8-2023

Auto-Segmentation In Pancreatic And Liver Radiation Therapy

Cenji Yu

Auto-segmentation in Pancreatic and Liver Radiation Therapy

by

Cenji Yu, B.S.

APPROVED:

_____

Laurence Court, Ph.D.

Advisory Professor

_____

Rachael Martin Palpeter, Ph.D.

_____

Ethan Ludmir, M.D.

_____

Carlos Cardenas, Ph.D.

_____

Tinsu Pan, Ph.D.

_____

Albert Koong, M.D.

APPROVED:

_____

Dean, The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences

Auto-segmentation in Pancreatic and Liver Radiation Therapy

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

Cenji Yu, B.S.

Houston, Texas

August 2023

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Laurence Court, for the support and encouragement I received throughout my PhD journey. I am forever grateful for your mentorship as a student. Thank you so much for the opportunity to contribute to a project that has a profound impact on the lives of so many patients.

I am also thankful to the members of my committee, Drs. Carlos Cardenas, Albert Koong, Ethan Ludmir, Tinsu Pan and Rachael Martin Paulpeter. Your dedication has played a pivotal role in steering me in the right direction. This project would have not been possible without your clinical and technical knowledge. I have learned so much from each and every one of you. Thank you so much for being my biggest advocate both professionally and personally.

Furthermore, I would like to thank all members the Court Lab. Your support helped me concur some of the toughest challenges during graduate school. I am truly grateful for your encouragement along the way. Thank you for offering such a warm and safe environment. My graduate school career would be a lot less fun without all of you.

Auto-segmentation in Pancreatic and Liver Radiation Therapy

Cenji Yu, B.S.

Advisory Professor: Laurence Court, Ph.D.

Background

Gastrointestinal cancers exhibit a high mortality rate compared to other cancer types. Among these, pancreatic cancer ranks as the fourth leading cause of cancer-related deaths worldwide. The five-year survival rate remains alarmingly low at a mere 9%. Hepatocellular carcinoma (HCC), another aggressive form of cancer, is rapidly becoming the primary cause of cancer-related deaths in the United States. The treatment of both liver cancer and pancreatic cancer heavily relies on a multidisciplinary approach. Innovative treatment strategies involving dose-escalated regimens, such as stereotactic body radiation therapy (SBRT), are emerging as an important pillar of the management of liver and pancreatic cancer. The success of these treatment modalities hinges upon the precise and standardized segmentation of organs-at-risk and target volumes to ensure the optimal quality of treatment plans.

Methods

We first developed an automated organs-at-risk segmentation tool for upper abdominal radiation therapy treatment. A dataset of 70 patients was collected and utilized as the training set and benchmark for our auto-segmentation tool. We employed the adaptive

nnU-Net architecture to develop a model ensemble capable of contouring various organs, including the duodenum, small bowel (ileum and jejunum), large bowel, liver, spleen, kidneys, and spinal cord. The performance of the segmentation tool was evaluated on 75 patients using both contrast-enhanced and non-contrast-enhanced CT images, employing a five-point Likert scale assessment by five experts from three different institutions. To capture contours requiring major edits, we developed a distance-based quality assurance (QA) system. This system identified CT scans that were likely to yield suboptimal contours requiring time-consuming major edits. Evaluation of the QA system was conducted on clinical CT scans, with the clinical review score serving as the ground truth. For target volume segmentation, we employed transformer-based architectures, leveraging self-supervised learning and uncertainty estimation techniques to enhance performance and allow for stylistic customization. A total of 3094 unlabeled CT scans from liver cancer patients, along with 5050 publicly available CT scans, were collected for self-supervised pretraining in liver tumor segmentation. The pretrained encoders were then utilized to optimize downstream liver tumor segmentation models, evaluating the impact of self-supervised learning on tumor segmentation performance. For pancreatic tumor segmentation, we developed an ensemble-based approach incorporating multiple segmentation styles. Probability thresholding was employed to generate the final segmentation, enabling customization according to clinicians' preferences.

Results

Our organs-at-risk segmentation tool achieved a clinical acceptance rate of over 90% for all organs except the duodenum, demonstrating its accuracy in delineation. Quantitative results were comparable to state-of-the-art methods, using a small but high-quality dataset.

The QA system achieved an AUC of 0.89 for capturing contours requiring major edits on randomly sampled clinical CT scans. In liver tumor segmentation, our study revealed that self-supervised learning demonstrated 4-5% performance improvement when diverse unlabeled data were used for pretraining. This finding highlights the importance of incorporating a wide range of data during the pretraining stage. For pancreatic tumor segmentation, our ensemble-based segmentation method proved highly effective. It provided pixel-by-pixel uncertainty estimates and allowed customization through probability thresholding. Our customized contours surpassed the performance of the state-of-the-art segmentation model, even when utilizing identical training data, pretraining techniques, and hyperparameters.

Conclusion

Our auto-segmentation system for organs-at-risk achieved high clinical acceptance rates in upper-abdominal radiation treatment. The accompanying QA tool effectively captured contours requiring major edits. Leveraging a wide range of unlabeled data in self-supervised learning improved the performance of our transformer-based segmentation system. Additionally, our uncertainty-guided segmentation network allowed customization and identification of low-confidence regions. Our suite of auto-segmentation tools for pancreatic and liver cancer radiation treatment has the potential to streamline clinical workflows while prioritizing patient safety.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Gastrointestinal (GI) cancers represent more than one-third (35%) of all cancer-related deaths[1]. Among them, liver cancer and pancreatic cancer are 4th and 7th leading cause of cancer mortality worldwide[23]. Treatment of both liver cancer and pancreatic cancer relies heavily on multidisciplinary approach[45]. Radiation therapy plays a crucial role in the multidisciplinary care of both diseases. In recent years, immobilization techniques and image guidance have significantly increased the quality of radiation treatment for these two types of cancers[45]. This has allowed clinicians to increase dose and introduce hypo-fractionated or stereotactic body radiation therapy (SBRT) for pancreatic and liver cancer patients. There has been increasing adoption of partial liver hypo-fractionated radiation therapy in the management of liver cancer. Excellent local control rates were observed in both hepatocellular carcinoma (HCC) and cholangiocarcinoma when treated with SBRT[6]. For pancreatic cancer, dose-escalation was shown to be effective and tolerable at doses of 25-35 Gy in 3-5 fractions[7]. These dose-escalated procedures, however, are challenging due to the proximity of surrounding critical structures. The pancreas, for example, is surrounded by radiosensitive serial organs such as duodenum, large bowels, small bowels and stomach. These organs are often protected by max dose constraints[8], leaving little margin for error in treatment planning. Accurate and consistent segmentations of organs-at-risk (OARs) and targets are thus essential to the safety of these hypo-fractionated treatments. In addition, for both pancreas and liver radiation treatment, at least nine OARs are required for treatment planning. These tasks are currently done manually by clinicians, occupying significant amount of time from their schedule[9]. Furthermore, there is a growing trend in academic centers towards adopting adaptive workflows for radiation therapy, aiming to achieve dose escalation to the target while minimizing damage to normal tissue[10]. Various treatment platforms now offer high-quality daily guidance images that

facilitate real-time adaptation. This is particularly crucial in gastrointestinal (GI) radiation treatment, where the movement of radiosensitive bowel structures due to peristalsis poses a challenge. However, implementing an adaptive workflow requires significant expert involvement while the patient is positioned on the treatment couch. Therefore, it is crucial to establish a fast and safe adaptive planning workflow that effectively utilizes the available imaging hardware. One of the most time-consuming steps in the adaptive workflow is contouring[11], highlighting the importance of accurate and robust segmentation of organs-at-risk and treatment targets. Precise and fast segmentation serves as the fundamental basis for an effective adaptive workflow.

Here, we propose a series of deep learning-based automation tools to streamline clinical radiation treatment processes. Deep learning has achieved significant progress in image segmentation in recent years[12]. Auto-segmentation tools driven by deep learning have seen fast adoption by clinics[13–16]. The U-Net[17] approach is widely used for auto-segmentation due to its efficient feature extraction and integration at multiple scales, yielding accurate results. However, it requires substantial data for optimal performance. Transformer-based architectures, originally designed for natural language processing, have shown impressive performance in computer vision tasks[18]. However, these state-of-the-art deep learning approaches require extra consideration when deployed in the clinic. An auto-segmentation tool needs to first exhibit outstanding performance upon clinician evaluation. The end users need to be extensively involved in the development of the tool to ensure clinical adoption. Furthermore, patient safety is paramount in clinical workflow. The proposed auto-segmentation tool requires extensive safety features to prevent subpar contours from entering the treatment planning workflow. The quality assurance of the auto-segmentation tool is thus essential to ensure safe clinical deployment.

This project's long-term goal is to provide accurate and robust segmentation for pancreatic and liver cancer radiation treatment without human intervention. Our objective is to create an auto-segmentation system for pancreatic and liver cancer using deep learning. The rationale behind our objective is two-fold: 1) Physicians from our clinic on average spend three hours delineating structures and target volumes 2) The resulting contours are susceptible to inconsistencies[9]. We aim to develop auto-segmentation tool that can streamline clinical workflow and standardize contouring practice. With the advancement of image guidance techniques, the workload required from clinicians is trending upwards. This automation project will hopefully alleviate clinicians from time and effort spent on repetitive tasks and help them better focus on patient care and research efforts. In addition, consistent treatment planning from an automated tool can be valuable in standardizing academic clinical trials as well as day-to-day patient care. We hope that our automation tools become the central piece of the efforts to standardize patient care in both community and academic settings.

Chapter 2: Specific Aims and Central Hypothesis

2.1 Introduction

We hypothesize that we can create a clinically robust auto-segmentation system for pancreatic and liver cancer radiation therapy with deep learning-based techniques that achieves 90% physician acceptance rate.

2.2 Specific Aims

Specific Aim 1:

Develop an auto-segmentation system for organs-at-risk contouring for pancreatic and liver cancer treatment. We will train convolutional neural networks (CNNs) to automatically

contour 7 OARs (duodenum, stomach, small bowel, large bowel, liver, spleen, kidney) on primary breath-hold CT images. We will introduce deep learning-based QA methods to detect contour errors to ensure patient safety. We hypothesize that CNN-based automated contouring can achieve 90% physician acceptance rate on organs-at-risk segmentation.

Specific Aim 2:

Develop an auto-segmentation system for target volume delineation for pancreatic and liver cancer treatment to enable automatic treatment planning. We will automatically segment gross tumor volume (GTV) on breath-hold contrast-enhanced CT images. We hypothesize that CNN-based automated target volume delineation can achieve 90% physician acceptance rate.

Chapter 3: Automation of Organs-at-risk Segmentation

This chapter is based upon the following article:

Yu C, Anakwenze CP, Zhao Y, Martin RM, Ludmir EB, S.Niedzielski J, Qureshi A, Das P, Holliday EB, Raldow AC, Nguyen CM, Mumme RP, Netherton TJ, Rhee DJ, Gay SS, Yang J, Court LE, Cardenas CE. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. Sci Rep. 2022;12(1). doi:10.1038/s41598-022-21206-3

reproduce the contribution in whole or in part in any printed volume (book or thesis) of which they are the author(s).

3.1 Introduction

Pancreatic cancer is one of the most aggressive tumor types, as it accounts for 3% of all cancers in the United States, as well as 7% of all cancer-related deaths[19]. Radiation therapy, along with chemotherapy, play a vital role in local tumor control for locally advanced pancreatic cancer[20]. Radiation treatment planning for pancreatic cancer is often complex with tight dose constraints[21]. This is a consequence of the pancreas being surrounded by highly radiosensitive and serial organs at risk (OARs) (duodenum, stomach, and small bowel) that require maximum dose constraints. However, OAR delineation in pancreatic and liver cancer is time consuming[22]. At our cancer center, pancreas radiation treatment requires delineation of 8 OARs: stomach, duodenum, large bowel, small bowel, liver, spleen, left kidney and right kidney. The average time spent on OAR delineation has been shown to be over 20 minutes[9]. For upper abdominal OAR delineation, reproducibility is a major challenge. Experts often have conflicting OAR delineations for the same patient, especially at the gastroesophageal junction[23]. Delineation of bowel structures (duodenum, large bowel and small bowel) is also susceptible to interobserver variability[9,24]. Margins reserved for motion management[25] and poor soft tissue contrast at the small/large bowel border[26] makes establishing the ground-truth for bowel structures difficult. It is often found in clinical practice that normal tissues extending (~1.0 cm) beyond the superior and inferior extent of the planning target volume (PTV) are not contoured on slices located outside of these margins. This is generally true for normal tissues that have maximum dose objectives where the whole volume is not needed for dose optimization[27], but this practice also introduces interobserver variability and clouds the establishment of the ground-truth.

Deep learning-based image segmentation has achieved expert level performance in both target and normal structure delineation when trained with large datasets[28,29,30]. It has also been shown to reduce contouring inconsistency in clinical trials and to provide more accurate dose metrics[31]. Among deep learning-driven approaches, U-Net derived models dominate in organ segmentation tasks in the abdomen[32,33] where public datasets are abundant (liver, spleen and kidney). For serial OARs (duodenum, stomach, and small bowel) in pancreatic cancer treatment, a few U-Net based models were developed on private datasets and achieved better results than alternative approaches such as fully convolutional network-based models[34]. Wang et al. explored the multi-planar fusion approach with 2D U-Nets predicting on both axial, sagittal and coronal views[26]. Liu et al. utilized a 3D self-attention U-Net to segment the OARs in pancreatic radiotherapy[35] and achieved state-of-the art performance. These specialized U-Net models from large academic institutions required extensive research expertise to develop. In addition, these models required at least 80 sets of complete patient contours for training and validation alone. Due to aforementioned inconsistencies in the clinical contours, extensive curation by experts is required before contours qualify for deep learning training. This expensive, time-consuming process[36] hinders the development and adoption of deep learning models outside of large academic institutions.

Recently, the self-configuring nnU-Net framework[37] has shown promising results in abdominal organ segmentation. This framework systematically configured U-Nets on the basis of distribution of spacings, median shape, and intensity distribution of the training CT images. The

framework is also exceedingly data efficient due to robust data augmentation methods. nnU-Net has shown promising results in abdominal organ segmentation tasks and won two of the five tasks in the CHAOS challenge[33]. This framework was thus chosen as our candidate for automating upper-abdominal OAR segmentation.

In summary, upper abdominal OAR contouring is time-consuming and susceptible to variabilities. Deep learning-based auto-segmentation provides a fast and consistent alternative to manual contouring. However, specialized U-Nets and large datasets are deemed essential to a robust deep learning auto-segmentation tool according to existing literature. These requirements confine the development of auto-segmentation tool to large academic centers with research expertise. In this study, we proposed using the streamlined nnU-Net framework to customize three-dimensional (3D) U-Nets that delineate eight OARs (stomach, duodenum, large bowel, small bowel, liver, spleen, left kidney and right kidney) simultaneously on contrast-enhanced and non-contrast-enhanced CT images. We hypothesized that with a small, but consistent, training set, the standard U-Net architecture could create clinically deployable models for upper-abdominal OAR segmentation. This study demonstrated clinical utility of the automatically generated segmentations through a robust evaluation via multi-observer rating of individual contours on 75 abdominal CT scans as well as quantitative evaluation on 30 CT scans. Our approach provided an easy-to-implement, data-efficient alternative for automating the clinical workflow of pancreatic radiation treatment, including adaptive radiation therapy. Our method utilized the least amount of data to achieve clinically acceptable qualitative results and competitive quantitative results compared to existing literature. In addition, we examined the organ-by-organ segmentation performance gain as we increased the number of patients in the

training dataset to provide insights on the amount of data required for training robust upper abdominal segmentation models for clinics interested in developing their own tools. We will release the entire training and testing dataset on TCIA to serve as additional resources for future abdominal organs auto-segmentation development.

3.2 Method

Imaging Data

Seventy patients were selected from patients with pancreatic cancer who were treated at The University of Texas MD Anderson Cancer Center from 2017 to 2020 under an IRB (institutional review board) approved protocol. CT images were acquired with the breath-hold technique on Philips Brilliance Big Bore (Philips Healthcare, Best, The Netherlands) CT simulators. CT scans had pixel sizes ranged from 0.98mm to 1.04 mm and slice thickness from 1mm to 2.5mm. Patients were scanned from 5 cm above the diaphragm to the iliac crest with intravenous contrast injection. The clinical OAR contours included the duodenum, small bowel, large bowel, stomach, liver, spleen, left kidney and right kidney.

Data Curation and Manual Segmentation

The duodenum, small bowel, and large bowel were manually delineated under physician supervision to increase consistency in normal tissue definition for these organs. To provide sufficient contextual information for the 3D U-Net models, bowel structures were extended along the z-axis and contoured throughout the entire scan. Stomach contours were trimmed to eliminate motion management margins. Liver, spleen and kidney contours were edited to ensure anatomical accuracy. All ground truth contours were reviewed and approved by a radiation

oncologist. Forty sets of contours were randomly selected for training and validation. The remaining thirty sets of contours were reserved as the held-out test set.

Data Preprocessing

To segment all OARs simultaneously, labels were compiled into a single segmentation map. When organ borders overlapped, the priority of the segmentation map was duodenum, small bowel, stomach, large bowel, liver, spleen and kidneys. Organs with the most stringent dose constraints were prioritized and overwrote organs with less stringent dose constraints. All images were resampled to 0.98mm × 0.98mm pixel size and 2.5mm slice thickness.

Model Training

The adaptive nnU-Net framework[38] was employed to customize 3D U-Nets for our dataset. 3D patches of image-label pairs were used for training. The patch size was 192×192×48. The 3D U-Net network depth was dynamically optimized by nnU-Net framework to ensure sufficient depth to fully utilize the large patch size. The training batch size was 2. The resulting U-Net architecture generated by the nnU-Net framework is shown in Figure 1.

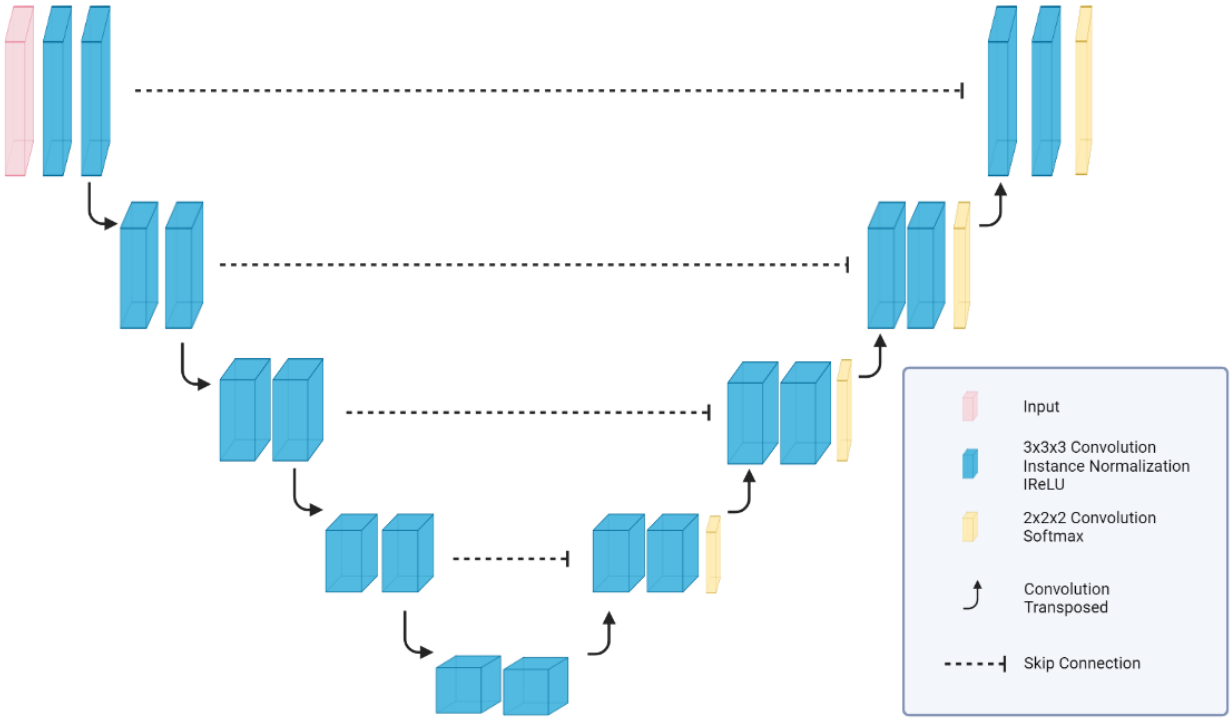


Figure 1. U-Net architecture customized by the nnU-Net framework based on the training dataset.

The loss function was a combination of Dice similarity coefficient (DSC) loss and cross-entropy loss. Training and testing were done on NVIDIA Tesla V100 GPUs with 32 GB VRAM. Training was stopped after 1000 epochs. To fully extract features from a small data set, five-fold cross-validation was used among the 40-patient dataset: 32 patients were used for training, and eight patients were used for validation in each fold (80-20 split). Five 3D U-Net models were trained, and the final prediction was produced by an ensemble of all five trained models from the cross validation. Training time for the U-Net ensemble was 36 hours when individual models were trained in parallel. Inference time using the U-Net ensemble for each patient was 8 minutes on average.

To evaluate performance gains as the size of training data expanded, additional model ensembles were also trained on an escalating number of patients. Subsets of 10, 15, 20, 25, 30, and 35 patients were randomly selected. The training-validation split for each set was also 80-20, which was identical to the final model ensemble. These six additional 3D U-Net ensembles were trained under the nnU-Net framework with identical training procedures.

Quantitative Evaluation

The final model ensembles from various sizes of the training data were evaluated on the held-out test set of thirty patients. The performance of the model ensembles was evaluated by the 3D DSC, 95% Hausdorff distance (HD95), and mean surface distance (MSD) between the predicted contours and the ground truth contours.

Qualitative Evaluation

An additional 75 patients simulated under the breath-hold protocol were randomly selected from the clinical database as an independent qualitative test set. Our center captures two non-contrast-enhanced and three to four contrast-enhanced CT images during simulation for patients who are suitable for CT imaging with a contrast agent. For each patient, one contrast-enhanced and one non-contrast-enhanced CT image were randomly selected as part of the qualitative analysis, resulting in a total of 150 patient CT images. The automatically generated contours on both contrast-enhanced and non-contrast-enhanced images were visually evaluated and scored using a five-point Likert scale as shown in Table 1 by five radiation oncologists from

three institutions and two countries. Each image was scored once by a radiation oncologist; and each organ was scored individually.

Likert Scale		Explanation for this study
5	Strongly agree	Use-as-is (i.e. clinically acceptable, and could be used for treatment without change)
4	Agree	Minor edits that are not necessary. Stylistic differences, but not clinically important. The current contours/plan are acceptable.
3	Neither agree or disagree	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.
2	Disagree	Major edits. This category indicates that the necessary edits are required to ensure appropriate treatment, and sufficiently significant that the user would prefer to start from scratch.
1	Strongly disagree	Unusable. This category indicates that the quality of the automatically generated contours or plan are so bad that they are unusable.

Table 1. Likert scale used by physicians to evaluate contours generated on contrast-enhanced and non-contrast-enhanced CT images

3.3 Results

Quantitative Evaluation

A summary of the quantitative evaluation (n = 30) is provided in Table 2. All automatically generated contours had a mean DSC value of 0.80 or higher when compared to the ground-truth contours. Solid organs such as liver, spleen and kidneys all achieved mean DSC

values ranging from 0.96 to 0.97. Radiosensitive hollow organs such as small bowel, large bowel and stomach achieved mean DSC values ranging from 0.89 to 0.92. Duodenum achieved a mean DSC of 0.80. For distance metrics, solid organs (liver, spleen and kidneys) had mean HD95 ranging from 2.21 to 2.51mm and mean MSD ranging from 0.61 to 1.07mm. Radiosensitive hollow organs (small bowel, large bowel and stomach) had mean HD95 ranging from 4.77 to 7.77mm and mean MSD ranging from 1.23 to 1.99mm. Duodenum had a mean HD95 of 12.34mm and mean MSD of 1.68mm.

DSC boxplots of all organs were shown in Figure 2. Auto-segmentation performance had more variability in hollow organs compared to solid organs. Outliers from small bowel and large bowel auto-segmentations were often caused by misidentification of small/large bowel in inferior regions of CT scans outside of treatment fields. Low DSC examples of duodenum were often caused by disagreements at the stomach/duodenum and duodenum/jejunum borders.

Organs	DSC		HD95 (mm)		MSD(mm)	
	Mean	SD	Mean	SD	Mean	SD
Duodenum	0.80	0.08	12.34	9.09	1.68	1.04
Small Bowel	0.89	0.05	7.77	8.90	1.99	2.10
Large Bowel	0.90	0.06	7.15	8.42	1.27	0.87
Stomach	0.92	0.03	4.77	2.98	1.23	0.78
Liver	0.96	0.01	3.56	1.71	1.07	0.49
Spleen	0.97	0.01	2.21	1.27	0.56	0.23
Kidney_R	0.96	0.01	2.51	1.29	0.59	0.18
Kidney_L	0.96	0.01	2.52	0.90	0.61	0.19
SpinalCord	0.76	0.15	42.52	38.62	10.57	10.49

Table 2. Mean Dice similarity coefficient (DSC), 95% Hausdorff distance (HD95), and mean surface distance (MSD) between ground truth and prediction results from our tool on contrast-enhanced CT images

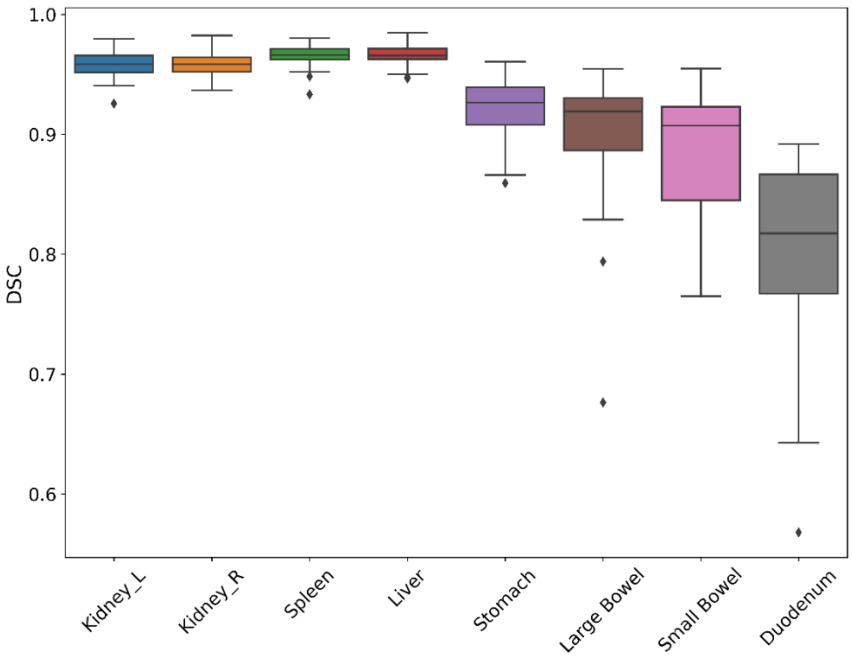


Figure 2. Box and whisker plots of Dice similarity coefficient (DSC) distance between ground-truth and automatically generated contours by our tool on contrast-enhanced CT images. The central line represents the median value. The border of the box represents the 25th and 75th percentiles. The outliers are represented by diamond markers.

In order to determine if 40 patients were sufficient for optimal model performance, the mean DSCs for the individual organs were also examined for an escalating number of patients. The result was plotted in Figure 3. The mean DSC increased as the size of the training dataset

increased. The mean DSCs of all organs tended to converge as the number of patients approached 40.

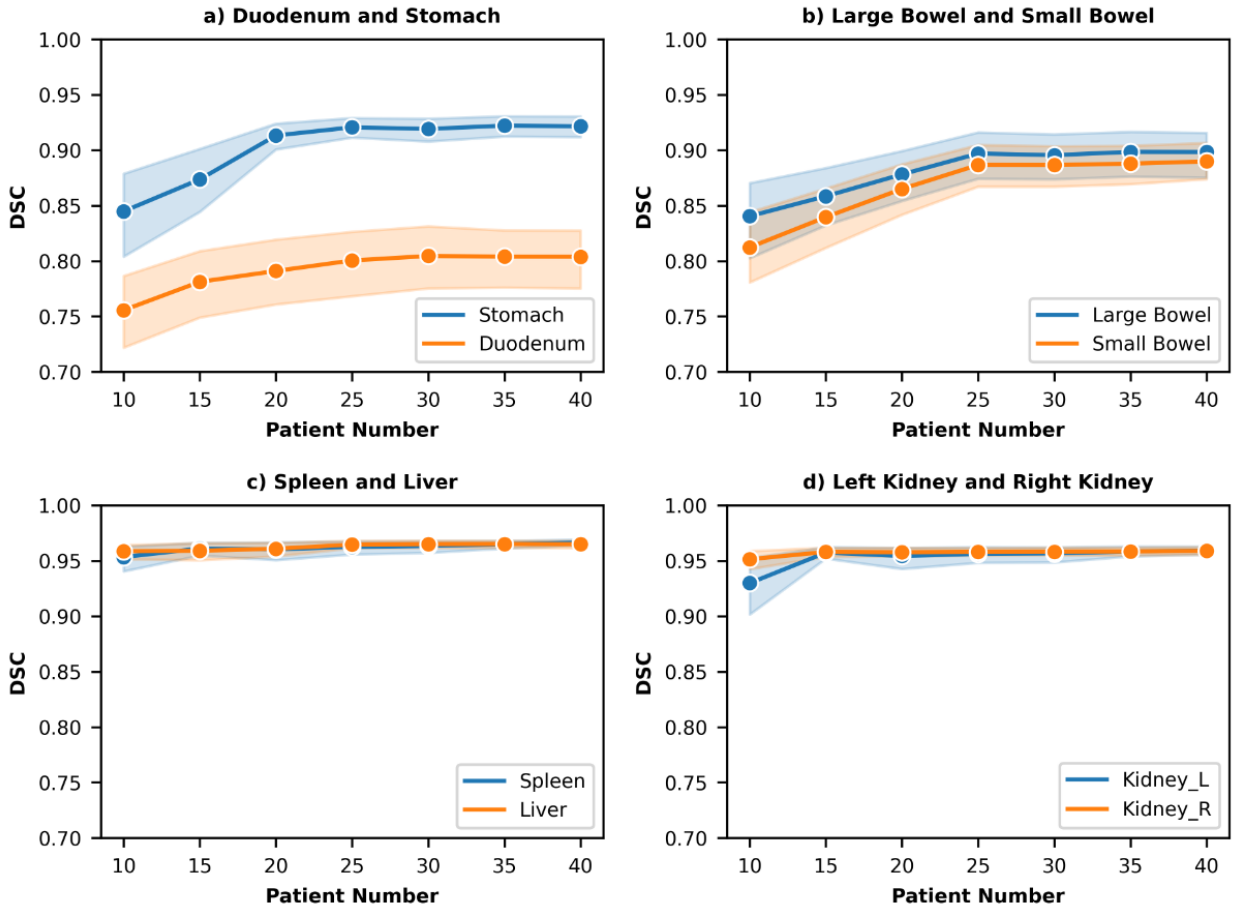


Figure 3. Mean DSC values between automatically generated contours and ground-truth contours increased as the number of patients in the dataset increased. The shadow represents the corresponding standard deviation for individual DSC values.

Qualitative Evaluation

The results from physicians' qualitative evaluations are shown below in Tables 3. Among the non-contrast-enhanced CT images, 85.3% of the duodenum contours, 92.0% of the small bowel contours, 93.3% of the stomach contours and more than 95% of the other organ contours received a score of 3 or greater, suggesting that these contours required only minor edits from physicians. More than 50% of the duodenum, small bowel, large bowel, and stomach contours as well as more than 85% of the spleen and kidney received a score of 4 or above.

	Non-contrast-enhanced CT Images				Contrast-enhanced CT Images			
	<3	≥3	≥4	5	<3	≥3	≥4	5
Duodenum	14.7%	85.3%	50.7%	18.0%	10.7%	89.3%	60.0%	22.0%
Small bowel	8.0%	92.0%	58.7%	28.0%	5.3%	94.7%	62.7%	30.0%
Large bowel	2.7%	97.3%	62.7%	28.0%	2.7%	97.3%	69.3%	30.0%
Stomach	6.7%	93.3%	62.7%	38.0%	4.0%	96.0%	66.7%	38.0%
Liver	4.0%	96.0%	77.3%	60.0%	2.7%	97.3%	84.0%	66.0%
Spleen	1.3%	98.7%	90.7%	86.0%	1.3%	98.7%	93.3%	86.0%
Kidney left	1.3%	98.7%	90.7%	70.0%	1.3%	98.7%	94.7%	74.0%
Kidney right	2.7%	97.3%	86.7%	66.0%	1.3%	98.7%	93.3%	72.0%

Table 3. Qualitative scores for contours generated on contrast-enhanced and non-contrast-enhanced CT images of 75 randomly selected patients

There was a small improvement in contour scores for auto-segmentations on contrast-enhanced CTs. 89.3% of the duodenum contours, 94.7% of the small bowel contours, and more than 95% of the other organ contours were scored as a 3 or greater. More than 60% of the duodenum, small bowel, large bowel, and stomach contours and more than 90% of the spleen and kidney scored a 4 or greater. Examples of automatically generated contours scored as 3,4 and 5 for duodenum, stomach and small bowel are shown in Figure 4.

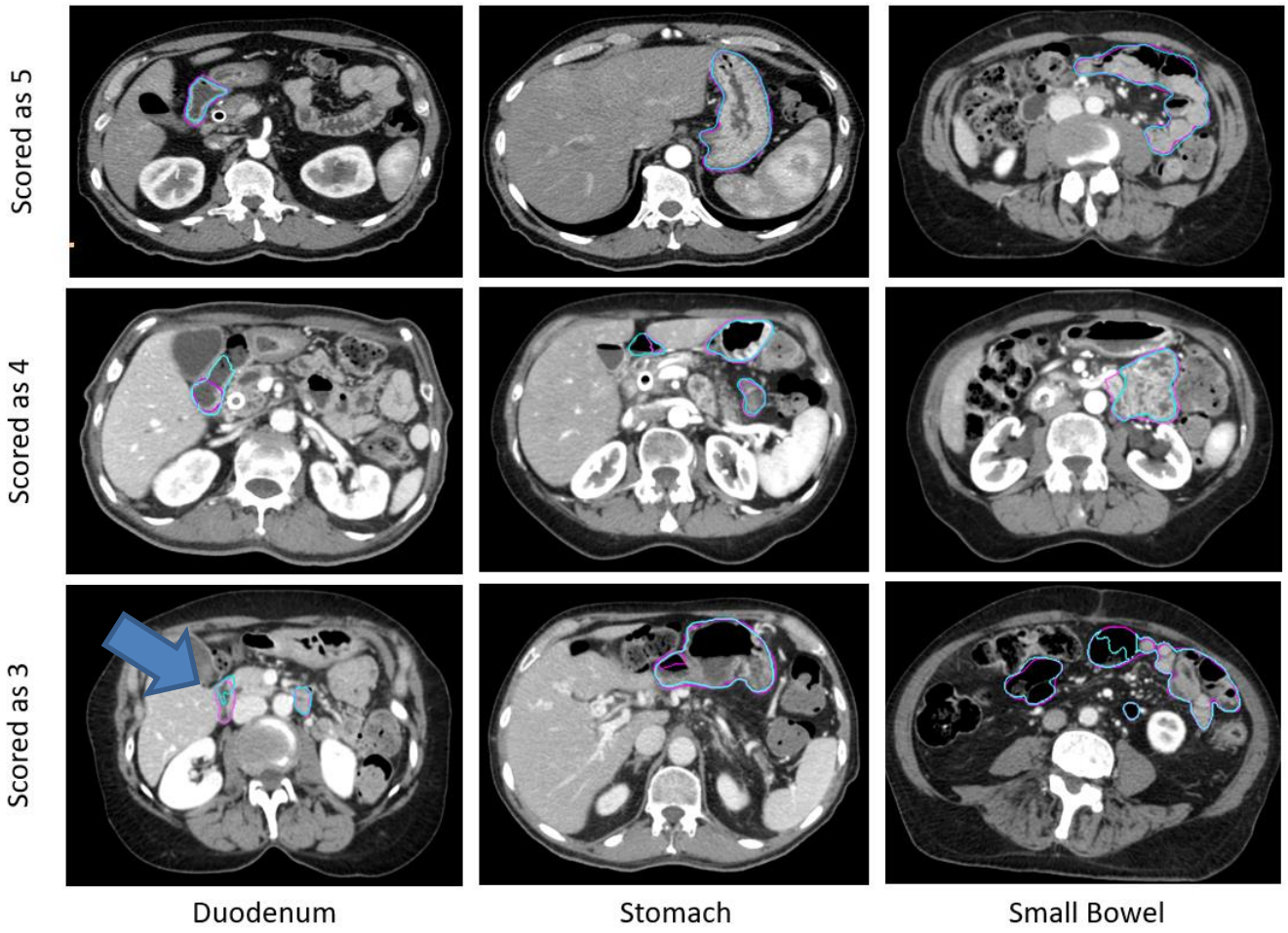


Figure 4. Representative contours of organs scored on a Likert scale as 5, 4, and 3 (top to bottom) by physicians. The ground truth contours are shown as purple in all images. The automatically generated contours are shown as cyan in all images. The arrow indicated a segment of under-contoured duodenum that required minor edits.

3.4 Discussion

We have developed a deep-learning-based tool for accurate and robust upper-abdominal OAR auto-segmentation. Our tool could simultaneously segment the duodenum, large bowel, small bowel, stomach, liver, spleen, and kidneys. Upon evaluation, the tool performed well in both quantitative and qualitative assessments. These tests were conducted on randomly selected

held-out test patients (30 and 75 patients for quantitative and qualitative assessments, respectively). Our qualitative assessment was conducted by five radiation oncologists from three different institutions. The tool achieved acceptable performance for clinical deployment, even though it was trained and validated with only 40 patients. Based on the results from this study, we have clinically implemented this auto-contouring system in the clinic at MD Anderson Cancer Center. In the future, we will make this auto-contouring tool available as part of the Radiation Planning Assistant[39] (rpa.mdanderson.org) to make this tool available to radiation oncology clinics in low- and middle-income countries.

Deep learning-based auto-segmentation approaches typically require a large amount of high-quality segmented datasets to achieve optimal performance[12]. In clinical scenario, the amount of high-quality labeled images is limited[40]. Creating high-quality contours suitable for deep learning training requires significant time resources and expertise[36,41]. A number of self-supervised deep learning approach were proposed by generating artificial data[42,43,44], but these approaches required technical expertise only available at large academic centers. Our findings offered an affordable, easy to implement approach to create auto-segmentation tools when public dataset is not available. The self-adaptive nnU-Net framework provided a standardized platform for U-Net architectures, allowing us to customize 3D U-Net ensembles that maximized the performance of the U-Net architecture. The qualitative evaluation provides evidence for the prowess of our tool. Automatically generated contours received a Likert score of 3 or above required only minor edits. Physicians deemed these contours beneficial to their segmentation workflow. Among 75 independent test patients, over 90% of the automatically generated contours received a Likert score of 3 or greater on most organs. For organs with poor soft tissue

boundaries such as the duodenum, 89.7% of CT contours only required minor edits for clinical use. Our results have shown that with a dataset of 40 patients, a standard 3D U-Net architecture could deliver automatically generated contours suitable for clinical deployment.

Clinical context of segmentation errors differentiated acceptable contours (Likert ≥ 4) from contours needed necessary minor edits (Likert = 3). Small contour errors may have significant clinical relevancy. For the duodenum contour scored as a 3 in Figure 4, the tool under-contoured a portion of the duodenum as shown by the arrow. The error shown was critical to patient safety because this segment of the duodenum was medially located and was close to the treatment target. Although most of the duodenum was properly contoured, the generated contour was scored as a 3 instead of a 4. The edit required from physicians, however, was marginal. Physicians were less concerned about absolute anatomical accuracy in other cases. For example, interobserver variability could be significant at the border of stomach and duodenum. The anatomical landmarks used to distinguish the two are subtle, often lacking a clear border. While the generated contour deviated drastically from the ground truth as shown in Figure 5, it was scored as a 4 and deemed acceptable for treatment planning by physicians. This was because the duodenum and stomach are often optimized to have the same maximum dose constraints ($D_{max} < 28Gy$).

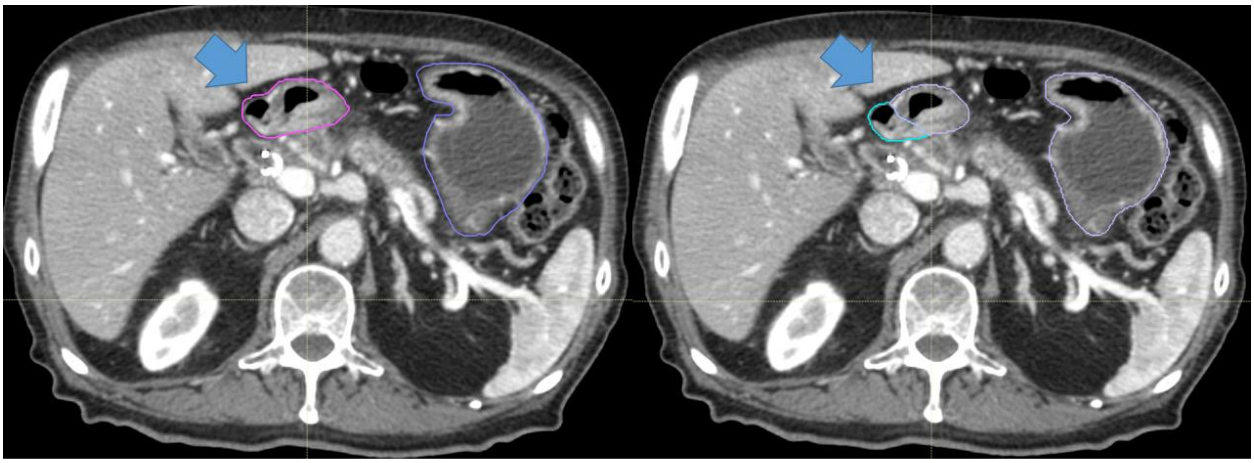


Figure 5. Representative ground-truth (left) and the automatically generated (right) contour of a patient's duodenum and stomach. These contours differed significantly, but because the duodenum and stomach are often optimized using the same dose constraints (i.e. Dmax < 28Gy), the contours were scored as a 4 and deemed acceptable for treatment planning.

Individual stylistic preferences differentiated use-as-is contours (Likert = 5) from the acceptable contours (Likert = 4). These stylistic preferences were the most prominent at the intersection of the duodenum and jejunum (contoured as part of the small bowel). The superior border of the fourth section of the duodenum had no visible border features on CT images. In Figure 6, the automatically generated contour was scored as a 4. The ground truth duodenum contour extended more superiorly compared to the automatically generated contour at the region indicated by the arrow. The varying cranial ends of duodenum contours were deemed as stylistic differences. The physicians were uncertain about the anatomical ground truth in the region. Since duodenum and small bowel were often optimized to have the same maximum dose constraints

(D_{max} < 28Gy), physicians decided that these differences had limited impact on treatment planning.

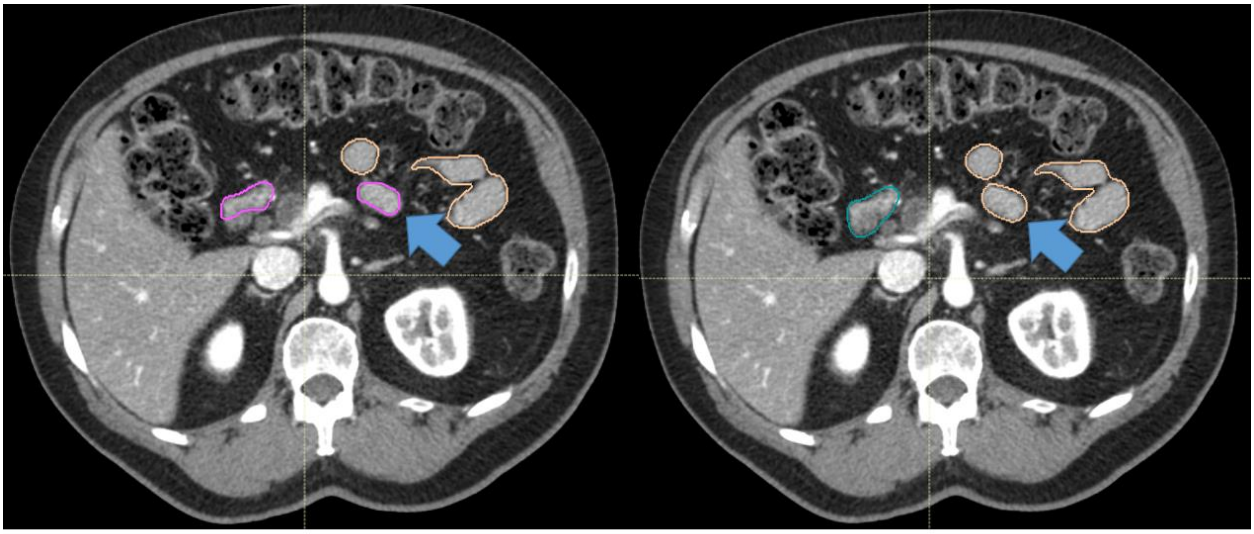


Figure 6. Representative ground truth (left) and the automatically generated (right) contours of a patient's duodenum and small bowel (jejunum). The ground truth is ambiguous at the transition from duodenum to small bowel (jejunum). The deviation from the ground truth was deemed as a stylistic difference.

Our quantitative results are comparable to those of state-of-the-art models trained with datasets of 80 patients or more for most organs. The DSC scores of the tool on small bowel, large bowel, stomach, spleen, liver, and kidney contours were within 0.01 of the current 3D state-of-the-art model (Liu et al.) as shown in Table 4. The MSDs were also comparable or smaller than the 3D state-of-the-art model shown in Table 5. Our tool, however, was trained and validated with a much smaller dataset of 40 patients. Our approach seemed to be more data efficient compared to the state-of-the-art approach. As data curation process is known to be time-

consuming and expensive, our method would allow easier development and adoption in the clinic.

Studies have suggested that 3D models demand too many parameters and required a large training dataset[45] to converge. Previous state-of-the-art approaches, such as organ-attention 2D deep networks with reverse connections by Wang et al., have been developed to segment 2D slices along axial, sagittal, and coronal views to reduce the number of trainable parameters[26]. Our tool outperformed the 2D-based multi-planar fusion approach in DSC for duodenum, small bowel and large bowel as shown in Table 4. We also achieved lower MSD for small bowel, large bowel, stomach and liver as shown in Table 5. When challenged with structures that span along the z-axis, 3D models were better equipped to segment these structures compared to 2D-based multi-planar fusion model due to its capability to capture anatomical context along the z-axis. Since only 40 patients were used for training and validation, our tool's 3D approach seemed to be more data efficient than the 2D multi-planar fusion approach as well.

	Ours (n=40)		Liu et al. (n=80)		Wang et al. (n=177)	
	Mean	SD	Mean	SD	Mean	SD
Duodenum	0.80	0.08	0.86	0.06	0.75	9.10
Small Bowel	0.89	0.05	0.89	0.06	0.80	10.20
Large Bowel	0.90	0.06	0.91	0.03	0.83	7.40
Stomach	0.92	0.03	0.93	0.03	0.95	2.60
Liver	0.96	0.01	0.96	0.01	0.98	0.70
Spleen	0.97	0.01	NA	NA	0.97	1.50
Kidney Right	0.96	0.01	0.95	0.02	0.98	2.10
Kidney Left	0.96	0.01	0.95	0.02	0.97	1.90

Table 4. Dice similarity coefficient comparison between our tool and other state-of-the-art upper-abdominal auto-segmentation models

	Ours (n=40)		Liu et al. (n=80)		Wang et al. (n=177)	
	Mean (mm)	SD (mm)	Mean (mm)	SD (mm)	Mean (mm)	SD (mm)
Duodenum	1.68	1.04	1.39	0.54	1.36	1.31
Small Bowel	1.99	2.10	1.99	1.08	3.01	3.35
Large Bowel	1.27	0.87	1.67	0.55	3.59	4.17
Stomach	1.23	0.78	1.77	1.19	1.68	1.55
Liver	1.07	0.49	1.45	0.80	1.23	1.52
Spleen	0.56	0.23	NA	NA	0.42	0.25
Kidney Right	0.59	0.18	1.05	0.86	0.45	0.89
Kidney Left	0.61	0.19	1.06	0.79	0.30	0.30

Table 5. Mean surface distance comparisons between our tool and other state-of-the-art upper-abdominal auto-segmentation models

The model performance progression with increasing patient number (Figure 3) gave us a better perspective on why our quantitative results were comparable to state-of-the-art models. For challenging hollow structures such as the stomach and duodenum, the 3D U-Net models initially gained performance as the patient number increased. The DSC curve started converging as we approached 25 patients. Similar trends were observed in the large bowel and small bowel DSCs. While the mean DSCs converged, the standard deviations were decreasing for the stomach, large bowel and small bowel. Prediction results were less variable with a larger training/validation dataset. For solid organs such as the spleen, liver, and kidney, DSC scores were above 90 even with only 10 patients. This data provides insights for clinics or individuals that are interested in developing their individual 3D U-Net models for upper-abdominal organ segmentation. When faced with the task of creating auto-segmentation tools with a limited annotation budget, our findings might be a guideline for budget allocation.

Our tool was developed and tested on the ground truth label delineated according to our institution's implementation of the RTOG guideline. While we introduced five radiation oncologists from three institutions to conduct qualitative evaluation, the test patients were from the same institution. With varying imaging protocols, image acquisition and reconstruction parameters, the model performance might suffer if the test patients were from various institutions from our experience[46]. In this case, small training samples might not be sufficient to guarantee great performance across varying patient cohorts. Further evaluation is needed to assess the model ensemble's performance on different patient populations.

For future work, automatic quality assurance of the generated contour, i.e. capturing clinically unusable contours, would also be a crucial addition to our automation tool. In addition, our center utilizes CT-on-rails image guided system for pancreatic radiation treatment. While our tool exhibited robust qualitative results on non-contrast-enhanced CT images, future work would include dose accumulation studies using automatically generated contours. This can pave the way for adaptive radiation therapy in pancreatic radiation treatment.

3.5 Conclusion

We proposed a simple but effective approach for developing a deep learning-based segmentation model for upper-abdominal OAR segmentation. Using only 40 patients, we trained a nnU-Net model to generate automatic contours that was able to produce clinically acceptable results on both contrast-enhanced and non-contrast-enhanced CT images. The results of the presented analysis led to the clinical deployment of this tool.

Chapter 4: Deep Feature-based Contour Quality Assurance for Auto-segmentation Models

4.1 Introduction

Pancreatic cancer is one of the most aggressive tumor types. It is the 7th leading cause of cancer mortality worldwide[3]. Management of this type of tumor requires multidisciplinary collaboration[47]. Hypo-fractionated treatment for pancreatic cancer is becoming more popular due to increasing adoption of image-guidance prior to treatment[7]. With escalated dose per fraction, more organs-at-risk (OARs) are required to be delineated on the simulation CT image to complement the more stringent dose constraints. At our institution, contours of duodenum, large bowel, small bowel (ileum and jejunum), liver, spleen, kidneys and spinal cord are required for hypo-fractionated pancreatic cancer treatment.

Deep learning-based auto-segmentation has dominated a variety of medical image segmentation challenges[48]. Recently, deep learning-based architectures were applied in delineating organs in the abdominal region[9,49]. These architectures achieved state-of-the-art performance on public datasets. Clinics are quickly adopting deep learning models for contouring in radiation treatment[14–16]. Auto-segmentation via deep learning, however, is still a data-driven approach. Deep learning-based techniques, therefore, are limited by models' training data. Out-of-distribution examples often lead to poor performance[50]. For auto-segmentation tools deployed in a clinical setting, the capability to handle out-of-distribution patient images is crucial. Clinicians often encounter patients with variable anatomy caused by poor NPO (nothing by mouth), ascites, or prior surgical procedures. The presence of these variations frequently results in a decline in the performance of auto-segmentation. High performance deep learning models, on the other hand, are often created with well-curated datasets[26,45]. These datasets

usually do not include abnormal anatomy as they serve as poor ground truths for model training and validation. Therefore, capturing out-of-distribution samples and requesting human intervention are crucial safeguards for auto-segmentation tools deployed in clinics.

In this study, we proposed a contour quality assurance approach by identifying out-of-distribution samples. We hypothesized that CT images that are drastically different from the training set would lead to degradation in the clinical utility of generated contours. By comparing distances between deep features of the training and testing datasets, we aimed to capture out-of-distribution patient images that were more likely to fail in the auto-segmentation workflow. We characterized the area under the curve (AUC) of the distance metric in detecting failed contour sets. Based on the AUC, we proposed an optimized threshold to flag patients that were unsuitable for the trained segmentation model.

4.2 Method

Thirty pancreatic cancer patients who underwent radiation treatment at MD Anderson Cancer Center were randomly extracted from the clinical database. A clinically validated nnU-Net model that has been deployed on more than 800 patients since 2021 was used to generate organs-at-risk contours on duodenum, small bowel, large bowel, stomach, liver, spleen and kidneys[16]. A radiation oncologist scored the predicted contours using a five-point Likert scale as identified abnormalities on the test cases. Contour sets were deemed as failed if any individual organ contour received a score below 3.

Since U-Net-based segmentation networks are composed of an encoder and a decoder. Our distance-based quality assurance method utilized the trained encoder of the nnU-Net model. Inspired by Gonzalez et al.[51], we first extracted deep features from the training dataset using the trained nnU-Net encoder. Features were sampled from the deepest layer of the trained encoder. Due to GPU memory constraint, training image patches of $192 \times 192 \times 48$ were fed through the model. The patch extraction process and patch sizes were identical to the preprocessing pipeline for our trained nnU-Net model to eliminate confounding factors. In order to summarize features extracted from all patches of training images, we reduced their dimensionalities via strided average pooling operations and vectorized the resulting matrices. We then estimated the mean μ and covariance matrix $\sum$ of all extracted training patches. A summary of this workflow was shown in Figure 7.

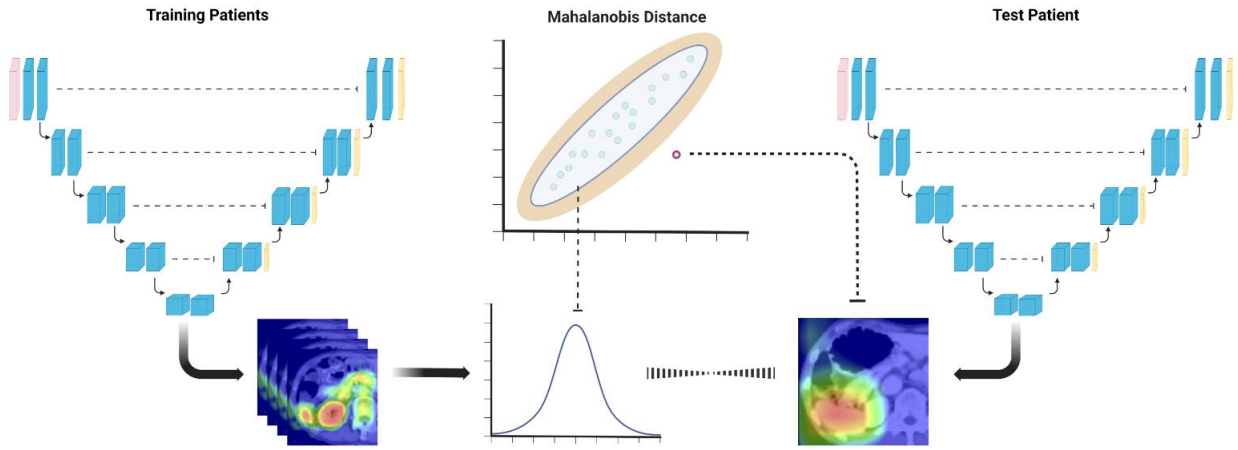


Figure 7. Quality assurance of deep learning auto-segmentation tool using deep features from the trained encoder.

During the assessment of a previously unseen test patient, the image completed one forward pass through the trained encoder with its features extracted on a patch-by-patch basis. These features were also averaged pooled and vectorized. We measured the Mahalanobis distance, D, between the resulting test image patch feature x to the Gaussian distribution of training features:

$$D = \sqrt{(x - \mu)^T \sum^{-1} (x - \mu)} \quad (1)$$

Here, mean μ and covariance matrix ∑ were previously estimated from the training set. Each test image patch would be assigned with a distance. To determine if the test image was out of distribution against the training set, we summed all distances across the entire image volume to provide an overall assessment of this patient. We repeated this process for all clinically scored test images to assess the effectiveness of our quality assurance approach.

4.3 Results

Among all test patients, eleven of the thirty test patients received a score of 1 or 2 on at least one organ in qualitative evaluations as shown in Table 6. These contour sets required major edits and were deemed as failed contours that required flagging. The Mahalanobis distance between test patients and training patients were shown in Figure 8. Patients likely to require major edits were successfully differentiated using this metric.

	Number of Patients
Major Edit (1-2)	11
Minor Edit (3)	8
Stylistic Edit or Use-as-is (4-5)	11

Table 6. Qualitative evaluation scores for contours automatically generated by nnU-Net model ensemble on 30 test patients. Patients required major edits on at least one organ were identified as true positive cases for our QA approach

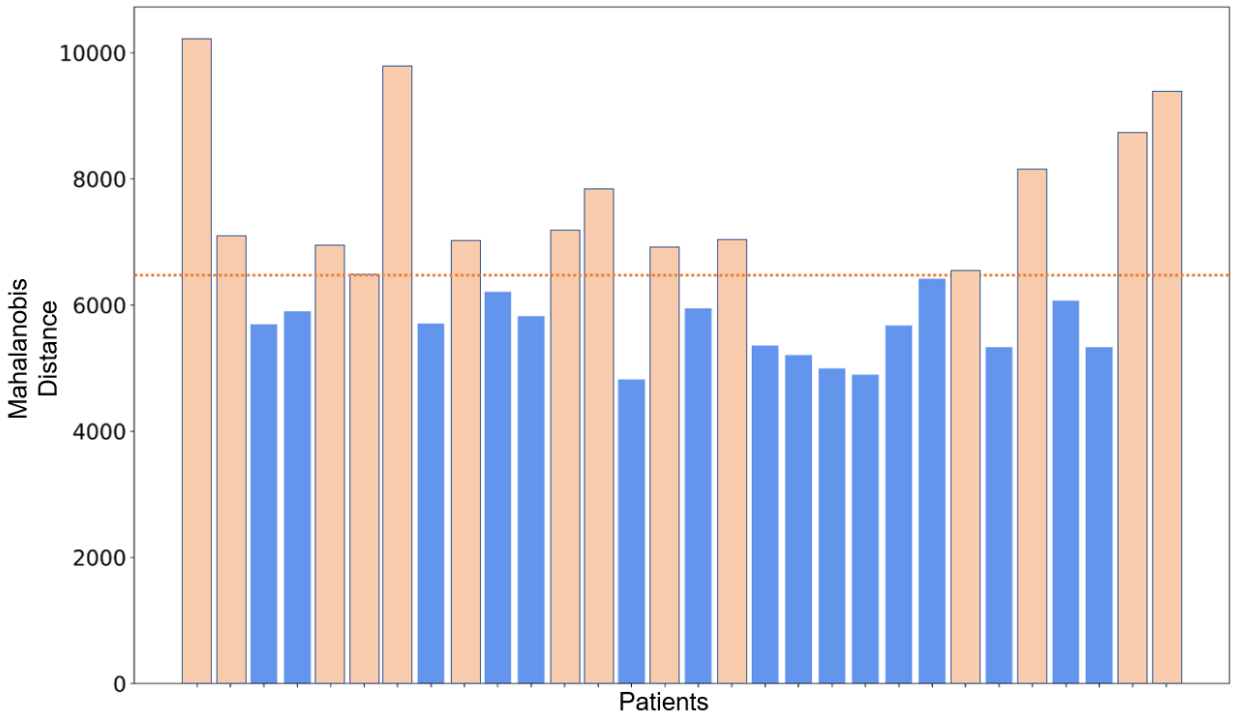


Figure 8. Mahalanobis Distances Between Test Patient Images and Training Patient Image Distribution. Flagged patients using the optimal threshold were shown in peach color.

The receiver operating characteristic (ROC) curve using the Mahalanobis distance to capture patients that require major edits was plotted in Figure 9. The area under curve value was 0.89. Using the optimal threshold indicated by ROC curve that maximized true positive rate and minimized false negative rate, the specificity and sensitivity of flagging failed contours were 0.91 and 0.79 respectively.

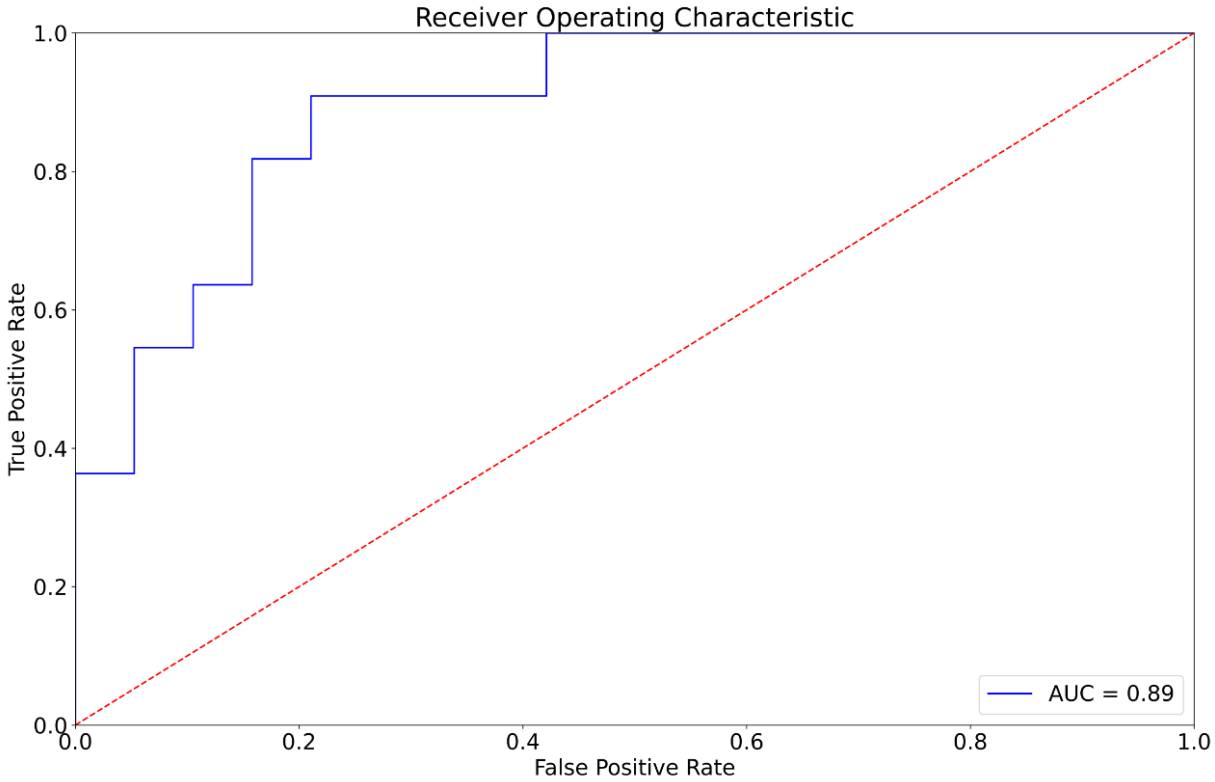


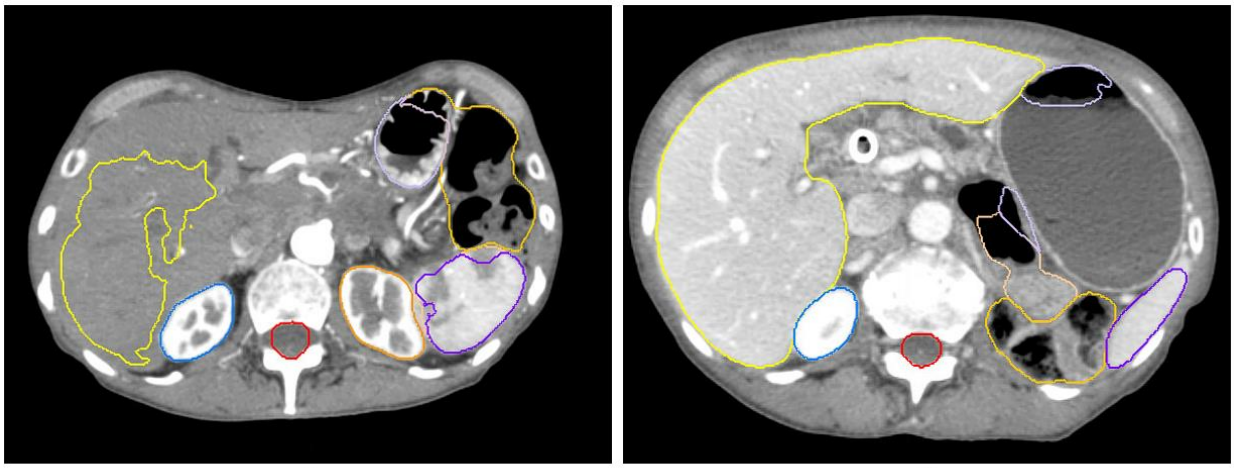
Figure 9. Receiver operating characteristic curve of our proposed contour QA method.

Our deep feature-based contour QA approach was able to successfully detect 10 out of 11 patient contour sets that required major edits on at least one organ, using the optimal threshold value. A total of 14 patients were identified and flagged as shown in figure 8. Among them, seven patients required major edits on multiple organs. Anatomical variations that tend to degrade contour quality were identified by our QA approach. One patient was flagged due to the poor NPO status that led to contour failures as shown in Figure 10b. Another patient was flagged due to ascites as shown in Figure 12a. By flagging these images, the approach could provide with an indication of which patients may not be suitable for the deep learning segmentation model.

Identifying misuse of the segmentation model on unapproved medical images was also a crucial component of our QA tool. At our center, we utilize both breath-hold and 4D-CT techniques for motion management in GI cancer. Most of our hypo-fractionated treatments are completed with the breath-hold technique. During the initial deployment of the abdominal OAR segmentation model, we considered contour generation on average CT as off-label use, as the model had not been validated for such images. Out of the 30 patients sampled for our test cohort, 6 had average CT images as planning images. Our approach identified all patients with average CT, demonstrating its effectiveness in detecting off-label use of a deep learning model. In terms of contour quality, we observed acceptable segmentation performance from 2 patients, as shown in Figure 11.

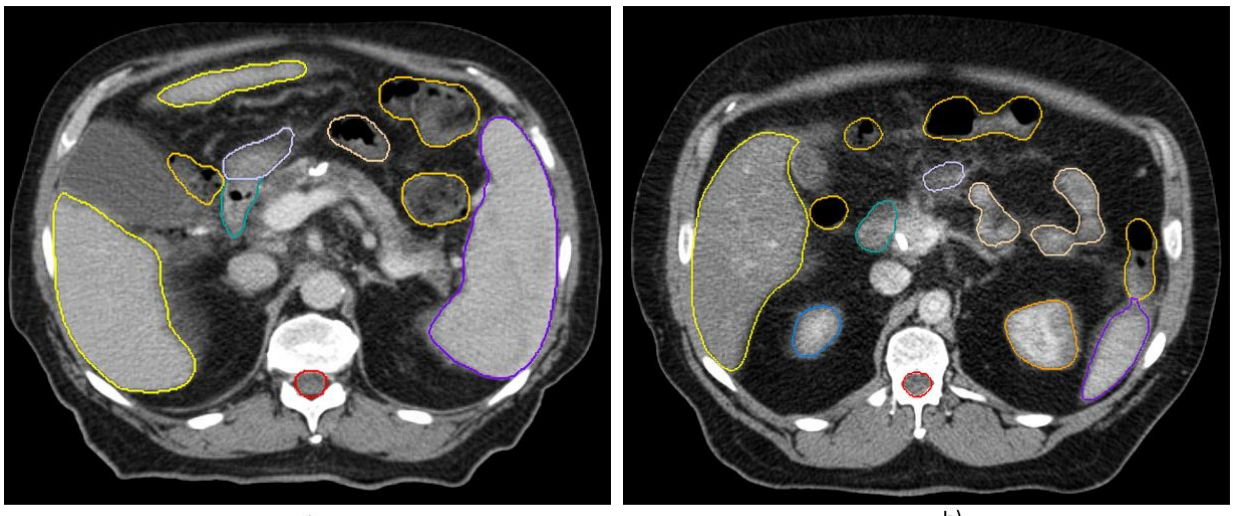
In total, four patients were identified as false positives, with two of these patients having average CT scans with acceptable clinical contours. The other two false positive patients were displayed in Figure 12. One patient had a single failed liver contour, which was missed by our

QA method and resulted in a false negative. Nonetheless, the remaining patients were correctly flagged, demonstrating the effectiveness in our contour QA approach.



a) b)

Figure 10. Patients with multiple organ contour failures were correctly flagged with our QA approach. All organs required major edits except for kidneys and spinal cord on patient a. Duodenum, small bowel and stomach required major edits in patient b due to poor NPO (nothing-by-mouth) status.



a) b)

Figure 11. Acceptable contours falsely flagged by our contour QA approach on patients planned on with average CT. These patients used 4D CT as their motion management and are common source of off-label use of our segmentation model.

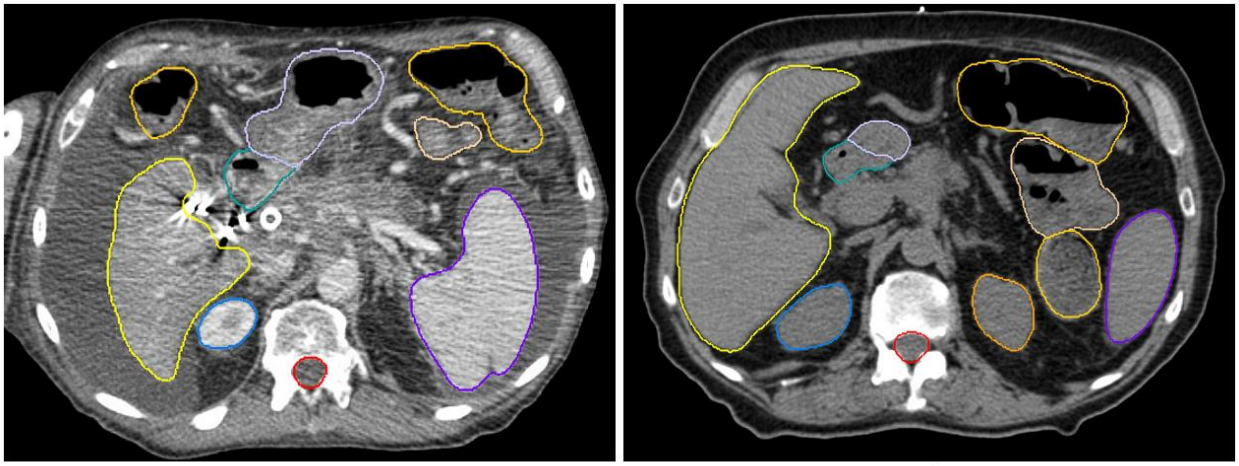


Figure 12. Falsely flagged patients by our contour QA approach. Patient a had ascites as well as metal artifacts. Patient b was planned on a non-contrast-enhanced CT due to contrast allergy The auto-segmentation model exhibited acceptable performance on these two challenging cases with varying imaging characteristics.

4.4 Discussion

As deep learning-based auto-segmentation enters clinics, the quality assurance of its performance becomes crucial in the safe deployment of deep learning models[52]. These auto-segmentation models, however, was often seen as black-box and lack interpretability[12]. Therefore, tracing and predicting failures of these models remains challenging. By extracting features from the deepest layers of our model, we probed the root causes of failure in deep learning models. In earlier work, we had noticed that DSC scores are not good indicators of clinical acceptability[53]. Here we used physician review to identify failures that were clinically relevant. With inputs from experts, we successfully captured out-of-distribution images that caused degradation in patient contour quality using deep features.

We achieved excellent discrimination by using Mahalanobis distance to differentiate out-of-distribution images that might lead to subpar segmentation performance. Existing contour QA approaches for deep learning auto-segmentation often require multiple models. A few studies have addressed the quality assurance of deep learning contours in the clinic. Isaksson et al. and Chen et al. aimed to predict Dice similarity coefficient between the ground truth and the generated contour given an image contour pair[54,55]. Rhee et al. proposed using independent segmentation models to capture failed contours[56,57]. These methods typically required new training dataset or separate models to capture failed contours and demanded expert input and are expensive to develop. Moreover, deep learning auto-segmentation models are subject to updates during their clinical lifecycle, often resulting from performance enhancement using additional data[58] or evolving clinical contouring guidelines[59]. A contour QA approach that requires new model development would need to undergo a new cycle of training and validation for a new

model. In contrast, our method required no additional data collection or model training. Since the developers of the segmentation model are familiar with the training dataset, our approach offers an intuitive solution to identify patients that are less suitable for the deployed model. By only requiring deep features from the encoders, our method can be customized to perform QA for different deep learning models. This feature makes it suitable for the constantly changing landscape of clinical practice. Furthermore, our approach can be easily incorporated alongside clinically validated nnU-Net models during the contour generation process, as it only takes 30 seconds per patient.

Our distance-based contour QA approach can be sensitive to the relationship between the distribution of the training and testing image set. If training samples were not representative of images encountered during clinical operation, the test samples might consistently fall out-of-distribution. As noted in Figure 11 and 12, models might encounter out-of-distribution images and still deliver clinically acceptable contours. Therefore, when models were deployed on patient cohort that were drastically different from the training set resulting from different imaging protocol, the Mahalanobis distance between training and testing images would be less indicative of contour quality. Our approach could generate excessive false positives in this scenario. Expanding the training dataset to include high-quality contours that are automatically generated from the segmentation model on the test cohort could potentially address this issue. In addition, our study was also constrained by the limited number of expert-scored contours. Organ-by-organ scores on complete contour sets evaluated by experts were time-consuming and expensive to obtain. Future work would include more patients in both training and testing cohort to show the prowess of our QA approach in a high-throughput environment.

4.5 Conclusion

In this study, we developed a contour QA approach that can be deployed alongside a validated deep learning auto-segmentation tool. By extracting features from the deepest layer of the neural network, we sought to identify out-of-distribution test samples from the model's perspective that are susceptible to contour failures. This approach can successfully identify patients unsuitable for the segmentation model by comparing deep features from the training set and the test sample. Using Mahalanobis distance as an indicator, our contour QA approach achieved an AUC of 0.89. Using this threshold, the specificity and sensitivity of flagging failed contours were 0.91 and 0.79 respectively. Our contour QA tool offered a fast and accurate solution for quality assurance of deep learning-based segmentation models.

Chapter 5: Transformer-based Liver Tumor Segmentation Driven by Self-supervised Learning

5.1 Introduction

Hepatocellular carcinoma (HCC) is the most common type of primary liver cancer, accounting for approximately 75% of cases[60]. It is a major global health problem, with over 800,000 new cases diagnosed each year and a high mortality rate[61]. The treatment of HCC is often multidisciplinary including surgery, ablation therapy and radiation therapy. With the advancement of image guidance techniques, stereotactic body radiation therapy (SBRT) has promoted radiation therapy as a more important pillar for liver cancer management[62]. These types of dose escalated radiation therapy require accurate delineation of targets to ensure patient safety. Gross tumor volume (GTV) segmentation is particularly important in the dose escalated treatment planning process. It serves as the foundation for treatment planning since other target volumes such as planning target volume (PTV) and planning organs-at-risk volume (PRV) were dependent on GTV delineation in liver SBRT treatment planning. Automating GTV delineation is crucial for adaptive treatment in dose-escalated radiation therapy[63,64]. It enables timely adjustments to the treatment plan based on changes in patient anatomy, ensuring accurate and effective radiation delivery. Furthermore, the utilization of auto-segmentation for GTV delineation in liver SBRT facilitates post-treatment analysis[65]. By collecting GTV volumes automatically, clinicians can more efficiently analyze treatment outcomes, such as local control rates and overall survival. This automated approach enables the aggregation of large datasets, allowing for comprehensive and statistically significant analyses of treatment response across a broader patient population.

Deep learning-based auto-segmentation has demonstrated significant potential across various segmentation tasks. Among the different architectures used, the U-Net[17] approach has emerged as the most popular and widely utilized method for auto-segmentation. It consists of an encoder path that captures the context and spatial information from the input image, followed by a decoder path that reconstructs the segmented output. This architecture's unique design allows for the efficient extraction and integration of features at multiple scales, enabling accurate and detailed segmentation results. The U-Net architecture, however, requires significant amounts of data to achieve state-of-the-art performance. Recently, the transformer-based architecture has achieved state-of the-art results in a variety of computer vision tasks[18]. It is based on the transformer architecture which was originally proposed for natural language processing tasks[66]. Unlike traditional convolutional neural networks (CNNs), which are designed to process image data in a hierarchical and localized manner, transformer-based models can capture long-range dependencies and global context information, making them well-suited for tasks that require a more holistic understanding of the input data. In transformer-based segmentation, the model takes an image as input and generates a pixel-wise segmentation mask as output. The key idea is to use self-attention mechanisms, which enable the model to attend to different parts of the input image and incorporate relevant information into the segmentation process. This allows the model to better handle complex and variable-sized objects, as well as to deal with occlusions and overlapping regions. Its ability to incorporate global context and long-range dependencies has made it capable of adopting self-supervised learning techniques from natural language processing[67].

Self-supervised learning addresses the limitations of traditional supervised learning methods that require extensive annotated data, which can be expensive to obtain. In contrast to

supervised learning, self-supervised learning leverages the intrinsic information present in unlabeled data to learn meaningful representations without the need for explicit annotations[68]. In the context of medical image segmentation, self-supervised learning methods aim to train models by predicting image properties or generating informative surrogate tasks that indirectly facilitate segmentation. By utilizing these surrogate tasks, the model can acquire robust and discriminative features, which can subsequently be applied to accurately segment medical images. One commonly employed approach in self-supervised learning is the use of pretext tasks[69]. These tasks involve generating auxiliary labels or transformations from the unlabeled data, such as image rotations, translations, or predicting missing parts of an image. By training the model to solve these pretext tasks, it learns to capture relevant contextual information and underlying structures within the data, which can be advantageous for subsequent segmentation tasks.

In this study, our aim was to investigate the influence of pretraining on segmentation performance in transformer-based auto-segmentation for liver tumors. Specifically, we employed the Swin-UNETR architecture[70], which combines an encoder-decoder design similar to U-Net with the Swin-transformer as the encoder and a CNN-based decoder. The utilization of the hierarchical Swin-transformer allowed for both local accuracy and self-supervised training. To evaluate the impact of self-supervised pretraining on transformer-based auto-segmentation, we curated a new set of unlabeled CT images tailored for liver tumor segmentation. Our objective was to examine whether utilizing an in-domain pretraining dataset can improve the performance of the downstream segmentation model.

5.2 Method

A comprehensive dataset comprising a total of 3093 CT scans was gathered of liver cancer patients treated at MD Anderson Cancer Center between the years 2011 and 2022. The

dataset encompassed portal-venous phase contrast-enhanced simulation scans and non-contrast-enhanced image guidance scans. Our unique motion management protocol involving CT-on-rails enabled the acquisition of a substantial number of unlabeled CT scans tailored for the pretraining of the liver tumor segmentation model. In addition, a publicly available dataset consisting of 5050 CT scans was collected to serve as a benchmark for the pretraining process. This supplementary dataset encompassed CT scans from various anatomical regions, including the head, neck, chest, abdomen, and pelvis. To ensure standardization and comparability, 95% of both datasets were allocated for training purposes, while the remaining 5% was set aside for validation. Prior to training, all CT scans underwent preprocessing steps, which involved intensity value clipping within the range of -1000 to 1000, followed by renormalization. Furthermore, the scans were resampled to achieve an isotropic voxel size of 1.0x1.0x1.0 mm. Subsequently, patches measuring 96x96x96 were extracted from the non-air regions of the CT scans, serving as the training input.

During the pre-training process, we optimized the Swin-transformer encoder by employing various proxy tasks for self-supervised representation learning. The main objective was to encode ROI-aware information of the human body. Inspired by previous studies on context reconstruction and contrastive encoding, we incorporated three proxy tasks. Firstly, we applied the cutout augmentation technique to randomly mask out ROIs in the sub-volume, compelling the model to regenerate the original patch. Secondly, we predicted the angle categories representing the rotation of the input sub-volume. Additionally, we leveraged self-supervised contrastive coding to enhance representation learning. This involved maximizing the mutual information between positive pairs (augmented samples from the same sub-volume) while minimizing it between negative pairs (views from different sub-volumes). The overall

training process involved minimizing the total loss function, which encompassed masked volume inpainting, 3D image rotation, and contrastive coding as the key objectives in pre-training the Swin-transformer encoder[70]. The Swin-transformer encoder was trained on an A100 GPU. The training process utilized a learning rate of 1e-4, which was decayed by a factor of 1e-5. A batch size of 4 was used in pre-training. The model was trained for a total of 450,000 iterations of the training process. The training duration for the pretraining phase was approximately 303 hours.

In the final training phase, we utilized the scans and segmentations from the liver task of the Medical Segmentation Decathlon Challenge[48]. The dataset consisted of 130 contrast-enhanced CT images acquired from a cohort of patients diagnosed with primary cancers, specifically colorectal, breast, and lung cancers, as well as metastatic liver disease originating from these primary cancers. To assess the performance of our model, we randomly selected 30 scans as the independent test set. From the remaining 100 scans, we performed random sampling to create multiple training and validation sets. Specifically, we randomly selected subsets of 20, 50, and 100 scans, which were then divided into an 80:20 ratio for training and validation, respectively. The same training/validation split was maintained during the training of the final liver tumor segmentation models to ensure a fair comparison. For preprocessing, each CT scan was resampled to an isotropic voxel size of 1.0 x 1.0 x 1.0 mm. From these resampled scans, we extracted patches with dimensions of 96 x 96 x 96, which served as the input for our model. To enhance the variability and robustness of the training data, we employed extensive data augmentation techniques, including random flipping, rotation, intensity scaling, and shifting. The training process was conducted on an A100 GPU for a total of 5000 epochs. Three different training configurations were explored: training from scratch, training with an encoder pretrained on the public CT dataset comprising 5050 CT scans, and training with an encoder pretrained the

private MDA dataset specific to liver cancer patients, consisting of 3094 CT scans. To ensure a fair comparison, we maintained identical hyperparameters across all configurations using the reported configuration in Tang.et al[70]. which achieved competitive results at the Medical Segmentation Decathlon[48]. The learning rate was set to 1e-5, and the batch size was set to 2. The Dice Similarity Coefficient (DSC) scores between the ground truth and the generated contours for both liver and tumor segmentation were measured to evaluate the performance of different training strategies.

5.3 Results

The quantitative evaluation of the Swin-UNETR segmentation model, trained with different pretraining configurations, is presented in Table 7. When trained with 20 patients, the encoder pretrained with the public unlabeled dataset achieved the best performance in liver segmentation. However, training from scratch yielded the best mean DSC scores for liver tumor segmentation. Upon closer analysis in Figure 13, we found that the encoder pretrained with liver cancer patients achieved the highest median DSC. Additionally, this configuration exhibited the narrowest range between the 25th and 75th percentiles of all DSC scores, coinciding with the smallest standard deviation among the three configurations. In the case of training with 50 patients, training the model from scratch demonstrated superior mean DSC performance in liver segmentation, while the encoder pretrained with the public dataset yielded the best mean DSC results for liver tumor segmentation. The encoder pretrained with liver cancer patients produced the highest median DSC. The utilization of pretraining led to improved segmentation outcomes for liver tumors when trained with 50 patients. However, this performance enhancement was not sustained when the training and validation set was expanded to include 100 patients. In this scenario, training the model from scratch exhibited the best performance in both liver and liver

tumor segmentation tasks, as indicated by the mean DSC. Furthermore, the public pretrained weight exhibited a narrower range between the 25th and 75th percentiles, while training from scratch yielded the highest median DSC.

Patient No.	Structure	Liver Pretraining		Public Pretraining		Train from Scratch	
		DSC	Std	DSC	Std	DSC	Std
20	Tumor	0.354	0.270	0.370	0.286	0.381	0.297
	Liver	0.921	0.074	0.913	0.066	0.920	0.083
50	Tumor	0.468	0.313	0.474	0.295	0.407	0.310
	Liver	0.934	0.081	0.940	0.046	0.940	0.037
100	Tumor	0.509	0.300	0.549	0.290	0.576	0.281
	Liver	0.950	0.035	0.948	0.039	0.951	0.035

Table 7. DSC scores between ground truth and contours generated by Swin-UNETR liver segmentation models using different pretraining strategies

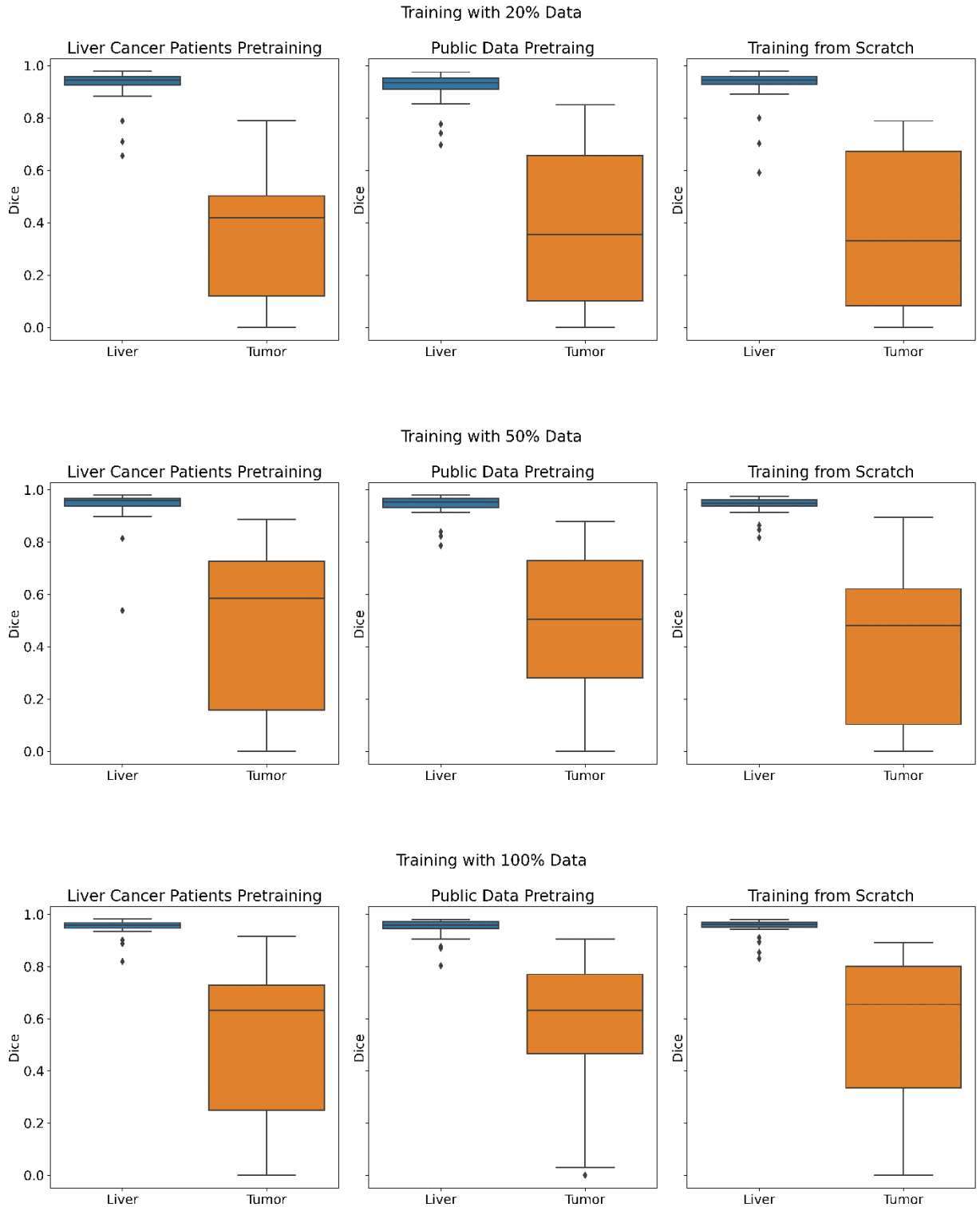


Figure 13. Boxplots of DSC scores between ground truth and contours generated by Swin-UNETR liver segmentation models using different pretraining strategies.

The qualitative assessment revealed impressive performance from all Swin-UNETR models in challenging cases. As depicted in Figure 14, we observed excellent tumor segmentation quality, particularly when the contrast timing was optimal, even in scenarios with a high number of tumors present. Although a few small lesions were missed, the overall performance remained respectable even when a small training set of patients was utilized. For tumors with ideal contrast, the impact of pretraining strategies was deemed minimal as all models managed to segment the majority number of lesions.

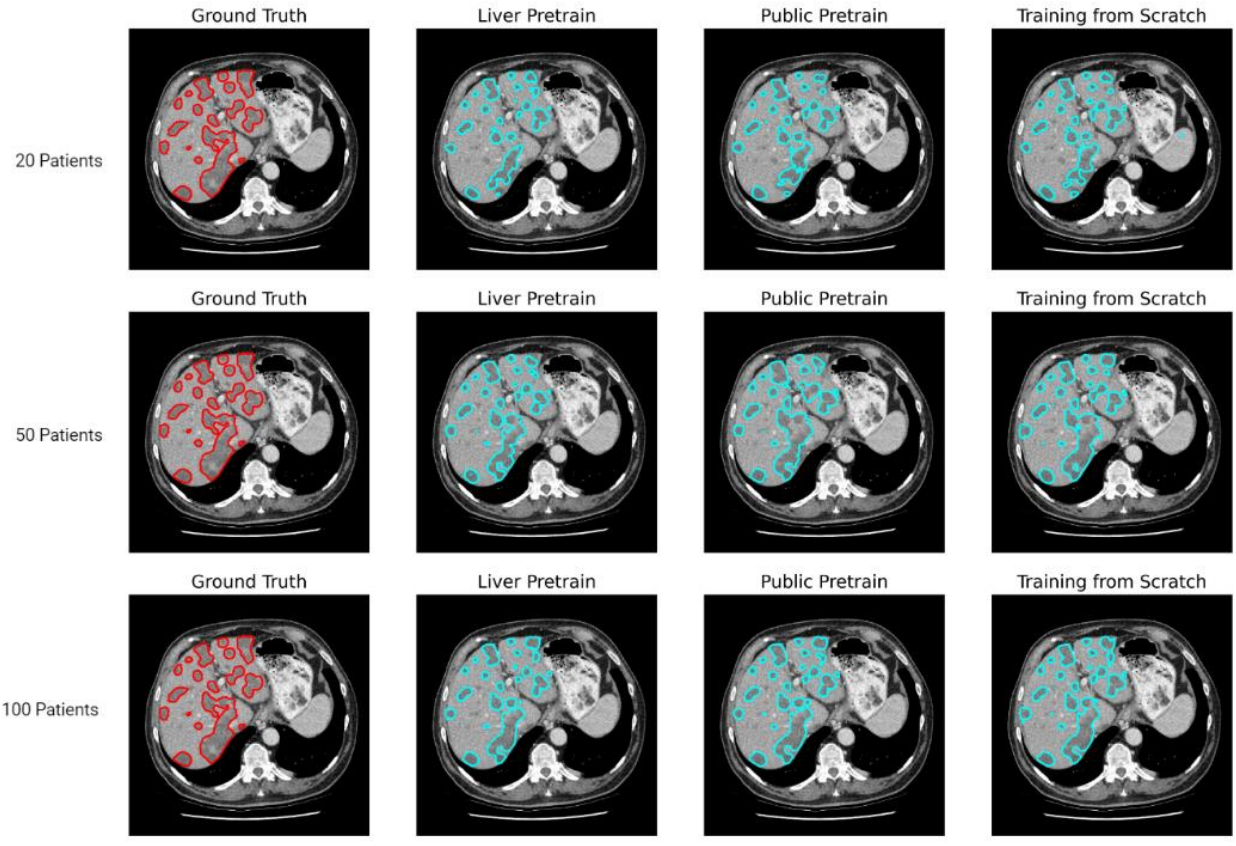


Figure 14. Generated tumor contours of Swin-UNETR models using different pretraining strategies and training dataset size for patients with multiple lesions. All models achieved respectable performance partially due to excellent contrast between tumor and liver parenchyma.

Figure 15 demonstrated a steady improvement in model performance as the number of available training patients increased. Initially, all models displayed hesitancy in segmenting tumors along the posterior inferior end of the liver when trained with a small dataset. Models optimized with pretraining exhibited an even more cautious approach in segmenting tumors, although their performance improved as the number of patients increased. The models' tendencies to under-segment uncertain regions resulted in reduced variations, indicated by a smaller range between the 25th and 75th percentile DSC scores. However, these models also yielded lower overall DSC scores when compared to models trained from scratch. The encoder pretrained with liver cancer patients did not exhibit superior performance compared to its counterpart pretrained with a public dataset. In fact, it performed worse than the model trained from scratch. These findings indicated that domain-specific pretraining data did not enhance, but rather deteriorated model performance for the task of liver tumor segmentation. This result highlighted the importance of proper model initialization prior to training.

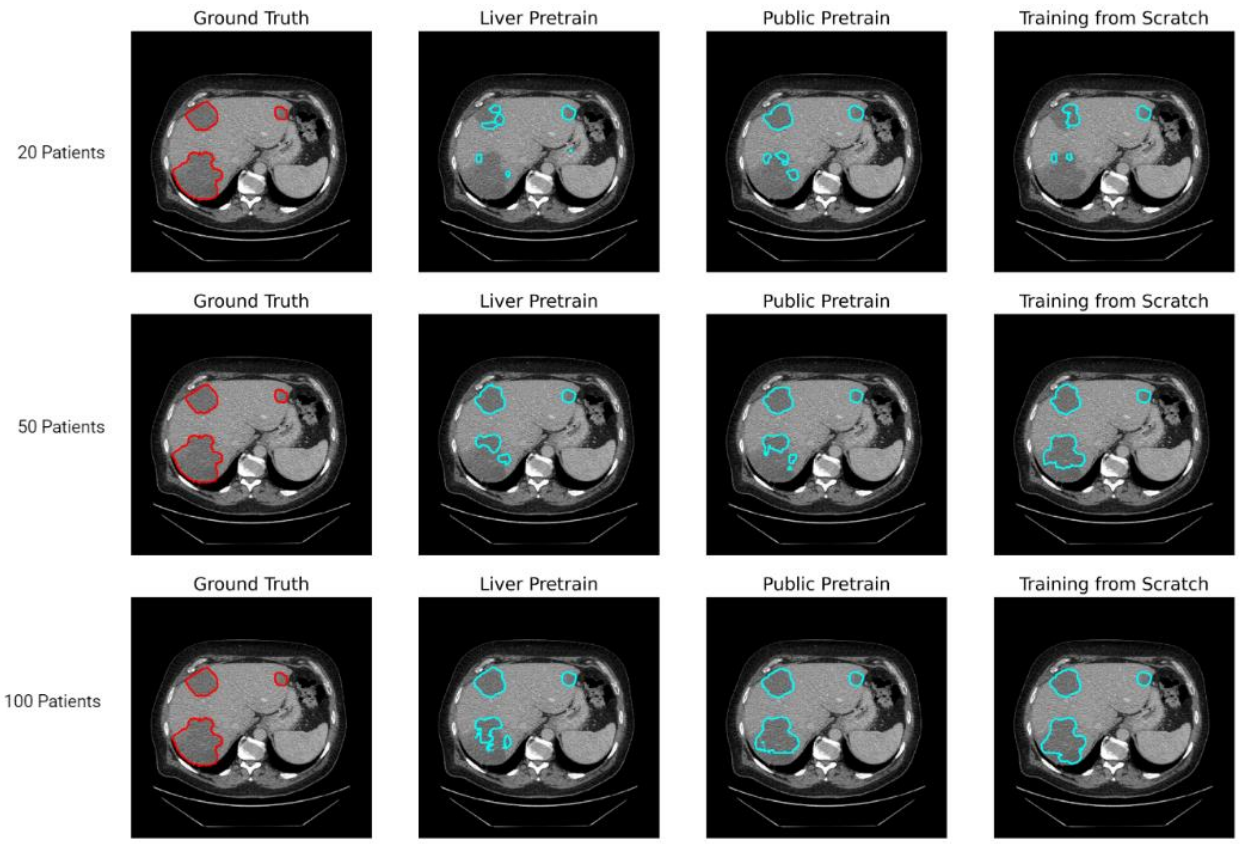


Figure 15. Generated tumor contours of Swin-UNETR models on portal venous phase CT images using different pretraining strategies and training dataset size. The segmentation quality increased as the dataset size increased. Encoder pretraining exhibited more conservative behavior for the lesion along the posterior inferior end of the liver.

5.4 Discussion

Deep learning-based auto-segmentation has emerged as the predominant approach for medical image segmentation, offering substantial advancements in accuracy and efficiency. The fundamental advantage of CNN-based architectures, such as U-Net[17], V-Net[71], or DeepLab[72], lies in their exceptional capacity to capture spatial dependencies and hierarchical representations.

Through multiple layers of convolution and pooling operations, these networks extract intricate features at varying scales, enabling them to capture fine-grained details and contextual information necessary for precise segmentation.

However, the success of CNN-based auto-segmentation heavily relies on the availability of large-scale annotated medical image datasets[12]. The scarcity of annotated medical image datasets poses a limitation on the generalizability of the models. Moreover, the variability in imaging protocols across different institutions can degrade model performance[73], particularly when models are optimized on publicly available datasets. Therefore, it is imperative to develop effective approaches that leverage unlabeled or weakly labeled data through self-supervised or semi-supervised learning techniques, which hold promise in mitigating these challenges.

With the emergence of transformer-based auto-segmentation approaches as a viable alternative in computer vision, self-supervised learning has gained significant attention in the field of medical image segmentation. Unlike their CNN counterparts, transformers are not limited by their receptive fields[74], enabling transformer-based encoders to reconstruct representations from corrupted image volumes. This capability provides transformers with an inherent advantage in various pretext tasks for self-supervised learning. By combining transformers with self-supervised learning, models can capture intrinsic patterns and structures in unlabeled data, allowing them to extract meaningful representations that can greatly benefit subsequent segmentation tasks[75]. This approach not only reduces the reliance on annotated data but also allows for the discovery of complex patterns and relationships derived from a much larger number of patients that may not be readily apparent in limited labeled datasets.

In this study, we aimed to investigate the impact of different pretraining datasets on the segmentation performance of our final model. The selection of the unlabeled data used for self-supervised pretraining plays a crucial role in shaping the segmentation model's capabilities. We curated a liver cancer CT dataset consisting of 3094 CT scans, specifically focusing on the anatomical structure and pathology relevant to the segmentation task, namely the liver and liver tumors. The intention behind this tailored dataset was to enable the pretrained model to acquire task-specific knowledge that could enhance the segmentation performance in the target domain. In contrast, the publicly available dataset, comprising 5050 CT scans, encompassed a broader spectrum of anatomies and pathologies. These datasets have been widely adopted for pretraining purposes in various medical imaging tasks. By comparing the performance of the model pretrained on our private liver cancer CT dataset with that of the model pretrained on the publicly available dataset, we can evaluate the benefits of tailored vs. generic pretraining data. This analysis sheds light on the extent to which task-specific knowledge acquired from a curated dataset can outperform the broader knowledge gained from a more diverse dataset.

The utilization of encoder weights generated from our specific in-domain liver cancer dataset did not result in superior performance in terms of mean DSC compared to pretrained weights optimized with publicly available CT scans from diverse anatomical regions, such as the head, neck, chest, abdomen, and pelvis. In models trained with 25 and 50 patients, the model pretrained with liver cancer patients exhibited the highest median DSC. This meant the distribution of DSC scores for this configuration was skewed towards higher values. Furthermore, both pretrained models exhibited a more conservative segmentation style during qualitative evaluation when encountering regions of uncertainty as shown in figure 15. This

tendency towards under-segmentation had a detrimental effect on the quantitative results of tumor segmentation, leading to inferior performance overall.

However, both pretrained models demonstrated impressive performance in tumor segmentation when the training set consisted of 50 patients. This particular training data size appears to be an optimal point for self-supervised pretraining to enhance the performance of the baseline model. Existing literature has also reported diminishing performance gains from self-supervised pretraining as the training set grows larger. Our study corroborated these findings, as the model trained from scratch achieved the best performance when all 100 patients were included in the training set. On the other hand, consistently, the pretrained models exhibited the smallest range between the 25th and 75th percentile DSC scores as shown in Figure 13. This observation further supported our qualitative observation that pretrained models exhibit less variability in the generated contours during qualitative evaluation. Adequate pretraining allowed the encoder to assimilate more data, resulting in more cautious models.

The utilization of liver cancer patients for pretraining did not yield noticeable benefits in liver tumor segmentation. Despite achieving the highest median Dice Similarity Coefficient (DSC) scores for both the 25 and 50 patient datasets, our liver cancer pretrained weights exhibited inferior performance overall compared to the publicly pretrained weights. The liver tumor segmentation model derived from the liver cancer pretrained weights did not exhibit enhanced robustness, even though the pretraining dataset was abundant and closely aligned with the downstream task. This unexpected finding could be attributed to two potential causes. Firstly, the model initialized with liver cancer pretrained weights might have suffered from overfitting. Given the similarity within the dataset (identical pathology, similar imaging protocol), the model might have acquired suboptimal representations due to the homogeneity of the dataset. In the

context of self-supervised learning in medical imaging, the objective is to construct a wholistic representation of cross-sectional anatomy. Hence, the more diverse the pretraining dataset, the higher the likelihood of achieving generalizability in the model encoder. However, generalizability was essential when the downstream task was drastically different from the pretraining population. It was surprising to observe that a homogeneous dataset failed to contribute to the model's ability to generalize effectively in an in-domain downstream segmentation task (liver cancer patients to liver tumor segmentation). The relationship between representation construction through self-supervised learning and downstream segmentation turned out to be more intricate than anticipated.

This led us to consider the second potential cause for the decline in performance. While we were able to monitor the validation loss during pretraining, there was no direct indication of how well it would translate to the downstream task of liver tumor segmentation. Assessing the training and data quality of pretraining was challenging until after the completion of training for the downstream segmentation model. Since the accomplishment of meta-tasks in self-supervised learning only had theoretical correlations with improved performance in the downstream model, real-time evaluation of pretraining quality prior to final testing was difficult. Even though we selected the pretrain weights with the lowest validation loss, there was no guarantee that this particular set of weights would yield the best results in the downstream liver tumor segmentation task. In addition, a number of hyperparameter combinations led to instability in training and the encoder failed to converge. Our final pretrained encoder was selected from a limited number of candidates based on validation loss and might have not been the best optimized option. Hence, we suspected that we were unable to fully extract the potential of our liver cancer dataset due to suboptimal pretraining.

Therefore, to ensure the efficacy of the pretrained model, it is essential to include variations and complexities representative of the target segmentation task within the pretraining dataset. This enables the model to acquire robust and discriminative features that exhibit effective generalization capabilities when applied to new data. Moreover, the quality of training during the pretraining stage demonstrated considerable variation. The process required extensive expert input to stabilize the training and avoid potential performance degradation. Proper handling of the training process is essential to ensure optimal model performance and prevent any adverse effects. These findings underscore the significance of diverse dataset selection and optimized training procedures in achieving superior performance in self-supervised learning-enhanced auto-segmentation. A meticulous approach that incorporates representative variations and expert-guided training strategies is crucial for maximizing the effectiveness and generalizability of pretrained models in medical image segmentation tasks.

In contrast, training the segmentation model from scratch yielded satisfactory segmentation performance and often surpassed the performance of both pretrained encoders in terms of mean DSC. These results indicated that while self-supervised pretraining shows potential, its effectiveness in enhancing performance can be variable. Merely conducting self-supervised training does not guarantee improved performance. The successful utilization of pretrained encoders relies on careful tuning not only of the pretrained encoder itself but also of the downstream segmentation model. Proper initialization and parameter optimization are critical for achieving optimal performance. Neglecting these aspects can lead to performance degradation rather than improvement. Careful attention should be given to ensure appropriate initialization and effective parameter tuning to maximize the benefits of self-supervised pretraining in the context of medical image segmentation.

5.5 Conclusion

Self-supervised pretraining with unlabeled data holds significant potential for enhancing medical image segmentation. The utilization of large and diverse datasets, including both publicly available and private datasets, enables models to acquire valuable representations that contribute to improved segmentation performance. However, the success of self-supervised pretraining relies on meticulous selection of the pretraining dataset and careful tuning of the encoder and the segmentation model. In this study, we utilized a large private unlabeled liver cancer CT dataset to pretrain our Swin-transformer encoder for liver tumor segmentation. We found that our unique dataset combined with self-supervised learning technique failed to enhance our segmentation results. We recommend a diverse and large unlabeled dataset for self-supervised pretraining instead of a domain specific dataset. In addition, training from scratch is also a viable option when sufficient labelled data are available. We hope that our work can contribute to the understanding of self-supervised learning in the field of medical image segmentation.

Chapter 6: Uncertainty-guided Pancreatic Tumor Auto-segmentation with Tversky Ensemble

6.1 Introduction

Pancreatic cancer remains a significant challenge in modern oncology, projected to become the second leading cause of death in the United States by 2030[76]. To optimize the management of this deadly disease, a multi-disciplinary approach is commonly employed, with radiation therapy serving as a critical component. However, standard doses of radiation therapy have been found to be inadequate for effective tumor control in pancreatic cancer patients[77]. Consequently, dose escalation has emerged as a prevailing strategy for treating inoperable locally advanced pancreatic cancer after systemic chemotherapy[47]. To implement these treatment modalities successfully, accurate identification of the pancreatic tumor is essential. Unfortunately, pancreatic tumors remain notoriously difficult to differentiate from the surrounding parenchyma, even for experienced clinicians. Currently, the clinical workflow in radiation therapy of pancreatic cancer involves contouring the tumor on the portal-venous phase of the contrast-enhanced CT scans. Given the significant level of inherent uncertainty, clinical notes are often necessary to achieve the desired level of accuracy in tumor contouring.

In recent years, deep learning-based auto-segmentation has emerged as the preferred method for biomedical image segmentation, owing to its remarkable performance in a broad spectrum of applications. Notably, the U-Net architecture and its variants have demonstrated exceptional efficacy in diverse image segmentation tasks. More recently, transformer-based architectures have gained considerable attention in computer vision tasks, with vision transformers (ViTs) exhibiting state-of-the-art performance in image classification[18]. In the

context of medical image segmentation, transformer-based architectures combined with self-supervision training techniques have shown superior results compared to U-Net in the medical segmentation decathlon[70]. Although deep learning-based approaches have shown remarkable performance, they often suffer from a tendency towards overconfidence in probability estimation[73]. This can be particularly challenging in segmentation tasks where ground truths are uncertain, as in the case of pancreas tumor segmentation. While segmentations are typically derived from probability maps, these maps often do not accurately reflect the true probability distribution.

To achieve real-world probability estimates using deep learning models, calibration is necessary to ensure the predicted probability map is accurate. Calibration techniques such as Monte-Carlo dropout[78] and test-time augmentation[79] are widely used to generate accurate uncertainty estimates in tandem with the segmentation results, which enables users to obtain interpretable outcomes. Another promising approach to address overconfidence in deep learning models is the use of deep ensembles[80]. This study highlights that models trained with different configurations can reach their conclusions in distinct ways. By averaging the output probability of high-performance segmentation models within an ensemble, a more robust probability map can be generated that reflects the consensus of expert models. Notably, calibration results are more precise when model configurations diverge[80]. Incorporating a diverse set of model configurations within an ensemble for segmentation can not only provide accurate uncertainty estimation but also enhance the model's segmentation performance.

In this study, we employed the state-of-the-art transformer-based architecture Swin-UNETR to perform pancreatic tumor segmentation[70]. Traditional segmentation methods with discrete output encounter challenges in cases where the ground truth is uncertain. To tackle this

problem, we incorporated Tversky losses[81] to generate models with different contouring styles, which were developed to accommodate stylistic differences from different experts. By incorporating different segmentation styles, we constructed a deep ensemble with varying segmentation tendencies to create a calibrated probability map for pancreas tumor segmentation. This probability map enabled us to generate segmentations that align with physician needs through thresholding. Moreover, we could mitigate segmentation errors by eliminating regions with high uncertainty via thresholding.

6.2 Method

Our study included a total of 282 portal-venous phase CT scans from the pancreas task of the Medical Segmentation Decathlon[48]. Both pancreatic masses (cyst or tumor) and parenchyma were delineated. To create an independent test set, 30 patients were randomly selected. The remaining 252 patients were divided into a training set (80%) and a validation set (20%). To fully leverage the entire training set, we utilized five-fold cross validation. The CT images were clipped from -87 to 199 HU and resampled isotropically at 1.0mm x 1.0mm x 1.0mm. Given that the Swin-UNETR architecture is 3D-based, we cropped images into 96 x 96 x 96 patches with an overlap of 50%. Additionally, data augmentation strategies such as random flip, rotation, intensity scaling, and shifting with varying probabilities were employed. The training of the model was conducted on a single A100 GPU for a total of 5000 epochs with a learning rate of $1e^{-4}$ and a batch size of 2. Each member of the ensemble required 151 hours to complete training. In order to ensure a fair comparison with the state-of-the-art Swin-UNETR models in the pancreas task, the preprocessing pipeline and hyperparameters were kept identical as reported in Tang et al.[49]. This was done to eliminate any potential confounding factors that could influence the performance comparison.

To integrate various segmentation styles into our ensemble, we utilized the Tversky loss layer during our training process. The baseline ensemble of Swin-UNETR models aims to minimize the Dice similarity coefficient during training, which assigns equal weight to false positives (FP) and false negatives (FN):

$$DSC = \frac{2TP}{2TP + FN + FP}$$

Tversky index, on the other hand, allows us to weigh FP and FN:

$$TI = \frac{TP}{TP + \alpha FN + \beta FP}$$

Here, α and β (α+β=1) controls the magnitude of the penalties for FN and FP. Through manipulating the Tversky index hyperparameters, we can customize the segmentation tendencies of our models. Models with an α greater than 0.5 have a tendency to under-segment as they penalize false negatives more heavily. Conversely, models with an α less than 0.5 tend to over-segment as they prioritize false positives. However, optimal and well-balanced segmentation is still maximally rewarded regardless of these tendencies.

Utilizing the Tversky loss, we can regulate each model's segmentation tendencies to imitate the contouring styles of multiple experts. To create a Tversky ensemble, we assigned unique α values to each of the five members. The ensemble was trained with α values of 0.1, 0.3, 0.5, 0.7, and 0.9, respectively. The introduction of distinct training-validation folds for each member introduced both data and stylistic variations into the ensemble, leading to improved uncertainty estimation.

The model's predictions were generated using sliding windows with a 50% overlap, and the mean probability of all members in the Tversky ensemble was utilized. The calibrated probability map for tumor prediction was directly extracted from the inference results. Probability thresholding was applied to produce the final segmentation, and for quantitative evaluations against the ground truths, we extracted eleven final segmentations by varying the threshold values ranging from 0.05 to 0.9 on the probability map.

6.3 Results

Table 8 presents the quantitative results of all the thresholded contours. Our findings indicated that thresholding the probability maps with a value of 0.05 yielded the highest Dice similarity coefficient (DSC) results, while contours thresholded with a probability value of 0.5 exhibited the lowest distance metric. Employing a less stringent probability threshold resulted in improved DSC values at the expense of increased variability in the generated contours, as reflected by the increasing distance metrics.

	DSC		HD95 (mm)		MSD(mm)	
	Mean	SD	Mean	SD	Mean	SD
0.05	0.47	0.33	14.43	14.19	4.92	9.54
0.1	0.46	0.33	14.80	14.45	4.93	9.57
0.2	0.45	0.33	14.94	14.63	4.95	9.62
0.3	0.44	0.34	15.38	15.91	5.65	10.64
0.4	0.44	0.34	15.55	16.05	5.61	10.46
0.5	0.43	0.34	14.03	10.98	4.07	6.13
0.6	0.42	0.34	14.11	10.99	4.09	6.12
0.7	0.41	0.34	14.16	11.03	4.10	6.11
0.8	0.40	0.33	14.29	11.07	4.11	6.10
0.9	0.39	0.33	14.52	11.11	4.11	6.08
Best	0.47	0.33	12.70	9.96	3.24	4.58
Swin-UNETR	0.43	0.34	13.40	10.08	3.83	4.12

Table 8. Contours created with varying thresholds of the probability map. Best results were created from selecting the contours with the lowest HD95 for each individual case. The Swin-UNETR results were from a 5fcv Swin-UNETR ensemble trained with DSC loss. The same data preprocessing and hyperparameter configuration won the pancreas task of the Medical Segmentation Decathlon.

In the clinical workflow, contours were thresholded in real-time, enabling physicians to select contours that align with their preferred stylistic preferences. To mimic the human-in-the-loop adaptation process, we selected the contours with the lowest 95th percentile Hausdorff distance (HD95) among the eleven probability thresholds from each patient for final quantitative evaluation. Our final quantitative results surpassed those of the Swin-UNETR configuration, which achieved state-of-the-art results in the pancreas task of the Medical Segmentation Decathlon challenge. The boxplots of the quantitative results were shown in Figure 16.

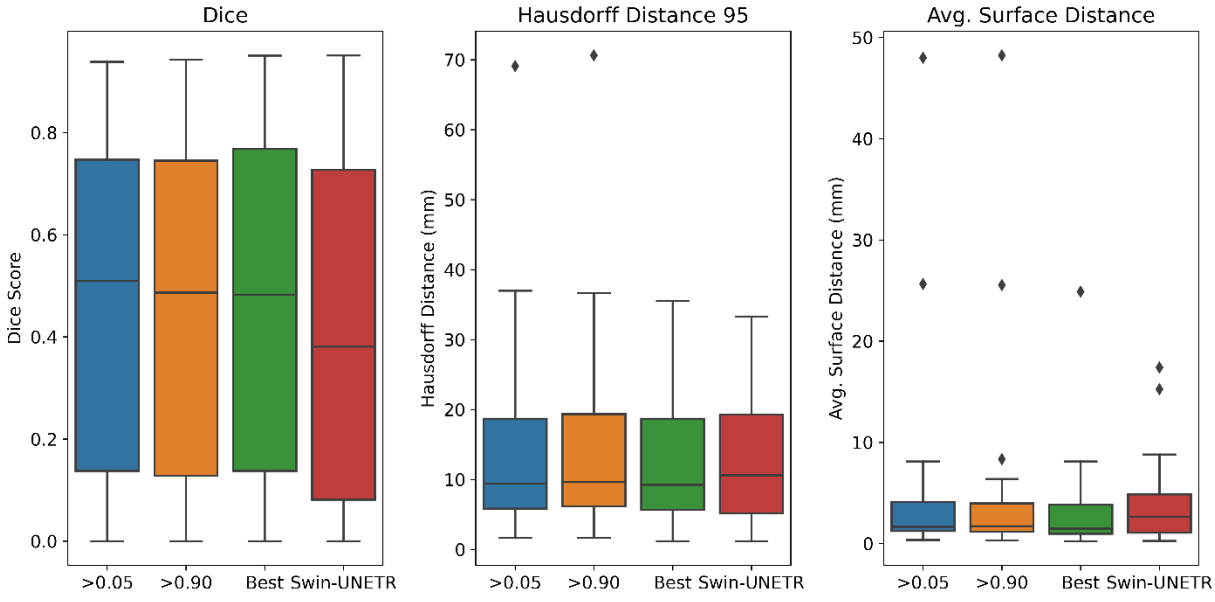


Figure 16. Quantitative results of automatically generated contours compared to ground truths. Contours were generated by thresholding the probability map with a variety of

values (0.05 and 0.9 as shown) and the contour with the lowest HD95 were chosen to serve as the best contour to compare against the Swin-UNETR ensemble.

We observed that our method achieved higher DSC scores while maintaining relatively low distance metrics across all thresholded contours. By mimicking expert input and selecting the contour with the lowest HD95 value, we obtained more competitive distance metric results compared to Swin-UNETR. Moreover, the mean and median DSC values remained competitive when the contours were selected by the lowest HD95 distance. These findings highlighted the effectiveness of incorporating customization into the auto-segmentation pipeline.

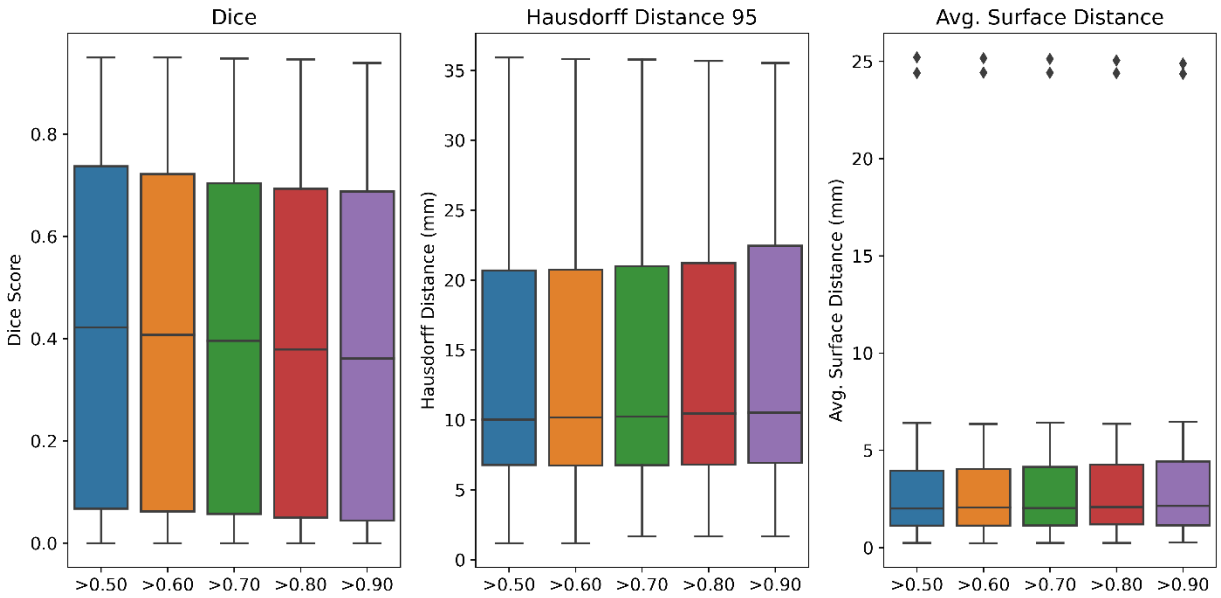


Figure 17. The trend of segmentation quality as thresholding value increases was shown in boxplots. The DSC scores between generated contours and ground truths decreased and the distance metric increased.

Upon closer examination of the segmentation performance with varying thresholding values, we observed a surprising trend. When employing an overly cautious segmentation

approach (high probability threshold), as depicted in Figure 17, the generated contours exhibited lower DSC scores, as anticipated. However, we also observed that this conservative contouring strategy led to poorer distance-based results. In contrast, an over-segmentation style yielded more favorable quantitative outcomes in both DSC and distance metrics in pancreatic tumor segmentation.

6.4 Discussion

Contouring pancreatic tumors on contrast-enhanced CT images is a challenging task, even for experienced radiologists. The hypodense areas on CT images often fail to capture all of the diseased areas, as experts tend to include areas that cannot be identified by imaging features alone when delineating tumors. This process introduces inter-observer variability since the imaging features at the tumor borders are faint[82]. Thus, this segmentation task requires high sensitivity and low specificity, as well as options for customization to accommodate clinician preferences. In deep learning-based auto-segmentation, uncertain ground truths can pose significant challenges for models to learn and validate due to the inherent variability in the training data[83]. Moreover, the inconsistent anatomical context provided by the training contours can hinder model convergence, further complicating the task.

In our study, we proposed an approach to address segmentation tasks with uncertain ground truths by utilizing ensemble-based uncertainty estimation techniques. Deep ensembles have demonstrated remarkable performance in uncertainty estimation tasks, and greater variability within the ensemble has been observed to improve the calibration of the pixelwise

probability map[80]. To introduce human-like variability and incorporate multiple segmentation styles into the consensus probability estimation, we employed the Tversky loss function to fine tune the contouring style of each individual model[81]. In addition to using different data folds, we tuned the Tversky hyperparameters to generate models with varying segmentation tendencies. This enabled the creation of multiple segmentations from a well-calibrated probability map that can be adjusted to the physician's preferences as shown in Figure 18. Our approach yielded superior quantitative results compared to the Swin-UNETR ensemble, which was trained and tested on the same dataset with identical cross-validation data folds. Both the Tversky ensemble and the Swin-UNETR ensemble were trained using the preprocessing and hyperparameters reported by the ensemble that achieved state-of-the-art performance in the pancreas task of the Medical Segmentation Decathlon.

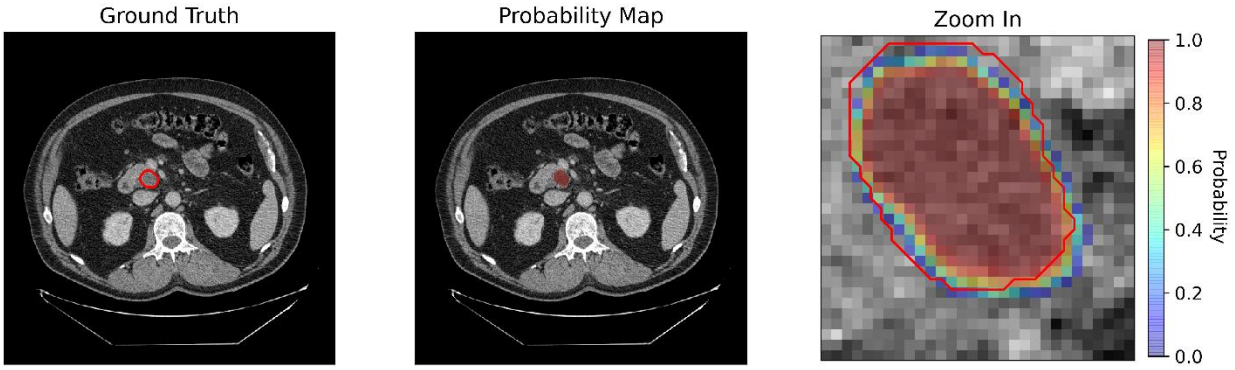


Figure 18: A sample probability map generated by Tversky ensemble. Final segmentations were derived from thresholding the probability map. On this CT slice, the probability map perfectly reflected the tumor volume while providing pixelwise uncertainty estimation.

The Dice similarity coefficient (DSC) results showed consistent improvement with lower probability thresholds as shown in figure 17. Upon qualitative observation, we found that the model consistently under-segmented the tumor compared to the ground truth. While the

generated segmentations captured the hypodense regions in the CT images, they failed to extrapolate to the surrounding diseased areas that were less prominent to the human eye. By lowering the probability threshold, the generated contours became more aggressive in delineating the uncertain regions at the tumor border. This resulted in a greater overlap with the ground truths labeled by experts, as depicted in figure 19, leading to improved quantitative performance. Selecting the contours based on the lowest HD95 distance further improved the distance metrics without compromising the DSC. By optimizing the thresholding strategy on a patient-by-patient basis, we retained aggressive segmentations that incorporated uncertain regions while eliminating erroneous regions with low confidence. In the clinical workflow, physicians could threshold the probability map in real-time to accommodate their preferences. Our approach offered a promising option for generating an accurate tumor segmentation on the contrast-enhanced CT in a timely manner.

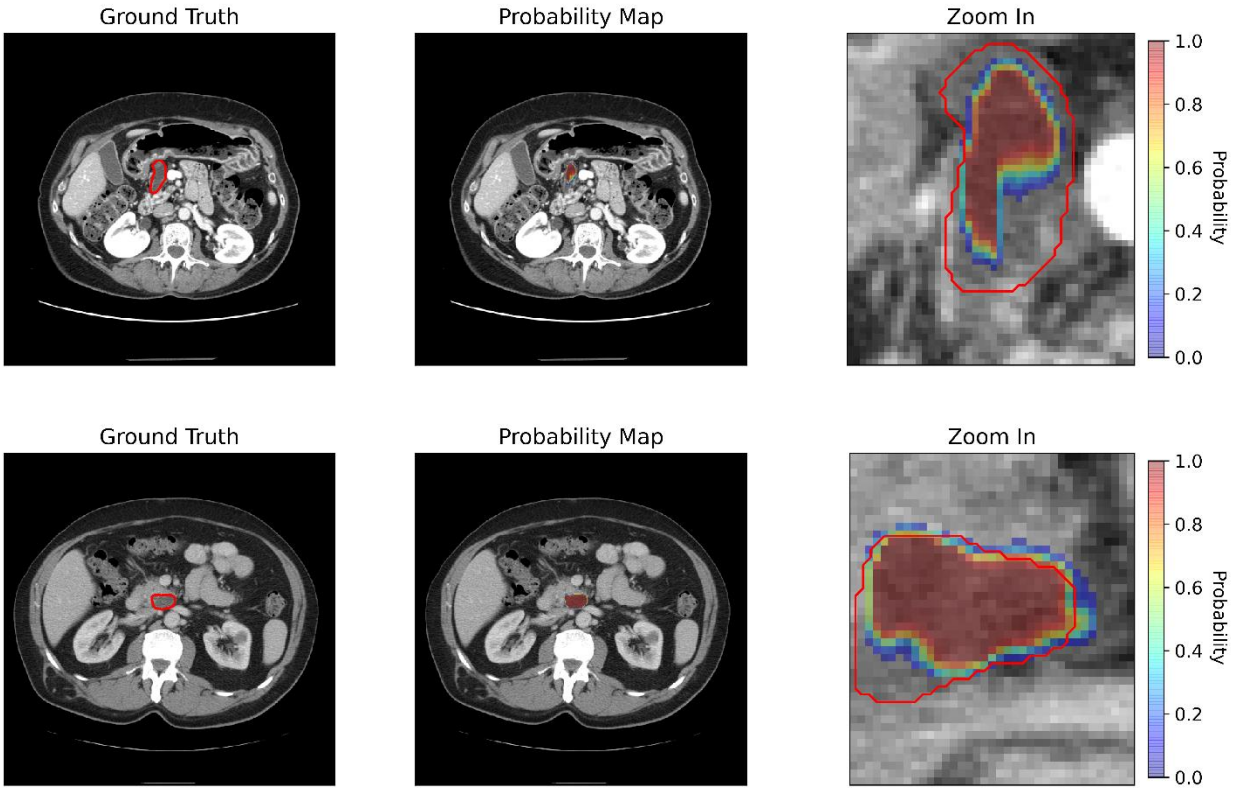


Figure 19: Deep learning models often suffered from under-segmentation of pancreatic tumors at the tumor border. Our Tversky ensemble allowed for the application of a more lenient thresholding, leading to better quantitative results.

While over-segmentation was preferred in pancreatic tumor segmentation due to the inherent uncertainty at tumor borders, incorporating low probability regions was not without its drawbacks. When lenient thresholding was applied, the ensemble could falsely identify tumors from benign anatomy, as illustrated in figure 20. This occurrence was common in auto-segmentation since pancreatic tumors often displayed low contrast compared to the surrounding tissue. False positives were frequently observed due to the presence of hypodense regions throughout the CT scans. In our post-processing step, we retained the largest connected component of the predicted contours, which could result in falsely labeled low probability regions becoming the larger connected component and leading to poor quantitative results. This perturbation to the distance metrics occurred when increasing the threshold from 0.4 to 0.5. However, the calibrated probability map offered an opportunity to detect some mis-contoured cases based on uncertainty estimates. Clinicians could visually identify regions with low confidence. If the initial probability map was found to be erroneous, they had the ability to eliminate falsely identified tumor regions by increasing the probability threshold, as demonstrated in figure 20. This feature allowed the model to maintain an aggressive approach in most cases to ensure optimal results, while producing accurate contours after human intervention when the Tversky ensemble was uncertain.

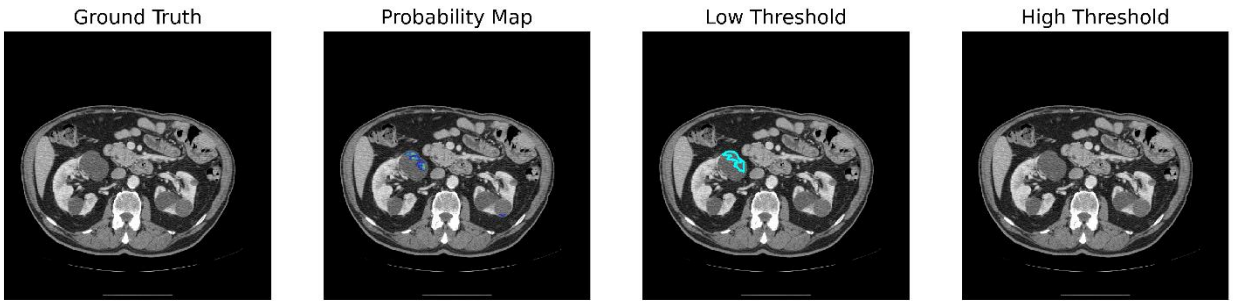


Figure 20: False positive regions eliminated via probability thresholding based on uncertainty context provided by Tversky ensemble.

Based on our observation that over-segmentation tendencies yielded contours that were closer to the ground truth, we aimed to investigate the feasibility of creating an over-segmenting ensemble through manipulation of the Tversky hyperparameters. We selected a Tversky α value of 1.0 to construct an ensemble that maximally rewarded over-segmentation. To ensure fairness in our comparative analysis, we again utilized identical data split, preprocessing techniques, and hyperparameters outlined by the state-of-the-art Swin-UNETR ensemble. We trained the over-segmenting ensemble using 5-fold cross-validation, specifically emphasizing over-segmentation characteristics with the largest α value possible. The resulting over-segmenting ensemble only achieved an average DSC of 0.40. This performance deterioration underscored the significance of the diversity introduced by the varying Tversky hyperparameters. It substantiated the crucial role of a well-calibrated probability map in achieving accurate segmentation. Directly tuning the model towards the desired behavior did not yield improvements in segmentation quality. By incorporating diverse segmentation tendencies within the Tversky ensemble, we successfully generated a probability map that was better calibrated. Consequently, this advancement facilitated more precise segmentation and offered opportunities for stylistic customization in our results.

Our proposed approach still required human intervention for the final contouring. This was due to the existing challenges in segmenting pancreatic tumors with deep learning approaches. The state-of-the-art approach achieved an average DSC of 0.43 in our test set, indicating that expert input remained necessary for achieving optimal plan quality in segmentation workflows. Despite outperforming the state-of-the-art model ensemble using identical preprocessing and hyperparameters, our tool was not yet capable of fully automating the segmentation of pancreatic tumors. Additionally, our post-processing pipeline, which retained the largest component, might introduce unintended variabilities when using low probability thresholds for aggressive contouring at the tumor border. Therefore, caution is advised when conducting thresholding to avoid compromising the accuracy and quality of the final segmentation results.

6.5 Conclusion

In this study, we employed an ensemble-based uncertainty estimation technique to facilitate the segmentation of pancreatic tumors. Given the inherent ambiguity of ground truth delineation, we adapted the Tversky loss function to account for a variety of contouring styles and generate a consensus probability map that can be fine-tuned by clinicians in line with their preferences, following model inference. By utilizing the same network architecture, data preprocessing pipeline, hyperparameters, and ensembling strategy as the state-of-the-art model, our approach outperformed its Swin-UNETR counterpart in the pancreatic tumor segmentation task of the Medical Segmentation Decathlon. Furthermore, our method provides pixel-wise uncertainty estimation, which enables clinicians to generate contours with greater confidence. We are optimistic that our Tversky ensembles can serve as an accurate and dependable solution for pancreatic tumor segmentation.

Chapter 7: Discussion

In this project, we aimed to develop automated segmentation tools for pancreatic and liver cancer radiation treatment. Manual segmentation, which is time-consuming, subjective, and susceptible to variability among observers, has been the conventional protocol prior to radiation treatment planning. However, recent advancements in deep learning-based auto-segmentation have demonstrated exceptional results across various medical segmentation challenges. By integrating deep learning into the treatment planning workflow, we can streamline and standardize clinical practices.

Throughout the development of various initial segmentation models, we observed that data quality played a pivotal role in determining the clinical performance of deep learning-based auto-segmentation tools. Consequently, our focus shifted towards curating a unique dataset specifically tailored to our contouring practice. We meticulously contoured the organs-at-risk on contrast-enhanced CT scans from a cohort of 70 patients. The compilation of this dataset served a dual purpose: it served as both the training set for our segmentation models and as a benchmark for evaluating their performance. Through U-Net-based segmentation architectures, we obtained exceptional quantitative and qualitative results from this carefully curated dataset. These outcomes have substantiated our hypothesis that, with a high-quality training set, a total of 40 patients would suffice to construct a robust and reliable auto-segmentation tool. This finding underscores the importance of data quality in facilitating the development of accurate and effective segmentation models. We decided to make these expertly reviewed data publicly available, with the aim of facilitating the development of new auto-segmentation models for clinics worldwide. To the best of our knowledge, this is the first highly-curated dataset

encompassing both small and large bowel segmentation, as it necessitated expert input for accurate delineation.

Since its implementation in 2021, our nnU-Net-based organs-at-risk segmentation tool has been widely utilized in our clinic, providing contour generation for more than 800 patients. The exceptional performance of the nnU-Net played a significant role in the success of this auto-segmentation tool. However, another crucial factor contributing to its success was the extensive collaboration between our team and the clinical department. The training dataset was recontoured under the guidance of our department's experienced physicians, resulting in generated contours that closely aligned with their stylistic preferences. Moreover, the clinical validation of our tool involved the active participation of experts from our institution, further ensuring its accuracy and reliability. This comprehensive collaboration and the involvement of domain experts have been instrumental in the seamless deployment of our organs-at-risk segmentation tool.

Our comprehensive qualitative evaluation, encompassing both contrast-enhanced and non-contrast CT images from a cohort of 75 patients, provided us with valuable insights into the clinical adoption of deep learning auto-segmentation techniques. While the majority of our generated contours demonstrated clinical acceptability, it is noteworthy that certain contours required major edits. These contours would need to be flagged prior to entering the treatment planning phase. Moreover, physicians identified a notable portion of contours that exhibited stylistic deviations from their preferences, indicating the need for further integration of diverse contouring styles into auto-segmentation models. Such integration is crucial to enhance clinician satisfaction and optimize the overall performance of the segmentation tool.

In order to address the patients requiring major edits, we implemented a quality assurance tool designed to identify out-of-distribution patients that could potentially degrade the performance of the segmentation process. This tool proved particularly effective due to the fact that our training dataset was sourced internally from our institution. The imaging protocol, including contrast timing and motion management, employed in the training CT images remained relatively consistent with the CT images used in the deployment of the model. Consequently, the detection of out-of-distribution samples was highly successful in distinguishing patients likely to exhibit poor segmentation performance. This approach contributed to the overall robustness and reliability of our segmentation models, enhancing their clinical utility.

Other contour quality assurance approaches were also considered. Our lab had experience in using a secondary contouring system to detect erroneous contours[56]. We considered using other deep learning architectures to serve as the secondary contouring system. However, the performance of our secondary contouring system was inferior to the nnU-Net as shown in table 9, 10 and 11. Using an inferior model as the secondary contouring system to capture failure contours could lead to a large number of false positives since contour differences were most likely caused by the failure of the secondary system. This posed an issue for structures prone to contouring errors such as the bowel structures. The transformer-based segmentation methods significantly underperformed at these structures. However, the development of the secondary contouring system for OARs led to the utilization of transformers in the target segmentation component of the project.

Organs	DSC		HD95 (mm)		MSD(mm)	
	Mean	SD	Mean	SD	Mean	SD
Duodenum	0.80	0.08	12.34	9.09	1.68	1.04
Small Bowel	0.89	0.05	7.77	8.90	1.99	2.10
Large Bowel	0.90	0.06	7.15	8.42	1.27	0.87
Stomach	0.92	0.03	4.77	2.98	1.23	0.78
Liver	0.96	0.01	3.56	1.71	1.07	0.49
Spleen	0.97	0.01	2.21	1.27	0.56	0.23
Kidney_R	0.96	0.01	2.51	1.29	0.59	0.18
Kidney_L	0.96	0.01	2.52	0.90	0.61	0.19
SpinalCord	0.76	0.15	42.52	38.62	10.57	10.49

Table 9. Mean Dice similarity coefficient (DSC), 95% Hausdorff distance (HD95), and mean surface distance (MSD) between ground truth and prediction results from nnU-Net on contrast-enhanced CT images

Organs	DSC		HD95 (mm)		MSD(mm)	
	Mean	SD	Mean	SD	Mean	SD
Duodenum	0.76	0.11	13.16	9.47	2.24	2.07
Small Bowel	0.85	0.07	10.42	8.64	2.62	1.87
Large Bowel	0.87	0.06	11.34	9.83	2.00	1.47
Stomach	0.90	0.04	6.45	4.46	1.62	1.06
Liver	0.96	0.01	7.29	19.09	1.86	4.19
Spleen	0.96	0.02	2.82	2.52	0.96	1.13
Kidney_R	0.96	0.01	2.69	1.64	0.63	0.24
Kidney_L	0.95	0.02	2.95	1.76	0.75	0.41

Table 10. Mean Dice similarity coefficient (DSC), 95% Hausdorff distance (HD95), and mean surface distance (MSD) between ground truth and prediction results from Swin-UNETR on contrast-enhanced CT images

Organs	DSC		HD95		MSD	
	Mean	Std	Mean	Std	Mean	Std
Duodenum	0.73	0.10	12.91	7.01	2.45	1.58
Small Bowel	0.82	0.08	14.68	8.58	3.62	2.17
Large Bowel	0.83	0.08	17.37	13.52	3.98	3.06
Stomach	0.86	0.08	9.61	8.87	2.42	2.47
Liver	0.96	0.03	7.62	20.32	2.30	6.02
Spleen	0.96	0.02	3.85	5.06	0.93	1.19
Kidney_R	0.96	0.02	2.86	1.81	0.72	0.47
Kidney_L	0.95	0.02	4.25	6.74	0.96	1.32
SpinalCord	0.75	0.14	42.82	37.15	10.50	9.81

Table 11. Mean Dice similarity coefficient (DSC) between ground truth and prediction results from UNETR with different configurations on contrast-enhanced CT images

In order to address diverse stylistic preferences in target volume contouring, our objective was to incorporate a broader spectrum of contouring styles into the process. It was observed that stylistic disagreements were more prevalent in the definition of target volumes compared to the contouring of organs-at-risk[82,84].84 The adoption of automated target volume generation using auto-segmentation techniques in clinical practice has been hindered by variations in stylistic preferences among clinicians, resulting in a lack of trust in the accuracy and reliability of the automated system. In order to mitigate these concerns, we employed the Tversky loss function to regulate the segmentation tendencies of deep learning models. By adjusting the hyperparameters of the Tversky loss layer, we could control the level of under-segmentation or over-segmentation of the tumor, thus providing flexibility in contouring preferences. Building upon the Swin-UNETR model ensemble, we created multiple models with varying Tversky hyperparameters. This ensemble of models enabled the incorporation of diverse stylistic preferences, resulting in a

calibrated probability map through consensus. By applying a threshold to the probability map, clinicians were able to identify contours that best aligned with their preferred contouring styles.

As we transitioned to target segmentation tasks, specifically the delineation of gross tumor volumes (GTVs), we observed that transformer-based segmentation networks had attained remarkable performance in various tumor segmentation challenges[4984]. Encouraged by these advancements, we opted to also employ the Swin-UNETR architecture for liver tumor segmentation. Leveraging our unique simulation protocols and motion management techniques, we acquired a substantial number of unlabeled abdominal images, which served as ideal candidates for self-supervised learning approaches facilitated by transformer-based architectures. During the development of liver GTV segmentation, we noted that while self-supervised learning had achieved state-of-the-art performance in diverse segmentation tasks, its successful deployment required intervention in both its training and pretraining. The choice of data used for pretraining had a significant impact on the final results, as it could potentially degrade the performance compared to training from scratch. Moreover, careful hyperparameter tuning of both the pretraining and training stages played a critical role in optimizing the network's performance. Our findings underscored the importance of thoughtful data selection for pretraining, as well as the careful tuning of hyperparameters during training, in order to maximize the performance and applicability of transformer-based networks in the field of medical image segmentation.

Throughout this comprehensive segmentation study, we underscored the essential contribution of multidisciplinary expert inputs in the development and deployment of auto-segmentation tools. The construction of a suitable dataset for deep learning development necessitated the integration of expert knowledge derived from medical training. Furthermore, the

effective handling of deep learning algorithms demanded substantial expertise in hyperparameter tuning. In the context of supervised learning, our findings consistently demonstrated that the nnU-Net showcased exceptional performance as an out-of-the-box solution. It provided medical professionals with a reliable framework to leverage their domain knowledge while yielding reliable results. Therefore, for clinical deployment, the nnU-Net remains a dependable choice to extract robust performance from diverse datasets.

As the complexity of the models increased, expertise in deep learning development became crucial for achieving optimal performance through exhaustive model tuning. Particularly when encoder pretraining was employed in transformer-based architectures, evaluating the training quality became a convoluted process. The recorded validation loss on the validation dataset provided a preliminary indication of the training status. However, unlike the validation loss of a segmentation network, these losses did not necessarily correlate with optimal performance in the downstream segmentation task. Consequently, selecting the most suitable pretrained encoder for downstream segmentation necessitated a trial-and-error approach as shown in our study. We also observed that the results of self-supervised pretraining were highly sensitive to hyperparameter settings. Careful selection of batch size, learning rate, and learning rate decay played a pivotal role in ensuring stable training. Hyperparameter settings beyond the acceptable range, such as larger batch size or higher learning rate, often resulted in overfitting and yielded suboptimal pretrained encoders and led to poor performance of the downstream segmentation model. Tuning the downstream segmentation model also involved a significant amount of trial and error. Transformer-based architectures encountered challenges during pancreas and liver tumor segmentation, as they suffered from gradient collapsing, leading to premature termination of the training process. Additionally, the memory consumption of

transformer-based architectures fluctuated depending on the combination of software being utilized, which posed limitations due to GPU memory size. Overcoming these issues required expert intervention and implementation in clinical settings could prove challenging. While transformer-based architectures provided an opportunity to leverage unlabeled data, their development necessitated substantial expertise in deep learning. The intricate process of model tuning, addressing gradient collapsing, and memory management required the input of experienced deep learning developers.

A unique aspect of our segmentation project was the topic of quality assurance. Quality assurance plays a vital role in clinical systems. In the context of auto-segmentation, we would like to prevent erroneous contours from entering treatment planning or treatment delivery. In this regard, careful consideration was given to the model's ability to forecast its confidence or lack of confidence on unseen samples, a task essential for reliable performance. Our proposed QA approach for OARs focused on using out-of-distribution[50] detection as a means of quality assurance, assuming that substantial deviations from encountered samples would lead to performance degradation[73]. However, this assumption presented challenges in clinical practice. Firstly, as clinical practices evolve, the data distribution encountered by the model may shift away from the training set[85], resulting in a high number of false positives and potential disregard of alarms by clinicians. Secondly, deep learning models have demonstrated remarkable tolerance to domain shifts, enabling our segmentation model to generate clinically acceptable results even in the presence of challenging cases involving ascites or out-of-field artifacts. Consequently, distribution-based quality assurance systems may trigger false alarms in such scenarios. We also investigated the application of uncertainty-based quality assurance systems[78,80]. The main advantage of such systems lies in their ability to provide pixel-wise uncertainty estimates. With

our Tversky ensemble, we were able to generate expressive probability map with a relatively small ensemble. However, we observed that these approaches typically required substantial calibration efforts to achieve reliable uncertainty estimation. Furthermore, the resulting uncertainty maps often exhibited a sharp probability gradient, resulting in more binary outcomes and reduced interpretability. Apart from ensemble-based techniques, these approaches often generate a large amount of data to perform uncertainty estimation. This poses challenge for clinical implementation. As the adoption of deep learning-based automation increases in clinical settings, the integration of quality assurance systems becomes crucial and warrants further investigation.

For future work, we believe that the field of medical image segmentation is entering a transformative phase. The availability of an unprecedented number of public datasets makes it considerably easier to develop models using existing out-of-the-box solutions such as nnUNet. Tools like TotalSegmenter[86] have demonstrated remarkable capability in accurately segmenting nearly all normal anatomical structures within the human body. By using a combination of public dataset and a small amount of private dataset, we would be able segment organs-at-risk and tumors with greater confidence. As these models continue to improve in performance, we anticipate increased adoption in clinical settings and a corresponding rise in throughput of auto-segmentation tools. Thus, it is crucial for us to remain vigilant against automation bias. Ensuring the quality assurance of the latest models remains an important and ongoing challenge for researchers worldwide. As automation progresses, clinicians would expect auto-segmentation suites to exhibit enhanced robustness. As developers, it is our responsibility to prioritize patient safety when designing these automation tools.

Chapter 8: Conclusion

In conclusion, our auto-segmentation system for organs-at-risk has demonstrated remarkable success in the context of upper-abdominal radiation treatment. Its high clinical acceptance rates indicate its reliability and accuracy in delineating critical structures. Furthermore, the accompanying QA tool has proven to be effective in identifying and capturing contours that require significant edits, enhancing the overall quality of the segmentation process.

By leveraging a diverse range of unlabeled data in our self-supervised learning approach, we have significantly improved the performance of our transformer-based segmentation system. This highlights the importance of incorporating a wide variety of data sources during the pretraining stage, allowing our model to learn robust and comprehensive representations of the target organs. Moreover, our uncertainty-guided segmentation network has provided valuable capabilities in terms of customization and identification of low-confidence regions. This feature allows clinicians to have better control over the segmentation process and make informed decisions based on the level of certainty in specific areas, ultimately leading to more precise and tailored treatment plans.

Overall, our suite of auto-segmentation tools for pancreatic and liver cancer radiation treatment holds great promise for streamlining clinical workflows and ensuring patient safety. With their impressive performance, these tools have the potential to revolutionize the field by providing efficient and accurate segmentation results, enabling clinicians to deliver targeted and personalized treatments while optimizing outcomes for cancer patients.

References

1. Arnold M, Abnet CC, Neale RE, Vignat J, Giovannucci EL, McGlynn KA, Bray F. Global Burden of 5 Major Types of Gastrointestinal Cancer. *Gastroenterology*. 2020;159(1):335-349.e15. doi:10.1053/j.gastro.2020.02.068

2. Huang J, Lok V, Ngai CH, Zhang L, Yuan J, Lao XQ, Ng K, Chong C, Zheng ZJ, Wong MCS. Worldwide Burden of, Risk Factors for, and Trends in Pancreatic Cancer. *Gastroenterology*. 2021;160(3):744-754. doi:10.1053/j.gastro.2020.10.007

3. Dasgupta P, Henshaw C, Youlden DR, Clark PJ, Aitken JF, Baade PD. Global Trends in Incidence Rates of Primary Adult Liver Cancers: A Systematic Review and Meta-Analysis. *Front Oncol*. 2020;10(February):1-17. doi:10.3389/fonc.2020.00171

4. Apisarnthanarax S, Barry A, Cao M, Czito B, DeMatteo R, Drinane M, Hallemeier CL, Koay EJ, Lasley F, Meyer J, Owen D, Pursley J, Schaub SK, Smith G, Venepalli NK, Zibari G, Cardenes H. External Beam Radiation Therapy for Primary Liver Cancers: An ASTRO Clinical Practice Guideline. *Pract Radiat Oncol*. 2022;12(1):28-51. doi:10.1016/j.prro.2021.09.004

5. Palta M, Godfrey D, Goodman KA, Hoffe S, Dawson LA, Dessert D, Hall WA, Herman JM, Khorana AA, Merchant N, Parekh A, Patton C, Pepek JM, Salama JK, Tuli R, Koong AC. Radiation Therapy for Pancreatic Cancer: Executive Summary of an ASTRO Clinical Practice Guideline. *Pract Radiat Oncol*. 2019;9(5):322-332. doi:10.1016/j.prro.2019.06.016

6. Ohri N, Tomé WA, Méndez Romero A, Miften M, Ten Haken RK, Dawson LA, Grimm J, Yorke E, Jackson A. Local Control After Stereotactic Body Radiation Therapy for Liver Tumors. *Int J Radiat Oncol Biol Phys*. 2021;110(1):188-195. doi:10.1016/j.ijrobp.2017.12.288

7. Bruynzeel AME, Lagerwaard FJ. The role of biological dose-escalation for pancreatic cancer. *Clin Transl Radiat Oncol*. 2019;18:128-130. doi:10.1016/j.ctro.2019.04.020

8. Goldsmith C, Price P, Cross T, Loughlin S, Cowley I, Plowman N. Dose-Volume Histogram Analysis of Stereotactic Body Radiotherapy Treatment of Pancreatic Cancer: A Focus on Duodenal Dose Constraints. *Semin Radiat Oncol*. 2016;26(2):149-156. doi:10.1016/j.semradonc.2015.12.002

9. Kim H, Jung J, Kim J, Cho B, Kwak J, Jang JY, Lee S wook, Lee JG, Yoon SM. Abdominal multi-organ auto-segmentation using 3D-patch-based deep convolutional neural network. *Sci Rep*. 2020;10(1):1-9. doi:10.1038/s41598-020-63285-0

10. Rhee DJ, Beddar S, Jaoude JA, Sawakuchi G, Martin R, Perles L, Yu C, He Y, Court LE, Ludmir EB, Koong AC, Das P, Koay EJ, Taniguichi C, Niedzielski JS. Dose Escalation for Pancreas SBRT: Potential and Limitations of using Daily Online Adaptive Radiation Therapy and an Iterative Isotoxicity Automated Planning Approach. *Adv Radiat Oncol*. 2023;8(4). doi:10.1016/j.adro.2022.101164

11. Fung NTC, Hung WM, Sze CK, Lee MCH, Ng WT. Automatic segmentation for adaptive planning in nasopharyngeal carcinoma IMRT: Time, geometrical, and dosimetric analysis. *Medical Dosimetry*. 2020;45(1):60-65. doi:10.1016/j.meddos.2019.06.002

12. Cardenas CE, Yang J, Anderson BM, Court LE, Brock KB. Advances in Auto-Segmentation. *Semin Radiat Oncol*. 2019;29(3):185-197. doi:10.1016/j.semradonc.2019.02.001

13. Cardenas CE, Beadle BM, Garden AS, Skinner HD, Yang J, Rhee DJ, McCarroll RE, Netherton TJ, Gay SS, Zhang L, Court LE. Generating High-Quality Lymph Node Clinical Target Volumes for Head and Neck Cancer Radiation Therapy Using a Fully Automated Deep Learning-Based Approach. *Int J Radiat Oncol Biol Phys*. 2021;109(3):801-812. doi:10.1016/j.ijrobp.2020.10.005

14. Netherton TJ, Rhee DJ, Cardenas CE, Chung C, Klopp AH, Peterson CB, Howell RM, Balter PA, Court LE. Evaluation of a multiview architecture for automatic vertebral labeling of palliative radiotherapy simulation CT images. *Med Phys*. 2020;47(11):5592-5608. doi:10.1002/mp.14415

15. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, Vedam S, Kry S, Brock KK, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N, Trauernicht C, Simonds H, Court LE. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*. 2020;47(11):5648-5658. doi:10.1002/mp.14467

16. Yu C, Anakwenze CP, Zhao Y, Martin RM, Ludmir EB, S.Niedzielski J, Qureshi A, Das P, Holliday EB, Raldow AC, Nguyen CM, Mumme RP, Netherton TJ, Rhee DJ, Gay SS, Yang J, Court LE, Cardenas CE. Multi-organ segmentation of abdominal structures from non-contrast and contrast enhanced CT images. *Sci Rep*. 2022;12(1). doi:10.1038/s41598-022-21206-3

17. Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Published online May 18, 2015. http://arxiv.org/abs/1505.04597

18. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houlsby N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. Published online October 22, 2020. http://arxiv.org/abs/2010.11929

19. Khalaf N, El-Serag HB, Abrams HR, Thrift AP. Burden of Pancreatic Cancer: From Epidemiology to Practice. *Clinical Gastroenterology and Hepatology*. 2021;19(5):876-884. doi:10.1016/j.cgh.2020.02.054

20. Moningi S, Dholakia AS, Raman SP, Blackford A, Cameron JL, Le DT, De Jesus-Acosta AMC, Hacker-Prietz A, Rosati LM, Assadi RK, Dipasquale S, Pawlik TM, Zheng L, Weiss MJ, Laheru DA, Wolfgang CL, Herman JM. The Role of Stereotactic Body Radiation Therapy for Pancreatic Cancer: A Single-Institution Experience. *Ann Surg Oncol*. 2015;22(7):2352-2358. doi:10.1245/s10434-014-4274-5

21. Brunner TB, Haustermans K, Huguet F, Morganti AG, Mukherjee S, Belka C, Krempien R, Hawkins MA, Valentini V, Roeder F. ESTRO ACROP guidelines for target volume definition in pancreatic cancer. *Radiotherapy and Oncology*. 2021;154:60-69. doi:10.1016/j.radonc.2020.07.052

22. Ahn SH, Yeo AU, Kim KH, Kim C, Goh Y, Cho S, Lee SB, Lim YK, Kim H, Shin D, Kim T, Kim TH, Youn SH, Oh ES, Jeong JH. Comparative clinical evaluation of atlas and deep-

learning-based auto-segmentation of organ structures in liver cancer. *Radiation Oncology*. 2019;14(1):1-13. doi:10.1186/s13014-019-1392-z

23. Jabbour SK, Hashem SA, Bosch W, Kim TK, Finkelstein SE, Anderson BM, Ben-Josef E, Crane CH, Goodman KA, Haddock MG, Herman JM, Hong TS, Kachnic LA, Mamon HJ, Pantarotto JR, Dawson LA. Upper abdominal normal organ contouring guidelines and atlas: A Radiation Therapy Oncology Group consensus. *Pract Radiat Oncol*. 2014;4(2):82-89. doi:10.1016/j.prro.2013.06.004

24. Lukovic J, Henke L, Gani C, Kim TK, Stanescu T, Hosni A, Lindsay P, Erickson B, Khor R, Eccles C, Boon C, Donker M, Jagavkar R, Nowee ME, Hall WA, Parikh P, Dawson LA. MRI-Based Upper Abdominal Organs-at-Risk Atlas for Radiation Oncology. *Int J Radiat Oncol Biol Phys*. 2020;106(4):743-753. doi:10.1016/j.ijrobp.2019.12.003

25. Reyngold M, Parikh P, Crane CH. Ablative radiation therapy for locally advanced pancreatic cancer: Techniques and results. *Radiation Oncology*. 2019;14(1):1-8. doi:10.1186/s13014-019-1309-x

26. Wang Y, Zhou Y, Shen W, Park S, Fishman EK, Yuille AL. Abdominal multi-organ segmentation with organ-attention networks and statistical fusion. *Med Image Anal*. 2019;55:88-102. doi:10.1016/j.media.2019.04.005

27. Murphy JD, Christman-Skieller C, Kim J, Dieterich S, Chang DT, Koong AC. A dosimetric model of duodenal toxicity after stereotactic body radiotherapy for pancreatic cancer. *Int J Radiat Oncol Biol Phys*. 2010;78(5):1420-1426. doi:10.1016/j.ijrobp.2009.09.075

28. Netherton TJ, Rhee DJ, Cardenas CE, Chung C, Klopp AH, Peterson CB, Howell RM, Balter PA, Court LE. Evaluation of a multiview architecture for automatic vertebral labeling of palliative radiotherapy simulation CT images. *Med Phys*. 2020;47(11):5592-5608. doi:10.1002/mp.14415

29. Rhee DJ, Jhingran A, Rigaud B, Netherton T, Cardenas CE, Zhang L, Vedam S, Kry S, Brock KK, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N, Trauernicht C, Simonds H, Court LE. Automatic contouring system for cervical cancer using convolutional neural networks. *Med Phys*. 2020;47(11):5648-5658. doi:10.1002/mp.14467

30. Gay SS, Yu C, Rhee DJ, Sjogreen C, Mumme RP, Nguyen CM, Netherton TJ, Cardenas CE, Court LE. A Bi-directional, Multi-modality Framework for Segmentation of Brain Structures. In: Shusharina N, Heinrich MP, Huang R, eds. *Segmentation, Classification, and Registration of Multi-Modality Medical Imaging Data*. Springer International Publishing; 2021:49-57.

31. Thor M, Apte A, Haq R, Iyer A, LoCastro E, Deasy JO. Using Auto-Segmentation to Reduce Contouring and Dose Inconsistency in Clinical Trials: The Simulated Impact on RTOG 0617. *Int J Radiat Oncol Biol Phys*. 2021;109(5):1619-1626. doi:10.1016/j.ijrobp.2020.11.011

32. Heller N, Isensee F, Maier-Hein KH, Hou X, Xie CM, Li FY, Nan Y, Mu G, Lin Z, Han M, Yao G, Gao Y, Zhang Y, Wang Y, Hou F, Yang J, Xiong G, Tian J, Zhong C, Ma J, Rickman JM, Dean JF, Stai B, Tejpaul R, Oestreich M, Blake PA, Kaluzniak H, Raza S, Rosenberg J, Moore K, Walczak E, Rengel Z, Edgerton Z, Vasdev R, Peterson MS, McSweeney S, Peterson SJ, Kalapara A, Sathianathen NJ, Papanikolopoulos N, Weight CJ. The state of

the art in kidney and kidney tumor segmentation in contrast-enhanced CT imaging: Results of the KiTS19 challenge. Medical Image Analysis. 2021;67:101821. doi:10.1016/j.media.2020.101821

33. Kavur AE, Gezer NS, Baris M, Aslan S, Conze PH, Groza V, Pham DT, Chatterjee S, Ernst P, Ozkan S, Baydar B, Lachinov D, Han S, Pauli J, Isensee F, Perkonigg M, Sathish R, Rajan R, Sheet D, Dovletov G, Speck O, Nürnberger A, Maier-Hein KH, Akar GB, Unal G, Dicle O, Selver MA. CHAOS Challenge - combined (CT-MR) healthy abdominal organ segmentation. Medical Image Analysis. 2021;69:101950. doi:10.1016/j.media.2020.101950

34. Gibson E, Giganti F, Hu Y, Bonmati E, Bandula S, Gurusamy K, Davidson B, Pereira SP, Clarkson MJ, Barratt DC. Automatic Multi-Organ Segmentation on Abdominal CT with Dense V-Networks. *IEEE Trans Med Imaging*. 2018;37(8):1822-1834. doi:10.1109/TMI.2018.2806309

35. Liu Y, Lei Y, Fu Y, Wang T, Tang X, Jiang X, Curran WJ, Liu T, Patel P, Yang X. CT-based multi-organ segmentation using a 3D self-attention U-net network for pancreatic radiotherapy. *Med Phys*. 2020;47(9):4316-4324. doi:10.1002/mp.14386

36. Tajbakhsh N, Jeyaseelan L, Li Q, Chiang JN, Wu Z, Ding X. Embracing imperfect datasets: A review of deep learning solutions for medical image segmentation. *Med Image Anal*. 2020;63:101693. doi:10.1016/j.media.2020.101693

37. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z

38. Isensee F, Jaeger PF, Kohl SAA, Petersen J, Maier-Hein KH. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. *Nat Methods*. 2021;18(2):203-211. doi:10.1038/s41592-020-01008-z

39. Court LE, Kisling K, McCarroll RE, Zhang L, Yang J, Simonds H, Du Toit M, Trauernicht C, Burger H, Parkes J, Mejia M, Bojador MR, Balter PA, Branco D, Steinmann A, Baltz GC, Anderson BDO, Ibbott GS, Jhingran A, Shaitelman SF, Bogler O, Schmeller K, Followill DS, Howell RM, Nelson CP, Peterson CB, Beadle BM. Radiation Planning Assistant - A Streamlined, Fully Automated Radiotherapy Treatment Planning System. Journal of Visualized Experiments. 2018;(134). doi:10.3791/57411

40. Siddique N, Paheding S, Elkin CP, Devabhaktuni V. U-net and its variants for medical image segmentation: A review of theory and applications. *IEEE Access*. Published online 2021:82031-82057. doi:10.1109/ACCESS.2021.3086020

41. Lugo-Fagundo C, Vogelstein B, Yuille A, Fishman EK. Deep Learning in Radiology: Now the Real Work Begins. *Journal of the American College of Radiology*. 2018;15(2):364-367. doi:10.1016/j.jacr.2017.08.007

42. Zhao A, Balakrishnan G, Durand F, Guttag J V., Dalca A V. Data augmentation using learned transformations for one-shot medical image segmentation. *Proceedings of the*

IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2019;2019-June:8535-8545. doi:10.1109/CVPR.2019.00874

43. Zhao Y, Rhee DJ, Cardenas C, Court LE, Yang J. Training deep-learning segmentation models from severely limited data. *Med Phys*. 2021;48(4):1697-1706. doi:10.1002/mp.14728

44. Sandfort V, Yan K, Pickhardt PJ, Summers RM. Data augmentation using generative adversarial networks (CycleGAN) to improve generalizability in CT segmentation tasks. *Sci Rep*. 2019;9(1):1-9. doi:10.1038/s41598-019-52737-x

45. Zhou Y, Wang Y, Tang P, Bai S, Shen W, Fishman EK, Yuille A. Semi-supervised 3D abdominal multi-organ segmentation via deep multi-planar co-training. *Proceedings - 2019 IEEE Winter Conference on Applications of Computer Vision, WACV 2019*. Published online 2019:121-140. doi:10.1109/WACV.2019.00020

46. Huang K, Rhee DJ, Ger R, Layman R, Yang J, Cardenas CE, Court LE. Impact of slice thickness, pixel size, and CT dose on the performance of automatic contouring algorithms. *J Appl Clin Med Phys*. 2021;22(5):168-174. doi:10.1002/acm2.13207

47. Palta M, Godfrey D, Goodman KA, Hoffe S, Dawson LA, Dessert D, Hall WA, Herman JM, Khorana AA, Merchant N, Parekh A, Patton C, Pepek JM, Salama JK, Tuli R, Koong AC. Radiation Therapy for Pancreatic Cancer: Executive Summary of an ASTRO Clinical Practice Guideline. *Pract Radiat Oncol*. 2019;9(5):322-332. doi:10.1016/j.prro.2019.06.016

48. Antonelli M, Reinke A, Bakas S, Farahani K, AnnetteKopp-Schneider, Landman BA, Litjens G, Menze BH, Ronneberger O, Zhang D, Van Ginneken B, Bilello M, Bilic P, Christ PF, G RK DO, Gollub MJ, Heckers S, Huisman HJ, Jarnagin WR, McHugo M, Napel S, Pernicka JSG, Rhode K, Tobon-Gomez C, Vorontsov E, Meakin JA, Ourselin S, Wiesenfarth M, Arbeláez P, Bae B, Chen S, Daza LA, Feng J, He B, Isensee F, Ji Y, Jia F, Kim N, Kim ID, Merhof D, Pai A, Park B, Perslev M, Rezaiifar R, Rippel O, Sarasua I, Shen W, Son J, Wachinger C, Wang L, Wang Y, Xia Y, Xu D, Xu Z, Zheng Y, Simpson AL, Maier-Hein L, Cardoso MJ. The Medical Segmentation Decathlon. Nature Communications. 2022;13(1). doi:10.1038/s41467-022-30695-9

49. Tang Y, Yang D, Li W, Roth HR, Landman B, Xu D, Nath V, Hatamizadeh A. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. Published online 2022:20698-20708. doi:10.1109/cvpr52688.2022.02007

50. Hsu YC, Shen Y, Jin H, Kira Z. Generalized ODIN: Detecting Out-of-Distribution Image without Learning from Out-of-Distribution Data. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Published online 2020:10948-10957. doi:10.1109/CVPR42600.2020.01096

51. González C, Gotkowski K, Fuchs M, Bucher A, Dadras A, Fischbach R, Kaltenborn IJ, Mukhopadhyay A. Distance-based detection of out-of-distribution silent failures for Covid-19 lung lesion segmentation. *Med Image Anal*. 2022;82(August):102596. doi:10.1016/j.media.2022.102596

52. Claessens M, Oria CS, Brouwer C, Ziemer BP, Scholey JE, Lin H, Witztum A, Morin O, Naqa I El, Van Elmpt W, Verellen D. Quality Assurance for AI-Based Applications in

Radiation Therapy. *Semin Radiat Oncol*. Published online October 1, 2022. doi:10.1016/j.semradonc.2022.06.011

53. Baroudi H, Brock KK, Cao W, Chen X, Chung C, Court LE, Basha MDE, Farhat M, Gay SS, Gronberg MP, Gupta AC, Hernandez S, Huang K, Jaffray DA, Lim R, Marquez B, Nealon KA, Netherton TJ, Nguyen CM, Reber B, Rhee DJ, Salazar RM, Shanker MD, Sjogreen C, Woodland M, Yang J, Yu C, Zhao Y. Automated Contouring and Planning in Radiation Therapy: What Is 'Clinically Acceptable'? Diagnostics. 2023;13(4):667. doi:10.3390/diagnostics13040667

54. Isaksson L, Summers P, Bhalerao A, Gandini S, Raimondi S, Pepa M, Zaffaroni M, Corrao G, Mazzola G, Rotondi M, Lo Presti G, Haron Z, Alessi S, Pricolo P, Mistretta FA, Luzzago S, Cattani F, Musi G, De Cobelli O, Cremonesi M, Orecchia R, Marvaso G, Petralia G, Jereczek-Fossa BA. Quality assurance for automatically generated contours with additional deep learning. Insights Into Imaging. 2022;13(1). doi:10.1186/s13244-022-01276-7

55. Chen X, Men K, Chen B, Tang Y, Zhang T, Wang S, Li Y, Dai J. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. *Front Oncol*. 2020;10. doi:10.3389/fonc.2020.00524

56. Rhee DJ, Cardenas CE, Elhalawani H, McCarroll R, Zhang L, Yang J, Garden AS, Peterson CB, Beadle BM, Court LE. Automatic detection of contouring errors using convolutional neural networks. *Med Phys*. 2019;46(11):5086-5097. doi:10.1002/mp.13814

57. Rhee DJ, Akinfenwa CPA, Rigaud B, Jhingran A, Cardenas CE, Zhang L, Prajapati S, Kry SF, Brock KK, Beadle BM, Shaw W, O'Reilly F, Parkes J, Burger H, Fakie N, Trauernicht C, Simonds H, Court LE. Automatic contouring QA method using a deep learning–based autocontouring system. *J Appl Clin Med Phys*. 2022;23(8). doi:10.1002/acm2.13647

58. Zhong Y, Yang Y, Fang Y, Wang J, Hu W. A Preliminary Experience of Implementing Deep-Learning Based Auto-Segmentation in Head and Neck Cancer: A Study on Real-World Clinical Cases. *Front Oncol*. 2021;11. doi:10.3389/fonc.2021.638197

59. Jabbour SK, Hashem SA, Bosch W, Kim TK, Finkelstein SE, Anderson BM, Ben-Josef E, Crane CH, Goodman KA, Haddock MG, Herman JM, Hong TS, Kachnic LA, Mamon HJ, Pantarotto JR, Dawson LA. Upper abdominal normal organ contouring guidelines and atlas: A Radiation Therapy Oncology Group consensus. *Pract Radiat Oncol*. 2014;4(2):82-89. doi:10.1016/j.prro.2013.06.004

60. Rawla P, Sunkara T, Gaduputi V. Epidemiology of Pancreatic Cancer: Global Trends, Etiology and Risk Factors. *World J Oncol*. 2019;10(1):10-27. doi:10.14740/wjon1166

61. Zhuo Y, Chen Q, Chhatwal J. Changing Epidemiology of Hepatocellular Carcinoma and Role of Surveillance. In: ; 2019:53-67. doi:10.1007/978-3-030-21540-8_3

62. Roberts HJ, Wo JY. Stereotactic body radiation therapy for primary liver tumors: An effective liver-directed therapy in the toolbox. *Cancer*. 2022;128(5):956-965. doi:10.1002/cncr.34033

63. Brock KK. Adaptive Radiotherapy: Moving Into the Future. *Semin Radiat Oncol*. 2019;29(3):181-184. doi:10.1016/j.semradonc.2019.02.011

64. Zhang Y, Liang Y, Hall WA, Paulson ES, Chen X, Erickson BA, Li A, Xu X, Lian C, Yap PT, Wang AZ, Chera BS, Shen C, Lian J. *A Generalizable Guided Deep Learning Auto-Segmentation Method of Pancreatic GTV on Multi-Protocol Daily MRIs for MR-Guided Adaptive Radiotherapy Prediction of Optimal Dosimetry for Intensity-Modulated Radiotherapy With a Cascaded Auto-Content Deep Learning Model*. Vol 111.

65. Park JE, Ham S, Kim HS, Park SY, Yun J, Lee H, Choi SH, Kim N. Diffusion and perfusion MRI radiomics obtained from deep learning segmentation provides reproducible and comparable diagnostic model to human in post-treatment glioblastoma. *Eur Radiol*. 2021;31(5):3127-3137. doi:10.1007/s00330-020-07414-3

66. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. Attention Is All You Need. Published online June 12, 2017. http://arxiv.org/abs/1706.03762

67. Xie Z, Lin Y, Yao Z, Zhang Z, Dai Q, Cao Y, Hu H. Self-Supervised Learning with Swin Transformers. Published online May 10, 2021. http://arxiv.org/abs/2105.04553

68. Krishnan R, Rajpurkar P, Topol EJ. Self-supervised learning in medicine and healthcare. *Nat Biomed Eng*. 2022;6(12):1346-1352. doi:10.1038/s41551-022-00914-1

69. Navarro F, Watanabe C, Shit S, Sekuboyina A, Peeken JC, Combs SE, Menze BH. Evaluating the Robustness of Self-Supervised Learning in Medical Imaging. Published online May 14, 2021. http://arxiv.org/abs/2105.06986

70. Tang Y, Yang D, Li W, Roth H, Landman B, Xu D, Nath V, Hatamizadeh A. Self-Supervised Pre-Training of Swin Transformers for 3D Medical Image Analysis. Published online 2021. http://arxiv.org/abs/2111.14791

71. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. Published online June 15, 2016. http://arxiv.org/abs/1606.04797

72. Chen LC, Papandreou G, Kokkinos I, Murphy K, Yuille AL. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Trans Pattern Anal Mach Intell*. 2018;40(4):834-848. doi:10.1109/TPAMI.2017.2699184

73. Mehrtash A, Wells WM, Tempany CM, Abolmaesumi P, Kapur T. Confidence Calibration and Predictive Uncertainty Estimation for Deep Medical Image Segmentation. *IEEE Trans Med Imaging*. 2020;39(12):3868-3878. doi:10.1109/TMI.2020.3006437

74. Li S, Sui X, Luo X, Xu X, Liu Y, Goh R. Medical Image Segmentation Using Squeeze-and-Expansion Transformers. Published online May 20, 2021. http://arxiv.org/abs/2105.09511

75. Atito S, Awais M, Kittler J. SiT: Self-supervised vIsion Transformer. Published online April 8, 2021. http://arxiv.org/abs/2104.03602

76. Rahib L, Smith BD, Aizenberg R, Rosenzweig AB, Fleshman JM, Matrisian LM. Projecting cancer incidence and deaths to 2030: The unexpected burden of thyroid, liver, and pancreas cancers in the united states. *Cancer Res*. 2014;74(11):2913-2921. doi:10.1158/0008-5472.CAN-14-0155

77. Hammel P, Huguet F, Van Laethem JL, Goldstein D, Glimelius B, Artru P, Borbath I, Bouché O, Shannon J, André T, Mineur L, Chibaudel B, Bonnetain F, Louvet C. Effect of chemoradiotherapy vs chemotherapy on survival in patients with locally advanced pancreatic cancer controlled after 4 months of gemcitabine with or without erlotinib the LAP07 randomized clinical trial. *JAMA - Journal of the American Medical Association*. 2016;315(17):1844-1853. doi:10.1001/jama.2016.4324

78. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. Published online June 6, 2015. http://arxiv.org/abs/1506.02142

79. Wang G, Li W, Aertsen M, Leuven KU, Deprest J, Ourselin S, Vercauteren T. *Test-Time Augmentation with Uncertainty Estimation for Deep Learning-Based Medical Image Segmentation*.

80. Fort S, Hu H, Lakshminarayanan B. Deep Ensembles: A Loss Landscape Perspective. Published online 2019:1-15. http://arxiv.org/abs/1912.02757

81. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Published online June 18, 2017. http://arxiv.org/abs/1706.05721

82. Wong J, Baine M, Wisnoskie S, Bennion N, Zheng D, Yu L, Dalal V, Hollingsworth MA, Lin C, Zheng D. Effects of interobserver and interdisciplinary segmentation variabilities on CT-based radiomics for pancreatic cancer. *Sci Rep*. 2021;11(1). doi:10.1038/s41598-021-95152-x

83. Baumgartner CF, Tezcan KC, Chaitanya K, Hötker AM, Muehlematter UJ, Schawkat K, Becker AS, Donati O, Konukoglu E. PHiSeg: Capturing Uncertainty in Medical Image Segmentation. Published online June 7, 2019. http://arxiv.org/abs/1906.04045

84. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, Lu L, Yuille AL, Zhou Y. *TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation*. https://github.com/Beckschen/

85. Srivastava S, Yaqub M, Nandakumar K, Ge Z, Mahapatra D. Continual Domain Incremental Learning for Chest X-Ray Classification in Low-Resource Clinical Settings. In: *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Vol 12968 LNCS. Springer Science and Business Media Deutschland GmbH; 2021:226-238. doi:10.1007/978-3-030-87722-4_21

86. Wasserthal J, Meyer M, Breit HC, Cyriac J, Yang S, Segeroth M. TotalSegmentator: robust segmentation of 104 anatomical structures in CT images. Published online August 11, 2022. http://arxiv.org/abs/2208.05868

Vita

Cenji Yu is a graduate student from The University of Texas MD Anderson Cancer Center UTHealth Houston Graduate School of Biomedical Sciences. During his undergraduate career, he studied physics and completed thesis research in molecular dynamics. He subsequently entered graduate school to pursuit a PhD in medical physics. After his PhD, Cenji will pursue a clinical physics residency and aim to become a board-certified medical physicist