Learning Clinically Acceptable Segmentation of Organs at Risk in Cervical Cancer Radiation Treatment from Clinically Available Annotations

Author name(s) withheld EMAIL(S) WITHHELD
Address withheld

Editors: Under Review for MIDL 2023

Abstract

A deep learning-based automatic segmentation method for medical images is truly helpful in saving the time and effort of clinicians if such a method can be developed without any extra efforts from clinicians in annotating the dataset. Following this motivation, we present an approach to learn a deep learning model for the automatic segmentation of organs at risk (OARs) in cervical cancer radiation treatment, while using only the clinically available dataset and annotations. We employ simple heuristics for automatic data cleaning to minimize the label noise present in the clinically available data and an uncertainty-guided teacher-student training strategy to utilize a large dataset containing missing annotations. The results of our ablation experiments empirically demonstrate the added value of different components of our proposed approach toward learning a robust deep learning model while avoiding the need for dedicated manual annotations. Further, the delineations generated from segmentation masks predicted by our best-performing model were equally clinically acceptable to manual delineations.

Keywords: organs at risk, segmentation, deep learning, missing annotations

1. Introduction

Automatic organ segmentation can save hours of manual work required in the process of radiation treatment planning. Despite the availability of a plethora of deep learning models for semantic segmentation and their surprisingly good performance on new unseen data, developing a deep learning model for a new problem is still full of challenges. At the core of these challenges is the need for a large annotated dataset for supervised learning of a neural network. In the medical imaging domain, building a large annotated dataset becomes even more difficult and costlier due to the underlying requirement of clinical expertise for annotating the data. Medical data is however continuously annotated by clinical experts for use during the radiation treatment planning process. An automatic segmentation method would truly be helpful to the medical field if these clinically available annotations can be used for developing such a method with no extra manual efforts from clinicians. A few studies have looked in this direction [REFs], but there is still a long way ahead due to the challenges that come with the clinically available data.

Clinically available data is acquired and annotated based on the underlying clinical problem, for example, in case of radiation treatment planning, if a tumor is in the lower abdominal region, a pelvic scan is performed, but if the tumor is located near a kidney, a whole abdominal scan is acquired. Similarly, the organs at risk (OARs) are annotated based

© 2022 Anonymous.

on the tumor location. Furthermore, since the data is accumulated over years, the acquisition protocols, delineation guidelines, and observers may change. These factors contribute to large variations in the images, partial annotations, and label noise.

In this work, we focus on the above mentioned issues. Specifically, we aim to develop a supervised deep learning model for segmentation of four OARs (bowel bag, bladder, hips, and rectum) in Computed Tomography (CT) scans for cervical cancer treatment planning. We exported a large dataset from the database of a hospital along with the clinically available annotations. We sought answers to the following questions:

1. Can we make use of the clinically available scans and annotations to develop a robust deep learning model while minimizing the adverse effects of large variance, label noise, and missing annotations on the test performance?
2. What is the test performance of a deep learning model learned from only clinically available data and annotations in terms of clinical acceptability?

For answering the first question, we developed heuristics to collect as many clinically available annotations as possible while avoiding ambiguities (2.2) and increase homogeneity in the available dataset (2.2). Further, We used a teacher-student strategy, where a teacher model was trained on a small, fully annotated dataset and used to impute the missing organ annotations in a large dataset. Then, we trained a student model on the full dataset, which included both the available annotations and the imputed annotations (2.3). We trained the teacher model to also estimate the epistemic uncertainty maps for each image (2.4) and used these maps to train the student model with an uncertainty-guided loss function (2.5).

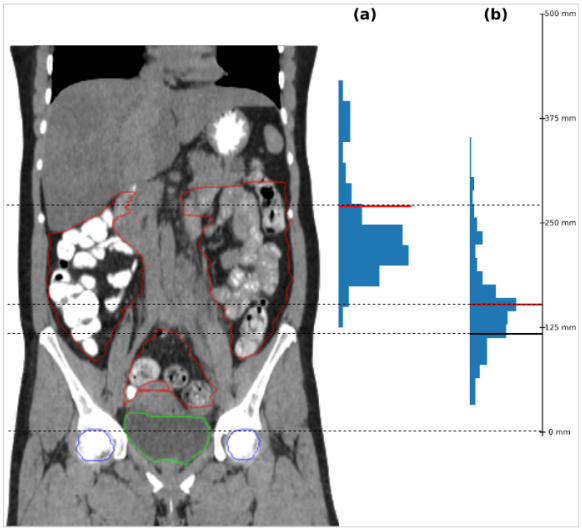
To answer the second question, we conducted a validation study with a small test dataset (3.2), in which we compared the clinical acceptability grades (provided by a radiation oncologist independently and blindly) of the delineations generated from our deep learning approach and the manual delineations available in the test datset. We also compared the agreement between deep learning model predictions and manual delineations with inter-observer agreement to assess the potential of clinical use.

2. Materials and Methods

2.1. Data

We retrospectively selected the CT scans of female patients who were treated for a tumor anywhere in the abdominal region from 2009 to 2019 in a local hospital. The CT scans with associated delineations were received in anonymized form through a data transfer agreement. A total of 1170 scans from 800 patients were received. These scans were used for training and validation. In addition, we used CT scans from 140 cervical cancer patients who received radiation treatment as test data.

The dataset was preprocessed to make it suitable for training a deep learning model. First the annotations for four OARs relevant for cervical cancer radiation treatment (bowel bag, bladder, hips, and rectum) were extracted by following the following steps: (1) collected all annotations except the ones related to the tumor and target volumes (annotations containing gtv, ptv, or ctv), (2) standardized different variations of organ labels (e.g., bowel, bowel bag, Bowel bag, bowel_bag, Bowel_bag were all considered bowel bag), (3) combined left and right hip annotations as a single organ, (4) removed voxels annotated as bladder or rectum from the bowel bag annotation to avoid ambiguous labeling in those voxels. Next,



The red lines indicate the thresholds to crop the FOV and delete the bowel bag annotations in the cranial direction. The black line in (b) shows the threshold for discarding the scans, where the bowel bag annotations did not cover the pelvic region. The dashed lines indicate the corresponding anatomy for each threshold.

Figure 1: The histograms of number of scans with respect to the distance from the most cranial point of hip annotations to (a) the cranial end of FOV of scan and (b) the cranial end of bowel bag annotation. A representative CT scan and reconstructed delineations in the coronal view are shown (red: bowel bag, green: bladder, blue: hips).

all the CT scans were resampled to 2.5mm × 2.5mm × 2.5mm voxel spacing. This resulted in a total of 186 scans that contained annotations for all the four OARs considered in this work (referred to as the fully annotated dataset, $\mathcal{D}_f$ in the rest of the manuscript). The remaining scans had missing annotations for at least one of the OARs (referred to as the partially annotated dataset, $\mathcal{D}_p$). The distribution of the available annotations in the data is described in Table 1.

Bowel bag	✓	✓		✓	✓	✓			✓	✓		✓			
Bladder	✓	✓	✓	✓		✓	✓				✓		✓		
Hip	✓	✓	✓		✓		✓	✓	✓					✓	
Rectum	✓		✓	✓	✓			✓		✓	✓				✓
No. of scans	186	349	263	198	201	366	480	285	365	213	832	383	1103	504	865
Datasets	$\mathcal{D}_f$	$\mathcal{D}_p$													

Table 1: Distribution of annotations in the training dataset. ✓: annotation present, $\mathcal{D}_f$: fully annotated dataset, $\mathcal{D}_p$: partially annotated dataset.

2.2. Automatic Data Cleaning

Since the data was accumulated over 10 years and the scans belonging to patients who were treated for a tumor anywhere in the abdominal region were included, it exhibited a large variance[1] in the transverse FOV of the scan (causing unnecessarily increase in background voxels and potentially less efficient training), and the cranial extent of bowel

1. Note that the variance in the image intensity patterns arising from anatomical variations and difference in contrast in organs could potentially be beneficial towards learning a robust model.

bag annotations (attributing to label noise). This is due to the fact that scans for patients with upper abdominal tumors included a chest and abdominal scan with annotation of only the cranial slices of the bowel bag, while the scans for patients with pelvic tumors included a pelvic scan with annotation of only the caudal part of the bowel bag.

To make the data more homogeneous such that the adverse effects of inefficient training and label noise could be avoided, we developed heuristics derived from the histograms of training data ($\mathcal{D}_f$) as follows. In Figure 1, the histograms of transverse FOV (Figure 1 (a)), distance between most cranial part of hip annotation and most cranial part of bowel bag annotation (Figure 1 (b)), from all the scans in $\mathcal{D}_f$ are shown. We analyzed the histograms and decided thresholds such that the histograms represented a unimodal distribution corresponding to the most frequently used scanning protocol and annotation style. The thresholds are marked in the figure. Based on these thresholds, the scans were cropped in the cranial direction to remove the chest region. The bowel bag annotations in abdominal region roughly above the lumbar 4 spinal segment were deleted. The scans that did not contain bowel bag annotations covering the entire pelvic region were discarded. These step resulted in a decrease in the size of $\mathcal{D}_f$ from 186 to 134. The resulting dataset of 134 scans is referred to as fully annotated auto-cleaned dataset, $\mathcal{D}_f^{clean}$ in the rest of the manuscript.

2.3. Self-supervised Training through Data Imputation

We used a teacher-student strategy, wherein a teacher model was trained using the small, fully annotated dataset ($\mathcal{D}_f^{clean}$). The predictions from the trained teacher model were used to impute a large dataset with missing annotations ($\mathcal{D}_p$). Then a student was trained with the entire dataset ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) containing imputed annotations. Alternative approaches include fine-tuning the teacher with large dataset containing imputed annotations [REF]. We chose a teacher-student strategy because existing literature suggests that training a student from scratch yields better results that fine-tuning the teacher [REF].

2.4. Uncertainty Estimation

Epistemic uncertainty refers to the lack of knowledge in a model about the underlying data and this informs about the reliability of a model's prediction for a specific sample. There exist many alternatives to estimate epistemic uncertainty in deep learning model predictions [REFs]. We used a K-head neural network, in which a single head was selected randomly for back propagation at each iteration during training. We selected the k-head approach because it allows independence between predictions from different heads with faster inference times as compared to MC dropout approach [REF]. Moreover, the memory overhead was not much compared to fully independent ensembles. We used the entropy of the mean prediction from k-heads as an estimate of epistemic uncertainty [REF].

2.5. Uncertainty Guided Training

We trained the student model with a cross-entropy loss weighted by uncertainty guided term (e^{-u}, where u is uncertainty in the teacher's predictions) at each voxel. The weight e^{-u} ensured a high weight on voxels where the uncertainty in the teacher's predictions was low and vice-versa. The loss at each voxel is defined as $L = e^{-u} y \cdot log(\hat{y})$, where, y is the

reference label, and $\hat{y}$ is the predicted probability. $u = 0$ at the voxels where annotations were clinically available. In this way, the student model would benefit from training with a large dataset while avoiding deterioration in the network performance due to uncertain label predictions from the teacher model.

2.6. Implementation details

We used the U-Net architecture [REF], which is well known for its performance on semantic, especially medical image segmentation tasks. The implementation was done in Python by using the PyTorch library [REF] and the training was done on NVIDIA RTX2080 GPUs. Other training hyperparameters were: network initialization = He norm, optimizer = Adam [REF], learning rate (LR) = $1e^{-3}$, weight decay = $1e^{-4}$, learning schedule = step LR with step size = $\frac{1}{3}\times$ total training steps, data augmentations = global brightness and contrast variations, random rotations.

3. Experiments

Method	Dice	Surface Dice	HD
U-Net + $\mathcal{D}_f$	83.30 ± 5.88		
U-Net + $\mathcal{D}_f^{clean}$	84.76		
K-head teacher + $\mathcal{D}_f^{clean}$ (*basic teacher*)	85.03		
K-head student + ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) (*basic student*)	86.77		
K-head teacher + $\mathcal{D}_f^{clean}$ + Aug. (*robust teacher*)	85.22		
K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in student training (*basic teacher + robust student*)	86.91		
K-head student + ($\mathcal{D}_f^{clean}$+$\mathcal{D}_p$) + Aug. in teacher and student training (*robust teacher + robust student*)	in queue		

Table 2: Mean ± standard deviation of test performance of the best models obtained from 5-fold cross-validation. Aug.: additional augmentations, HD: Hausdorff distance

3.1. Ablation

We conducted ablation experiment to look into the effect of the different components mentioned in the above sections, on the test performance. Specifically, we compared the test performance of the following models: U-Net trained with small, fully annotated dataset

before automatic data cleaning ($\mathcal{D}_f$), U-Net trained with small, fully annotated dataset after automatic data cleaning ($\mathcal{D}_f^{clean}$), k-head U-Net teacher model trained with $\mathcal{D}_f^{clean}$ (referred to as *'basic teacher'*), k-head U-Net student model trained with the large dataset ($\mathcal{D}_f^{clean} + \mathcal{D}_p$) with uncertainty guided loss (referred to as *'basic student'*).

Earlier studies suggest that training a noisy student (noise in the data or model) on a large dataset with imputed annotations yields performance gain [REF]. To test this, we employed the following additional data augmentations as a part of noise in the data: left right flipping, masking an organ with a random intensity to simulate contrast, global elastic deformations, and elastic deformations centered in either bowel bag or bladder as additional augmentations. We compared the performance of three models: a teacher model trained with $\mathcal{D}_f^{clean}$ + additional augmentations (referred to as 'robust teacher'), a student model trained with $\mathcal{D}_f^{clean} + \mathcal{D}_p$ + additional augmentation using the imputed annotations from 'basic teacher' (referred to as *'robust student + basic teacher'*), and a student model trained with $\mathcal{D}_f^{clean} + \mathcal{D}_p$ + additional augmentation using the imputed annotations from 'robust teacher' (referred to as *'robust student + robust teacher'*).

The mean and standard deviation of the test performance from the best models obtained after a 5-fold cross-validation is reported in Table 2. The automatic data cleaning had a clear impact on the test performance, which is mainly due to an increase in recall and precision of bowel bag segmentation. The automatic data cleaning increased the recall of bowel bag from X to X, and precision of bowel bag from X to X. Furthermore, learning from a large dataset containing partial annotations with the proposed teacher-student strategy and uncertainty-guided training provided a mean gain of X% in recall, X% in precision, and X% in Dice scores. Adding noise in the data through additional augmentations provided X% gain in mean dice in a teacher model and X% gain in mean dice in a student model. Few representative examples from *'basic teacher + robust student'* are shown in Figure 2 for reader's perusal.

3.2. Clinical Acceptability Test

We conducted a validation study to assess the clinical acceptability of the automatic OAR segmentations. We used the best *'basic teacher + robust student'* model (based on cross-validation performance) from first data-split, to predict OAR segmentations for a subset of 4 scans, which were used to generate automatic contour delineations. We showed both the clinically available delineations (used as ground truths for calculating test performance) and the automatic delineations to a radiation oncologist (henceforth referred to as 'clinical expert'), who was blind to the method of generating the delineations. The delineations were presented on 2D transverse slices spaced at 10 mm distance to make it similar to clinical scenario where the contours are delineated on 2D transverse slices. The clinical expert optionally inspected the delineations and scans in coronal and sagittal view also to ensure comprehensiveness. The clinical expert graded each delineation for its clinical acceptability according to a 4 point Likert scale: 1 = acceptable as it is, 2 = acceptable but marginally deviating from exact anatomical definition (subjective to observer), 3 = acceptable with minor corrections (e.g., only adapting a small part of the contour) because either a part of the organ is not delineated or a peripheral tissue is included in the delineation, 4 =

Figure 2: Representative examples.

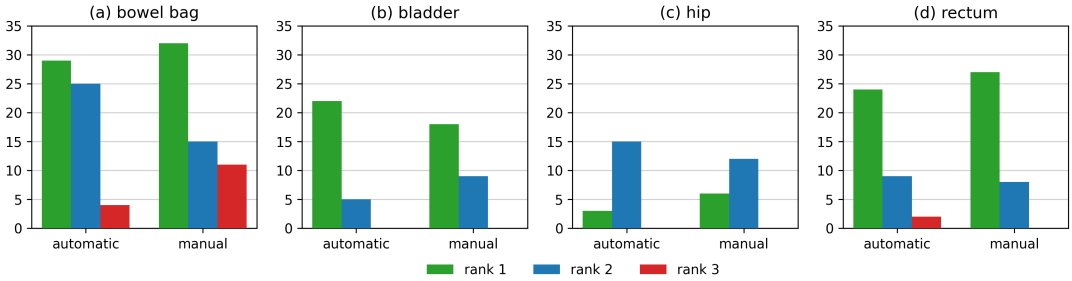


Figure 3: Ranking comparison

not acceptable because a correction involving both deletion as well as delineation of an additional contour is required.

The clinical acceptability grading for the automatic and manual delineations for all the OARs is shown in Figure 3. A chi-squared test of goodness of fit indicated that the histograms of clinical acceptability grades of the automatic delineations were significantly different from manual delineations for bowel bag ($\chi^2(1, N{=}58) = 11.402, p = 0.003$). However, it was difficult to identify whether manual or automatic delineations were overall more clinically acceptable, because the automatic delineations were graded 1 but also 3 fewer times compared to automatic delineations. The clinical acceptability grades for automatic and manual delineations were not significantly different for bladder ($\chi^2(1, N{=}27) = 2.667, p = 0.102$), and hips delineations ($\chi^2(1, N{=}18) = 2.250, p = 0.134$). For rectum delineations, the chi-squared test statistics could not be obtained because the frequency counts corresponding to grade 3 were less than 5, however, it is apparent from the Figure that the frequency counts in each category were similar for both automatic and manual delineations.

Observer 1 & Observer 2	xx	xx
Observer 1 & Automatic	xx	xx
Observer 2 & Automatic	xx	xx

Table 3: Comparison of mean test performance of 8 scans with inter-observer agreement

3.3. Inter-Observer Agreement vs. Auto-Manual Agreement

In a subset of 8 scans in the test data, the manual delineations of the OARs were available from two observers. In these scans, we calculated dice coefficient as a measure of the agreement between delineation pairs of observer 1 and observer 2 ($D_{1,2}$), observer 1 and automatic ($D_{1,auto}$), observer 2 and automatic ($D_{2,auto}$). The mean $\pm$ standard deviation of dice coefficients are reported in Table 3. The Friedman test between distributions of $D_{1,2}, D_{1,auto}, D_{2,auto}$ did not show a significant main effect (TEST STAT), indicating that the automatic delineations generated from predicted segmentations by our deep learning model had as much agreement to a clinician's delineation as two clinicians agreed among each-other. This highlights the potential of clinical acceptability of the OAR delineations predicted by our deep learning model.

4. Discussion and Conclusions

We investigated the possibility of using a clinically available dataset of the abdominal region to learn a deep learning model for automatic OAR segmentation for cervical cancer radiation treatment. To the best of our knowledge, this is one of the few works in the direction of utilizing the clinically available annotations for learning a deep learning model and the first work to attempt this for OARs in the female pelvic region. The presented work is a step towards providing true help to the hospital through an automatic OAR segmentation method, which did not require clinicians' time for annotating a large dataset during its development also. Learning automatic segmentation of OARs in the female pelvic region (e.g., bowel bag and rectum) is challenging even in the supervised learning paradigm using curated and manually annotated data specific for the learning task. Our presented work achieved it with clinically available data and annotations, which contained high variance in the data and labels, and missing annotations. The obtained segmentations were of clinically acceptable quality, which is encouraging.

Limitations of our work include ablation study involving only a single run (i.e., network initialization), and lack of experiments with different semantic segmentation architectures. Both the decisions were consciously taken to find sensible results despite the expensive nature of training deep neural networks on large data. Earlier studies have shown that different network initialization does not influence the neural network performance as much as using a different data split [REF]. Therefore, we believe that the presented performance differences are representative of the true performance differences. Moreover, we expect that with a different architecture also, we should gain similar insights. In future, we aim to extend the current work to add the capability of segmenting more OARs relevant to cervical cancer radiation treatment planning e.g., sigmoid and anal canal.

In conclusion, we demonstrated that training a deep learning model without using curated and specifically annotated medical imaging data, but with a capability of predicting clinically acceptable segmentations is possible. Apart from saving clinicians' time, our proposed approach leads to faster development time because of using the readily available data and increased test performance because of the increased dataset size.

Acknowledgments

Acknowledgments withheld.

References