, reducing the number of false positives helped improve the accuracy of the contour QA model.

#### 4.2.5 Error detection model for the bony structures

The failure cases were much rarer for the six bony structures (Spinal cord, Femurs, Pelvis, Sacrum, L4, and L5) and the discrepancy between clinically acceptable and unacceptable contours were more substantial than that of the soft-tissue structures. We used 87 auto-contours for each structure from the 49 internal and 38 external CT scans without manually generated contours. We adapted the contour QA technique used for the head-and-neck auto-contouring structures for the RPA system [21], as this technique can be applied with a very limited number of failure cases. In this QA technique, the distributions of DSC and HD<sub>100</sub> were plotted for each structure and the thresholds for both DSC and HD<sub>100</sub> were determined based on the visual inspection of the plots.

### 4.3 Results

#### 4.3.1 Soft-tissue structures

##### 4.3.1.1 Single-metric analysis

The average accuracy of the 3-fold cross-validation results is shown in Figure 13 for the 11 quantitative metrics that were tested individually with an SVM algorithm using a linear kernel with various penalty parameters,  $C$ , from 1 to 50. Overall, the SDSC<sub>1</sub>, SDSC<sub>2</sub>, and SDSC<sub>3</sub> were the most accurate indicators in detecting contouring errors. The penalty parameters  $C = 10$  gave the best results for these 3 metrics on average, although there were no substantial differences between the penalty parameters between 3 and 50. Therefore, we presented all the accuracies, including the multi-metric cases, with the penalty parameter  $C = 10$ .

The highest accuracy result was higher than 0.9 for the UteroCervix ( $0.91 \pm 0.05$  with SDSC\_1), the Nodal CTV ( $0.90 \pm 0.03$  with SDSC\_1 and SDSC\_2), the Bladder ( $0.92 \pm 0.03$  with SDSC\_3), the Rectum ( $0.94 \pm 0.04$  with SDSC\_1), and the Kidneys ( $0.97 \pm 0.03$  with SDSC\_2) and almost 0.9 for the PAN ( $0.89 \pm 0.04$  with SDSC\_3).

The accuracy decreased as the tolerance for the surface DSC increased after 3 mm. DSC and MSD also accurately predicted the clinical acceptability of the contours. On the other hand, HD\_100 and HD\_95 were not as accurate as the other metrics.

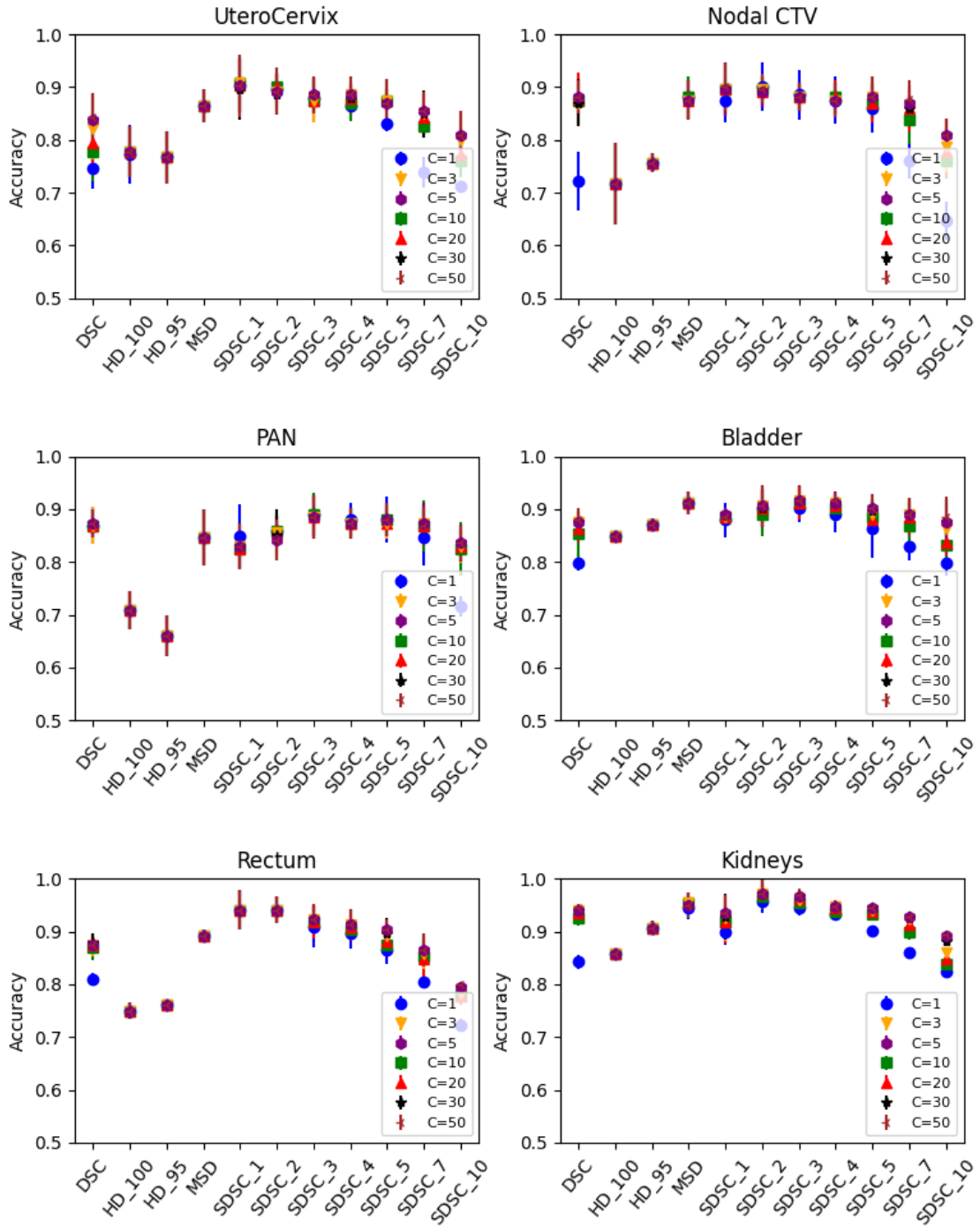


Figure 13. Average accuracies of the contour QA model with an individual metric for each structure with various penalty parameters, C. The error bar represents  $\pm 1$  standard deviation from three-fold cross-validation.

To investigate the stability of each metric in response to small changes in the threshold, we calculated the average of the thresholds for four major metrics (DSC, SDSC\_1, SDSC\_2, and SDSC\_3) over the structures and calculated the accuracy using the average thresholds on each structure (0.75, 0.30, 0.54, and 0.69 for DSC, SDSC\_1, SDSC\_2, and SDSC\_3, respectively). The changes in the thresholds could exceed 20%, but the overall accuracy barely changed, as shown in Table 6 and Table 7.

Table 6. Changes in accuracy when applying the average threshold of various structures instead of optimal thresholds for each structure.

Change in accuracy (% $\Delta$ Threshold)	UteroCervix	CTVn	PAN	Bladder	Rectum	Kidneys
<b>DSC</b>	0.82 $\rightarrow$ 0.79 (4.5%)	0.87 $\rightarrow$ 0.88 (1.5%)	0.87 $\rightarrow$ 0.76 (17.0%)	0.88 $\rightarrow$ 0.86 (3.7%)	0.88 $\rightarrow$ 0.88 (9.8%)	0.94 $\rightarrow$ 0.87 (11.4%)
<b>SDSC_1</b>	0.91 $\rightarrow$ 0.91 (9.5%)	0.90 $\rightarrow$ 0.91 (2.0%)	0.83 $\rightarrow$ 0.81 (18.3%)	0.89 $\rightarrow$ 0.90 (7.5%)	0.94 $\rightarrow$ 0.93 (8.8%)	0.93 $\rightarrow$ 0.92 (2.9%)
<b>SDSC_2</b>	0.89 $\rightarrow$ 0.89 (0.2%)	0.90 $\rightarrow$ 0.89 (0.2%)	0.86 $\rightarrow$ 0.79 (20.2%)	0.91 $\rightarrow$ 0.90 (0.4%)	0.94 $\rightarrow$ 0.93 (10.8%)	0.97 $\rightarrow$ 0.93 (20.5%)
<b>SDSC_3</b>	0.88 $\rightarrow$ 0.88 (0.0%)	0.74 $\rightarrow$ 0.74 (1.2%)	0.88 $\rightarrow$ 0.80 (21.7%)	0.92 $\rightarrow$ 0.90 (8.1%)	0.92 $\rightarrow$ 0.92 (8.9%)	0.96 $\rightarrow$ 0.93 (15.4%)

Table 7. Overall accuracies, sensitivities, and specificities with maximized accuracy through the SVM, fixed sensitivity of 0.90, and fixed sensitivity of 0.95 when surface DSC with a thickness of 2 mm was used.

SDSC_2	maximize accuracy			Sensitivity $\geq 0.90$			Sensitivity $\geq 0.95$		
	accuracy	sensitivity	specificity	accuracy	sensitivity	specificity	accuracy	sensitivity	specificity
<b>UteroCervix</b>	0.89	0.79	0.94	0.90	0.90	0.90	0.86	0.95	0.81
<b>CTVn</b>	0.90	0.78	0.97	0.80	0.91	0.74	0.72	0.96	0.59
<b>PAN</b>	0.86	0.68	0.95	0.67	0.90	0.56	0.62	0.95	0.46
<b>Bladder</b>	0.91	0.79	0.97	0.85	0.90	0.83	0.79	0.95	0.72
<b>Rectum</b>	0.94	0.86	0.97	0.89	0.90	0.88	0.79	0.96	0.72
<b>Kidney</b>	0.97	0.90	0.99	0.97	0.90	0.99	0.97	0.95	0.97

To evaluate the performance more comprehensively, the ROC curves were generated on each metric and each structure and AUCs were calculated, as shown in Table 8. Again, SDSC\_1, SDSC\_2, or SDSC\_3 was the best metric to predict the clinical acceptability of contours, and HD\_100 or HD\_95 was not a good indicator. The ROC curves for SDSC\_2, the best indicator to detect contouring errors based on the AUCs, were presented in Figure 14.



Table 8. AUCs of each structure and each metric. 95% confidence interval (CI) for AUCs were derived with the bootstrapping method with n=2000

<b>AUC (95% CI)</b>	<b>UteroCervix</b>	<b>CTVn</b>	<b>PAN</b>	<b>Bladder</b>	<b>Rectum</b>	<b>Kidneys</b>
<b>DSC</b>	0.92 (0.89 – 0.94)	0.92 (0.89 – 0.95)	0.86 (0.82 – 0.89)	0.92 (0.90 – 0.94)	0.92 (0.89 – 0.94)	0.97 (0.95 – 0.99)
<b>HD_100</b>	0.85 (0.81 – 0.88)	0.75 (0.71 – 0.79)	0.75 (0.70 – 0.80)	0.93 (0.90 – 0.95)	0.81 (0.76 – 0.84)	0.91 (0.88 – 0.93)
<b>HD_95</b>	0.87 (0.83 – 0.89)	0.83 (0.79 – 0.86)	0.70 (0.65 – 0.74)	0.96 (0.94 – 0.97)	0.83 (0.80 – 0.86)	0.95 (0.92 – 0.97)
<b>MSD</b>	0.93 (0.91 – 0.95)	0.92 (0.89 – 0.94)	0.84 (0.80 – 0.88)	0.97 (0.96 – 0.98)	0.92 (0.89 – 0.94)	0.96 (0.93 – 0.98)
<b>SDSC 1 mm</b>	0.96 (0.94 – 0.97)	0.93 (0.90 – 0.95)	0.90 (0.87 – 0.93)	0.95 (0.93 – 0.97)	0.96 (0.94 – 0.98)	0.95 (0.92 – 0.97)
<b>SDSC 2 mm</b>	0.96 (0.94 – 0.97)	0.93 (0.91 – 0.95)	0.89 (0.86 – 0.92)	0.96 (0.94 – 0.97)	0.96 (0.95 – 0.98)	0.97 (0.95 – 0.99)
<b>SDSC 3 mm</b>	0.95 (0.93 – 0.96)	0.93 (0.90 – 0.95)	0.87 (0.83 – 0.91)	0.97 (0.96 – 0.98)	0.95 (0.92 – 0.97)	0.97 (0.95 – 0.99)
<b>SDSC 4 mm</b>	0.93 (0.91 – 0.95)	0.92 (0.89 – 0.94)	0.85 (0.80 – 0.89)	0.97 (0.95 – 0.98)	0.93 (0.90 – 0.96)	0.96 (0.94 – 0.98)
<b>SDSC 5 mm</b>	0.92 (0.89 – 0.94)	0.91 (0.88 – 0.94)	0.83 (0.79 – 0.88)	0.97 (0.95 – 0.98)	0.92 (0.88 – 0.94)	0.95 (0.93 – 0.97)
<b>SDSC 7 mm</b>	0.90 (0.87 – 0.93)	0.89 (0.86 – 0.92)	0.81 (0.76 – 0.85)	0.96 (0.94 – 0.97)	0.89 (0.85 – 0.92)	0.94 (0.92 – 0.96)
<b>SDSC 10 mm</b>	0.88 (0.85 – 0.92)	0.85 (0.81 – 0.88)	0.80 (0.75 – 0.84)	0.91 (0.88 – 0.94)	0.85 (0.81 – 0.89)	0.91 (0.88 – 0.94)

## SDSC\_2

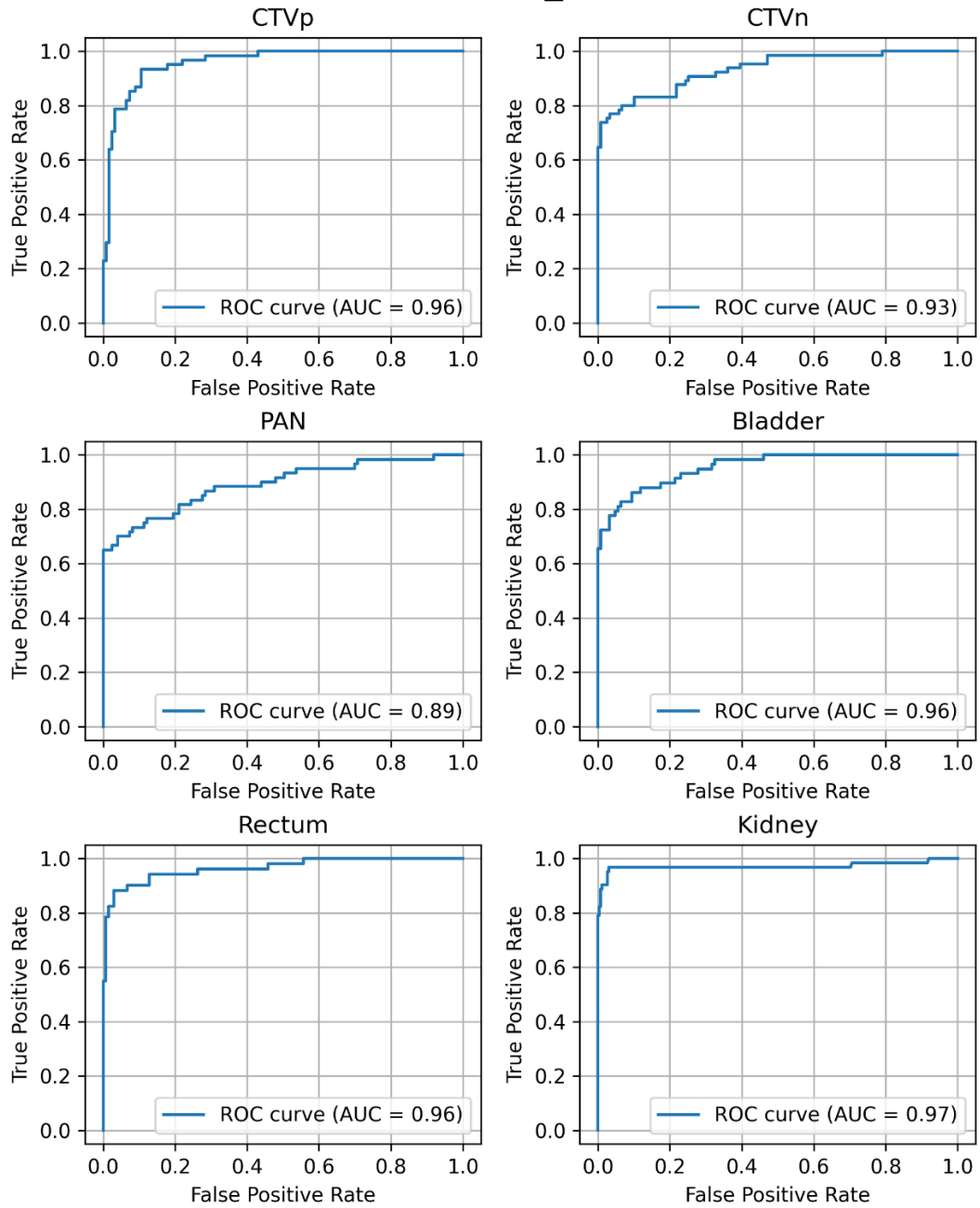


Figure 14. The ROC curves with a surface DSC with a tolerance of 2 mm, the best metric to predict the clinical acceptability of the automatically generated contours.

#### 4.3.1.2 Multi-metric analysis

We chose combinations of two (DSC\_HD) and four (Four\_metrics) widely used similarity metrics for contouring studies and the top three and top five most effective surface DSC metrics (Three\_SDSC and Five\_SDSC) from the single metric analysis. Furthermore, the top five, seven, and nine most effective metrics from the single metric analysis (Five\_metrics, Seven\_metrics, and Nine\_metrics), and all 11 metrics (All\_metrics) were tested in the multi-metric analysis.

The SVM with four kernels on different combinations of metrics were trained; the results are shown in Figure 15. Most of the kernels had similar performance, but the sigmoid kernel substantially underperformed compared to the other kernels. On average, the model performance with the radial basis function and polynomial kernels fluctuated more with the choice of the metrics than was observed with the linear kernel.

The highest accuracy with the linear kernel was higher than 0.9 for the UteroCervix ( $0.90 \pm 0.02$  with Five\_metrics), the Bladder ( $0.92 \pm 0.02$  with Four\_metrics), the Rectum ( $0.95 \pm 0.03$ , with Five\_metrics), and the Kidneys ( $0.97 \pm 0.02$  with Five\_SDSC) and just below 0.9 for the Nodal CTV ( $0.89 \pm 0.04$  with Three\_SDSC) and the PAN ( $0.88 \pm 0.03$  with Nine\_metrics). The overall accuracies, sensitivities of detecting erroneous contours, and specificities for the single- and multi-metric analyses with the linear kernel are presented in Table 9.

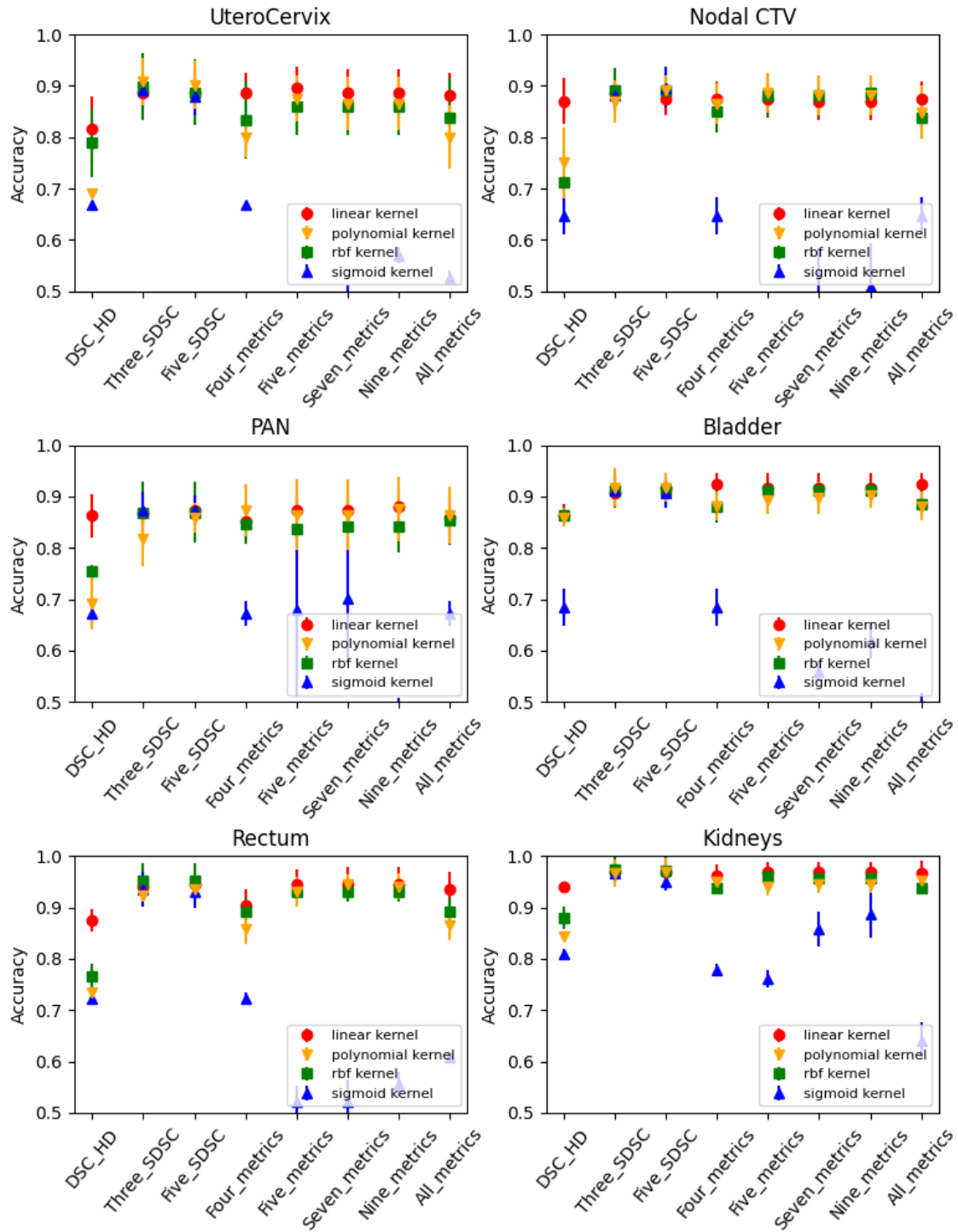


Figure 15. Average accuracies of SVM model with multiple metrics for each structure. The error bar represents  $\pm 1$  standard deviation. Four different kernels (linear, polynomial, radial basis function (rbf), and sigmoid) were tested.

Table 9. Overall accuracies, sensitivities, and specificities from the single-metric and multi-metric analyses, when SVM was used with the linear kernel.

<b>Single-metric (accuracy/sensitivity/specificity)</b>	<b>UteroCervix</b>	<b>CTVn</b>	<b>PAN</b>	<b>Bladder</b>	<b>Rectum</b>	<b>Kidneys</b>
DSC	0.82/0.59/0.94	0.87/0.67/0.98	0.87/0.68/0.97	0.88/0.69/0.96	0.88/0.70/0.94	0.94/0.71/1.00
HD_100	0.78/0.46/0.94	0.72/0.33/0.93	0.71/0.32/0.90	0.85/0.66/0.93	0.75/0.22/0.96	0.86/0.47/0.95
HD_95	0.77/0.46/0.92	0.76/0.53/0.89	0.66/0.00/0.93	0.87/0.72/0.94	0.76/0.33/0.93	0.91/0.65/0.97
MSD	0.87/0.74/0.93	0.88/0.73/0.96	0.85/0.65/0.95	0.91/0.78/0.98	0.89/0.78/0.93	0.95/0.82/0.98
SDSC 1 mm	0.91/0.82/0.95	0.90/0.76/0.97	0.83/0.72/0.89	0.89/0.76/0.95	0.94/0.87/0.97	0.93/0.73/0.97
SDSC 2 mm	0.89/0.79/0.94	0.90/0.78/0.97	0.86/0.68/0.95	0.91/0.79/0.97	0.94/0.86/0.97	0.97/0.90/0.99
SDSC 3 mm	0.88/0.74/0.94	0.88/0.74/0.96	0.89/0.71/0.98	0.92/0.78/0.99	0.92/0.81/0.97	0.96/0.82/0.99
SDSC 4 mm	0.89/0.77/0.94	0.88/0.73/0.96	0.87/0.69/0.97	0.91/0.76/0.99	0.91/0.77/0.97	0.95/0.74/1.00
SDSC 5 mm	0.88/0.74/0.94	0.88/0.70/0.98	0.87/0.67/0.98	0.89/0.69/0.99	0.89/0.69/0.97	0.94/0.71/1.00
SDSC 7 mm	0.85/0.67/0.94	0.87/0.64/0.99	0.87/0.68/0.98	0.89/0.68/0.99	0.86/0.59/0.96	0.92/0.61/1.00
SDSC 10 mm	0.80/0.53/0.94	0.79/0.41/1.00	0.83/0.58/0.96	0.86/0.57/1.00	0.78/0.31/0.96	0.86/0.29/1.00
<b>Multi-metric (accuracy/sensitivity/specificity)</b>						
DSC + HD_100	0.82/0.61/0.92	0.87/0.67/0.98	0.86/0.68/0.96	0.86/0.69/0.94	0.88/0.68/0.95	0.94/0.71/1.00
Three_SDSC	0.89/0.77/0.94	0.89/0.76/0.96	0.87/0.71/0.95	0.91/0.77/0.98	0.94/0.87/0.97	0.97/0.87/0.99
Five_SDSC	0.89/0.77/0.94	0.88/0.73/0.96	0.87/0.69/0.97	0.92/0.80/0.98	0.95/0.87/0.98	0.97/0.89/0.99
Four_metrics	0.89/0.80/0.93	0.88/0.73/0.96	0.85/0.64/0.96	0.92/0.82/0.98	0.90/0.77/0.96	0.96/0.87/0.98
Five_metrics	0.90/0.82/0.94	0.88/0.73/0.96	0.87/0.71/0.96	0.92/0.80/0.98	0.95/0.87/0.98	0.97/0.89/0.99
Seven_metrics	0.89/0.80/0.93	0.87/0.73/0.95	0.87/0.69/0.97	0.92/0.80/0.98	0.95/0.87/0.98	0.97/0.89/0.99
Nine_metrics	0.89/0.80/0.93	0.87/0.73/0.95	0.88/0.69/0.98	0.92/0.80/0.98	0.95/0.87/0.98	0.97/0.89/0.99
All_metrics	0.88/0.82/0.91	0.88/0.73/0.96	0.86/0.66/0.96	0.92/0.82/0.98	0.94/0.83/0.98	0.97/0.87/0.99

#### 4.3.2 Bony structures

DSC and HD\_100 thresholds were visually determined based on the distributions in Figure 16. The DSC/HD\_100 thresholds were 0.90/10.0 mm for the femurs, 0.80/10.0 mm for the spinal cord, 0.85/15.0mm for the pelvis, 0.85/25.0 mm for the sacrum, 0.85/7.5 mm for the L4, and 0.80/10.0 mm for the L5. As there were no failure cases for the spinal cord, the thresholds were determined based on those for the cervical spinal cord from the previous study with the cervical spinal cord [21] and the distributions of the clinically acceptable cases.

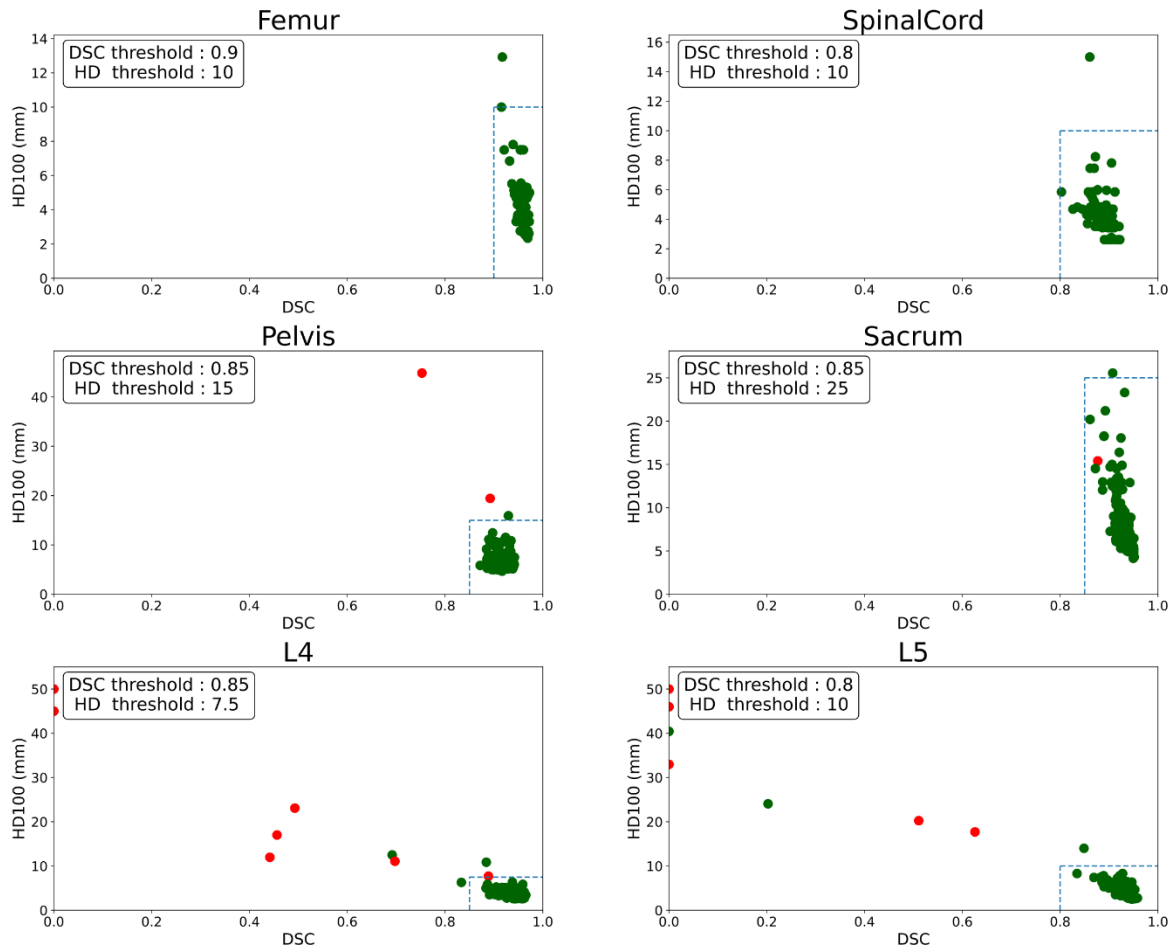


Figure 16. Distributions of DSC and HD<sub>100</sub> for the clinically acceptable (green markers) and unacceptable (red markers) cases for the 6 bony structures. The thresholds between the clinically acceptable and unacceptable contours were manually determined based on the distributions.

#### 4.4 Discussion

In this study, we demonstrated that errors in automatically generated contours can be detected by comparing the contours with other automatically generated contours. By

choosing appropriate similarity metrics and using the SVM classification algorithm, we were able to achieve an accuracy higher than 0.9 for most of the structures.

In the single-metric analysis, we found that the volumetric DSC and MSD are effective indicators for determining the similarity between the two contours in terms of error detection. On the other hand, the HDs (HD\_100 and HD\_95) often failed to detect contouring errors in our approach, demonstrating that most realistic contouring errors were not caused by a substantial failure in a single or small part of the contour and possibly explaining why the surface DSC was very effective at detecting contouring errors. As the surface DSC only uses the volume of the shell, the metric may indicate the overall similarity between the two contours near the surface. MSD is similar, but any small discrepancy between the two contours in each calculation point can contribute to the MSD. On the other hand, the surface DSC is more effective as the user can choose the tolerance value and anything below the tolerance will not contribute to reducing the surface DSC. For the surface DSC with a tolerance higher than 5 mm, however, the accuracy decreased substantially; thus, it is recommended to use the surface DSC with a tolerance of less than 5 mm to compare two contours in future studies.

The non-linear kernels (radial basis function, polynomial, or sigmoid) in the multi-metric analysis did not improve the model's accuracy. Since we had slightly more than 100 data points for training and dozens for the validation of each structure, using the more sophisticated, non-linear kernels could have resulted in overfitting of data points. We believe that the contour QA model with the linear kernel had more consistent



results over different combinations of metrics because the linear kernel itself operated as a regularization parameter in the SVM as a result of its inflexible shape.

Overall, using combinations of multiple metrics did not improve the accuracy of detecting contouring errors compared to the most accurate single-metric case. This could be because we did not have enough data to fine-tune the thresholds to substantially improve the results of the single-metric approach. In addition, because most of the metrics are already strongly correlated with each other, the classification model might have not been able to learn useful information from the additional metrics. In any case, using a single metric to flag incorrect contours performed as accurate as or even more accurate than did combinations of multiple metrics and makes it easier for users to interpret the results. Furthermore, considering the variations in the sizes and shapes of the structures used in this study, the single-metric approach should be feasible for most of the structures in various treatment sites. Therefore, we believe that the single-metric approach, especially using the surface DSC metric, is the best approach to detect contouring errors utilizing two auto-contouring systems and is expandable to other treatment sites. Although preliminary, this work indicates that a SDSC\_2 threshold of 0.54 may be a reasonable starting point for a wide variety of structures.

As shown in Table 9, the sensitivities were usually lower than the specificities. As described above, clinically acceptable contours were much more common than were clinically unacceptable contours in the auto-contouring systems we used. We intentionally added manually generated clinically unacceptable contours to reduce data imbalance, but the ratio between the clinically acceptable and unacceptable contours

was still 2 to 1 or even 3 to 1. Consequently, the SVM algorithm was likely to sacrifice sensitivity to increase specificity, which helped to improve the overall accuracy because of the imbalance in data. Fortunately, the thresholds in the SVM are easier to interpret and adjust than are those from other machine learning algorithms, such as neural networks; thus, we can fine-tune the thresholds through ROC analysis to achieve the desired sensitivity in the current model without sacrificing the overall accuracy much, as shown in Table 6 and Table 7.

As the majority of the clinically unacceptable contours were manually introduced to mimic a potential error, the distribution of errors and the derived accuracies presented in this study might not fully reflect the actual performance of the model on auto-contours. Yet the manually generated erroneous contours were less dramatically failed and much closer to clinically acceptable contours than clinically unacceptable auto-contours based on our observations. Therefore, the performance of the contour QA model was unlikely to be overestimated, but rather underestimated by adding these manual contours.

We predicted the verification auto-contours on its training dataset to reduce the false positives. Also, as the majority of the unacceptable contours were manually introduced to mimic a potential error that can be made by a human or a deep learning algorithm, the distributions of the metrics and the derived accuracies in this study might not fully reflect the actual performance of the contour QA model on the auto-contouring systems. Yet, the false positives confuse the contour QA model to determine the accurate thresholds, because even if the QA model correctly predicts that one of the auto-contours was clinically unacceptable, the model prediction would be considered to

be incorrect. Furthermore, in our preliminary study, the metric points corresponding to both the clinically acceptable and unacceptable contours near the thresholds were not sufficient, as shown in Figure 17. The thresholds could have been chosen anywhere in between the red and blue dashed lines on the right figure in Figure 17 when the manual contours were not included. Therefore, predicting the verification auto-contours on its training dataset and adding the manual contours helped us developing more robust contour QA models.

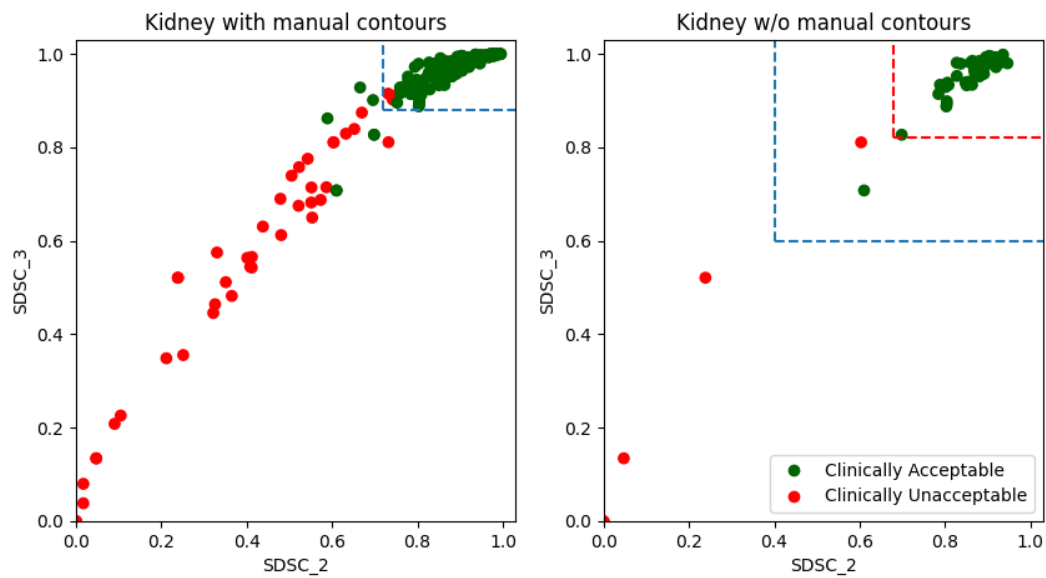


Figure 17. The surface DSC distributions of the clinically acceptable and unacceptable kidney contours with (left) and without (right) the manually generated contours. The thresholds can be confidently determined with the manual contours, whereas the threshold can be anywhere between the blue and red dashed lines without the manual contours due to an insufficient amount of data.

For the bony structures, we have visually checked all the false positives (i.e. green markers outside the thresholds), false negatives (i.e. red markers inside the thresholds), and most of the true positives and true negatives near the thresholds. The false positives were almost always due to the failure in the secondary auto-contouring system when the contours from the primary auto-contouring system were acceptable. Again, this does not deviate from the central hypothesis since one of the contours was still wrong, although the specificity of the model would become lower. All the false negatives or the true positives near the thresholds were due to the small missing parts of the structures. This means that the physicians were not satisfied with these contours, but the contours would be very unlikely to affect the quality of the final bony structure-based 4-field box plans. The auto-planning algorithm for the bony structure-based 4-field box plans uses the projected 2D contours to determine the beam apertures, and small flaws in the 3D contour do not usually affect the projected contours substantially.

Due to the lack of failure cases, the sensitivities and specificities were not able to be calculated using an independent test dataset for the bony structures. We could only approximate the performance based on the training dataset, and the specificities are over 96% for all structures and the sensitivities are 100% except for the sacrum where there is only one failure case that was incorrectly classified. With more clinical cases acquired from the RPA system in the future, we will be able to validate the performance of the models on the bony structures.

#### 4.5 Conclusion

We demonstrated that the discrepancy between two independently generated auto-contours is a strong indicator of an error in one of the contours. The most accurate similarity metric to detect contouring errors was surface DSC with a tolerance of 1, 2, or 3 mm. With this approach, we were able to achieve the average error detection sensitivity higher than 0.9 while the average specificity was higher than 0.8 for the targets and critical structures in the female pelvis. This method can be used to automatically detect errors in auto-contours to reduce the risks associated with the use of automated radiotherapy tools.

## Chapter 5 : Automated radiation treatment planning for cervical cancer radiation treatment

Several paragraphs in the introduction of this chapter are partially based on the following article:

Rhee, D.J., Jhingran, A., Kisling, K., Cardenas, CE., Simonds, H., and Court, L.E. (2020), Automated Radiation Treatment Planning for Cervical Cancer. *Seminars in Radiation Oncology*. <https://doi.org/10.1016/j.semradonc.2020.05.006>

### 5.1 Introduction

The prescription dose can be delivered to the target with various delivery techniques. Physicians choose the optimal technique based on patients' conditions and the capability of the medical linac they possess and other factors such as patient throughput and resources to do patient-specific QA. For external beam radiotherapy for cervical cancer, the common beam delivery techniques are 2D 4-field box, 3D conformal radiotherapy (3D-CRT), IMRT, and VMAT, and several studies have been conducted to automate such techniques and validate the performance of the automation tools.

The automation algorithm of the 2D 4-field-box technique for cervical cancer was developed by Kisling et al. [7]. The beam apertures were determined based on the bony landmarks in 2D projected CT scans for each gantry angle, and the bony structures were automatically contoured using a multi-atlas-based auto-contouring system. IMRT or VMAT plans can be automatically generated using the commercially

available knowledge-based planning (KBP) software programs, such as RapidPlan (Varian Medical Systems, Palo Alto, CA) and Erasmus-iCycle (Elekta AB, Stockholm, Sweden). The performance of KBP models for cervical cancer has been validated in many research studies. Ma et al. [64] tested an IMRT RapidPlan model for postoperative cervical cancer patients and showed that planning target volume coverage was within 1% and critical organ dose metrics were within 4% of manual plan results. Li et al. [65] and Tinoco et al. [66] showed that IMRT and VMAT RapidPlan models for cervical cancer patients are better than or equal to clinical plans. Sharfo et al. [67] showed that, for patients with cervical cancer, their dual-arc VMAT Erasmus-iCycle model created plans that were equivalent to or better than manually generated dual-arc VMAT and 9-beam IMRT. Thus, an automatically generated IMRT or VMAT plan for cervical cancer made using KBP techniques will be clinically acceptable if the user can provide high-quality plans for model training.

In this study, we developed algorithms that can automatically generate 2D 4-field-box and 3D-CRT plans for cervical cancer. We also automated the field-in-field algorithm to improve the quality of these plans. The VMAT technique was automated with the RapidPlan software. Unlike most of the auto-planning studies mentioned above, where the plans were generated using manual contours, we combined the auto-planning algorithms with the auto-contouring system in Chapter 3 to fully automate the radiotherapy plan generation process for cervical cancer with minimal human input. The quality of the plans was evaluated by multiple physicians from various countries. The fully automated radiotherapy planning system for cervical cancer is implemented in the RPA system to aid under-resourced clinics in low- and middle-income countries.

## 5.2 Methods

We developed the auto-planning systems for cervical cancer with 3 different treatment techniques: 4-field-box with bony landmarks (4-field-box), 3D conformal radiation therapy (3D-CRT), and volumetric modulated arc therapy (VMAT). The auto-planning systems were developed to treat cervical cancer patients with an intact uterus and without vaginal or PAN involvement. The systems were integrated with the auto-contouring system described in Chapter 3 to create fully automated radiotherapy planning for cervical cancer on CT images. The users only need to upload the CT images, select ITV, PTV, and beam aperture margins, and prescribe dose.

### 5.2.1 4-field-box plans with bony landmarks

The beam apertures for the 4-field-box plans were determined based on the algorithm from Kisling et al.'s study [7], [68]. In this study, the 3D contours of the pelvic bone, the sacrum, the left and right femurs, and the L4 and L5 vertebral bodies were generated on CT images, and the contours were projected in 0°, 90°, 180°, and 270° gantry angles. In each projection angle, certain bony landmarks from the projected contours were detected, and the beam apertures were shaped based on these bony landmarks.

In clinical practices, a calculation point is often used to normalize the plan. This point is arbitrarily determined by the clinician based on the patient geometry and initial dose distribution. As the automation of determining the calculation point is challenging, we defined a volume called “synthetic PTV” and normalize the plan with this volume. The synthetic PTV was defined based on the beam apertures as shown in Figure 18.



All the plans were normalized such that 100% of the prescription dose covers 97% of the synthetic PTV.

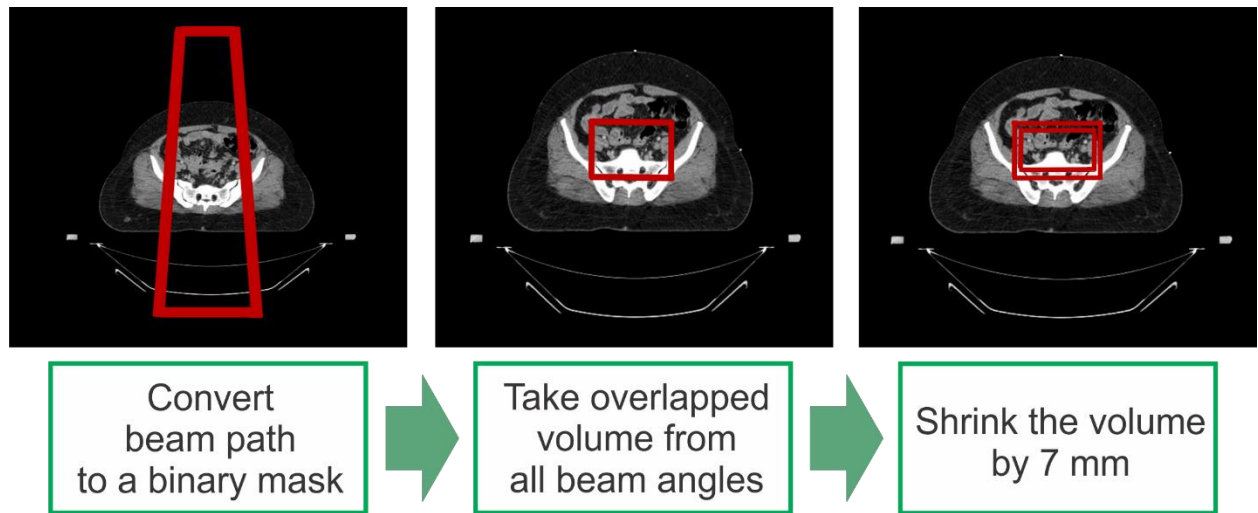


Figure 18. Synthetic PTV structure was defined based on the beam apertures for the 4-field-box plans. Firstly, the beam path from each beam angle was converted into a 3D binary mask, and then the volume overlapped by each mask was defined to be the region of hot spot detection (RHD). Finally, the synthetic PTV was created from 7 mm shrinkage from the RHD, the overlapped volume.

### 5.2.2 3D-CRT plans with the CTV contours

The beam apertures for the 3D-CRT plans were determined based on the projected PTV contours from 0°, 90°, 180°, and 270° gantry angles. The primary and the nodal CTV contours were generated from the auto-contouring system, and then the PTV was derived from the CTVs with the basic image-guided radiotherapy (IGRT)

margins for the primary CTV described in the GEC-ESTRO EMBRACE II protocol (10 mm in anterior, posterior and superior directions, 5 mm in lateral directions) [54] and 5 mm PTV margin. Finally, a 7 mm uniform margin was applied to the projected PTV to determine the beam shape at each gantry angle. The step-by-step process of generating the 3D-CRT plan is demonstrated in Figure 19. The plans were normalized such that 100% of the prescription dose covers 95% of the PTV.

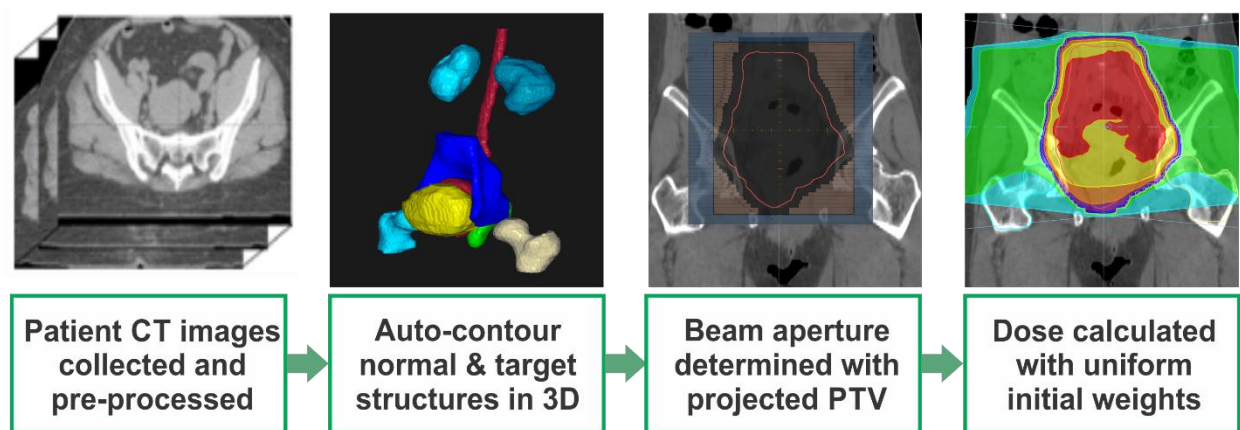


Figure 19. Workflow of the automated 3D-CRT system for cervical cancer. The PTV was derived from the automatically generated CTVs. The beam apertures were determined with a user-defined uniform margin (7 mm in this study) around the projected PTV. The dose was calculated with a pre-defined MU.

### 5.2.3 Field-in-field technique

Even with correctly generated beam apertures for the 4-field-box or 3D-CRT plans, the plans might not be clinically acceptable due to hot spots. In this study, hot spots

were defined to be any volume larger than 2 cc receiving more than 107% of the prescription dose. To reduce the hot spots and make the plans clinically more acceptable, we automated the field-in-field (FIF) technique, often used in clinics to reduce hot spots in these plans.

The FIF technique was automated by mimicking human planners as described in Figure 20. Firstly, with the given beam apertures, the dose was calculated with arbitrarily defined uniform MUs in the treatment planning system of choice, which was Eclipse in this study. Then, the optimal MUs were found with the optimization algorithm described in Chapter 5.2.3.1, and the plan was normalized to achieve the coverage we wanted for each treatment technique. If hot spots exist with the optimized plan, a sub-field was created to block the hot spots in the left of the right lateral beam alternatively, as shown in Figure 21. The plan was re-optimized with the updated sub-field and checked if the hot spots were removed. The FIF algorithm stopped if hot spots no longer existed or the number of sub-fields exceeded 6.



### 5.2.3.1 MU optimization algorithm

To find the optimized MU with the given beam apertures, we created an objective function, and then find the minimum point of the objective function using a convex optimization algorithm. To define the objective functions for the MU optimization algorithm, the PTV and the region of hotspot detection (RHD) were defined for each plan. The PTV was the synthetic PTV in Figure 18 for the 4-field-box plans and the PTV derived from the automatically generated CTVs for the soft-tissue-based 3D-CRT plans. The RHD was defined to be the overlapping volume from all beam paths, as demonstrated in Figure 18, to include all the regions that hot spots could exist. Therefore, the hot spots were only scanned in the RHD volume. This speeded up the optimization algorithm by restricting the number of voxels that were scanned and optimized in the objective function. The PTV and RHD for each treatment technique were demonstrated in Figure 22.

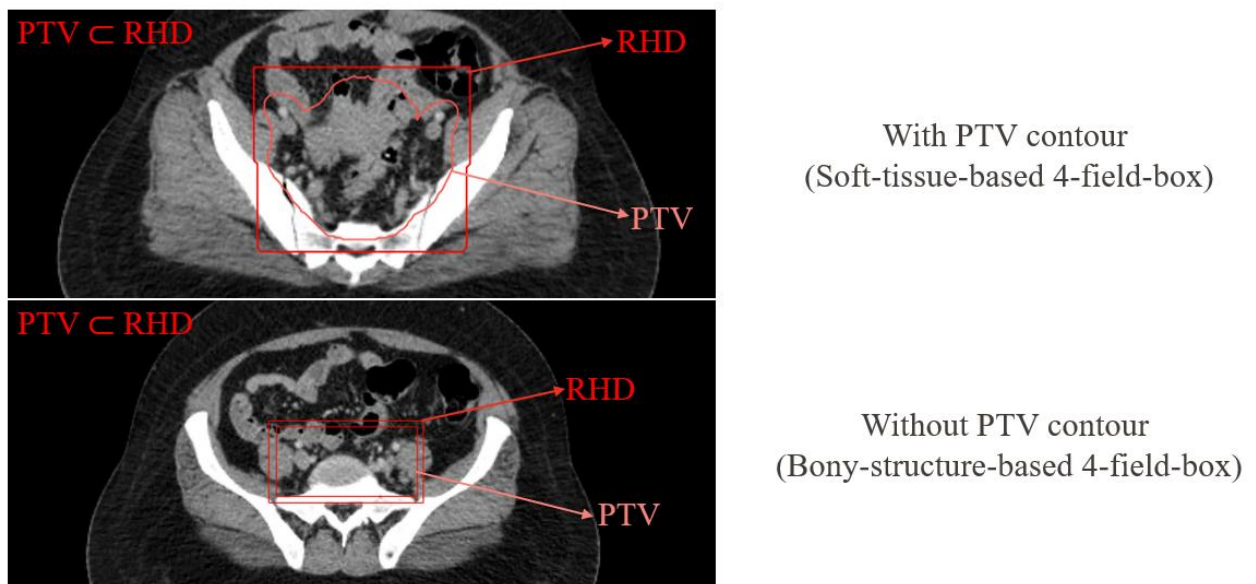


Figure 22. PTV and RHD definitions for the FIF optimization algorithm.

With the given volumes, the objective function  $\Omega$  takes MU from each beam as an input and is defined as

$$\begin{aligned} \Omega(MU) = & w_{max} * \sum (\max(d_{RHD} - d_{RX} * 1.07, 0))^2 \\ & + w_{min} \sum (\max(d_{RX} - d_{PTV}[0: 0.95 * N_{PTV}], 0))^2 \end{aligned} \quad (5)$$

where  $d_{RX}$  is the prescription dose,  $d_{RHD}$  and  $d_{PTV}$  are the dose arrays from RHD and PTV in descending order, respectively,  $N_{PTV}$  is the number of voxels in PTV, and  $w_{min}$  and  $w_{max}$  are the weighting factors for each term. In this particular equation, the objective is penalized when any volume receives more than 107% of the prescription dose and more than 5% of the PTV receives less than 100% of the prescription dose. These criteria can be modified easily by changing the parameters in the configuration file of the FIF automation program.

We used the trust-region optimization method [69] to find the minimum value of the objective function as this method allowed us to set the boundary conditions of the input MUs. We set the MU range to be  $[7 MU, \infty)$ , and the maximum number of iterations for the optimization algorithm was 1000.

#### 5.2.4 VMAT

We trained a Varian RapidPlan model to automatically generate the VMAT plans for cervical cancer. To train the RapidPlan model, we collected 97 VMAT plans, where 42 of them were with PAN involvement and 55 of them were without PAN involvement. The RapidPlan model can treat up to 3 dose levels, and use 6 MV photon beams, 3 full arcs, and 3 collimator angles including  $10^\circ$ ,  $90^\circ$ , and  $350^\circ$ . For planning objectives, we

used bladder, bowel space, femoral heads, kidneys, liver, rectum, spinal cord, and bone marrow contours, and the bone marrow contour was defined to be the summation of the pelvis, sacrum, femoral heads, and L5 vertebral body. The planning objectives of the PTV were set to achieve 95% of the PTV covered by 100% of the prescription dose, while the maximum dose was less than 107% of the prescription dose.

#### 5.2.5 Plan review

The radiotherapy plans were generated on 35 CT scans from 3 different South African hospitals with the 3 planning techniques. In total, 5 experienced radiation oncologists (3 physicians from South Africa, 1 physician from the United States (MD Anderson), and 1 physician from the United Kingdom) scored these plans and each plan for each technique was scored by two of these physicians. The plans were assigned to each physician in a way that physicians only reviewed the techniques they were experienced in using clinically. We used the Likert scale, a 5-scale scoring system, to evaluate the plans as defined in Table 10.

Table 10. Likert scale to score automatically generated radiotherapy plans

<b>Score</b>		<b>Description</b>
<b>5</b>	<b>Strongly agree</b>	Use-as-is. Clinically acceptable and the plans could be used for treatment without change.
<b>4</b>	<b>Agree</b>	Minor edits that are not necessary. Stylistic change preferred, but not clinically important. The current plans are clinically acceptable.
<b>3</b>	<b>Neither agree nor disagree</b>	Minor edits that are necessary. Minor edits are those that the review judges can be made in less time than starting from scratch or are expected to have minimal effect on treatment outcome.
<b>2</b>	<b>Disagree</b>	Major edits. The necessary edits are required to ensure appropriate treatment, and sufficiently significant that the user would prefer to start from scratch.
<b>1</b>	<b>Strongly disagree</b>	Unusable. The quality of the automatically generated plans is so bad that they are unusable.

We showed a few plans to each physician first to see if they were satisfied with the plan quality. If the plans were consistently scored low for the same reasons (e.g. plan being too hot or too cold), we adjusted all the plans by changing the FIF parameters or renormalization based on the physician's preference. For example, for the 3D-CRT plans, one of the reviewers found the plans to be too hot and wanted 95% of the PTV was covered by 95% of the prescription dose, instead of 100% of the prescription dose. After adjusting the plans based on their preferences, we asked them to review the plans from the beginning.



### 5.3 Results

The overall review results are shown in Table 11. The reviewer number in the table was randomly assigned to indicate that two physicians review each set of plans independently.

Table 11. Physician scoring results for each technique with each review session. The plan criteria for the coverage and the maximum dose were presented. The reviewer numbers are arbitrarily assigned.

Treatment Techniques	Reviewer #	Coverage (Rx/PTV)	Max dose	# of plans in each score				
				5	4	3	2	1
4-field-box	1	100%/97%	107%	9	20	5	0	1
	2	100%/97%	105%	28	4	2	0	1
3D-CRT	1	100%/95%	107%	27	7	1	0	0
	2	95%/99%	105%	3	32	0	0	0
VMAT	1	100%/95%	107%	16	15	3	1	0
	2	95%/99%	107%	35	0	0	0	0

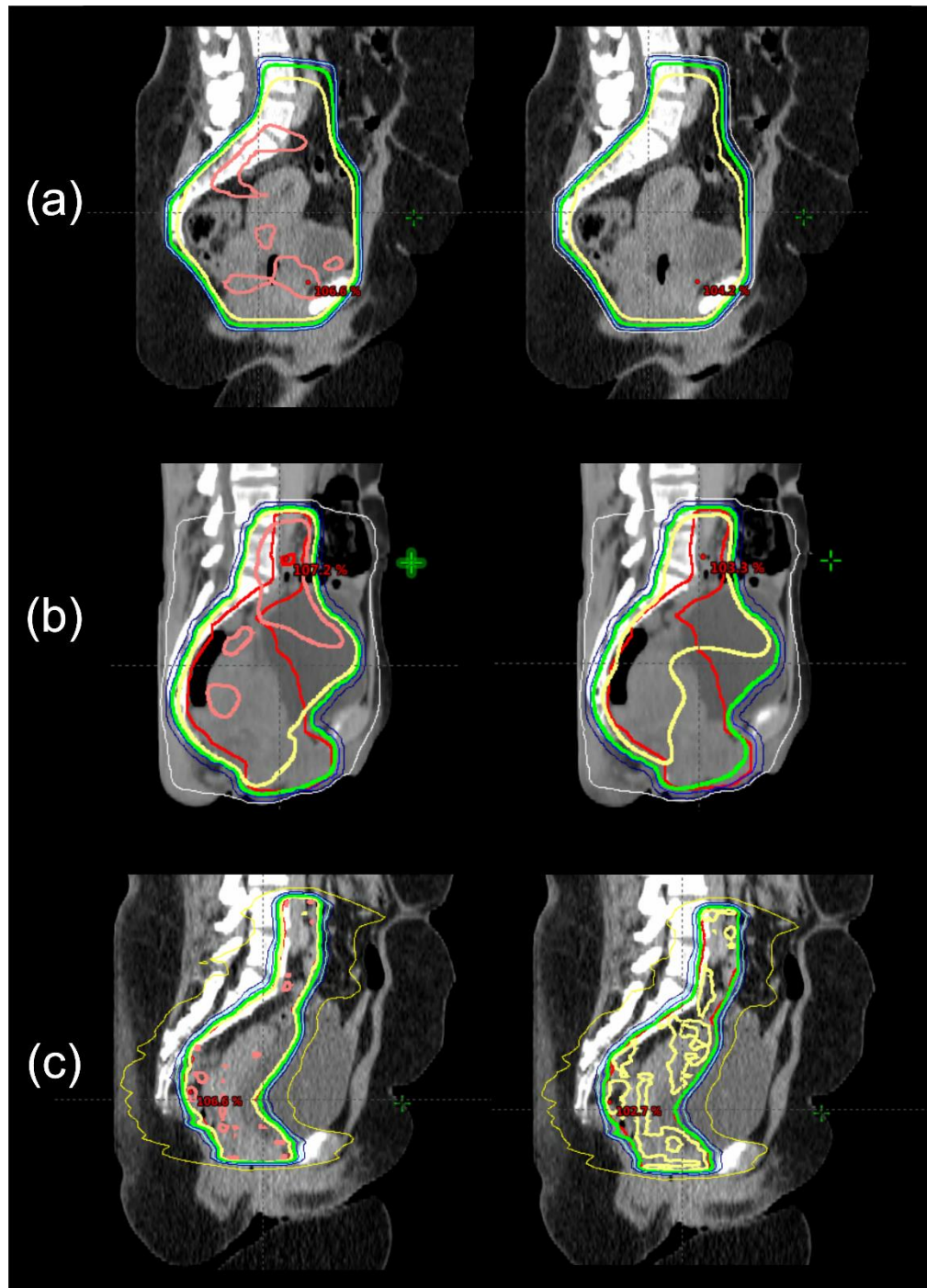


Figure 23. Demonstration of the customized plans for Reviewer #1 (left) and Reviewer #2 (right) for (a) 4-field-box, (b) 3D-CRT, and (c) VMAT techniques. The thick lines represent the PTV (red), 105% isodose line (pink), 100% isodose line (yellow), and 95% isodose line (green).

### 5.3.1 4-field-box

Reviewer #1 was satisfied with the original plans where the maximum dose was less than 107% of the prescription dose. This reviewer scored 29 plans to be clinically acceptable without modification (score  $\geq 4$ ), 5 plans to be minor edits required, and 1 plan to be unusable. The plans scored as 3 were all due to the excessive bowel dose. The plan that was scored 1 was due to the failure in generating correct L4 contours, and therefore, the superior borders of the beam apertures were incorrectly defined.

As Reviewer #2 wanted the dose above 105% to be minimized and not to be in the rectum, we changed the hotspot criteria in the FIF objective function from 107% to 105%. Consequently, most of the 105% isodose lines from the original plan were removed from the updated plan, as shown in Figure 23 (a), although the number of sub-fields was usually increased with the updated plan. This reviewer scored 32 plans to be clinically acceptable without modification. Two plans were scored 3 due to insufficient PTV coverage. The plan that was scored 1 was the same plan that was identified by the first reviewer.

### 5.3.2 3D-CRT

Reviewer #1 was satisfied with the original plan where 95% of the PTV was covered by 100% of the prescription dose. This reviewer scored 34 plans to be clinically acceptable without modification. One patient was scored 3 because some of the important parts of the PTV were not covered by the prescription dose.

Reviewer #2 found the original plans to be too hot and preferred to have almost all (99%) of the PTV covered by 95% of the prescription dose. Furthermore, this reviewer

wanted to remove all 70% isodose lines outside the main treatment region (i.e. RHD). We adjusted all the plans by changing the parameters in the FIF algorithm based on these criteria, and the updated plans were cooler than the original plans, as shown in Figure 23 (b). After the adjustment, this reviewer scored all 35 plans to be clinically acceptable without modification (score  $\geq 4$ ). The reviewer gave most of the plans 4, mostly because the PTV in the most inferior slice was not fully covered by 95% of the prescription dose.

### 5.3.3 VMAT

Reviewer #1 was satisfied with the original plan where the planning objectives were set to have 95% of the PTV covered by 100% of the prescription dose. This reviewer checked overall dose distribution instead of checking certain dose metrics for the OARs. The reviewer scored 31 plans to be clinically acceptable without modification (score  $\geq 4$ ). The plans scored 2 and 3 were due to insufficient PTV coverage.

Reviewer #2 preferred the PTV to be covered by 95% of the prescription dose. We renormalized the plans such that 99% of the PTV was covered by 95% of the prescription dose to meet Reviewer #2's preferences. This renormalization usually made the plans cooler than the original plans, as shown in Figure 23 (c). The dose metrics for the OARs including the bladder, bowel, femurs, and rectum were also considered although some of the metrics for the bowel and rectum were ignored when the PTV was substantially overlapped with these structures. This reviewer scored all 35 plans as 5.

## 5.4 Discussion

Overall, 87%, 99%, and 94% of the automatically generated plans were evaluated to be clinically acceptable without modification (score  $\geq 4$ ) for 4-field-box, 3D-CRT, and VMAT plans, respectively. Although we had to adjust the plans to meet each physician's preference, this adjustment can be easily achieved by renormalizing the plan in the TPS or changing parameters in the configuration files in the FIF automation program. For the plans scored 3 or lower, we had feedback from the reviewers and carefully reviewed them again. The thresholds used for the final QA model is presented in **Error! Reference source not found..**

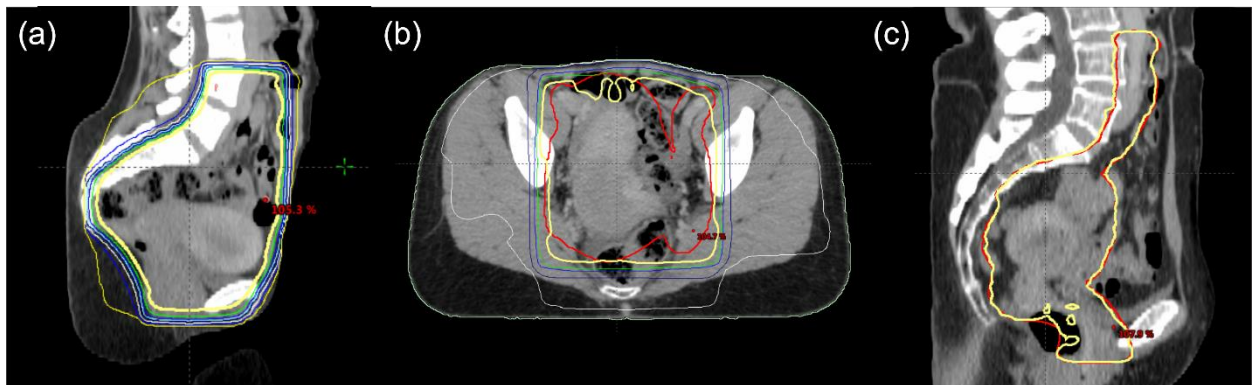


Figure 24. Examples of the plans scored 3 or lower. The thick yellow lines represent 100% isodose line and the thick red lines represent the PTV. (a) 4-field-box plan scored 3 because of the excessive dose to the bowel. (b) 3D-CRT plan scored 3 as the PTV (red) was not fully covered by 100% isodose line near the bowel. (c) VMAT plan scored 2 as the PTV (red) was not fully covered by 100% isodose line near the rectum.

#### 5.4.1 4-field-box

Five plans were scored 3 by Reviewer #1 for the 4-field-box plans, and they were all because of the excessive dose to the bowel. We found that 4 out of the 5 patients were underweight, as shown in Figure 24 (a), and the remaining patient was not underweight but wide laterally and narrow in the anterior-posterior direction. These patients have very limited space between the target (i.e. uterus and lymph nodes) and the bowel. The beam apertures for the 4-field-box plans were determined based on the bony landmarks, not the soft-tissue structures. These bony landmarks were chosen not to miss the targets for most of the patients, and therefore, the method to determine the beam shapes was not optimized to spare the dose to the bowel. As a result, the beam apertures of the 4-field-box plans are likely to excessively cover the bowel region for underweight patients. This is the limitation of the 4-field-box approach, not the automation system, so we would not count these cases in our performance evaluation.

There was a single case where the 4-field-box plan was scored as 1 and it was due to the incorrectly generated L4 contour. In our preliminary study, we asked physicians to score the automatically generated beam apertures on 103 patients. About 3% (3/103) of the beam apertures were evaluated as clinically unacceptable and all of the failures were due to the incorrectly created L4 contours. Therefore, the overall acceptance rate for the automatically generated 4-field-box plans was around 97% from these two studies.

#### 5.4.2 3D-CRT

Almost all the plans were scored greater than or equal to 4 for the 3D-CRT plans. Reviewer #2 scored one of the plans to be 3, and it was because the PTV was not

covered well and the 100% isodose line was not smooth, as shown in Figure 24 (b). For this patient, achieving good PTV coverage and the smooth isodose line was challenging as some parts of the PTV were too closely located to the surface of the body and the bowel was filled with gas.

Reviewer #1 scored most plans as 4 because the most inferior slices were not fully covered by the 95% isodose line. Similarly, the plans that were scored 4 by Reviewer #2 were mainly because the most inferior slices were not fully covered by the 100% isodose line. As the 3D-CRT plans were naively normalized by the PTV, the dose near the edges of the beam apertures, especially the most superior and inferior slices, were often colder than the rest of the PTV. We will further improve the FIF automation algorithm by modifying the objective function so that the PTV in the most inferior slice can be fully covered by the desired isodose line.

#### 5.4.3 VMAT

For the VMAT plans, Reviewer #1 scored 4 of the plans either 2 or 3. These plans failed to have good PTV coverages, as shown in Figure 24 (c). Three of them did not have good PTV coverage near the rectal region because of the gas-filled rectum like the case in Figure 24 (c). As mentioned from Chapter 3, the CT scans with gas-filled rectums were not supposed to be used for radiation treatment planning according to GEC-ESTRO EMBRACE II protocol [54], so we believe that the actual clinical acceptance rate for the automatically generated VMAT plans can be higher if the clinical protocol was strictly followed.

## 5.5 Conclusion

We have demonstrated that the auto-planning system, combined with the auto-contouring system, can generate clinically acceptable plans with three different beam delivery techniques for cervical cancer radiation treatment. The plans should be optimized to meet each user's preference. More than 90% of the automatically generated plans were clinically acceptable for all three techniques. The auto-planning system has been implemented into the RPA to aid under-resourced hospitals in low- and middle-income countries.



## Chapter 6 : Discussion and Conclusion

### 6.1 Project Summary

The main goal of this project is to automate the radiation treatment planning process for cervical cancer with an intact uterus and without vaginal or PAN involvement. To achieve this goal, we first developed an auto-contouring system for multiple CTVs and normal structures in the female pelvis in Aim 1. Then, we developed auto-planning systems for 3 different planning techniques for cervical cancer in Aim 2. We evaluated the quality of the automatically generated plans from the auto-contouring and auto-planning systems to ensure that most of the auto-plans are clinically acceptable. Furthermore, for safety purposes, we developed the automatic contour QA system to detect potential errors in the automatically generated plans in Aim 3.

In Aim 1, we developed the auto-contouring system for 3 CTVs and 12 normal structures for cervical cancer radiation treatment. We trained CNN-based classification and segmentation architectures using 2254 CT scans from MD Anderson to automate the contouring process. We evaluated the auto-contours on 30 CT scans from 3 hospitals in South Africa and showed that the CNN-based auto-contouring system can achieve clinically acceptable contours for 79% of the CTVs and 97% of the OARs. The quantitative study using the similarity metrics such as DSC and HD showed that our auto-contouring system performed comparably to or outperformed state-of-the-art auto-contouring systems developed by other groups.

In Aim 2, the 3 planning techniques for cervical cancer, 2D 4-field-box, 3D-CRT, and VMAT, were automated and combined with the auto-contouring system in Aim 1. To reduce the hotspots for 2D 4-field-box and 3D-CRT auto-plans, we also automated

the FIF algorithm. Thirty-five plans were generated on the CT scans from 3 South African hospitals for each planning technique. Each plan was evaluated by two out of the five physicians from South Africa, the United States, and the United Kingdom. Overall, 97%, 99%, and 94% of the 2D 4-field-box, 3D-CRT, and VMAT plans were clinically acceptable, respectively. We have implemented our auto-contouring and auto-planning systems to the Radiation Planning Assistant (RPA) to accelerate the radiation treatment planning process in hospitals in low- and middle-income countries.

In Aim 3, we developed the automatic contour QA method to detect incorrectly generated auto-contours. We compared each contours from two independently developed auto-contouring systems using 11 similarity metrics and predicted that at least one of the auto-contours was incorrectly generated if a similarity metric was above or below the thresholds. We discovered that surface DSC with a tolerance of 2mm was the best metric to identify errors in contours. Our model was able to detect 90% of the incorrectly generated contours when the average specificity was 82%.

## 6.2 Study Limitations and Future Directions

Although we have accomplished the main goals of the project, there is still room for improvement of the overall system. Firstly, we have noticed that the auto-contouring system was susceptible to the high-density materials in the bowel, as shown in Figure 7 and Figure 11. Although the high-density materials in the bowel are not commonly seen in the patients at MD Anderson, approximately 5 to 10% of the African patients had this problem based on the CT scans from the partner hospitals in Africa owing to their diets.

We will include these patients in our training datasets when we update the auto-contouring system.

In the auto-contouring system, a set of CNN-based classification and segmentation architectures was trained per each structure to maximize the number of training datasets. This approach was effective to optimize the performance of the auto-contouring system with the given number of clinical data but inevitably made the system slower than the model that can simultaneously predict multiple structures. Once we collect enough datasets that contain all the CTV and OAR contours through the RPA system, we will train the new models that can predict categorized structures (e.g. bony structures, CTVs) simultaneously to speed up the plan generation process.

For the auto-planning system, we have learned that different physicians have different preferences on the PTV coverage and tolerance for hot spots. We showed that highly customized plans were able to be achieved by changing the parameters in the FIF automation algorithm and/or re-normalizing the plans based on each physician's preference. However, it might not be ideal to offer too many options to the users, as the users will be likely to make more mistakes when more options are available in the system. We could potentially offer a customized planning option with an official commission process led by us. However, from the RPA maintenance perspective, this could be problematic as we should repeat the commissioning process whenever a new option is introduced. Therefore, we will investigate whether various preferences can be eventually converged on a couple of options and if so, we will make these options available in the RPA system and ask the users to choose the closest options from their clinical practices.

The limitation of the automatic contour QA study was that the manually generated contours were used to develop and evaluate the performance of the contour QA model. Ideally, the QA model and its performance should be measured solely with the automatically generated contours, but the manual contours were added to the dataset to make the model more robust, as shown in Figure 17. Once we collect more clinically unacceptable contours through the RPA system, we will re-evaluate and potentially re-train the contour QA model.

Another limitation in our study was that we did not have enough clinically unacceptable auto-plans. One of the goals of this study was to investigate whether we could automatically detect errors in auto-plans. However, as we had only a few incorrectly generated plans in each technique, we were not able to establish a good plan QA model nor test the performance of the QA model. Instead, we developed the automatic contour QA method in Aim 3. According to the TG-275 report, incorrect target and OAR contours were the #1 and #7 highest risks of failure modes for external beam radiation treatment [16]. Other high-risk failure modes were mostly caused by human errors such as miscommunication about pacemakers or pregnancy, improper PTV margins, or wrong fractionation or prescription dose. In the RPA system, the QA methods for these errors were already implemented; asking the user to double-check if the patient was pregnant and warning the users if exceptional numbers for margins, fractionations, or prescription doses were entered. Furthermore, all the plans that were scored low with clinically acceptable contours were owing to insufficient PTV coverage or excessive hot spots, and we will flag these cases based on the final dose distributions of the plans. Therefore, we believed that the uncertainties in detecting

clinically unacceptable plans are mostly originated from the uncertainties in detecting clinically unacceptable contours, and thus the performance of the contour QA system is highly correlated with that of the final plan QA system. In the future, we will collect more clinically unacceptable plans through the RPA system, find other failure patterns from these plans, and develop an additional plan QA method with these patterns to make the RPA system safer.

### 6.3 Conclusion

In this study, we developed a fully automated radiation treatment planning system for cervical cancer for the under-resourced hospitals in low- and middle-income countries and potentially for the hospitals in developed countries. This was accomplished by developing the CNN-based auto-contouring system for the CTVs and OARs for cervical cancer and developing the auto-planning system with 3 different beam delivery techniques, including 2D 4-field-box, 3D-CRT, and VMAT. Overall, 97%, 99%, and 94% of the automatically generated 2D 4-field-box, 3D-CRT, and VMAT plans were clinically acceptable, respectively. Our hypothesis that 90% of the automatically generated plans for cervical cancer are clinically acceptable was proven by this study. Furthermore, to reduce the risk of clinically unacceptable plans used in clinics, we developed the automatic contour QA method that can detect the incorrectly generated auto-contours. With this method, 90% of the incorrectly generated contours were identified when the specificity was 82%. Therefore, another hypothesis of our study that we can detect 90% of the clinically unacceptable plans with 80% of the specificity was proven in this work.

## Appendix A

The best thresholds for major metrics for each structure to distinguish clinically acceptable and unacceptable contours. The actual thresholds used for the RPA system (Surface DSC with 2mm tolerance) are highlighted in bold.

	<b>UteroCervix</b>	<b>CTVn</b>	<b>PAN</b>	<b>Bladder</b>	<b>Rectum</b>	<b>Kidneys</b>
<b>DSC</b>	0.787	0.763	0.642	0.780	0.684	0.848
<b>HD100</b>	26.8	31.8	27.2	19.4	35.4	17.7
<b>SDSC_1</b>	0.328	0.303	0.251	0.321	0.273	0.306
<b>SDSC_2</b>	<b>0.534</b>	<b>0.536</b>	<b>0.445</b>	<b>0.537</b>	<b>0.483</b>	<b>0.673</b>
<b>SDSC_3</b>	0.694	0.686	0.570	0.755	0.637	0.820

## Bibliography

- [1] J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, and F. Bray, "Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012.," *Int. J. cancer*, vol. 136, no. 5, pp. E359-86, Mar. 2015, doi: 10.1002/ijc.29210.
- [2] American Cancer Society, "Cancer treatment and survivorship facts and figures 2016-2017," *Am. Cancer Soc.*, 2016, doi: 10.3322/caac.21235.
- [3] O. Balogun, D. Rodin, W. Ngwa, S. Grover, and J. Longo, "Challenges and Prospects for Providing Radiation Oncology Services in Africa.," *Semin. Radiat. Oncol.*, vol. 27, no. 2, pp. 184–188, Apr. 2017, doi: 10.1016/j.semradonc.2016.11.011.
- [4] J. Van Schelt, D. L. Smith, N. Fong, D. Toomeh, P. A. Sponseller, D. W. Brown, M. W. Macomber, N. A. Mayr, S. Patel, A. Shulman, G. V Subrahmanyam, K. N. Govindarajan, and E. C. Ford, "A ring-based compensator IMRT system optimized for low- and middle-income countries: Design and treatment planning study.," *Med. Phys.*, vol. 45, no. 7, pp. 3275–3286, Jul. 2018, doi: 10.1002/mp.12985.
- [5] D. W. Brown, A. Shulman, A. Hudson, W. Smith, B. Fisher, J. Hollon, Y. Pipman, J. Van Dyk, and J. Einck, "A framework for the implementation of new radiation therapy technologies and treatment techniques in low-income countries," *Phys. Medica*, vol. 30, no. 7, pp. 791–798, Nov. 2014, doi: 10.1016/J.EJMP.2014.07.004.

- [6] B. R. Page, A. D. Hudson, D. W. Brown, A. C. Shulman, M. Abdel-Wahab, B. J. Fisher, and S. Patel, "Cobalt, Linac, or Other: What Is the Best Solution for Radiation Therapy in Developing Countries?," *Int. J. Radiat. Oncol.*, vol. 89, no. 3, pp. 476–480, Jul. 2014, doi: 10.1016/J.IJROBP.2013.12.022.
- [7] K. Kisling, L. Zhang, H. Simonds, N. Fakie, J. Yang, R. McCarroll, P. Balter, H. Burger, O. Bogler, R. Howell, K. Schmeler, M. Mejia, B. M. Beadle, A. Jhingran, and L. Court, "Fully Automatic Treatment Planning for External-Beam Radiation Therapy of Locally Advanced Cervical Cancer: A Tool for Low-Resource Clinics," *J. Glob. Oncol.*, no. 5, pp. 1–9, 2019, doi: 10.1200/JGO.18.00107.
- [8] K. Kisling, L. Zhang, S. F. Shaitelman, D. Anderson, T. Thebe, J. Yang, P. A. Balter, R. M. Howell, A. Jhingran, K. Schmeler, H. Simonds, M. du Toit, C. Trauernicht, H. Burger, K. Botha, N. Joubert, B. M. Beadle, and L. Court, "Automated treatment planning of postmastectomy radiotherapy," *Med. Phys.*, vol. 46, no. 9, pp. 3767–3775, 2019, doi: 10.1002/mp.13586.
- [9] F. Milletari, N. Navab, and S. A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," *Proc. - 2016 4th Int. Conf. 3D Vision, 3DV 2016*, pp. 565–571, 2016, doi: 10.1109/3DV.2016.79.
- [10] E. Shelhamer, J. Long, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, 2017, doi: 10.1109/TPAMI.2016.2572683.
- [11] C. E. Cardenas, B. M. Anderson, M. Aristophanous, J. Yang, D. J. Rhee, R. E. McCarroll, A. S. R. Mohamed, M. Kamal, B. A. Elgohari, H. M. Elhalawani, C. D.



- Fuller, A. Rao, A. S. Garden, and L. E. Court, "Auto-delineation of oropharyngeal clinical target volumes using 3D convolutional neural networks," *Phys. Med. Biol.*, vol. 63, no. 21, p. 215026, Nov. 2018, doi: 10.1088/1361-6560/aae8a9.
- [12] B. Ibragimov and L. Xing, "Segmentation of organs-at-risks in head and neck CT images using convolutional neural networks," *Med. Phys.*, vol. 44, no. 2, pp. 547–557, Feb. 2017, doi: 10.1002/mp.12045.
- [13] C. E. Cardenas, J. Yang, B. M. Anderson, L. E. Court, and K. B. Brock, "Advances in Auto-Segmentation," *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 185–197, 2019, doi: <https://doi.org/10.1016/j.semradonc.2019.02.001>.
- [14] H. Vorwerk, K. Zink, R. Schiller, V. Budach, D. Böhmer, S. Kampfer, W. Popp, H. Sack, and R. Engenhart-Cabillic, "Protection of quality and innovation in radiation oncology: The prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study)," *Strahlentherapie und Onkol.*, vol. 190, no. 5, pp. 433–443, May 2014, doi: 10.1007/s00066-014-0634-0.
- [15] V. A. Andrianarison, M. Laouiti, O. Fargier-Bochaton, G. Dipasquale, X. Wang, N. P. Nguyen, R. Miralbell, and V. Vinh-Hung, "Contouring workload in adjuvant breast cancer radiotherapy.," *Cancer Radiother.*, vol. 22, no. 8, pp. 747–753, Dec. 2018, doi: 10.1016/j.canrad.2018.01.008.
- [16] E. Ford, L. Conroy, L. Dong, L. F. de Los Santos, A. Greener, G. Gwe-Ya Kim, J. Johnson, P. Johnson, J. G. Mechalakos, B. Napolitano, S. Parker, D. Schofield, K. Smith, E. Yorke, and M. Wells, "Strategies for effective physics plan and chart review in radiation therapy: Report of AAPM Task Group 275," *Med. Phys.*, vol.

n/a, no. n/a, doi: 10.1002/mp.14030.

- [17] J. Yang, B. Haas, R. Fang, B. M. Beadle, A. S. Garden, Z. Liao, L. Zhang, P. Balter, and L. Court, "Atlas ranking and selection for automatic segmentation of the esophagus from CT scans," *Phys. Med. Biol.*, vol. 62, no. 23, pp. 9140–9158, 2017, doi: 10.1088/1361-6560/aa94ba.
- [18] J. Yang, Y. Zhang, L. Zhang, and L. Dong, "Automatic Segmentation of Parotids from CT Scans Using Multiple Atlases," *Med. Image Anal. Clin. A Gd. Chall.*, 2010.
- [19] M. W. Macomber, M. Phillips, I. Tarapov, R. Jena, A. Nori, D. Carter, L. Le Folgoc, A. Criminisi, and M. J. Nyflot, "Autosegmentation of prostate anatomy for radiation treatment planning using deep decision forests of radiomic features," *Phys. Med. Biol.*, vol. 63, no. 23, p. 235002, Nov. 2018, doi: 10.1088/1361-6560/aaeaa4.
- [20] S. Nikolov, S. Blackwell, R. Mendes, J. De Fauw, C. Meyer, C. Hughes, H. Askham, B. Romera-paredes, A. Karthikesalingam, C. Chu, D. Carnell, C. Boon, D. D. Souza, S. A. Moinuddin, K. Sullivan, D. R. Consortium, H. Montgomery, G. Rees, R. A. Sharma, M. Suleyman, T. Back, J. R. Ledsam, and O. Ronneberger, "Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy," *Prepr. ArXiv*, pp. 1–31, 2018.
- [21] D. J. Rhee, C. E. Cardenas, H. Elhalawani, R. McCarroll, L. Zhang, J. Yang, A. S. Garden, C. B. Peterson, B. M. Beadle, and L. E. Court, "Automatic detection of contouring errors using convolutional neural networks," *Med. Phys.*, vol. 46, no.

- 11, pp. 5089–5097, Sep. 2019, doi: 10.1002/mp.13814.
- [22] W. Zhu, Y. Huang, L. Zeng, X. Chen, Y. Liu, Z. Qian, N. Du, W. Fan, and X. Xie, “AnatomyNet: Deep learning for fast and fully automated whole-volume segmentation of head and neck anatomy,” *Med. Phys.*, vol. 46, no. 2, pp. 576–589, 2019, doi: 10.1002/mp.13300.
- [23] C. E. Cardenas, R. E. McCarroll, L. E. Court, B. A. Elgohari, H. Elhalawani, C. D. Fuller, M. J. Kamal, M. A. M. Meheissen, A. S. R. Mohamed, A. Rao, B. Williams, A. Wong, J. Yang, and M. Aristophanous, “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.*, vol. 101, no. 2, pp. 468–478, 2018, doi: <https://doi.org/10.1016/j.ijrobp.2018.01.114>.
- [24] X. Feng, K. Qing, N. J. Tustison, C. H. Meyer, and Q. Chen, “Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images,” *Med. Phys.*, vol. 46, no. 5, pp. 2169–2180, May 2019, doi: 10.1002/mp.13466.
- [25] S. Wang, M. Zhou, O. Gevaert, Z. Tang, D. Dong, Z. Liu, and T. Jie, “A multi-view deep convolutional neural networks for lung nodule segmentation,” in *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, 2017, pp. 1752–1755. doi: 10.1109/EMBC.2017.8037182.
- [26] C. McIntosh, I. Svistoun, and T. G. Purdie, “Groupwise Conditional Random

- Forests for Automatic Shape Classification and Contour Quality Assessment in Radiotherapy Planning,” *IEEE Trans. Med. Imaging*, vol. 32, no. 6, pp. 1043–1057, 2013, doi: 10.1109/TMI.2013.2251421.
- [27] T. Lustberg, J. Van Soest, M. Gooding, D. Peressutti, P. Aljabar, J. Van Der Stoep, W. Van Elmpt, and A. Dekker, “Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer,” *Radiother. Oncol.*, vol. 126, no. 2, pp. 312–317, 2018, doi: 10.1016/j.radonc.2017.11.012.
- [28] H. R. Roth, H. Oda, X. Zhou, N. Shimizu, Y. Yang, Y. Hayashi, M. Oda, M. Fujiwara, K. Misawa, and K. Mori, “An application of cascaded 3D fully convolutional networks for medical image segmentation,” *Comput. Med. Imaging Graph.*, vol. 66, pp. 90–99, 2018, doi: <https://doi.org/10.1016/j.compmedimag.2018.03.001>.
- [29] P. Hu, F. Wu, J. Peng, Y. Bao, F. Chen, and D. Kong, “Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 12, no. 3, pp. 399–411, Mar. 2017, doi: 10.1007/s11548-016-1501-5.
- [30] X. Zhou, R. Takayama, S. Wang, T. Hara, and H. Fujita, “Deep learning of the sectional appearances of 3D CT images for anatomical structure segmentation based on an FCN voting method,” *Med. Phys.*, vol. 44, no. 10, pp. 5221–5233, Oct. 2017, doi: 10.1002/mp.12480.
- [31] X. Yang, L. Yu, L. Wu, Y. Wang, D. Ni, J. Qin, and P.-A. Heng, “Fine-grained Recurrent Neural Networks for Automatic Prostate Segmentation in Ultrasound

Images,” Dec. 2016, Accessed: May 02, 2018. [Online]. Available:  
<http://arxiv.org/abs/1612.01655>

- [32] C. Liu, S. J. Gardner, N. Wen, M. A. Elshaikh, F. Siddiqui, B. Movsas, and I. J. Chetty, “Automatic Segmentation of the Prostate on CT Images Using Deep Neural Networks (DNN).,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 104, no. 4, pp. 924–932, Jul. 2019, doi: 10.1016/j.ijrobp.2019.03.017.
- [33] S. Kazemifar, A. Balagopal, D. Nguyen, S. McGuire, R. Hannan, S. Jiang, and A. Owrangi, “Segmentation of the prostate and organs at risk in male pelvic CT images using deep learning,” *Biomed. Phys. Eng. Express*, vol. 4, no. 5, p. 55003, 2018, doi: 10.1088/2057-1976/aad100.
- [34] A. Balagopal, S. Kazemifar, D. Nguyen, M.-H. Lin, R. Hannan, A. Owrangi, and S. Jiang, “Fully automated organ segmentation in male pelvic CT images,” *Phys. Med. Biol.*, vol. 63, no. 24, p. 245015, 2018, doi: 10.1088/1361-6560/aaf11c.
- [35] A. L. Breto, O. Zavala-Romero, D. Asher, J. B. Baikovitz, J. Ford, R. Stoyanova, and L. Portelance, “A Deep Learning Pipeline for per-Fraction Automatic Segmentation of GTV and OAR in cervical cancer,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 105, no. 1, p. S202, Sep. 2019, doi: 10.1016/j.ijrobp.2019.06.267.
- [36] Z. Liu, X. Liu, B. Xiao, S. Wang, Z. Miao, Y. Sun, and F. Zhang, “Segmentation of organs-at-risk in cervical cancer CT images with a convolutional neural network,” *Phys. Medica*, vol. 69, pp. 184–191, 2020, doi: <https://doi.org/10.1016/j.ejmp.2019.12.008>.

- [37] K. H. Cha, L. M. Hadjiiski, R. K. Samala, H.-P. Chan, R. H. Cohan, E. M. Caoili, C. Paramagul, A. Alva, and A. Z. Weizer, "Bladder Cancer Segmentation in CT for Treatment Response Assessment: Application of Deep-Learning Convolution Neural Network-A Pilot Study," *Tomogr. (Ann Arbor, Mich.)*, vol. 2, no. 4, pp. 421–429, Dec. 2016, doi: 10.18383/j.tom.2016.00184.
- [38] R. Cheng, H. R. Roth, N. Lay, L. Lu, B. Turkbey, W. Gandler, E. S. McCreedy, T. Pohida, P. A. Pinto, P. Choyke, M. J. McAuliffe, and R. M. Summers, "Automatic magnetic resonance prostate segmentation by deep learning with holistically nested networks.," *J. Med. imaging (Bellingham, Wash.)*, vol. 4, no. 4, p. 41302, Oct. 2017, doi: 10.1117/1.JMI.4.4.041302.
- [39] K. Men, J. Dai, and Y. Li, "Automatic segmentation of the clinical target volume and organs at risk in the planning CT for rectal cancer using deep dilated convolutional neural networks," *Med. Phys.*, vol. 44, no. 12, pp. 6377–6389, Dec. 2017, doi: 10.1002/mp.12602.
- [40] S. Trebeschi, J. J. M. van Griethuysen, D. M. J. Lambregts, M. J. Lahaye, C. Parmar, F. C. H. Bakers, N. H. G. M. Peters, R. G. H. Beets-Tan, and H. J. W. L. Aerts, "Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR," *Sci. Rep.*, vol. 7, no. 1, p. 5301, Jul. 2017, doi: 10.1038/s41598-017-05728-9.
- [41] D. Karimi, G. Samei, C. Kesch, G. Nir, and S. E. Salcudean, "Prostate segmentation in MRI using a convolutional neural network architecture and training strategy based on statistical shape models.," *Int. J. Comput. Assist.*

- Radiol. Surg.*, vol. 13, no. 8, pp. 1211–1219, Aug. 2018, doi: 10.1007/s11548-018-1785-8.
- [42] M. N. N. To, D. Q. Vu, B. Turkbey, P. L. Choyke, and J. T. Kwak, “Deep dense multi-path neural network for prostate segmentation in magnetic resonance imaging.,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 13, no. 11, pp. 1687–1696, Nov. 2018, doi: 10.1007/s11548-018-1841-4.
- [43] *Disease Control Priorities, Third Edition (Volume 1): Essential Surgery*. 2015. doi: 10.1596/978-1-4648-0346-8.
- [44] L. E. Court, K. Kisling, R. McCarroll, L. Zhang, J. Yang, H. Simonds, M. du Toit, C. Trauernicht, H. Burger, J. Parkes, M. Mejia, M. Bojador, P. Balter, D. Branco, A. Steinmann, G. Baltz, S. Gay, B. Anderson, C. Cardenas, A. Jhingran, S. Shaitelman, O. Bogler, K. Schmeller, D. Followill, R. Howell, C. Nelson, C. Peterson, and B. Beadle, “Radiation Planning Assistant - A Streamlined, Fully Automated Radiotherapy Treatment Planning System.,” *J. Vis. Exp.*, no. 134, p. e57411, 2018, doi: 10.3791/57411.
- [45] L. T. Chuang, S. Feldman, C. Nakisige, S. Temin, and J. S. Berek, “Management and Care of Women With Invasive Cervical Cancer: ASCO Resource-Stratified Clinical Practice Guideline.,” *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, vol. 34, no. 27. United States, pp. 3354–3355, Sep. 2016. doi: 10.1200/JCO.2016.68.3789.
- [46] International Atomic Energy Agency, “Management of Cervical Cancer: Strategies for Limited-resource Centres - A Guide for Radiation Oncologists,”

2013. <https://www.iaea.org/publications/8738/management-of-cervical-cancer-strategies-for-limited-resource-centres-a-guide-for-radiation-oncologists>
- [47] C. Szegedy, S. Ioffe, and V. Vanhoucke, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *CoRR*, vol. abs/1602.0, 2016.
- [48] E. Smistad, A. Østvik, B. O. Haugen, and L. Løvstakken, "2D left ventricle segmentation using deep learning," in *2017 IEEE International Ultrasonics Symposium (IUS)*, 2017, pp. 1–4. doi: 10.1109/ULTSYM.2017.8092573.
- [49] T. J. Sørensen, *A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on Danish commons*. København: I kommission hos E. Munksgaard, 1948.
- [50] D. P. Kingma and J. Ba, "Adam: {A} Method for Stochastic Optimization," *CoRR*, vol. abs/1412.6, 2014.
- [51] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," Feb. 2018, Accessed: Oct. 21, 2019. [Online]. Available: <http://arxiv.org/abs/1802.02611>
- [52] Ö. Çiçek, A. Abdulkadir, S. S. Lienkamp, T. Brox, and O. Ronneberger, "3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation," Jun. 2016, Accessed: Oct. 21, 2019. [Online]. Available: <http://arxiv.org/abs/1606.06650>



- [53] J. Yang, A. Amini, R. Williamson, L. Zhang, Y. Zhang, R. Komaki, Z. Liao, J. Cox, J. Welsh, L. Court, and L. Dong, "Automatic contouring of brachial plexus using a multi-atlas approach for lung cancer radiation therapy," *Pract. Radiat. Oncol.*, vol. 3, no. 4, pp. e139–e147, 2013, doi: <https://doi.org/10.1016/j.prro.2013.01.002>.
- [54] R. Pötter, C. Haie-Meder, E. Van Limbergen, I. Barillot, M. De Brabandere, J. Dimopoulos, I. Dumas, B. Erickson, S. Lang, A. Nulens, P. Petrow, J. Rownd, and C. Kirisits, "Recommendations from gynaecological (GYN) GEC ESTRO working group (II): Concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy—3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology," *Radiother. Oncol.*, vol. 78, no. 1, pp. 67–77, Jan. 2006, doi: [10.1016/J.RADONC.2005.11.014](https://doi.org/10.1016/J.RADONC.2005.11.014).
- [55] M. Buda, A. Maki, and M. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, Oct. 2017, doi: [10.1016/j.neunet.2018.07.011](https://doi.org/10.1016/j.neunet.2018.07.011).
- [56] N. Heller, N. Sathianathan, A. Kalapara, E. Walczak, K. Moore, H. Kaluzniak, J. Rosenberg, P. Blake, Z. Rengel, and M. Oestreich, "The kits19 challenge data: 300 kidney tumor cases with clinical context, ct semantic segmentations, and surgical outcomes," *arXiv Prepr. arXiv1904.00445*, 2019.
- [57] R. E. McCarroll, B. M. Beadle, P. A. Balter, H. Burger, C. E. Cardenas, S. Dalvie, D. S. Followill, K. D. Kisling, M. Mejia, K. Naidoo, C. L. Nelson, C. B. Peterson, K. Vorster, J. Wetter, L. Zhang, L. E. Court, and J. Yang, "Retrospective Validation and Clinical Implementation of Automated Contouring of Organs at Risk in the

- Head and Neck: A Step Toward Automated Radiation Treatment Planning for Low- and Middle-Income Countries,” *J. Glob. Oncol.*, no. 4, pp. 1–11, 2018, doi: 10.1200/JGO.18.00055.
- [58] H.-C. Chen, J. Tan, S. Dolly, J. Kavanaugh, M. A. Anastasio, D. A. Low, H. Harold Li, M. Altman, H. Gay, W. L. Thorstad, S. Mutic, and H. Li, “Automated contouring error detection based on supervised geometric attribute distribution models for radiation therapy: A general strategy,” *Med. Phys.*, vol. 42, no. 2, pp. 1048–1059, 2015, doi: 10.1118/1.4906197.
- [59] C. B. Hui, H. Nourzadeh, W. T. Watkins, D. M. Trifiletti, C. Alonso, S. W. Dutta, and J. V Siebers, “Automated OAR Anomaly and Error Detection Tool in Radiation Therapy,” *Int. J. Radiat. Oncol. • Biol. • Phys.*, vol. 99, no. 2, pp. E554–E555, Oct. 2017, doi: 10.1016/j.ijrobp.2017.06.1932.
- [60] B. Rigaud, B. M. Anderson, Z. H. Yu, M. Gobeli, G. Cazoulat, J. Söderberg, E. Samuelsson, D. Lidberg, C. Ward, N. Taku, C. Cardenas, D. J. Rhee, A. M. Venkatesan, C. B. Peterson, L. Court, S. Svensson, F. Löfman, A. H. Klopp, and K. K. Brock, “Automatic Segmentation Using Deep Learning to Enable Online Dose Optimization During Adaptive Radiation Therapy of Cervical Cancer,” *Int. J. Radiat. Oncol.*, 2020, doi: <https://doi.org/10.1016/j.ijrobp.2020.10.038>.
- [61] B. E. Boser, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classifiers,” in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 1992, pp. 144–152. doi: 10.1145/130385.130401.

- [62] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995, doi: 10.1007/BF00994018.
- [63] M. A. Hearst, S. T. Dumais, E. Osuna, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intell. Syst. their Appl.*, vol. 13, no. 4, pp. 18–28, 1998, doi: 10.1109/5254.708428.
- [64] C. Ma and F. Huang, "Assessment of a knowledge-based RapidPlan model for patients with postoperative cervical cancer," *Precis. Radiat. Oncol.*, vol. 1, no. 3, pp. 102–107, doi: 10.1002/pro6.23.
- [65] N. Li, R. Carmona, I. Sirak, L. Kasaova, D. Followill, J. Michalski, W. Bosch, W. Straube, L. K. Mell, and K. L. Moore, "Highly Efficient Training, Refinement, and Validation of a Knowledge-based Planning Quality-Control System for Radiation Therapy Clinical Trials.," *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 97, no. 1, pp. 164–172, Jan. 2017, doi: 10.1016/j.ijrobp.2016.10.005.
- [66] M. Tinoco, E. Waga, K. Tran, H. Vo, J. Baker, R. Hunter, C. Peterson, N. Taku, and L. Court, "RapidPlan development of VMAT plans for cervical cancer patients in low- and middle-income countries.," *Med. Dosim.*, pp. 5086–8097, 2019, doi: 10.1016/j.meddos.2019.10.002.
- [67] A. W. M. Sharfo, S. Breedveld, P. W. J. Voet, S. T. Heijkoop, J.-W. M. Mens, M. S. Hoogeman, and B. J. M. Heijmen, "Validation of Fully Automated VMAT Plan Generation for Library-Based Plan-of-the-Day Cervical Cancer Radiotherapy.," *PLoS One*, vol. 11, no. 12, p. e0169202, 2016, doi: 10.1371/journal.pone.0169202.

- [68] K. Kisling, "DEVELOPMENT OF AUTOMATED RADIOTHERAPY TREATMENT PLANNING FOR CERVICAL AND BREAST CANCER FOR RESOURCE-CONSTRAINED CLINICS," *PhD diss., Univ. Texas MD Anderson Cancer Cent. UT Heal. Grad. Sch. Biomed. Sci.*, May 2019.
- [69] A. R. Conn, N. I. M. Gould, and P. L. Toint, *Trust Region Methods*. Society for Industrial and Applied Mathematics, 2000. doi: doi:10.1137/1.9780898719857.

## Vita

Dong Joo Rhee was born in South Korea on May 15, 1989, the son of Tack Gyoo Rhee and Sun Hee Kim. After completing his work at 3 different High Schools in South Korea, Canada, and the UK, he entered Imperial College London, in the UK. He received the degree of Bachelor of Science with a major in physics from Imperial College London in October 2011. He then entered Duke University in Durham, North Carolina, and received the degree of Master of Science with a major in Medical Physics in May 2013. For the next three years, he worked as a researcher at Dongnam Institute of Radiological and Medical Sciences (DIRAMS). In August of 2017, he entered The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences.

Permanent address:

933-3 Madu-dong,

Ilsandong-gu, Goyang-Si,

Gyeonggi-do, South Korea, 10411