Deep learning for automatic segmentation of tumors on MRI

Rodríguez Outeiral, R.

Citation
Rodríguez Outeiral, R. (2024, June 25). *Deep learning for automatic segmentation of tumors on MRI*. Retrieved from https://hdl.handle.net/1887/3765390

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3765390

Note: To cite this publication please use the final published version (if applicable).

DEEP LEARNING FOR AUTOMATIC SEGMENTATION OF TUMORS ON MRI

Roque Rodríguez Outeiral

Deep learning for automatic segmentation of tumors on MRI

Copyright 2024 © Roque Rodríguez Outeiral
All rights reserved. No part of this thesis may be reproduced, stored in a retrieval system or transmitted in any form or by any means without permission of the author.

ISBN: 978-94-6506-083-5
Printing: Ridderprint | www.ridderprint.nl
Layout and design: Finn McAleer | www.ridderprint.nl

The work described in this thesis was performed at The Netherlands Cancer Institute.

DEEP LEARNING FOR AUTOMATIC SEGMENTATION OF TUMORS ON MRI

Proefschrift

ter verkrijging van
de graad van doctor aan de Universiteit Leiden,
op gezag van rector magnificus prof.dr.ir. H. Bijl,
volgens besluit van het college voor promoties
te verdedigen op dinsdag 25 Juni 2024
klokke 15.00 uur

door

Roque Rodríguez Outeiral
geboren te Vigo
in 1994

Promotor

prof. dr. U. A. van der Heide

Co-promoters

dr. R. Simões	The Netherlands Cancer Institute
dr. T. M. Janssen	The Netherlands Cancer Institute

Promotiecommissie

prof. dr. ir. M. Staring	
prof. dr. ing. C. Hurkmans	Universiteit Eindhoven
prof. dr. ir. P. M. A. van Ooijen	Universiteit Groningen
prof. dr. C. L. Creutzberg	

TABLE OF CONTENTS

Chapter 1 Introduction **7**

Chapter 2 Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning **15**

Chapter 3 Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging **31**

Chapter 4 Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy **47**

Chapter 5 A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations **61**

Chapter 6 General discussion **77**

Chapter 7 References **89**

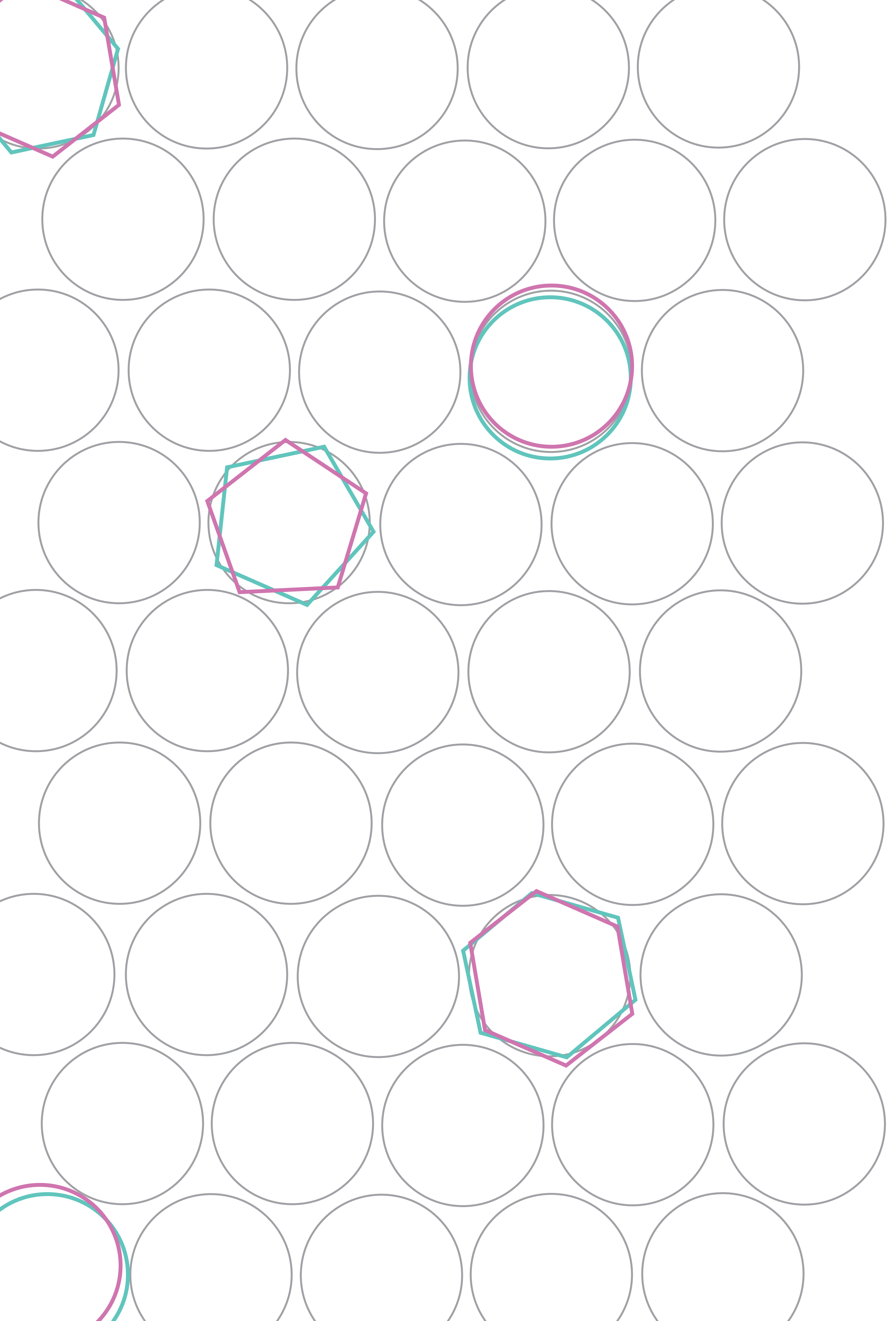
Appendices Summary **100**

Samenvatting **102**

Acknowledgements **105**

List of publications **107**

Curriculum vitae **109**



Chapter 1

Introduction

TUMOR SEGMENTATION IN RADIOTHERAPY

Cancer is the second leading cause of death worldwide, accountable for nearly 10 million deaths in 2020 [1]. Radiotherapy (RT) is one of the main modalities of cancer treatment, being indicated as part of the treatment for approximately half of the patients [2–4]. RT consists of the use of high doses of ionizing radiation to damage the DNA of the cancer cells, leading to their death. Simultaneously, it is intended that the least amount of radiation possible is given to the surrounding healthy tissue. Therefore, precise localization of the tumor and of the surrounding organs is critical for the RT treatment planning.

To visualize a patient's anatomy, an image of the patient is acquired using common imaging modalities such as computed tomography (CT), magnetic resonance (MRI), or positron emission tomography (PET). On these images, the pixels corresponding to the relevant structures are identified and labeled. This process is known as segmentation, delineation or contouring. The structures to segment are the target of the RT treatment and the organs to spare (known as organs at risk - OARs). Once these segmentations are made, the dose intended to each region of the patient anatomy can be calculated. Given that the segmentation step is one of the first steps of the RT workflow, it can influence the quality of the entire treatment [5].

The primary target of the RT treatment is typically the visible extent of the tumor on the image of the patient, known as gross tumor volume or GTV. In addition, to address any hidden microscopic disease that is not visible on the imaging, an extra margin is defined, known as clinical target volume or CTV. Furthermore, a safety margin is often defined to account for any geometrical uncertainties that may arise from the whole treatment pipeline, the planning target volume or PTV. In certain tumor sites, such as cervix or rectum, the CTV is defined following anatomical structures rather than as a margin around the tumor. Arguably, both the tumor or targets that are not defined as anatomical structures can be considered harder to segment than those anatomical structures, because their locations, shapes and sizes tend to be more variable.

In the current clinical practice, the segmentation of the tumor is carried out manually by physicians, which presents two issues. The first issue is that manual segmentation of the tumors is a time-consuming process, often regarded as one of the most time-intensive steps in the RT workflow [6] or even referred to as a bottleneck of the entire pipeline [7]. This is partly due to the complexity of the tumor segmentation process: it involves the physician scrolling through the 3D image of the patient and annotating the points that correspond to the tumor in a slice-by-slice manner. Additionally, physicians often integrate information from various sources, including clinical reports, other images, and clinical examinations of the patient.

Nowadays, it is aimed to deliver the RT treatment more accurately by considering anatomical changes that happen between RT planning and delivery or even during the course of the treatment [8], thereby potentially improving local control and reduce treatment-related toxicities. This approach is referred to as 'adaptive RT' or ART and it can be carried out in two ways: offline or online. Online ART (OART) specifically addresses changes that occur on the day of treatment [9]. A key step of OART is the daily (re-)segmentation of the relevant structures for the treatment plan, a process in which both the patient and the entire treatment staff are waiting while these segmentations are performed. Given the increased interest in OART, the clinical burden of segmentation is expected to increase even more in the coming years.

The second issue with manual segmentation of the tumors is that it depends on the observer. Inter-observer variability has been reported in a myriad of tumor sites, such as head and neck [10,11], esophagus [12], lung [13,14], rectum [15], or cervix [16]. Potential sources of this variability include, but are not limited to, the lack of clear boundaries of the tumor, the lack of standardized delineation guidelines or differences in experience among different observers. These deviations on the segmentations of the targets can affect the RT outcome in terms of RT-related toxicity, overall survival and local recurrence of cancer [17].

DEEP LEARNING FOR AUTOMATIC SEGMENTATION OF TUMORS IN RADIOTHERAPY

Deep learning techniques, specifically convolutional neural networks (CNNs), are currently considered state of the art in computer vision tasks, including segmentation. Therefore, CNNs have been the preferred method in the recent years for automatically segmenting the structures required for RT treatment, namely the targets and the OARs. For the case of OARs, CNNs have already been shown to outperform previous methods of automatic segmentation for several anatomical sites [15–18]. Furthermore, they are already incorporated in the software released by several companies [22–24].

For the case of the targets, commercial solutions are only available to segment the CTV in a few cases, when it is defined as an anatomical structure, like the prostate, the head and neck lymph nodes or the vessels of the pelvis [24,25]. Automatic segmentation of the tumors is, however, still uncommon and their clinical adoption is non-existent. Promising results have been reported [26,27], often achieving arguably high performance (Dice Scores between 0.73-0.94), but the segmentations resulting from these methods still can be unacceptable in some of the cases. Given that the tumors receive the highest dose of radiation during treatment, errors in their automatic segmentation could result in unacceptable treatment plans for clinical practice [17].

Typically, the set up for automatic segmentation of tumors is done in a fully supervised manner [26,28]. During a training phase, the CNN is shown images of different patients (i.e. input images) and asked to provide a segmentation of the tumor. In each training iteration, the proposed automatic segmentation is compared to the manual segmentation made by the doctors, and the error is quantified. The function that quantifies this error is known as loss or error function. This error function is used to modify the parameters of the CNN, until the proposed segmentations are similar to the manual segmentations. Important lines of research in the field of automatic segmentation of medical structures focus on finding the best input images to the network to achieve the most accurate segmentations [29] and determining the optimal loss functions to train the CNNs [30] .

Another line of research related to the automatic segmentation of tumors is the automation of the quality assurance of the automatic segmentations. Because these segmentations are not yet perfect [26,28,31], they still need to be checked by physicians, which hinders the promise of automatic segmentations for time saving. Recently, several works have focused on developing uncertainty maps, that show where the CNN is uncertain about its predictions [32–35]. These uncertainty maps have been shown to correlate with the areas where the network failed [36,37], and could thereby be used to flag the segmentations that would require a check.

AUTOMATIC SEGMENTATION OF TUMORS ON MRI

Due to its superior soft tissue contrast to other image modalities, MRI is a desirable image modality for RT treatment planning purposes. Moreover, each MRI sequence enables visualization of different tissue types, even with the possibility of imaging biological and functional processes through quantitative MRI techniques. This level of versatility renders MRI a valuable imaging modality and distinguishes it from other imaging techniques like CT. Furthermore, the acquisition of MRI does not involve the exposure of the patient to ionizing energy, as opposed to CT or PET. In recent years, systems that combine a MRI scanner and a RT linear accelerator for RT treatment delivery have been made commercially available [38], such as the Elekta Unity (Elekta AB, Stockholm, Sweden) MR-Linac system or the ViewRay MRIdian system (Viewray Inc., Oakwood, OH). Thus, there is an increased interest in MRI-only pipelines for RT treatment planning.

Partly due to its relative novelty in the field of RT as main imaging modality, the research on automatic segmentation of tumors on MRI is even less spread than on CT or PET-CT [26,28]. The lack of large MRI-based datasets, which are needed to train the data-hungry deep learning methods, can also explain the scarcity of literature on the topic. One notable exception is the case of brain tumors, where MRI is the dominant image modality [39]. Two tumor sites that showcase the potential of MRI as main imaging modality for automatic tumor segmentation are the oropharyngeal cancer, due to the high prevalence of soft tissue in the area; and the cervical cancer, for which MRI-guided radiotherapy is the standard treatment option [40].

One of the main treatment modalities for oropharyngeal cancers is RT, being indicated in more than 70% of the patients [41]. A challenge in RT for this tumor site is the proximity of many OARs, which makes the accurate segmentation of the tumor all the more critical to avoid potential damage to those organs. Furthermore, high interobserver variability has been reported for this tumor site, reaching differences in the volumes delineated by the different observers of up to a factor of 10 [11]. Automatic segmentation of the tumors is therefore a desirable solution for this tumor site [42], not only to decrease the clinical burden when treating this tumor but also to provide more consistent segmentations among different physicians or even hospitals.

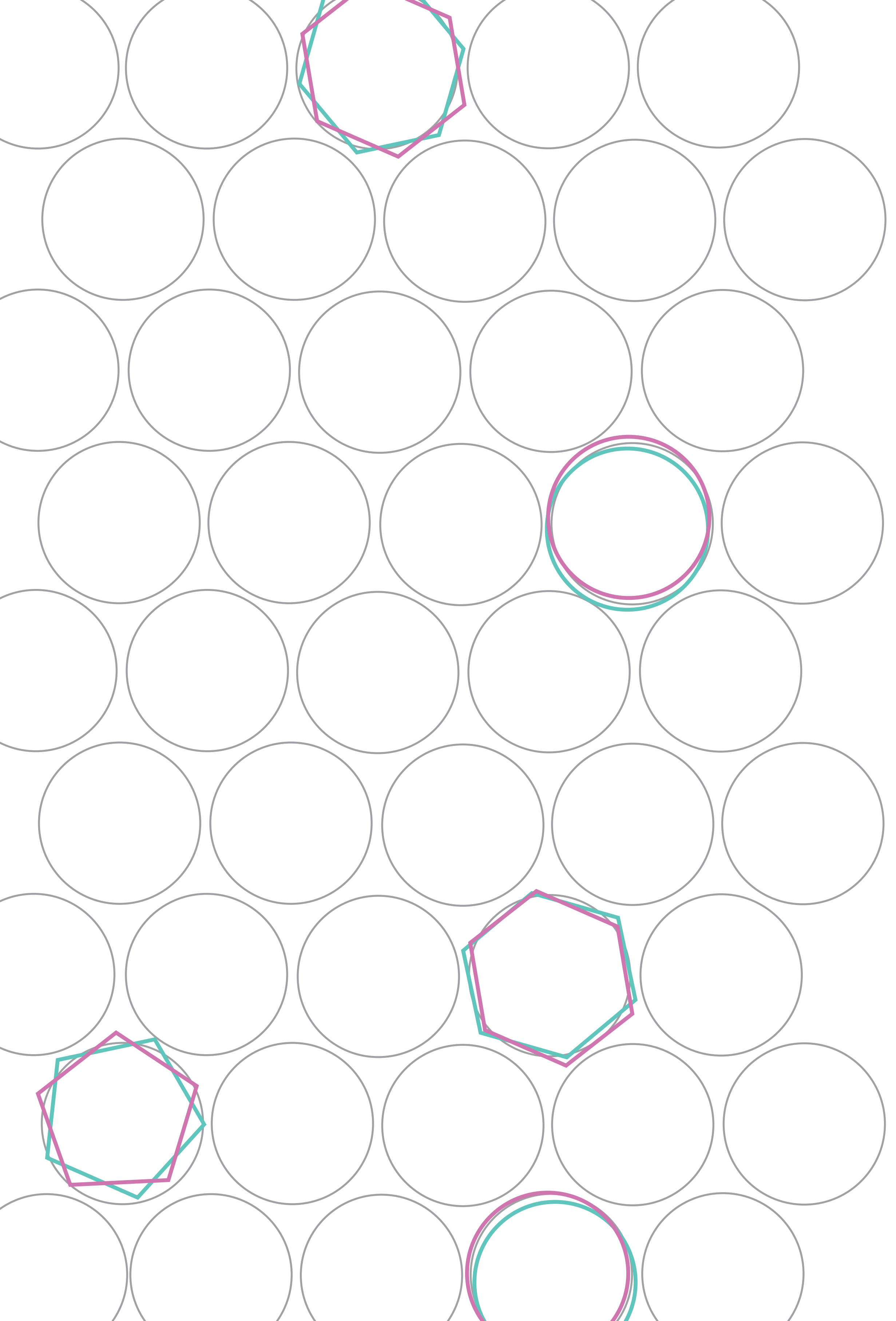
Locally advanced cervical cancer is typically treated with a combination of external beam radiotherapy (EBRT), concomitant chemotherapy, and 3 to 4 fractions of brachytherapy (BT) [43]. BT consists of the placement of radioactive sources directly into or next to the tumor. For the case of cervical cancer, this is achieved with an applicator which is inserted into the patient before the BT treatment. Afterwards, the MRI images are acquired, and the patient must remain immobilized in bed while the necessary structures (such as target volumes and organs at risk) are manually segmented, and a treatment plan is devised. This is highly uncomfortable for the patient and logistically complex, which makes automatic segmentation even more critical than for other RT treatment modalities.

AIM AND THESIS OUTLINE

Despite the versatility of MRI in visualizing anatomical structures and tumor tissue, there is a lack of research on automatic segmentation of tumors on MRI. Current deep learning methods still fail to produce accurate tumor segmentations in certain cases, particularly on MRI. Physicians would still need to manually review and potentially correct the automatic segmentations, which limits the promised time-savings and clinical applicability of these methods. Given the complexity of the task, we hypothesized that training a CNN with a considerably large dataset would achieve clinically acceptable automatic segmentations. However, in the medical field, gathering extensive datasets is challenging due to the expensive and logistically complex data acquisition process, as well as the need for highly skilled professionals to manually label the data. The goal of this thesis was to develop automatic segmentation techniques for tumors in MRI images that deliver clinically acceptable segmentations using clinical MRI datasets. Additionally, we aimed to automate the quality assurance of the automatic segmentations, thereby maintaining the time-saving benefits even in cases where the network fails. The different automatic segmentation methods were applied in two different tasks: the automatic segmentation of the oropharyngeal primary tumor in multiparametric diagnostic MRI images (chapters 2 and 3) and the automatic segmentation of the cervical cancer gross tumor volume (GTV) in the MRI images of the BT treatment images (chapters 4 and 5).

In the oropharyngeal cancer segmentation task, we investigated three different strategies to introduce prior information in the neural network training design. Firstly, we studied the effect of using different anatomical MRI sequences as input to the network (Chapter 2), similarly to how radiologists use the different sequences to delineate the tumor themselves. Secondly, we investigated the use of different loss functions (Chapter 3). Finally, we studied the effect of reducing the context around the tumor, first by proposing a semi-automatic approach in which human observers approximately located the tumor and a CNN segmented the tumor (Chapter 2) and then fully automatizing it into a two-stage approach, that split the task in detection and segmentation (Chapter 3).

In the cervical cancer segmentation task, we validated the quality of the segmentations obtained with a state-of-the-art segmentation framework not only geometrically but also with dose-volume parameters and investigated whether there were differences for clinically relevant parameters, such as volume or tumor stage (Chapter 4). Furthermore, even though the proposed method performed adequately on average, the network still failed in some cases. We identified a metric to predict the quality of the automatic segmentations (Chapter 5). Such a metric can potentially flag segmentations that would require a manual check and could therefore help with the clinical implementation of the automatic segmentation methods.



Chapter 2

Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning

Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA.

Phys Imaging Radiat Oncol. 19:39–44. (2021)

ABSTRACT

Background and purpose

Segmentation of oropharyngeal squamous cell carcinoma (OPSCC) is needed for radiotherapy planning. We aimed to segment the primary tumor for OPSCC on MRI using convolutional neural networks (CNNs). We investigated the effect of multiple MRI sequences as input and we proposed a semi-automatic approach for tumor segmentation that is expected to save time in the clinic.

Materials and methods

We included 171 OPSCC patients retrospectively from 2010 until 2015. For all patients the following MRI sequences were available: T1-weighted, T2-weighted and 3D T1-weighted after gadolinium injection. We trained a 3D UNet using the entire images and images with reduced context, considering only information within clipboxes around the tumor. We compared the performance using different combinations of MRI sequences as input. Finally, a semi-automatic approach by two human observers defining clipboxes around the tumor was tested. Segmentation performance was measured with Sørensen–Dice coefficient (Dice), 95th Hausdorff distance (HD) and Mean Surface Distance (MSD).

Results

The 3D UNet trained with full context and all sequences as input yielded a median Dice of 0.55, HD of 8.7 mm and MSD of 2.7 mm. Combining all MRI sequences was better than using single sequences. The semi-automatic approach with all sequences as input yielded significantly better performance ($p<0.001$): a median Dice of 0.74, HD of 4.6 mm and MSD of 1.2 mm.

Conclusions

Reducing the amount of context around the tumor and combining multiple MRI sequences improves the segmentation performance. A semi-automatic approach is accurate and clinically feasible.

INTRODUCTION

Worldwide, there are more than 679,000 new cases of head and neck cancer (HNC) per year and 380,000 of those cases result in death [44]. Radiotherapy (RT) is indicated for 74% of head and neck cancer patients, and up to 100% in some subsites [41]. Tumor delineation is needed for RT planning. In clinical practice, tumor contouring is done manually, which is time consuming and suffers from interobserver variability. Thus, accurate automatic segmentation is desirable.

Convolutional neural networks (CNNs) are considered the current state of the art for computer vision techniques, such as automatic segmentation. Specifically for tumor segmentation, promising results have been obtained for various tumor sites such as brain [45], lung [46], liver [47] and rectum [48].

For HNC, previous literature [49,50] focused on the segmentation of other RT-related target volumes rather than the primary tumor and without special focus on any particular HNC subsite, such as nasopharyngeal or oropharyngeal cancer. However, anatomy and imaging characteristics of tumors and their surrounding tissue vary greatly across subsites. Nasopharyngeal tumors are bounded by the surrounding anatomy and thus they present with lower spatial variability. Men et al [51] proposed an automatic segmentation method for nasopharyngeal primary tumors. To the best of our knowledge, no studies have been published on automatic segmentation of primary tumors in oropharyngeal squamous cell cancer (OPSCC). Tumors in this category are quite variable in shape, size and location compared to other subsites in head and neck cancer and their delineation suffers from high interobserver variability [11].

The modalities of choice in other works for HNC automatic segmentation are PET and/or CT [49,50]. PET presents low spatial resolution and only shows the metabolically active part of the tumor while CT has low soft tissue contrast. MRI is now becoming a modality of interest in RT and provides improved soft tissue contrast compared to other modalities, being better suitable for oropharyngeal tumor segmentation. In line with this, previous works have suggested that the use of MRI for head and neck cancer delineation provides unique information compared to PET/CT or CT [52].

We investigated the effect on segmentation performance of different MRI sequences and its combination as inputs to the model. We hypothesized that by decreasing the amount of context around the tumor, thereby simplifying the task, the performance of the segmentation model would improve. Hence, we proposed a semi-automatic approach in which a clipbox around the tumor is used to crop the input image. We demonstrated its clinical applicability by having two observers (including one radiation oncologist) manually selecting the clipbox. The aim of this study was to develop a CNN model for segmenting OPSCC on MRI images.

MATERIALS AND METHODS

Data

A cohort of 171 patients treated at our institute between January 2010 and December 2015 was used for this project. Mean patient age was 60 (Standard deviation ± 7 years) and 62% of the patients were male. Further details on tumor stage and HPV status can be found in the Supplemental Material (table S.1). All patients had histologically proven primary OPSCC and pre-treatment MRI, acquired for primary staging. The institutional review board approved the study (IRBd18047). Informed consent was waived considering the retrospective design. Any identifiable information was removed.

All MRI scans were acquired on 1.5T (n=79) or 3.0T (n=92) MRI scanners (Achieva, Philips Medical System, Best, The Netherlands). The imaging protocol included: 2D T1-weighted fast spin-echo (T1w), 2D T2-weighted fast spin-echo with fat suppression (T2w) and 3D T1-weighted high-resolution isotropic volume excitation after gadolinium injection with fat suppression (T1gd). Further details on the MRI protocols are given in the Supplemental Material (table S.2). The primary tumors were manually contoured in 3D Slicer (version 4.8.0, https://www.slicer.org/) by one observer with 1 year of experience (P.B.). Afterwards, they were reviewed and adjusted, if needed, by a radiologist with 7 years of experience (B.J.). All tumor volumes were delineated on the T1gd but observers were allowed to consult the other sequences.

For the experimental set-up, we split the data set in three subsets: training set (n=131), validation set (n=20) and test set (n=20). The test set was not used for training or hyper-parameter tuning. We stratified the three subsets for tumor volume, subsite, and aspect ratio since these features are likely relevant for segmentation. Subsites were defined as tonsillar tissue, soft palate, base of tongue and posterior wall. Aspect ratio was defined as the ratio between the shortest and the longest axis of the tumor. All images were resampled to a voxel size of 0.8 mm x 0.8 mm x 0.8 mm.

Model architecture

The UNet architecture was chosen as the basis for our experiments because of the promising results on segmentation of medical structures [47,53–56]. Given the 3D nature of the images, we chose a 3D UNet as the architecture in this work [53,57]. We used Dice as loss function [58], the Adam optimizer [59] and early stopping. Dropout and data augmentation were used for regularization. Further details on the training procedure can be found in the Supplemental Material (Tables S.3. and S.4.).

Fully automatic approach

We trained the 3D UNet using the full 3D scans. We studied the effect of incorporating multiple MRI sequences into the training by introducing the available MRI sequences as input channels. Five networks were trained for the following MRI sequences and combinations thereof: T1w, where the tumor is hypo-intense but homogeneous; T2w, where the tumor is hyper-intense; T1gd, since the tumor presents with clearer boundaries; combining T1gd and T2w, and combining all sequences together (T1gd, T2w and T1w), to explore all the available information.

Semi-automatic approach

We proposed a semi-automatic approach in which we trained the networks with only the information within a clipbox around the tumor instead of with the full image as input.

During training, the clipbox was computed from the tumor delineations. First, the bounding box was calculated (i.e. the minimal box around the tumor). Then, random shifts of up to 25 mm were applied to all of the six directions to make clipboxes of different sizes and allow off-centered positioning of the tumors. We considered that shifts of more than 25 mm would represent unrealistic errors during clipbox selection. Examples of inputs possibly seen by the network are shown in Figure 1.

To study the clinical feasibility of this semi-automatic approach, two human observers were asked to manually select a clipbox around the tumor for each test set patient. The clipboxes were selected using 3D Slicer on the T1gd with access to the other sequences. The first observer (P.B.) had delineated the tumors two years earlier. The second observer was a radiation oncologist with 16 years of experience (A.A.) and had no information about the tumor delineations. To mitigate the risk of the observers defining too small clipboxes, cropping the tumor, the clipboxes were dilated 5 mm so as to ensure that they encompass the tumors. We consider it unlikely that a human observer would crop the tumor by more than 5 mm.

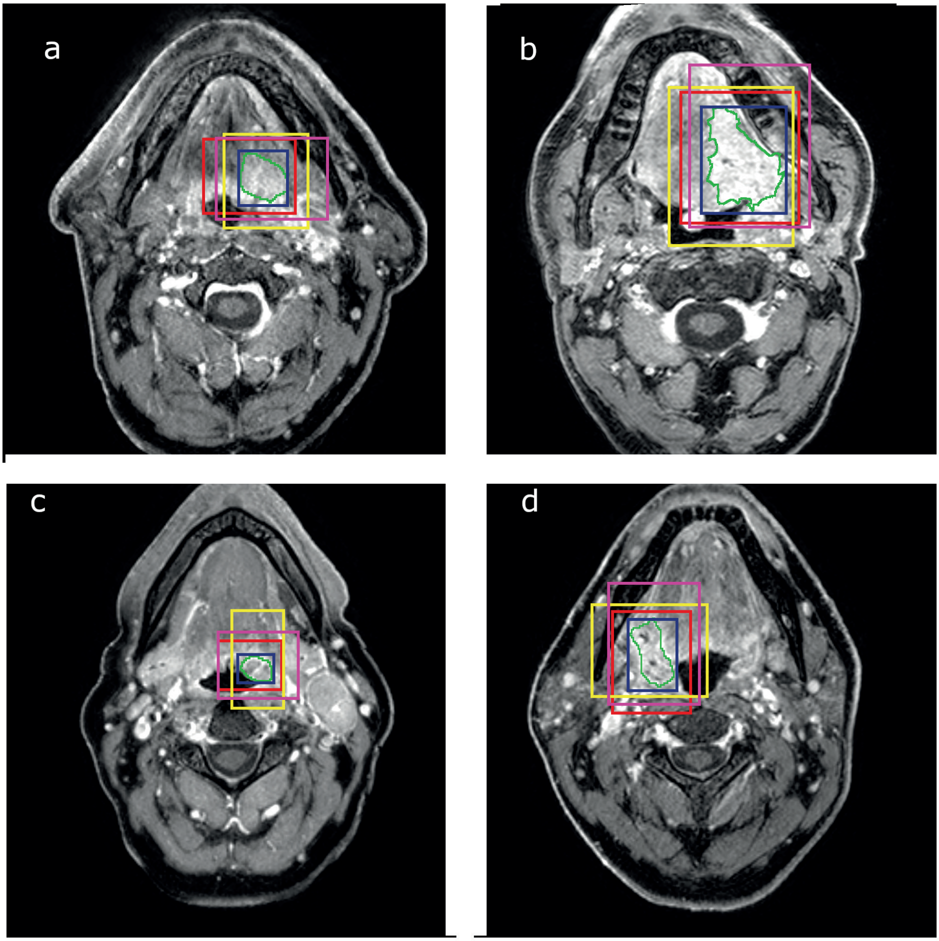


Figure 1. Original MRI image with the manual segmentation (green) of the oropharyngeal tumor. The blue boxes are the bounding boxes of the tumor. The rest of the boxes are used as inputs to the network during training.

Experiments

For the fully automatic approach, the performance of the networks trained with different sequences (T1w, T2w, T1gd, T1gd/T2w, and all sequences combined) was compared for the patients on the separate test set.

Because of memory constraints, scans were resized to a lower resolution by a factor of ~2.5 to 1.9 mm x 1.9 mm x 1.9 mm. Thus, even the smallest tumors were seen by the network. As a control experiment, to assess the impact of the resulting loss of resolution, we additionally trained a 2D UNet with full resolution axial slices. We checked for significant differences in performance of both approaches.

For the semi-automatic approach, one network was trained with all the sequences as input. The results with the clipboxes of the two observers were compared to the fully automatic approach experiment when combining all sequences as input (baseline).

To evaluate the robustness of the semi-automatic approach to off-centered tumors inside the clipboxes, we presented the trained model with increasingly shifted versions of the clipboxes, starting from the bounding box. The artificially induced shifts were applied in the 6 possible directions of the clipbox and expressed as two metrics: the centroid displacement and the relative difference in clipbox diagonal length before and after the shifts.

Statistics

To confirm that the three subsets were balanced in subsite, volume and aspect ratio, we used a Kruskal-Wallis test for continuous variables (volume and aspect ratio) and a chi-square test for independence for the categorical data (subsite).

Automatic contours were compared against the delineations from the human experts using common segmentation metrics: Sørensen–Dice coefficient (Dice), 95th Hausdorff Distance (HD) and Mean surface distance (MSD), implemented using the Python package from DeepMind (https://github.com/deepmind/surface-distance). Differences among experiments were assessed by the Wilcoxon signed-ranked test. P-values below 0.05 were considered statistically significant. Statistical analyses were performed with the SciPy package (version 1.1.0) and Python 3.6. Other relevant libraries can be found in the Supplemental Material (Table S.5.). The code is publicly available and can be found in: https://github.com/RoqueRouteiral/oroph_segmentation.git

RESULTS

Summary of tumor characteristics

Tumor characteristics (location, volume and aspect ratio) of our cohort are described in Table S.6. No significant differences were found in the distributions of subsite, volume and aspect ratio between the training, validation and test sets.

Fully automatic approach

As shown in Figure 2, combining all MR sequences resulted in the best performance, with a median Dice of 0.55 (range 0-0.78), median 95th HD of 8.7 mm (range 2.8-84.8

mm) and median MSD of 2.7 mm (range 1.0-26.8 mm), and the least variability among patients. The control experiment showed that by training a 2D UNet with full resolution scans the results were not significantly better than when using its 3D counterpart (Table S.7).

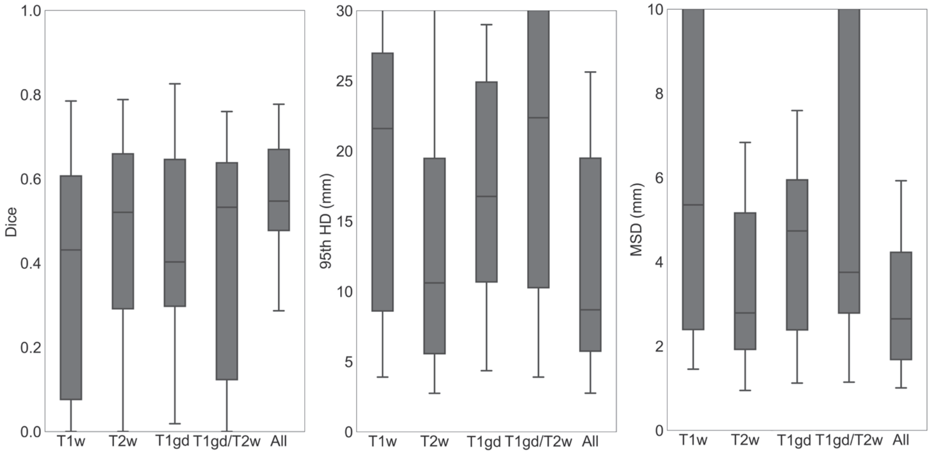


Figure 2. Segmentation performance in terms of Dice, 95th HD and MSD for the 3D. The different boxes show different MRI sequences as input: T1w(T1 weighted), T2w (T2 weighted), T1gd (T1 3D after gadolinium injection), T1gd and T2w (T1 3D after gadolinium injection and T2 weighted) and combining all sequences (All). The box includes points within the interquartile range (IQR) while the whiskers show points within 1.5 times the IQR.

Semi-automatic approach

In figure 3, it is observed that the semi-automatic approach using the boxes of the first observer achieved a median Dice score of 0.74 (range 0.32-0.80), HD of 4.6 mm (range 2.2 mm – 10.5 mm) and MSD of 1.2 mm (range 0.6 mm- 2.9 mm). For the second observer, the network achieved a median Dice score of 0.67 (range 0.28 – 0.87), HD of 7.2 mm (range of 3.0 mm – 19.9 mm) and MSD of 1.7 mm (range of 0.9 mm – 4.9 mm).

The semi-automatic approach significantly outperformed the fully automatic approach in all of the metrics for the first observer (p<0.001) and in Dice and MSD for the second observer (p<0.01). These results were expressed for 19 out of the 20 patients in the test set (also for the fully automatic approach - equivalent to "All' in figure 2), as one of the observers did not detect one of the tumors when asked to draw the clipbox.

The average time to draw the boxes was of 7.5 minutes per patient for the first observer and 2.8 minutes for the second observer.

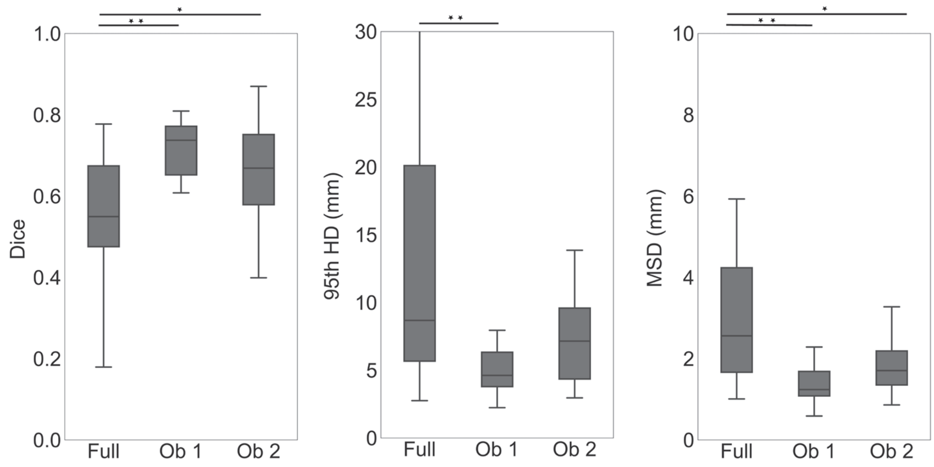


Figure 3. Segmentation performance of the semi-automatic approach with boxes drawn by two human observers. We compare the semi-automatic results (Ob 1 and Ob 2) to the fully automatic approach (baseline, Full). The box includes points within the interquartile range (IQR) while the whiskers show points within 1.5 times the IQR. Significance is represented as one star (*) for p<0.01 and two stars(**) for p<0.001.

Robustness to shifts

Figure 4 shows the segmentation performance of the network trained for the semi-automatic approach as a function of the artificially induced shifts applied to the tumor within the clipbox. For centroid displacements below 20 mm and diagonal length differences of between 25 mm and 60 mm the Dice was consistently greater than 0.70, the HD was lower than 6.5 mm and the MSD was lower than 1.7 mm.

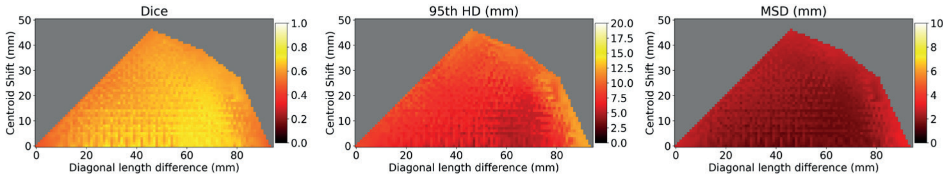


Figure 4. Robustness analysis. Segmentation performance in terms of median Dice, 95th HD and MSD for the semi-automatic approach as a function of the tumor centroid displacement and the clipbox diagonal length difference. The grey areas correspond to undetermined values due to the geometric constraints (i.e. no combination of shifts can achieve those values of centroid displacement and diagonal length difference).

Qualitative results

Figures 5a and 5b show examples in which the shape of the semi-automatic approach output and ground truth segmentation agreed while the fully automatic approach oversegmented (a) or undersegmented (b) the tumor. Figure 5c shows a case where the segmentation by the network trained with the fully automatic approach showed a similar shape to the ground truth segmentation but there were additional false positive volumes on the image.

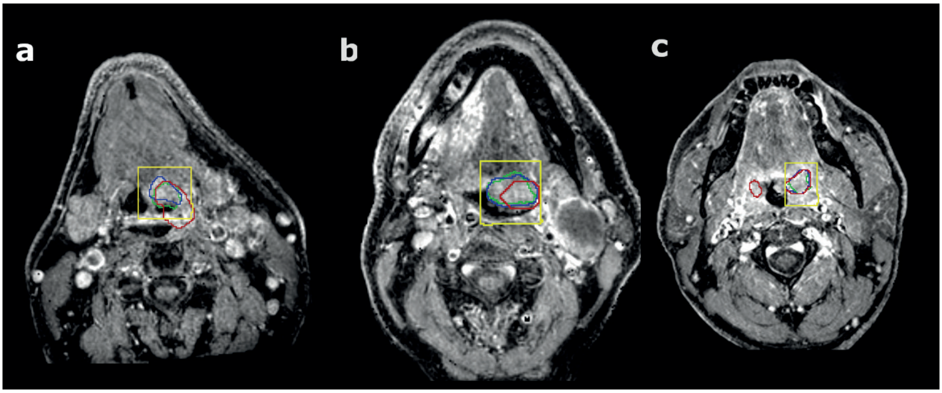


Figure 5. Comparison of the oropharyngeal segmentations in three different patients (a, b, c) trained with the fully automatic approach (red), with the semi-automatic approach (blue) and the manual delineation (green). The yellow boxes are the boxes drawn by the observer.

DISCUSSION

It was shown that using multiple MRI sequences yielded better results compared to using a single sequence as input. We also showed that decreasing the amount of context given to the CNN improved the segmentation performance. Finally, we proposed a functional semi-automatic approach that outperformed the fully automatic baseline and that was robust to clipbox selection errors, suggesting its potential clinical applicability.

Our network resulted in worse performance in terms of Dice compared to other tumor sites as reported by Sahiner et al [60], where the authors provide a comparison of CNN segmentations for different tumor/lesions (Dices: 0.51-0.92). However, lower performance for oropharyngeal tumor segmentation is consistent with what is known about the inter-observer variability for this subsite: Blinde et al [11] have shown differences in volume of up to 10 times among observers when segmenting OPSCC on MR, indicating the complexity of this task even for human observers. In this study, the mean Dice between our observers was 0.8. However, this number is an overestimation of the interobserver variability, considering that one of the observers corrected the other's delineation.

No significant differences were found between training the network with full context in 3D compared to its 2D counterpart. This shows that reducing the resolution due to memory constrains in the 3D case is not critical for the segmentation performance when the full image is used as input.

When restricting the context, the network outperformed significantly the full context approach for all metrics. This means that local textural differences between tumor and immediate surrounding tissues are sufficient for delineation.

Using clipboxes drawn by human observers demonstrates the feasibility of a semi-automatic approach for OPSCC primary tumor segmentation. Additionally, these boxes were drawn by two independent observers with different backgrounds and levels of expertise, suggesting that the method is not highly sensitive to the observer. This is supported by the results of our robustness analysis, which showed that when training with shifted versions of the clipbox, the networks were fairly robust to these shifts. More concretely, the network was robust centroid displacements below 20 mm and diagonal length differences of between 25 mm and 60 mm, which we consider a fair estimate of the maximum error an observer can make when selecting the clipbox.

A fully manual segmentation can take from 30 minutes to almost 2 hours (depending on the shape and size of the tumor), The average time between our two observers for the semi-automatic approach can take an average of 5 minutes (average of our two observers). Although after the proposed semi-automatic approach, some manual adaptations may be needed by a radiation oncologists to make the contours clinically acceptable, the overall process is expected to be less labor-intensive. Additionally, in the clinic it would be possible to use software designed to draw the clipboxes faster. Consequently, a functional semi-automatic system is not only feasible in terms of segmentation performance but also relevant for speeding up the radiotherapy workflow.

There are limitations in this study. First, given the high interobserver variability of OPSCC delineation, we are likely training the network with imperfect ground truths. However, we palliated the possible errors on the delineations by having the second observer correcting the first observer's delineation. Secondly, we used a standard 3D UNet in our studies. Despite the extensive literature on deep learning architecture modifications, investigating the best architecture for this task is outside of our scope. Thirdly, our results would need validation with an independent cohort in a multi-center study. Furthermore, the scan protocols were not standardized in our dataset. Arguably, that makes the network robust to such differences (e.g. TR/TE), given that the network has learnt from a diverse dataset.

Finally, our work can still be improved by adding other MRI sequences into the training (such as DWI) or by fully automatizing our semi-automatic approach, but we leave that as future work.

There is an increasing interest in the literature about differences on the tumors depending on their HPV status. According to Bos et al. [61], HPV positive tumors present on MRI post contrast with rounder shapes, lower maximum intensity values, and texture homogeneity. One strength of our work is that we include both HPV positive and HPV negative tumors in the training set, making the networks able to segment both subtypes of OPSCC. To check that the network is not biased to the HPV status, we compared the performance of the network stratified per HPV status and found non-significant results. We also did not find any relationship between performance and size.

In conclusion, this is the first study of primary tumor segmentation in the OPSCC site on MRI images with CNNs to the best of our knowledge. We trained a standard 3D UNet architecture using full MRI images as input. We showed that combining MRI sequences is beneficial for OPSCC segmentation with CNNs. Additionally, the CNN trained with reduced context around the tumor outperformed the fully automatic baseline and approaches that of other tumor sites reported in the literature. Hence, our proposed semi-automatic approach can save time in the clinic while achieving competitive performance and being robust to the choice of observer and manual clipbox selection errors.

SUPPLEMENTAL MATERIAL

Table S.1. Tumor stage and HPV status.

	N patients (percentage in %)
Tumor staging	
T1	27 (15.79)
T2	64 (37.43)
T3	42 (24.56)
T4	38 (22.22)
N-stage	
N0	34 (19.88)
N1	26 (15.20)
N2	107 (62.57)
N3	4 (2.33)
Subsite	
Tonsillar tissue	96 (56.14)
Soft palate	16 (9.36)
Base of tongue	55 (32.16)
Posterior wall	4 (2.34)
HPV status	
Positive	73 (42.69)
Negative	76 (44.44)
Unknown	22 (12.87)

Table S.2. Overview of MRI sequences. The MRI sequences are 2D T1 weighted (T1w), 2D T2 weighted with fat suppression (T2w) and 3D T1 weighted after gadolinium injection with fat suppression (T1gd).

	TR (ms)	**TE (ms)**	**Pixel size (mm)**	**Slice thickness (mm)**
T1w	180-892	2.3-10	0.417- 0.9375	3-5
T2w	1963-6880	20-90	0.417- 0.9	3-5
T1gd	4.3-10	1.7-4.6	0.197- 0.976	0.8-1

Training details

Table S.3. 3D UNets – For full, reduced context and semi-automatic approach.

Hyper-parameter / set-up	
Optimizer	Adam
Loss function	Dice
Initial learning rate	0.001
Learning rate scheduler	Multiply by 0.5 if validation loss does not decrease in ten epochs by an amount of 0.001
Batch size	1* (*Batch normalization with running mean and variance during inference time, because of stability issues during training with batch size of 1)
Dropout	0.2 in bottleneck convolutions
Data augmentation (only full context)	Horizontal flip (in coronal view) with a chance of 0.5 in every epoch. Random elastic deformation Using elasticdeform library (https://github.com/gvtulder/elasticdeform) Random rotations between -10 and 10 degrees in every epoch.

Table S.4. Control experiment with 2D UNet. Hyper-parameter / set up.

Hyper-parameter / set-up	
Optimizer	Adam
Loss function	Dice
Initial learning rate	0.001
Learning rate scheduler	Multiply by 0.5 if validation loss does not decrease in ten epochs by an amount of 0.001
Batch size	16
Dropout	0.2 in bottleneck convolutions
Data augmentation	Horizontal flip (in coronal view) with probability of 0.5 in every epoch. Random elastic deformation using the elasticdeform library (https://github.com/gvtulder/elasticdeform) Random rotations between -10 and 10 degrees in every epoch. Given the differences in histograms of the axial slices, the images were preprocessed by whitening and normalization. All zero slices were removed from the 2D dataset

Table S.5. Library versions.

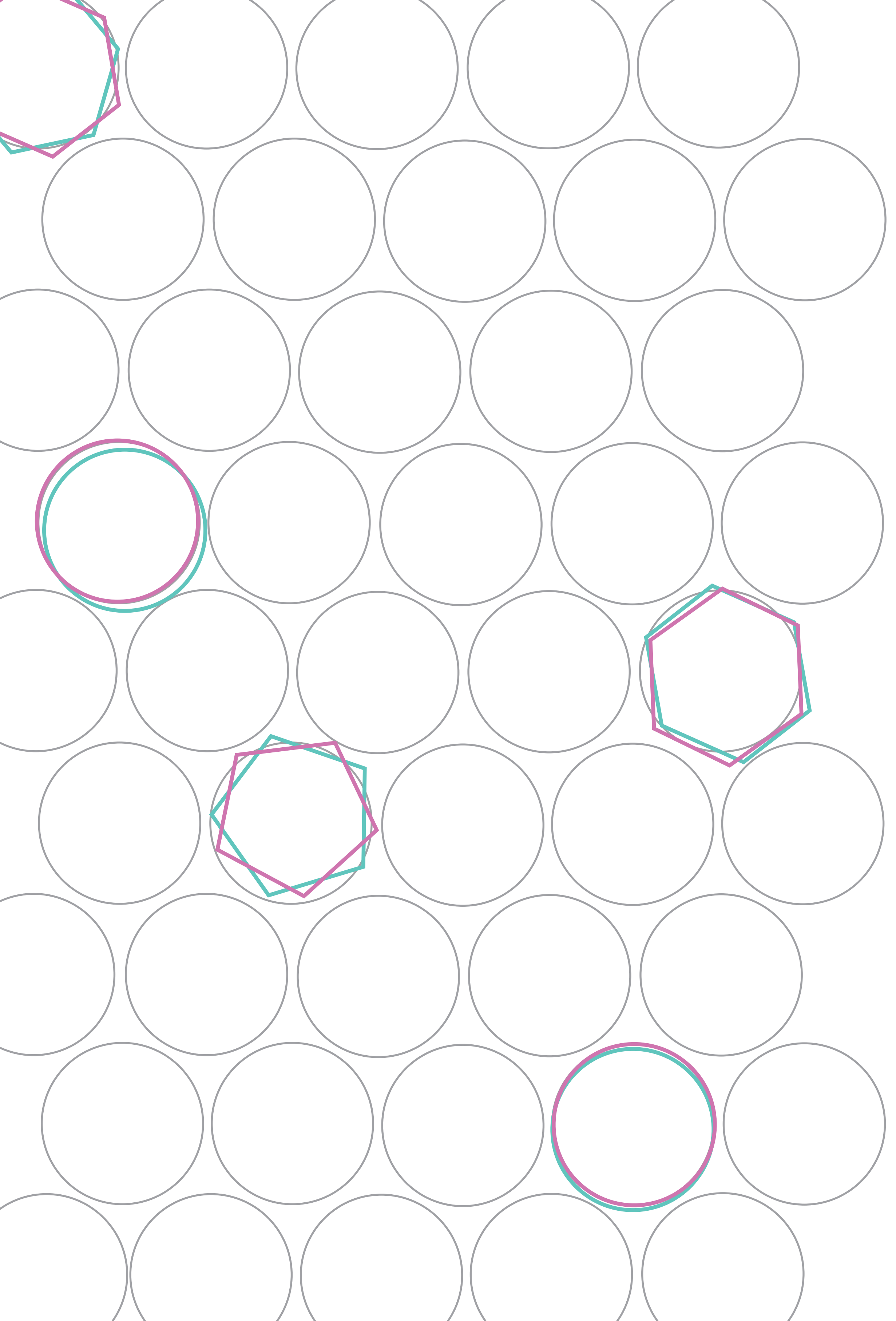
Library/Application	Version
Python	3.6.5
Pytorch	0.4.1
Cuda	9.2

Table S.6. Overview of tumor characteristics per data set.

	Training set	Validation set	Testing set	p-value
Patients (n)	131	20	20	
Subsite				0.700
Tonsillar tissue	76 (58%)	10 (50%)	10 (50%)	
Soft palate	10 (8%)	3 (15%)	3 (15%)	
Base of tongue	41 (31%)	7 (35%)	7 (35%)	
Posterior wall	4 (3%)	0 (0%)	0 (0%)	
Volume (cm^3)				0.553
Median	6.93	7.33	6.87	
Range	[0.27,67.2]	[0.51, 41.7]	[0.46, 17.17]	
Aspect ratio (%)				0.589
Median	52.22	53.87	54.77	
Range	[17.68 , 90]	[41.47, 64.36]	[47.63, 84.25]	

Table S.7. Control experiment with 2D UNet. Median values of the segmentation metrics for the control experiment done with 2D UNet and full resolution axial slides as input. P values refer to the comparison with the 3D network trained with full context and with all sequences. P value for significance after Bonferroni correction: $p<0.0041$. The MRI sequences are 2D T1 weighted (T1w), 2D T2 weighted with fat suppression (T2w) and 3D T1 weighted after gadolinium injection with fat suppression (T1gd).

Full context 2D	Dice (p-value)	HD [mm] (p-value)	MSD [mm] (p-value)
T1gd	0.51 (0.08)	25.48 (0.07)	3.53 (0.19)
T2w	0.45 (0.04)	22.54 (0.005)	3.92 (0.025)
All	0.62 (0.49)	7.93 (0.55)	2.13 (0.16)



Chapter 3

Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging

Rodríguez Outeiral R, Bos P, van der Hulst HJ, Al-Mamgani A, Jasperse B, Simões R, et al. Phys Imaging Radiat Oncol. 23;144–9 (2022)

ABSTRACT

Background and purpose

Contouring of the oropharyngeal primary tumor is currently done manually which is time-consuming. Autocontouring techniques based on deep learning methods are a desirable alternative, but these methods can render suboptimal results when the structure to segment is considerably smaller than the rest of the image. The purpose of this work was to investigate different strategies to tackle the class imbalance problem in this tumor site.

Materials and methods

A cohort of 230 oropharyngeal cancer patients treated between 2010 and 2018 was retrospectively collected. The following magnetic resonance imaging (MRI) sequences were available: T1-weighted, T2-weighted, 3D T1-weighted after gadolinium injection. Two strategies to tackle the class imbalance problem were studied: training with different loss functions (namely: Dice loss, Generalized Dice loss, Focal Tversky loss and Unified Focal loss) and implementing a two-stage approach (i.e. splitting the task in detection and segmentation). Segmentation performance was measured with Sørensen–Dice coefficient (Dice), 95th Hausdorff distance (HD) and Mean Surface Distance (MSD).

Results

The network trained with the Generalized Dice Loss yielded a median Dice of 0.54, median 95th HD of 10.6 mm and median MSD of 2.4 mm but no significant differences were observed among the different loss functions (p-value > 0.7). The two-stage approach resulted in a median Dice of 0.64, median HD of 8.7 mm and median MSD of 2.1 mm, significantly outperforming the end-to-end 3D U-Net (p-value < 0.05).

Conclusions

No significant differences were observed when training with different loss functions. The two-stage approach outperformed the end-to-end 3D U-Net.

INTRODUCTION

Radiotherapy is one of the common treatment options for head and neck cancer patients [4,41]. One key step of the radiotherapy workflow is tumor contouring. While contouring of organs at risk is increasingly being automated in clinical practice, tumor contouring is still done manually. This is time consuming and suffers from high interobserver variability [11].

Deep learning methods, particularly Convolutional Neural Networks (CNNs), are the current state of the art for automatic segmentation of medical images. Several review papers have been published on deep learning applied to radiotherapy and automatic segmentation is often discussed as one of the main applications [31,60,62,63]. For the particular case of head and neck cancer, various works have focused on the automatic segmentation of organs at risk with deep learning [64], some of them achieving clinically acceptable performance and being commercially available [18]. For the case of tumor contouring, the literature is more scarce and those algorithms are still not implemented in the clinic.

In our previous work [65], we segmented the oropharyngeal primary tumor on magnetic resonance imaging (MRI) and showed that combining multiple anatomical MRI sequences improved the segmentation performance compared to single-sequence. We also proposed a semi-automatic approach that improved the segmentation performance by splitting the segmentation task in manual detection and segmentation. To the best of our knowledge, there is only one other work where the authors segmented the oropharyngeal primary tumor on MRI [42]. The authors studied the impact of combining different anatomical (T1 weighted and T2 weighted) and quantitative images (ADC, Ktrans and ve) as input channels to a CNN and showed that combining anatomical sequences significantly improved the performance.

A known issue in the field of deep learning for medical image segmentation is class imbalance, meaning that the structure to be segmented is present in a smaller amount of voxels compared to the rest of the image. Class imbalance can result in suboptimal solutions because the network is exposed to proportionally less relevant information during the training process. Several works in the field of medical image segmentation have focused on this problem, either by modifying the input data to the network [66,67] or by defining different loss functions [68–70]. This problem is even more critical in the case of tumor segmentation, given that tumors tend to be smaller than other structures and they are heterogeneous in their location, shape and size. This is also the case for the oropharyngeal primary tumor.

Several loss functions have been designed with the aim of tackling class imbalance, such as the Generalized Dice loss [58],the Focal loss [68], the (focal) Tversky loss [69,71] and the Unified Focal loss [70]. Although the choice of the loss function can be critical for the training of a CNN, comprehensive loss function comparisons for specific tumor sites or anatomies are not commonly performed. Ma et al. [30] showed that the influence in performance of the loss function varies greatly depending on the segmentation task. To the best of our knowledge, this has not been studied yet in the particular case of oropharyngeal cancer segmentation.

Other works have implemented two-stage approaches (i.e. detection and segmentation) that resulted in more accurate segmentations than their one-stage counterparts [72–74]. By locating the tumor first, the context around the tumor is reduced. Consequently, two-stage approaches are a possible way of tackling class imbalance. The semi-automatic approach from our previous work [65] consisted of having human observers outlining a box around the tumor to provide a first approximation of the tumor location and consequently ease the segmentation task. However, the semi-automatic approach still needed manual intervention. The implementation of a two-stage approach will also allow us to fully automate the semi-automatic approach proposed in our previous work [65].

The aim of this study was to investigate two different strategies for tackling the class imbalance problem for oropharyngeal primary tumor segmentation: training with different loss functions and implementing a fully automatic two-stage approach.

MATERIALS AND METHODS

Data

A cohort of 230 patients treated at our institute between January 2010 and May 2018 was used for this project. The mean age of the patients was 61 years (standard deviation ± 7 years) and 66 % of the patients were male. Further details on tumor stage and HPV status can be found in the Supplemental Material (table S.1). All patients had histologically proven primary oropharyngeal squamous cell carcinoma and received a pre-treatment MRI for primary staging. The institutional review board approved the study (IRBd18047). Informed consent was waived by the institutional review board considering the retrospective design. The cohort was extended from our previous work [65]. A total of 59 new patients were included.

The scans were acquired on 1.5T (n=108) or 3.0T (n=122) MRI scanners (Philips Medical System, Best, The Netherlands). The imaging protocol included: 2D T1-weighted fast spin-echo, 2D T2-weighted fast spin-echo with fat suppression, 3D T1-weighted high-resolution isotropic volume excitation after gadolinium injection with fat suppression Further details on the MRI protocols are given in the Supplemental Material (table S.2). The primary tumors were manually contoured in 3D Slicer (version 4.8.0, https://www.slicer.org/) by one observer with 1 year of experience (P.B. or H.H.). Afterwards, they were reviewed and adjusted, if needed, by a radiologist with 7 years of experience (B.J.).

All tumor volumes were delineated on the T1gd but the observers were allowed to consult the other sequences.

For the experimental set-up, the data set was split in three subsets: a training set (n=190), a validation set (n=20) and a test set (n=20). The test set was not used for training or hyper-parameter tuning. We stratified the three subsets for tumor volume, subsite, and aspect ratio since these features are likely relevant for segmentation. Subsites were defined as tonsillar tissue, soft palate, base of tongue and posterior wall. The aspect ratio was defined as the ratio between the shortest and the longest axis of the tumor. All images were resampled to a voxel size of 0.8 mm x 0.8 mm x 0.8 mm.

Baseline model architecture

The 3D U-Net architecture [53,57] was used as the basis for our experiments. The Adam optimizer [59] and early stopping were used for training. Dropout and data augmentation were used for regularization. Further details on the training procedure can be found in table S.4. and in the code which is publicly available in: https://github.com/RoqueRouteiral/oroph_segm_ts.

Training with different loss functions.

We trained the 3D U-Net with four different loss functions: Dice loss [75], Generalized Dice loss[58], Focal Tversky loss [71] and Unified Focal loss [70]. For the particular case of the Unified Focal loss, Yeung et al. [70] showed that the choice of the γ hyperparameter can affect the performance. Consequently, we trained four networks with the Unified Focal loss for different values of its hyperparameter γ (γ = [0.2, 0.4, 0.6, 0.8]). We compared the segmentation performance of all the networks among each other.

Two-stage approach

In our previous work, we demonstrated that the segmentation of the oropharyngeal primary tumor was more accurate when the input image was manually cropped with a clipbox around the tumor before being fed to a segmentation network.

In this work, we fully automated this two-stage approach (figure 1). The first stage consisted of roughly detecting the tumor by automatically selecting a clipbox around it. In the second stage, this clipbox was used to crop the image which was then used as input to a segmentation network. The loss function chosen for both stages was the Generalized Dice loss function. The loss was backpropagated through each network separately.

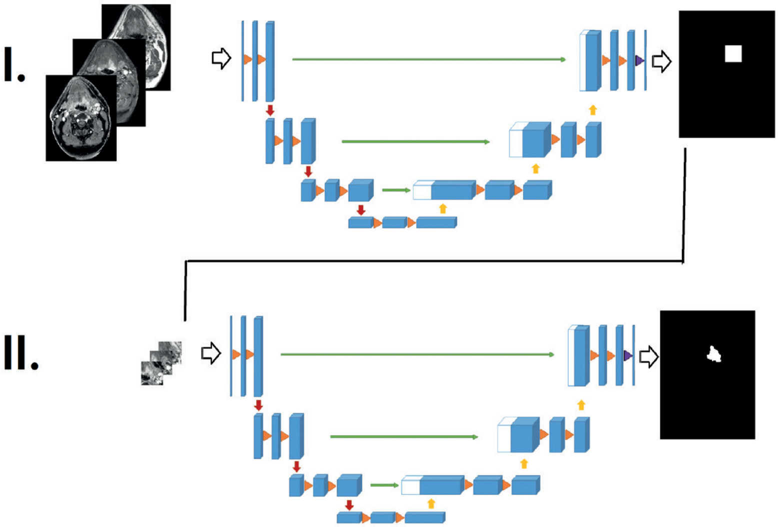


Figure 1. Overview of the two-stage approach.

For the detection stage, a 3D U-Net was trained using the bounding box of the tumor as ground truth. At inference time, the output of the detection was computed as the bounding box of the output.

For the segmentation stage, the same architecture as in our previous work was used [65]. This segmentation network was trained with only the information contained inside the clipboxes. In every training iteration, the clipboxes were randomly shifted by an amount of up to 25 mm in different directions to make the network robust to possible displacements in the detection. At inference time the input images were cropped by the clipboxes defined by the detection network. Similarly to our previous work, the clipboxes were dilated by 5 mm.

Statistics

To confirm that the three subsets were balanced in subsite, volume and aspect ratio, a Kruskal-Wallis test was used for continuous variables (volume and aspect ratio) and a chi-square test for independence for the categorical data (subsite).

Predicted segmentations and the segmentations from the human observers were compared for the patients on the unseen test set. Common segmentation metrics were used: Sørensen–Dice coefficient (Dice), 95th Hausdorff Distance (HD) and Mean surface distance (MSD). The metrics were implemented using the Python package from DeepMind (https://github.

com/deepmind/surface-distance). For the two-stage approach, the detection was evaluated by measuring the absolute mean shift in all 6 directions between the tumor bounding box and the detected clipbox for the patients on the unseen test set. The average shift of the boxes for the observers from our previous work was used for comparison [65]. Differences among the loss function experiments were assessed by the Friedman test whereas the two-stage approach experiments were assessed by the Wilcoxon signed-ranked test. P-values below 0.05 were considered statistically significant.

All networks were retrained four times. Reported results are the mean of the results of the four versions of each network. We opted for this approach over N-fold cross-validation to account for the random initialization of the network while ensuring proper stratification in the three sets for all the folds.

RESULTS

Summary of tumor characteristics

Table S.3 shows the tumor characteristics (location, volume and aspect ratio) of our cohort. No significant differences were found in the distributions of subsite, volume and aspect ratio between the training, validation and test sets.

Training with different loss functions

When comparing the performance of the networks trained with different loss functions no significant differences were found (p-value > 0.25 for the three metrics). Lower variance in the MSD and Dice can be observed for the network trained with the Generalized Dice loss (figure 2). The network achieved a median Dice of 0.54, median 95th HD of 10.6 mm and median MSD of 2.4 mm. Non-significant differences were observed when training the network with different γ values for the Unified Focal loss (Figure S.1).

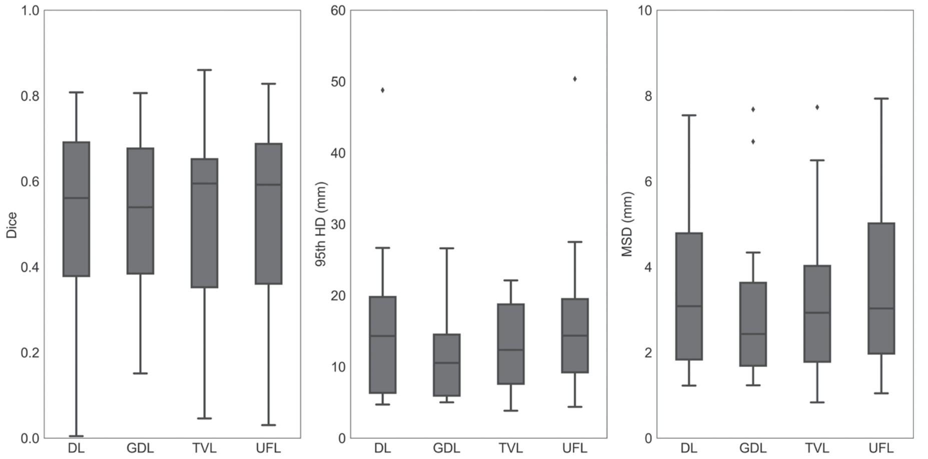


Figure 2. Segmentation performance of the 3D U-Net trained with different loss functions: Dice Loss (DL), Generalized Dice Loss (GDL), Tversky Loss (TVL) and Unified Focal Loss (UFL).

Two-stage approach

The mean shift for the detection network was of 8.9 mm (Table 1) and no significant differences were found when comparing to the detection of observer 2 from our previous work (p-value = 0.40). Significant differences were found when comparing the detection of this work to the detection of the observer 1 from out previous work (p-value<0.001). When separating the mean shift per direction, we observed a mean shift of 10.0 mm in the cranial-caudal direction, 8.4 mm in the medial-lateral direction and 7.7 mm in the dorsal-ventral direction.

Table 1. Detection and segmentation performance of the two-stage approach and comparison to results of the previous work [65].

	Detection	**Segmentation**		
	Avg. shift (mm) – [SD]	**Dice**	**HD (mm)**	**MSD (mm)**
This work				
3D end-to-end UNet	--	0.54	10.6	2.4
Two-stage approach	8.7 [8.2]	0.64	8.7	2.1
Previous work				
Semi-automatic approach (Obs. 1)	3.0 [3.9]	0.74	4.6	1.2
Semi-automatic approach (Obs. 2)	8.9 [6.9]	0.67	7.2	1.7

The segmentation results of the two-stage approach were significantly better for Dice (p-value = 0.03) and MSD (p-value = 0.02) than the results of the end-to-end 3D UNet (Table 1). The fully automated two-stage approach yielded a median Dice of 0.64, median HD of 8.7 mm and median MSD of 2.1 mm. One patient was missed in the detection of the two-stage approach for one of the folds, and thus removed from that fold for the analysis.

Qualitative results

Examples of segmentations obtained by the end-to-end 3D U-Net, the two-stage approach and ground truth segmentation are shown in Figure 3. The end-to-end 3D U-Net approach oversegmented (Figure 3a-c) the tumor, where the two-stage approach showed better segmentation comparison to the ground truth. Figure 3b shows cases where the segmentation end-to-end 3D U-Net rendered additional false positive structures on the image.

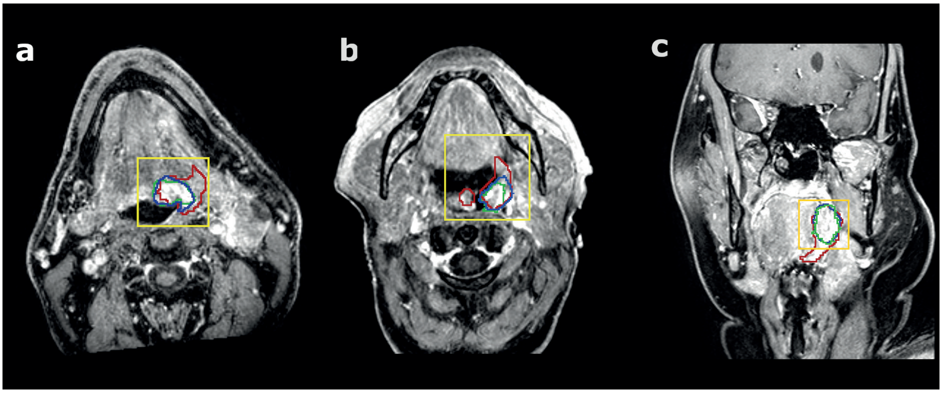


Figure 3. Comparison of the oropharyngeal segmentations in three different patients (a, b, c) trained with the end-to-end 3D U-Net (red), with the two-stage approach (blue) and the manual delineation (green). The yellow boxes are drawn by detection network from the two-stage approach. All the images correspond to the 3D T1gd sequence.

DISCUSSION

This work investigated two different strategies to tackle the class imbalance problem for the task of oropharyngeal primary tumor segmentation: training with the different loss functions and implementing a two-stage approach. Additionally, the proposed two-stage approach fully automated the semi-automatic approach described in our previous work [65].

When training the networks with different loss functions, no significant improvements were observed in the segmentation metrics. Hyperparameter tuning for the γ hyperparameter of the Unified Focal loss did not yield significantly better results either. This result is consistent with the work of Ma et al. [30], where they concluded that Dice-related losses are often optimal for medical image segmentation tasks. Additionally, it is also in line with the conclusions described by Isensee et al. and their proposed "no new Net" (nnU-Net) [76]. They showed that a tailored-to-task method configuration is more relevant than specific setup choices when designing a segmentation deep learning pipeline.

The two-stage approach achieved significantly better results compared to the conventional end-to-end approach. The high complexity of the task may make the end-to-end training of the network suboptimal, while focusing on two simpler tasks can render better results. In our previous work [10], a semi-automatic approach in which an observer selected a clipbox around the tumor was implemented. When comparing the current detection results to the semi-automatic approach of our previous work, we noted that one of the observers (Obs. 1) selected a tighter box (although all the tumors were included inside the clipboxes) compared to that of our two-stage approach which resulted in significantly different detection performance. However, we did not observe significant differences with the detection performance of the semi-automatic approach for the other observer (Obs. 2), showing that a fully automatic two-stage approach can be a feasible alternative to a semi-automatic approach. Also, the time spent on delineating in the clinical practice is aimed to be as low as possible. We reported in our previous work that the time spent on drawing the boxes was lower for observer 2 than for observer 1, making the delineations of observer 2 a more realistic representation of what is expected in the clinic. In the present work, the whole pipeline is automated, which can save time in the clinic. That said, further efforts in improving the detection are of interest to improve the segmentation performance of the two-stage approach.

The literature on automatic segmentation for the oropharyngeal tumor on MRI is scarce and its aims are heterogeneous. Besides our previous work [65], only Wahid et al. [42] have focused on the segmentation of this tumor site on MRI. Their work focused on studying the value of multiparametric MRI on the segmentation performance, both for qualitative and quantitative imaging. Other works focused on the automatic segmentation on multiparametric MRI of the head and neck cancer in general, rather than on the particular subset of oropharyngeal cancer: Bielak et al. [77] used diffusion weighted imaging while Schouten et al.[78] proposed a multiview CNN architecture. To the best of our knowledge, only our work is focused on tackling the class imbalance problem for head and neck cancer segmentation on MRI, and particularly for the oropharyngeal subsite.

In 2020, the first head and neck tumor segmentation challenge, known as HECKTOR challenge, was launched [79]. The main subsite of the challenge was the oropharyngeal tumor and the winner of the challenge achieved a mean Dice of 0.76, but the image modalities used were PET/CT. Additionally, Ren et al. [80] compared the use of PET/CT/MRI as different input image combinations for the automatic segmentation of head and neck GTV and observed that, when including PET, the segmentation performance improved. Considering all the above, it is possible that PET is a useful modality for the task of head and neck tumor segmentation. However, the differences in resolution between imaging modalities may be reflected in the detail of the manual ground truth delineations used for training and evaluation. Potentially, this can also explain the difference in performance of the MRI-based task. That said, we argue that the strategies to tackle class imbalance in this work can be useful in the development of autocontouring tools for the case of oropharyngeal cancer.

This study has limitations. Firstly, there is a high interobserver variability on this tumor subsite, especially in case of tonsillar fossa and base of tongue tumor which are rich in lymphatic tissue, so it is possible that the ground truth delineations used in this work are partially biased. However, one observer corrected the other's delineation, reducing this observer variation. Secondly, validation of our results is still needed with an independent cohort in a multi-center study. Thirdly, the performance could also be improved by making different decisions on the training setup, such as using larger batch sizes or non downsampled data, but other strategies to mitigate memory limitations would be needed. Finally, there is a certain variability in the scan protocols. However, variability in the training set can be desirable as it makes the network robust to protocol differences.

In conclusion, the loss functions designed to tackle class imbalance performed comparably among each other. The approach of splitting the problem into localization and segmentation outperformed the end-to-end network, proving an effective strategy to mitigate the class imbalance problem in oropharyngeal cancer segmentation.

SUPPLEMENTAL MATERIAL

Table S.1. Tumor stage and HPV status.

	N patients (percentage in %)
Tumor staging	
T1	38 (16.52)
T2	85 (36.96)
T3	52 (22.61)
T4	55 (23.91)
N-stage	
N0	44 (19.13)
N1	33 (14.35)
N2	149 (64.78)
N3	4 (1.74)
Subsite	
Tonsillar tissue	121 (52.61)
Soft palate	21 (9.13)
Base of tongue	80 (34.78)
Posterior wall	8 (3.48)
HPV status	
Positive	106 (46.09)
Negative	101 (43.91)
Unknown	23 (10.0)

Table S.2. Overview of MRI sequences. The MRI sequences are 2D T1 weighted (T1w), 2D T2 weighted with fat suppression (T2w), 3D T1 weighted after gadolinium injection with fat suppression (T1gd).

	TR (ms)	**TE (ms)**	**Pixel size (mm)**	**Slice thickness (mm)**
T1w	[180 – 890]	[2 - 10]	[0.4 – 0.9]	[3.0 – 5.0]
T2w	[2500 – 16000]	[20 – 90]	[0.4 – 0.9]	[3.0 – 5.0]
T1gd	[4 – 10]	[2 – 5]	[0.2– 1.0]	[0.8 – 2.0]

Table S.3. Overview of tumor characteristics per data set. [*]p-value was calculated using chi-square test. [**] p-value was calculated using Kruskal-Wallis test

	Training set	**Validation set**	**Testing set**	**p-value**
Patients (n)	190	20	20	
Subsite				0.757*
Tonsillar tissue	93 (53.2 %)	10 (50.0 %)	10 (50.0 %)	
Soft palate	14 (7.90 %)	3 (15.0 %)	3 (15.0 %)	
Base of tongue	63 (34.7 %)	7 (35.0%)	7 (35.0 %)	
Posterior wall	7 (4.20 %)	0 (0 %)	0 (0 %)	
Volume (cm^3)				0.465**
Median	7.54	7.40	6.88	
Range	[0.23,71.54]	[0.51, 41.6]	[0.46, 17.20]	
Aspect ratio (%)				0.350**
Median	52.17	53.54	55.61	
Range	[16.21, 90.67]	[41.53, 64.35]	[47.47, 84.28]	

Table S.4. Training details of the networks of the paper.

Hyper-parameter / set-up	
Optimizer	Adam
Loss function	Dice Loss / Unified Focal Loss/ Generalized Dice Loss/ Focal Tversky Loss
Initial learning rate	0.001
Learning rate scheduler	Multiply by 0.5 if validation loss does not decrease in ten epochs by an amount of 0.001
Batch size	1* (*Batch normalization with running mean and variance during inference time, because of stability issues during training with batch size of 1)
Dropout	0.2 in bottleneck convolutions
Data augmentation (only full context)	Horizontal flip (in coronal view) with a chance of 0.5 in every epoch. Random elastic deformation Using elasticdeform library (https://github.com/gvtulder/elasticdeform) Random rotations between -10 and 10 degrees in every epoch. Downsampled by a factor of 2.5
Shifts (only second stage of two-stage approach)	For second stage of the training of the two stage approach, random shifts of the tumor within the box of up to 25 mm.

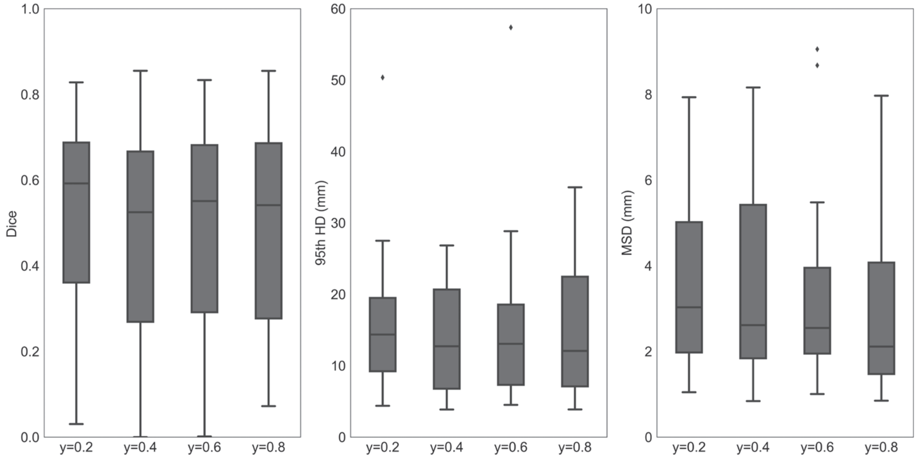
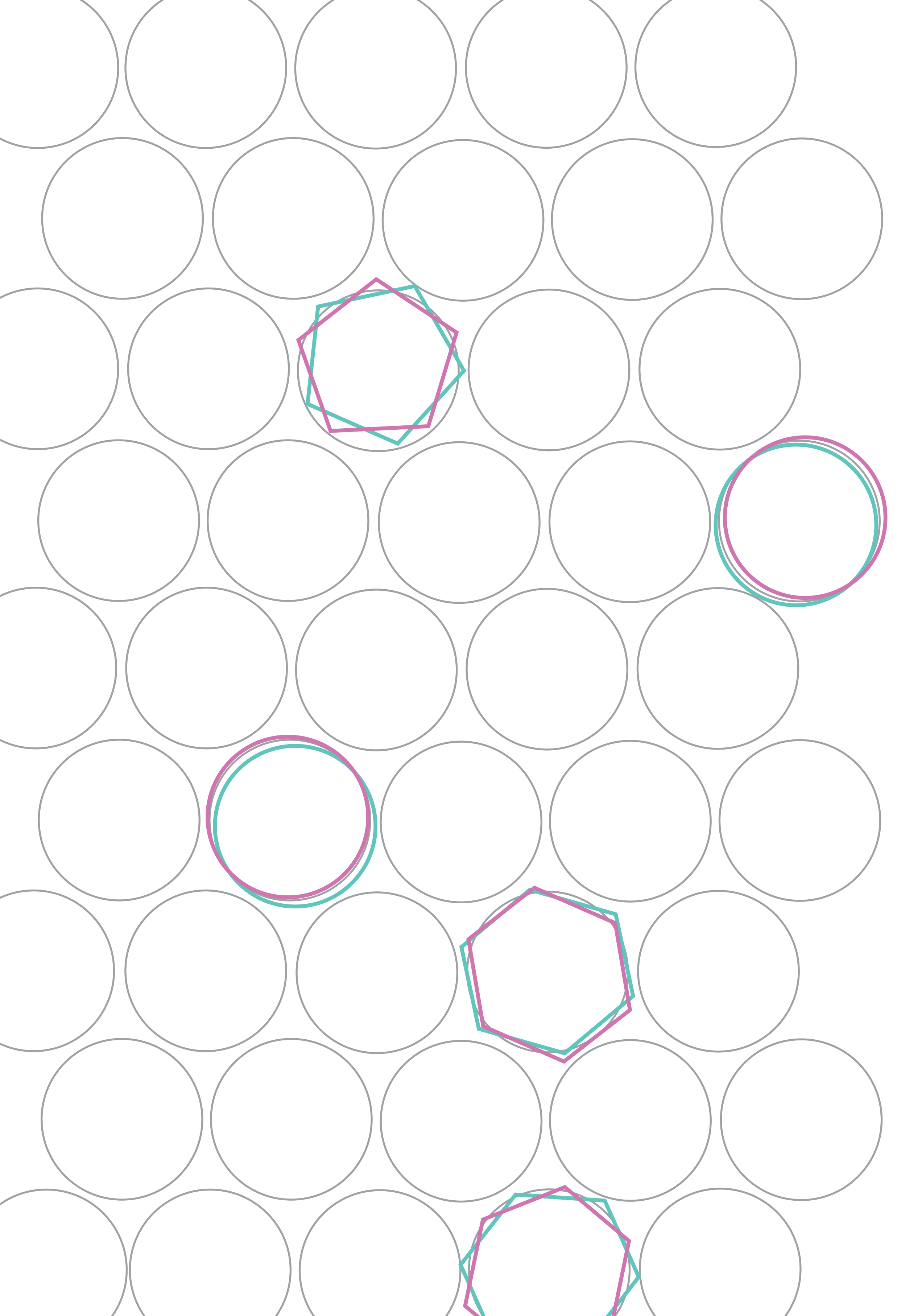


Figure S.1. Segmentation performance of the network trained with Unified Focal Loss for different values of γ.



Chapter 4

Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy

Rodríguez Outeiral R, González PJ, Schaake EE, van der Heide UA, Simões R.

Radiat Oncol. 1:18. (2023)

ABSTRACT

Background and purpose

Segmentation of the Gross Tumor Volume (GTV) is a crucial step in the brachytherapy (BT) treatment planning workflow. Currently, radiation oncologists segment the GTV manually, which is time-consuming. The time pressure is particularly critical for BT because during the segmentation process the patient waits immobilized in bed with the applicator in place. Automatic segmentation algorithms can potentially reduce both the clinical workload and the patient burden. Although deep learning based automatic segmentation algorithms have been extensively developed for organs at risk, automatic segmentation of the targets is less common. The aim of this study was to automatically segment the cervical cancer GTV on BT MRI images using a state-of-the-art automatic segmentation framework and assess its performance.

Materials and Methods

A cohort of 195 cervical cancer patients treated between August 2012 and December 2021 was retrospectively collected. A total of 524 separate BT fractions were included and the axial T2-weighted (T2w) MRI sequence was used for this project. The 3D nnU-Net was used as the automatic segmentation framework. The automatic segmentations were compared with the manual segmentations used for clinical practice with Sørensen–Dice coefficient (Dice), 95th Hausdorff distance (95th HD) and mean surface distance (MSD). The dosimetric impact was defined as the difference in D98 (ΔD98) and D90 (ΔD90) between the manual segmentations and the automatic segmentations, evaluated using the clinical dose distribution. The performance of the network was also compared separately depending on FIGO stage and on GTV volume.

Results

The network achieved a median Dice of 0.73 (interquartile range (IQR) = 0.50 - 0.80), median 95th HD of 6.8 mm (IQR = 4.2 – 12.5 mm) and median MSD of 1.4 mm (IQR = 0.90 - 2.8 mm). The median ΔD90 and ΔD98 were 0.18 Gy (IQR = -1.38 – 1.19 Gy) and 0.20 Gy (IQR =-1.10 – 0.95 Gy) respectively. No significant differences in geometric or dosimetric performance were observed between tumors with different FIGO stages, however significantly improved Dice and dosimetric performance was found for larger tumors.

Conclusions

The nnU-Net framework achieved state-of-the-art performance in the segmentation of the cervical cancer GTV on BT MRI images. Reasonable median performance was achieved geometrically and dosimetrically but with high variability among patients.

INTRODUCTION

For locally advanced cervical cancer the standard of care consists of external beam radiotherapy (EBRT), followed by 3 to 4 fractions of brachytherapy (BT) and concomitant chemotherapy [43]. A key step in both EBRT and BT treatment planning is the segmentation of organs at risk and target volumes. This is mostly performed manually, which is time consuming and suffers from the inherent bias of the observer. To circumvent these issues, automatic segmentation is being widely investigated in the field of radiotherapy [31,60,63]For the case of BT, the need for automatic segmentation is even more critical due to the time constraints of the workflow. At each fraction of BT treatment, the applicator is inserted surgically in the patient, after which the MRI images are acquired. The patient then needs to wait, immobilized in bed, while the needed structures (namely organs at risk and target volumes) are manually delineated and a treatment plan is made. The Gynecological (GYN) GEC-ESTRO working group defines the target volumes of interest for BT treatment planning for this cervical cancer as the Gross Tumor Volume (GTV), the high risk Clinical Target Volume (HR-CTV) and the intermediate risk Clinical Target Volume (IR-CTV) [81] and they are currently segmented by radiation oncologists. Automatic image segmentation methods are expected to reduce the clinical workload as well as patient burden.

Automatic segmentation of the targets volumes is still uncommon and it is mostly limited to positron emission tomography (PET) and/or computed tomography (CT) and rarely to magnetic resonance imaging (MRI) [9]. For the particular case of cervical cancer on BT images, automatic segmentation of the organs at risk has been widely investigated [20,26,82–84] but literature on the automatic segmentation of the targets, and especially the GTV, is more scarce [82–84]. Zhang et al. [84] and Wong et al. [83] developed automatic segmentation tools that segmented the HR-CTV (on CT and MRI images, respectively) but to the best of our knowledge, only Yoganathan et al. [82] have studied the automatic segmentation of the gross tumor volume (GTV) on BT MRI images. While they demonstrated that automatic segmentation of the GTV is possible in principle, the cohort was rather small with only 39 patients, resulting in a relatively weak performance with Sørensen–Dice coefficients (Dice) between 0.57 to 0.62. Furthermore, the segmentation architectures used in their project were based on the ResNet50 architecture[85], which is no longer considered state of the art.

A current state-of-the-art framework for the automatic segmentation of medical structures is the nnU-Net ('no-new-U-Net') [76]. The nnU-Net is a deep learning-based framework which automatically configures the parameters needed for training. It has been shown to outperform other approaches on 23 public datasets used on segmentation competitions.

The aim of this study was to assess the quality of the automatic segmentations of the cervical cancer GTV on BT MRI images. We used two methods to determine to what extent the automatic segmentations corresponded to the clinical segmentations performed by an expert radiation oncologist. First, the geometrical correspondence of the automatic and expert delineation was determined using Dice Similarity Coefficient (Dice), 95th Hausdorff Distance (95th HD) and mean surface distance (MSD). Then, to find if the observed geometrical differences between the delineations would have dosimetric consequences, we determined dose-volume parameters D90 and D98 for the automatic and expert delineations using the clinical dose distribution.

MATERIALS AND METHODS

Data

A cohort of 195 histologically proven cervical cancer patients treated in our institution between August 2012 and December 2021 was retrospectively collected. The average age was 53 (standard deviation of 15 years) and tumor stage ranged from IB to IV according to the International Federation of Gynecology and Obstetrics (FIGO) staging [86]. The treatment consisted of external beam radiotherapy (156 patients with 23 x 2 Gy and 39 patients with 25 × 1.8 Gy) followed by BT (3 x 7 Gy) and combined with chemotherapy (cisplatin 40 mg/m^2, weekly). The institutional review board approved the study (IRBd20276). Informed consent was waived considering the retrospective design.

A total of 524 separate BT fractions were included in this work. For each BT fraction, MRI images of the patient with applicator in place were acquired using a 1.5T (104 scans) or 3T (442 scans) Philips Medical Systems MRI scanner. Axial T2-weighted (T2w) turbo spin-echo images were used (TR =[3500-13300 ms], TE = [100 - 120 ms]) with a pixel spacing of 0.39 mm x 0.39 mm (442 scans) or 0.63 mm x 0.63 mm (104 scans) and a slice thickness of 3 mm. The GTV, as segmented for treatment planning by a radiation oncologist on each available MRI, was available as ground truth.

The data set was split into three subsets at the patient level: training set (117 patients, 314 images), validation set (39 patients, 104 images) and test set (39 patients, 106 images). The three subsets were stratified according to FIGO stage [86], because it is a relevant clinical parameter used to describe gynecological tumors.

Network architecture and training procedure

The nnU-Net framework was used in this work. This framework automatically configures the parameters needed for preprocessing, network architecture and training for each specific task. The loss function was a combination of the Dice loss [75] and cross entropy loss. We used the stochastic gradient descent (SGD) optimizer with learning rate scheduler and early stopping based on the validation loss as criterion to choose the best model. Dropout, data augmentation and weight decay were used as regularization techniques. Further details on the training procedure can be found in the Additional file 1 (available online: https://static-content.springer.com/esm/art%3A10.1186%2Fs13014-023-02283-8/MediaObjects/13014_2023_2283_MOESM1_ESM.docx) .

Experiment overview

The automatic segmentations were compared to the manual segmentations of the GTV that were performed by a radiation oncologist for treatment planning for the patients on the separate test set. The automatic segmentations were compared to the manual segmentations using common segmentation metrics: Dice, 95th HD and MSD, which were implemented using the Python package by DeepMind (https://github.com/deepmind/surface-distance). The segmentation results were additionally compared among patients with different FIGO stage and GTV volume. For the volume analysis, the patients of the test set were allocated to four volume ranges containing the same number of images in each bin.

Attention maps were computed for four different examples to highlight which parts of the input image were relevant for the network to decide on a segmentation. The attention maps were then qualitatively compared to the binary segmentations to investigate if the over-/under-segmentations of the network were on specific areas, therefore highlighting anatomically challenging regions. The attention maps were defined as the activations of the last layer of the nnU-net (i.e. before binarizing).

To assess if the differences between the automatic segmentations and manual segmentations would result in differences in dose-volume parameters, we calculated the D98 and the D90 for both segmentations on the clinical dose distribution used for the treatment. These dose parameters were chosen in accordance with the Embrace II guidelines [1]. The values for the manual segmentations represent the actual treatment parameters for the patients. The dosimetric impact of using automatically segmented structures was defined as the difference between these parameters compared to the clinical values (ΔD90 and ΔD98). The dosimetric impact was also reported as a relative measure by dividing the absolute difference on the dose parameters by the dose parameter on the manual segmentation (ΔD90rel and ΔD98rel). The dosimetric results were also compared for patients with different FIGO stage and GTV volume.

Statistics

The chi-square test for independence was used to confirm that the training, validation and test sets were balanced in terms of FIGO stage. The Kruskal-Wallis H test was used to assess differences among patients of different FIGO stage and GTV volume. If significant differences were found, Dunn's test with Bonferroni correction was used for the post-hoc analysis. A p-value of 0.05 was considered statistically significant. The SciPy Python package (version 1.5.4) and Python 3.9 were used for the statistical analysis.

RESULTS

Patients' characteristics of our cohort are described in Table 1. No significant differences were found in the distributions of FIGO stage or volume among the training, validation and test sets. The results are shown for 105 out of the 106 cases of the test set. The remaining case corresponded to a patient that had her uterus removed which resulted in a deviating anatomy unseen by the trained network.

Table 1. Patients' characteristics in the training, validation and test sets.

	Training	**Validation**	**Test**
Total	117	39	39
Age (years)			
Mean	53	52	56
Standard deviation	14	15	17
FIGO stage			
FIGO I	12 (10.2 %)	6 (15.4 %)	6 (15.4 %)
FIGO II	70 (59.8 %)	22 (56.4 %)	22 (56.4 %)
FIGO III	21 (18.0 %)	8 (20.5 %)	8 (20.5 %)
FIGO IV	9 (7.7 %)	3 (7.7 %)	3 (7.7 %)
Unknown	5 (4.3 %)	0	0
Volume at first BT fraction			
Less than 2.8 cc	22 (18.8 %)	8 (12.8 %)	5 (20.5 %)
between 2.8 and 4.3 cc	7 (6.0 %)	7 (17.9 %)	7 (17.9 %)
between 4.3 and 12.1 cc	53 (42.3 %)	12 (30.8 %)	13 (33.3 %)
More than 12.1 cc	35 (29.9 %)	12 (30.8 %)	14 (35.9 %)
Histopathological type			
Squamous cell carcinoma	97 (82.9 %)	31 (79.5 %)	33 (84.6 %)
Adenocarcinoma	16 (13.7 %)	6 (15.4 %)	4 (10.3 %)
Adeno-squamous cell carcinoma	(0.9 %)	1 (2.6 %)	1 (2.6 %)
Non specified/unknown	3 (2.5 %)	1 (2.5 %)	1 (2.5 %)
External beam radiotherapy scheme			
23 x 2 Gy	95 (81.2 %)	29 (74.4 %)	32 (82.0%)
25 × 1.8 Gy	22 (18.8 %)	10 (25.6 %)	7 (18.0 %)

The network achieved a median Dice of 0.73 (interquartile range (IQR) = 0.50 - 0.80), median 95th HD of 6.8 mm (IQR = 4.2 mm - 12.5 mm) and median MSD of 1.4 mm (IQR = 0.9 mm - 2.8 mm). When stratifying for FIGO stage (Figure 1 - top) no significant differences were found among the different subgroups. When comparing for GTV volume (Figure 1 - bottom) significant differences were found for the case of Dice (p-value < 0.001) but not for the distance-based metrics.

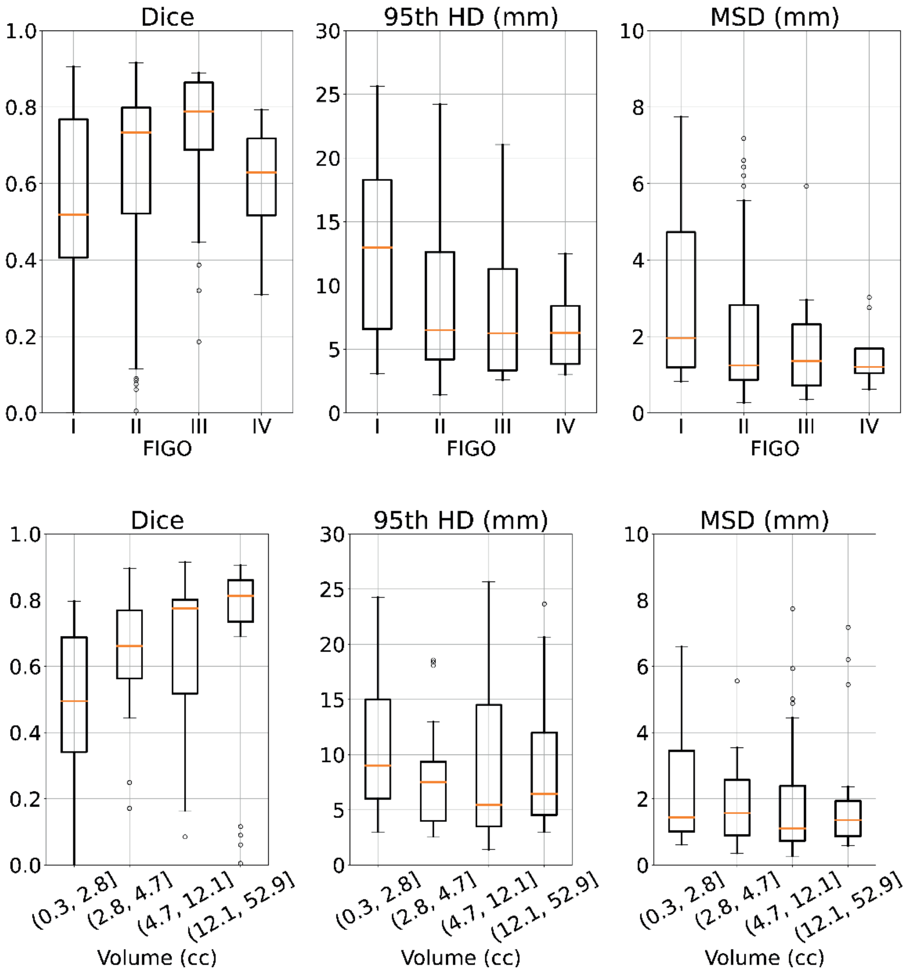


Figure 1. Geometric comparison by FIGO stage (top) and by volume (bottom). Segmentation performance in terms of Dice, 95th HD and MSD and stratified by FIGO stage (I-IV) and volume.

Four examples of automatic segmentations are shown in Figure 2 (a,c,e,g) with the corresponding attention maps (Figure 2-b,d,f,h). Even though the network under-/over-segmented the GTV for the last three cases (Figure 2-c,e,g), the error was in the area surrounding the applicator which is an area that is irradiated anyway. In the case 2c, the applicator was segmented by the network but not by the clinician while in the cases 2e and 2f, the clinicians segmented the applicator but the network did not. Furthermore, the attention map highlighted the undersegmented area of the last case (Figure 2h), meaning that the network looked at that area when deciding the segmentation.

The median D90 and D98 received by the manually segmented GTV were 12.5 Gy (IQR = 11.1 – 15.5 Gy) and 10.6 Gy (IQR = 9.4 – 13.1 Gy), respectively, in line with Embrace guidelines and the GYN GEC-ESTRO recommendations [87]. The resulting ΔD90 and ΔD98 were 0.18 Gy (IQR = -1.38 – 1.19 Gy) and 0.20 Gy (IQR = -1.10 – 0.95 Gy), respectively. The median ΔD90rel and ΔD98rel relative differences were 9.6 % (IQR = 4.2 - 19.28%) and 8.8 % (IQR = 0.15 - 92.5 %), respectively. When stratifying for FIGO stage (Figure 3 - top), no significant differences were observed among the different subgroups per FIGO stage. When comparing the results for GTV volume (Figure 3 - bottom), a significantly reduced ΔD90 and ΔD98 (p-value < 0.01) was found between the smaller tumors (0.3 - 2.8 cc) and the largest tumors (12.1 – 52.9 cc).

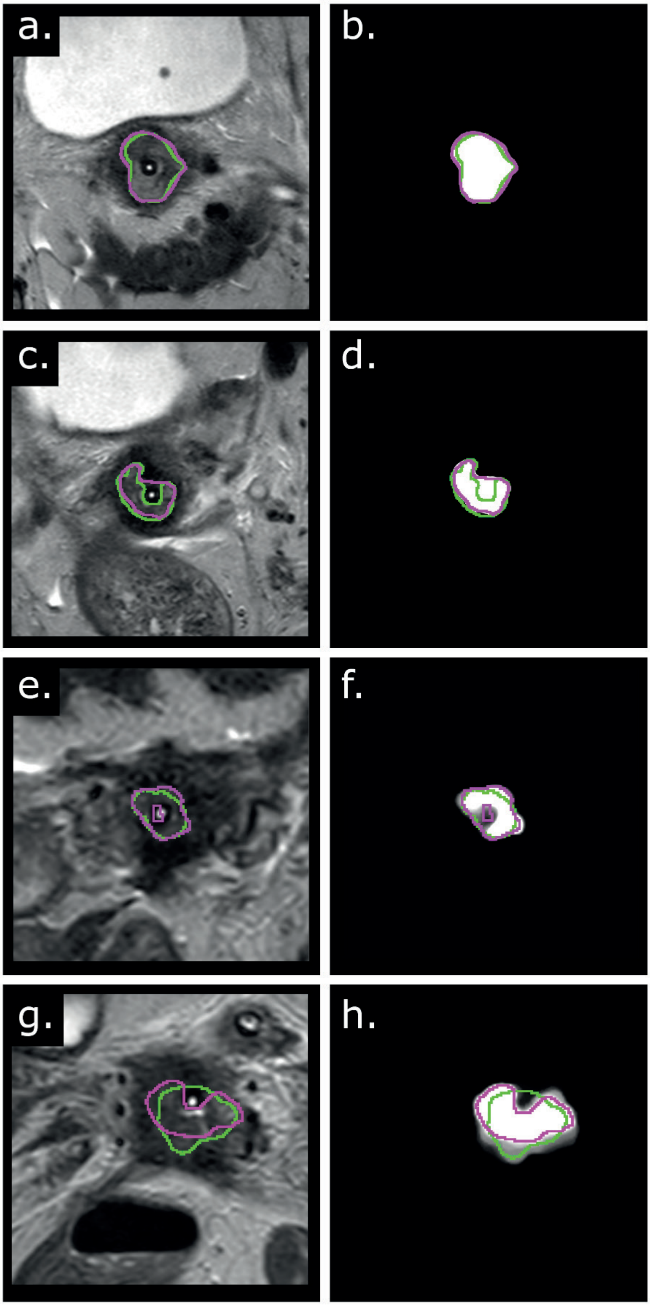


Figure 2. Qualitative results and attention maps. (Left) Examples of the automatic contours (pink) and the manual clinical contour (green) on four different patients. (Right) The corresponding attention maps for the same patients. The examples are sorted by decreasing Dice.

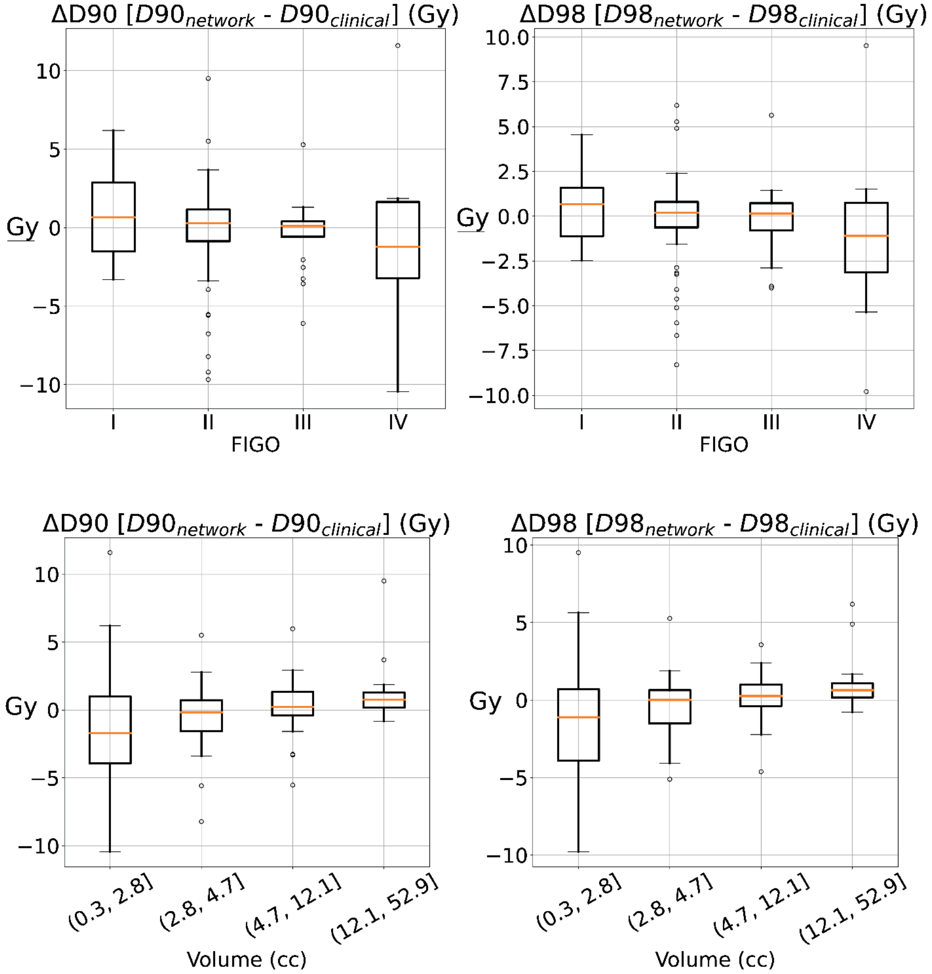


Figure 3. Dosimetric comparison by FIGO stage (top) and by volume (bottom). Dosimetric impact in terms of ΔD90 and ΔD98 stratified by FIGO stage (I-IV) and volume.

DISCUSSION

In this study we investigated the performance of a state-of-the-art automatic framework to segment the cervical cancer GTV on brachytherapy MR images. We used a cohort of patients that for their treatment were segmented manually by a radiation oncologist and compared these manual segmentations to the automatic segmentations. The comparison was performed geometrically and the impact of differences between automatic and manual delineations on dose-volume parameters of the clinical dose distribution was evaluated.

We achieved improved geometric performance when compared to previously published literature and automatic segmentations yielded a ΔD90 and ΔD98 of less than 0.25 Gy. No significant differences in geometric and dosimetric performance were observed when comparing for FIGO stage. When comparing per volume, decreased performance was observed for smaller tumors both for the Dice coefficient and dosimetrically.

To the best of our knowledge, only Yoganathan et al. [82] studied the automatic segmentation of cervical cancer GTV on brachytherapy images. In their work, they implemented and compared four different CNNs for the segmentation of the targets and organs at risk. The geometric performance of our model was considerably higher than what the authors obtained with the models trained with the axial T2w sequence (Dice: 0.56, 95th HD 9.7 mm). Their models were based on the ResNet and Inception architectures, while we use the nnU-Net, currently the state-of-the-art architecture for medical image segmentation. Additionally, we used a larger cohort.

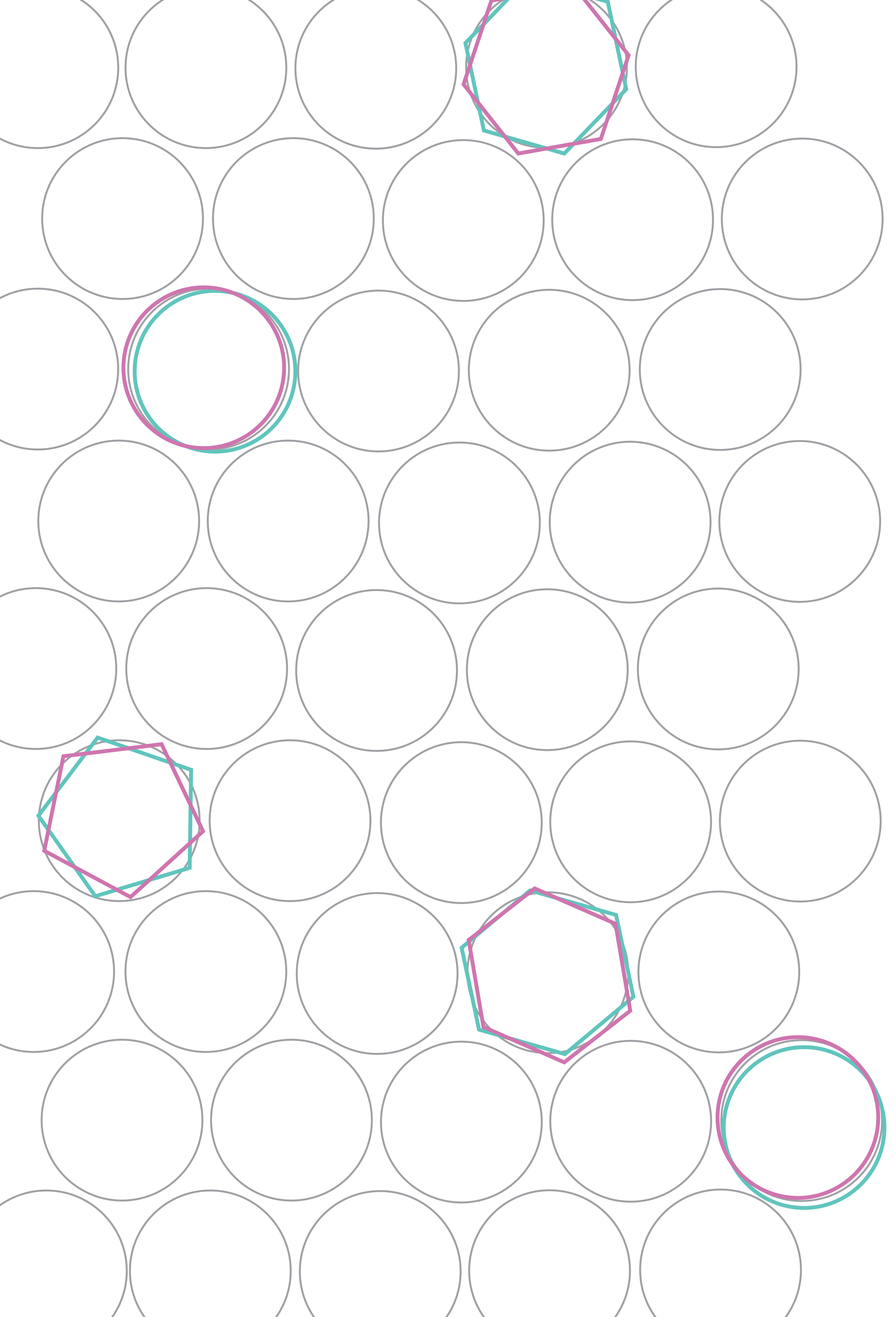
We observed that the median relative ΔD90rel and ΔD98rel were lower than 10%. Hellebust et al. [88] showed that the relative ΔD90 between different observers was 9.4% for the GTV, meaning that the average difference dosimetric difference between observers is comparable to using the automatic segmentation tool. However, in some of the cases the dosimetric difference was large. These large differences in dosimetric performance can be partially explained by the marked steepness of the brachytherapy dose distributions, which results in that small geometric errors can lead to large differences in dose parameters.

When comparing the results per FIGO stage, no significant differences were found between the different FIGO stages for neither the geometric nor the dosimetric comparisons. A priori, we would have expected the performance to be different between tumors of different FIGO stages because FIGO stage is an important clinical parameter to describe gynecological tumors. One possible reason is that the FIGO stage is defined at the time of diagnosis and consequently does not take into account the regression of the tumor during the external beam radiotherapy treatment, potentially reducing the differences between FIGO stages. On the other hand, when stratifying per GTV volume, significant differences were found for the Dice and for the dosimetric comparisons. For the Dice, the explanation can be rather trivial, because the Dice is defined as the overlap between the two structures and it therefore favors the bigger structures. However, for the dosimetric impact, larger tumors had lower ΔD90 and ΔD98, and less variability, which suggests that smaller tumors may require more accurate automatic segmentation methods than larger tumors.

This work has the following limitations. Firstly, even though our cohort includes a large amount of patients, patients from only one center were included and fewer patients were included for FIGO I and IV. A multi-center validation study is therefore desirable.

Secondly, the GTV segmentations used for training and evaluation were manually segmented for clinical practice with the treatment plan in mind, meaning that although the segmentations were clinically acceptable, they may contain geometric errors. These geometric errors could potentially lead the network to reproduce these errors and therefore partially bias our geometric analysis. However, we presented the results in terms of dosimetric impact as well and showed that the dosimetric impact of the automatic segmentations is comparable to that derived from the interobserver variability. Finally, the scope of this work was limited to the GTV automatic segmentation while the HR-CTV and the IR-CTV are also needed for treatment planning. The definition of those structures is intrinsically related to the information of the image before external been radiotherapy (5 weeks before BT) and not only to the information present in the BT image. Therefore in this work we focused solely on the structure that can be found in the BT image.

In this study we evaluated a state-of-the-art framework for the automatic segmentation of the cervical cancer GTV. The quality of the automatic segmentations improved with respect to previously published works. The automatic segmentations yielded similar dose-volume parameters as the manual segmentations used clinically and differences were comparable to the interobserver variability reported in literature.



Chapter 5

A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations

Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, Simões R.

Phys Imaging Radiat Oncol. 28. (2023)

ABSTRACT

Background and purpose

Existing methods for quality assurance of the radiotherapy auto-segmentations focus on the correlation between the average model entropy and the Dice Similarity Coefficient (DSC) only. We identified a metric directly derived from the output of the network and correlated it with clinically relevant metrics for contour accuracy.

Materials and Methods

Magnetic Resonance Imaging auto-segmentations were available for the gross tumor volume for cervical cancer brachytherapy (106 segmentations) and for the clinical target volume for rectal cancer external-beam radiotherapy (77 segmentations). The nnU-Net's output before binarization was taken as a score map. We defined a metric as the mean of the voxels in the score map above a threshold (λ). Comparisons were made with the mean and standard deviation over the score map and with the mean over the entropy map. The DSC, the 95th Hausdorff distance, the mean surface distance (MSD) and the surface DSC were computed for segmentation quality. Correlations between the studied metrics and model quality were assessed with the Pearson correlation coefficient (r). The area under the curve (AUC) was determined for detecting segmentations that require reviewing.

Results

For both tasks, our metric (λ=0.30) correlated more strongly with the segmentation quality than the mean over the entropy map (for surface DSC, r>0.65 vs. r<0.60). The AUC was above 0.84 for detecting MSD values above 2 mm.

Conclusions

Our metric correlated strongly with clinically relevant segmentation metrics and detected segmentations that required reviewing, indicating its potential for automatic quality assurance of radiotherapy target auto-segmentations.

INTRODUCTION

Target segmentation is a crucial part of the radiotherapy (RT) workflow. In clinical practice, this step is typically done manually by radiation oncologists, which is time consuming and suffers from inter- and intra- observer variability. In particular in online adaptive treatment settings, the time pressure is high because both the patient and the staff involved in the RT treatment are waiting while the segmentations are performed. With the aim of saving time in the clinic, automatic segmentation algorithms based on convolutional neural networks have been investigated for gross tumor volumes (GTVs) in a variety of tumor sites, such as brain [89,90], head and neck [42,65,80], rectum [48] and cervix [82,91]; and clinical target volumes (CTVs) such as cervical cancer CTV [92] and prostate cancer CTV [93,94].

Although segmentation algorithms are reaching a reasonable performance [27,31,95], they still produce faulty segmentations in some cases. To identify whether automatically generated segmentations are acceptable for clinical use, it is necessary for a clinician to verify them. This limits the time gains of automatic segmentation methods. Therefore, there is a need to recognize automatically in which cases the automatic segmentations need correction. In the context of RT, automatic quality assurance (QA) of the automatic segmentations is a topic of interest nowadays, as showcased in recent reviews [96,97].

Deep learning networks for auto-segmentation typically predict a score that correlates with the probability that a voxel belongs to the structure to be segmented. Only at the last step, voxel scores are thresholded into a binary segmentation mask. These score maps are often converted into uncertainty maps by applying the entropy operator [32,33,98]. It has been shown qualitatively that incorrect areas of the automatic segmentations cover areas of high network entropy [36,37,99]. Once an entropy map is computed, the mean over all the voxels [32,33,98] is often used as a metric for QA of auto-segmentations. Alternatively, a common approach for QA of auto-segmentations consists of developing machine learning models that directly predict segmentation quality [100,101].

Up to now, the QA metrics are typically correlated only with the Dice Similarity Coefficient (DSC) [32–34,36,102]. Although DSC is a common metric of segmentation performance, it presents several drawbacks. By construction, it is volume-dependent since it overestimates the performance for large structures. Additionally, it has been shown to correlate poorly with clinically relevant endpoints in RT planning, such as dose/volume metrics [103] and the expected editing time [104]. Distance-based metrics, such as the 95th Hausdorff distance (95th HD), the mean surface distance (MSD) and the surface DSC, suffer less from these drawbacks and are recommended to be reported together with the DSC [104,105].

We hypothesize that the commonly used entropy operator may overshadow relevant information that is contained in the score maps. The aim of this study was to identify a quality metric that can be generated directly from the output of the network, and which correlates with clinically relevant distance-based metrics. We additionally assessed the capability of the proposed metric to identify automatic segmentations that would need review.

MATERIALS AND METHODS

Data

Two cohorts were retrospectively collected and used in this study. One cohort consisted of a total of 195 histologically proven cervical cancer patients treated in our institution between August 2012 and December 2021. Further details on patient characteristics and their treatment are described in Table S1. The institutional review board approved the study (IRBd20276). Informed consent was waived considering the retrospective design of the study.

A total of 524 separate MRI images of the patients with the brachytherapy applicator in place were included in this work. These images were acquired using a 1.5T (104 scans) or 3T (442 scans) Philips MRI scanner. Axial T2-weighted (T2w) turbo spin-echo images were used (TR =[3500-13300 ms], TE = [100-120 ms]) with a pixel spacing of 0.39 mm x 0.39 mm (442 scans) or 0.63 mm x 0.63 mm (104 scans) and a slice thickness of 3 mm. The GTV, as segmented for treatment planning by a radiation oncologist on each available MRI, was available as ground truth.

The other cohort used in this study consisted of a total of 30 patients with intermediate risk or locally advanced rectal cancer treated in our institution. Further details on patient characteristics are described in Table S2. All patients in the study were enrolled in the Momentum prospective registration study (NCT04075305) and gave written informed consent for the retrospective use of their data.

For this cohort, a total of 483 EBRT images were considered. All the fractions were carried out on the Unity MR-Linac (Elekta AB, Stockholm). Axial T2-weighted (T2w) turbo spin-echo images were used (TR=1300 ms, TE=128 ms) with a pixel spacing of 0.57 mm x 0.57 mm (349 images) or 0.87 mm x 0.87 mm (134 images) and a slice thickness of 1.20 mm (155 images), 1.8 mm (134 images) or 2.4 mm (194 images). In our institution, the radiation therapists (RTTs) have been trained and certified to segment the CTV for the MRI-guided online adaptive RT workflow of the rectal cancer treatment. Therefore, the CTV used as ground truth was segmented by a RTT on each available MRI for clinical practice. The CTV segmentations were also verified by a radiation oncologist with over 10 years of experience.

Segmentation framework and training scheme.

In previous studies, we used the nnU-Net framework [76] to segment the cervical cancer GTV [91] and the rectal cancer mesorectum CTV [106]. In the current work, we used a 5-fold cross validation scheme to retrain the networks and assess the robustness of the quality metrics to changes in the training set composition. The training sets were the same as those described in previous articles [91,106], with 156 patients (418 images) for the cervical cancer cohort and 25 patients (406 images) for the rectal cancer cohort. For both cohorts, the 3D variant of the nnU-Net was used.

Score map definition.

The score map was defined as the voxelwise softmax scores of the last layer of the network of the target segmentation channel before binarization (as depicted in Figure 1). This strategy was chosen because it can be applied to any trained network without requiring changes to the architecture or training procedure.

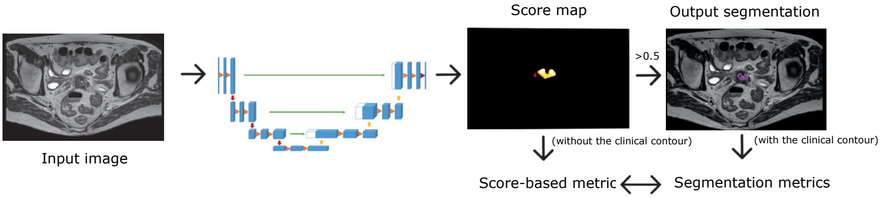


Figure 1. Workflow of the study design.

The score maps were created for the test sets described in previous studies [91,106], which included 39 patients (106 images) for the cervical cancer cohort and five patients (77 images) for the rectal cancer cohort. We further subdivided these sets at the scan level into a validation set for parameter optimization and a final test set for evaluating the quality metrics. For the cervical cancer GTV segmentation task, the final validation and test sets each included 53 images. The analyses were done for 52 out of the 53 cases of the test set. The remaining case corresponded to a patient who had her uterus removed which resulted in a variation in anatomy unseen by the trained network. The final validation and test sets for the rectal cancer CTV segmentation task included 39 and 38 images, respectively. Note that the term “score maps” is referred to as “attention maps” in our previous work [91].

Score-based metrics.

We defined a metric (High Score or HiS metric) as the mean of the score map values that were higher than a threshold λ. By thresholding the score map and retaining only the high score voxels, we aimed to remove information that is unimportant for the flagging of potentially incorrect segmentations, as very low values on the score map are expected both in correct and incorrect segmentations.

The mean and the standard deviation (STD) were computed over the non-zero values of the score map to represent the overall score and its variability, respectively. Additionally, the mean over the entropy map was computed for direct comparison with other studies [32–34,98].

For each value of λ, the difference in correlation with respect to the performance of the mean over all values of the score map (i.e. $\lambda=0$) was determined. The optimal value of λ was determined empirically as the value at which the HiS correlated best with the MSD in the validation set, in the range (0,0.45) with steps of 0.05. The MSD was chosen to determine the optimal threshold because it is a distance-based metric and therefore more suitable for RT applications (unlike the DSC), it evaluates the whole contour (unlike the 95th HD, which focuses on the gross errors) and it has no hyperparameters (unlike the surface DSC).

Statistics

The correlation between the metrics and the segmentation performance was assessed with the Pearson correlation coefficient (r) and with the Spearman correlation coefficient. To check the assumption of linearity for Pearson, residual plots were computed. To study the robustness of each metric to the training set composition, the correlations were computed separately for the score maps resulting from each of the five training folds. The mean and the standard deviation of the r were computed over all folds.

To assess the capability of the metrics to distinguish between segmentations that require reviewing and those that can be left unchecked, the area under the curve (AUC) was determined for detecting segmentations that exceeded a specified MSD or 95th HD threshold. The code and additional training details are available in: github.com/RoqueRouteiral/his_qa.

RESULTS

For both segmentation tasks and for the four segmentation metrics, the largest improvement of the proposed HiS metric with respect to the mean (Δr) occurred for λ<0.10, as depicted in Figure 2. Moreover, for λ>0.10, Δr remained fairly stable. For the case of the MSD, the largest correlations were found for λ=0.35 and λ=0.25 for the cervical and rectal cancer target segmentation tasks, respectively. We took the average between these two values, λ=0.30, in the subsequent analyses. The computed residual plots (Figure S1) show that the points were randomly scattered around the horizontal axis, confirming the assumption of linearity between the performance metrics and the HiS.

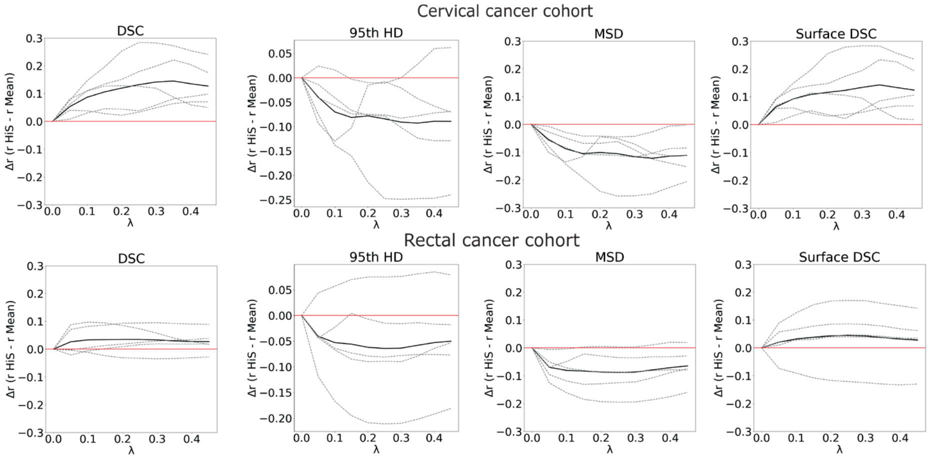


Figure 2. Difference in Pearson correlation coefficient (Δr) with the segmentation metrics between the HiS metric and the mean over the score map as a function of the parameter λ. The bold line is the average Δr among the five folds. The dashed lines represent the Δr for each of the five folds.

Table 1 shows the correlation between the studied metrics and the segmentation quality metrics for the test sets of both cohorts. For the segmentations of the cervical cancer GTV, the HiS achieved a mean r of 0.79 with DSC, -0.60 with 95th HD, -0.66 with MSD and 0.67 with surface DSC. For the segmentations of the rectal cancer CTV, the HiS yielded a mean r of 0.76 with DSC, -0.53 with 95th HD, -0.74 with MSD and 0.62 with surface DSC. For both tasks, the HiS correlated more strongly with the segmentation quality metrics than the rest of the score-based metrics. The only exception was the STD in the case of the cervical cancer task, which correlated as strongly as the HiS and the surface

DSC. The HiS also correlated more strongly with all the segmentation metrics for both tasks with the Spearman correlation coefficient (Table S3).

Table 1. Pearson correlation coefficients (mean ± standard deviation among folds) between the metrics and the segmentation performance metrics. Bold letters indicate the highest correlation among the different metrics.

	DSC	**95th HD**	**MSD**	**Surface DSC**
Cervical cancer cohort				
Mean	0.72 ± 0.10	-0.53 ± 0.16	-0.57 ± 0.13	0.60 ± 0.1
STD	0.68 ± 0.06	-0.53 ± 0.14	-0.64 ± 0.13	**0.70 ± 0.1**
Mean (over entropy map)	0.43 ± 0.14	-0.38 ± 0.09	-0.43 ± 0.11	0.43 ± 0.15
HiS (λ = 0.30)	**0.79 ± 0.05**	**-0.60 ± 0.13**	**-0.66 ± 0.10**	0.67 ± 0.06
Rectal cancer cohort				
Mean	0.60 ± 0.03	-0.42 ± 0.10	-0.61 ± 0.06	0.50 ± 0.08
STD	-0.32 ± 0.11	0.22 ± 0.18	0.35 ± 0.15	-0.27 ± 0.17
Mean (over entropy map)	-0.74 ± 0.06	0.47 ± 0.08	0.69 ± 0.07	-0.58 ± 0.09
HiS (λ = 0.30)	**0.76 ± 0.08**	**-0.53 ± 0.07**	**-0.73 ± 0.09**	**0.62 ± 0.10**

As an illustration, Figure 3 shows the scatter plots between the HiS metric and the segmentation metrics obtained in one of the five folds of the trained auto-segmentation networks. Note that the range of HiS values is task-dependent and the values are therefore not directly comparable between the two tasks. Figure 4 illustrates the segmentations and score maps with one example from each auto-segmentation task. For the cervical cancer case (Figure 4, left), the HiS metric was relatively high for this cohort (HiS=0.76). Indeed, the segmentation performance was high (MSD=0.78 mm), with the main error at the location of the applicator channel. For the rectal cancer example (Figure 4, right), the HiS value was relatively low for this cohort (HiS=0.89). This case corresponded to a target that was oversegmented by the network, resulting in poor performance (MSD=3.6 mm), as expected.

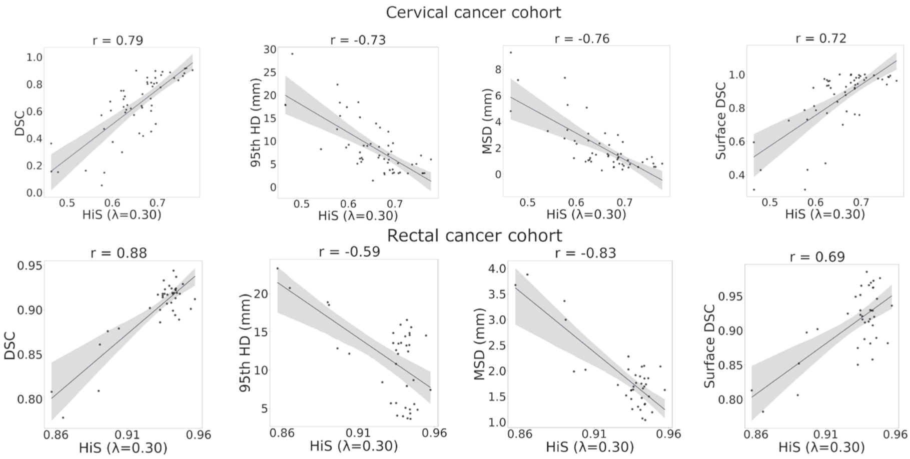


Figure 3. Scatter plots between the segmentation metrics and the HiS metric for the cervical cancer cohort (top) and the rectal cancer cohort (bottom). The translucent band corresponds to the 95 % confidence interval for the estimated regression, computed via bootstrap.

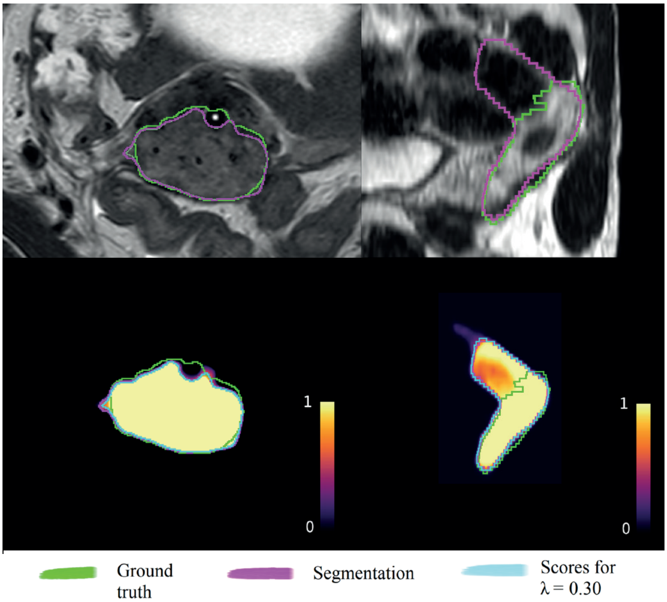


Figure 4. Examples of the segmentations and the correspondent score maps for a cervical cancer case (left, HiS = 0.76) and a rectal cancer case (right, HiS = 0.89). The input images for the segmentation framework, the ground truth segmentation (green) and the automatic segmentation (pink) are depicted on the top row. The corresponding score maps are depicted on the bottom row. The blue line encompasses the voxels for which the score values are higher than λ = 0.3.

The capability of the studied metrics to detect segmentations that require reviewing is illustrated in Figure 5, which shows the AUC for detecting segmentations that exceed varying MSD and 95th HD threshold values. The proposed HiS metric achieved higher AUC values than the other baselines metrics for most MSD and 95th HD values, for both auto-segmentation tasks. In particular, for the cervical cancer cohort, the AUC varied between 0.82 and 0.94 for detecting cases for MSD values above 1 mm. For the rectal cancer cohort, the AUC varied between 0.84 and 0.99 for detecting cases with an MSD above 2 mm.

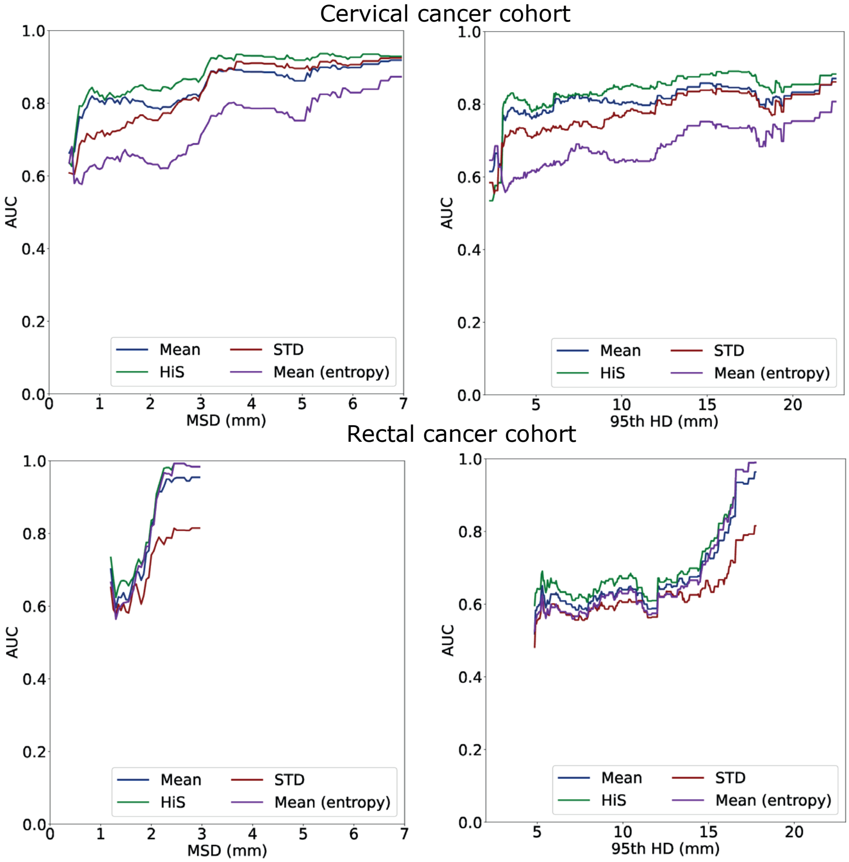


Figure 5. AUC for detecting segmentations exceeding a specified MSD (left) or 95th HD (right).

For each task, the AUC was reported between the minimum and maximum values of the obtained MSD and 95th HD over all folds, because the sensitivity and specificity are only defined in these ranges. For the cervical cancer task, these ranges were 0.4 mm to 7.0 mm for the MSD and 2.6 mm to 22.5 mm for the 95th HD. For the rectal cancer task, the ranges were 1.2 to 3.0 mm for the MSD and 4.8 mm to 17.8 mm for the 95th HD.

DISCUSSION

In this work we proposed a simple metric based on the network output for automatic QA of auto-segmentations of RT target volumes. This metric averages all score values above a threshold of 0.3. We showed that it correlated strongly with the segmentation performance metrics for two different auto-segmentation tasks. The correlations were strong not only for the DSC but also for the more clinically relevant distance-based metrics. Our proposed metric outperformed the often used mean value of the entire entropy map in the distinction between segmentations that require reviewing and those that can be used without an extra manual check.

The strongest correlations between the proposed metric and the segmentation performance occurred for λ values above 0.1, suggesting that the lowest score values are not very representative of the segmentation performance. Furthermore, it was observed that the choice of λ was not critical for values above 0.2.

Despite the high correlations between the proposed metric and the segmentation quality, similar HiS values corresponded to a large range of values on the segmentation quality metrics, suggesting that the HiS might not always be an accurate surrogate of the segmentation performance. Other works have shown similar behavior in their correlation plots [33,100]. The aim of this metric, however, is to flag cases that need reviewing, not to predict the segmentation performance. This was demonstrated with the high AUC values achieved by the metric.

Previous studies have qualitatively related the uncertain areas with the segmentation errors [36,99]. Metrics that show qualitatively where the local edits should be performed could aid clinicians during editing and should therefore be investigated in future work. We speculate that the proposed metric could also be used to select the voxels that are more likely wrong in the segmentation. From our results, we can infer that voxels from the score map that are below the λ=0.10 threshold did not contribute to the correlation with the segmentation performance. This suggests that those voxels are not relevant for a potential correction. Clinicians could then use this information as an aid to edit the segmentation.

Pearson's correlation coefficient has been used in previous works to study the correlation between the segmentation performance and QA metrics [34,102]. Its application assumes linearity between the two variables. Furthermore, outliers can skew its evaluation. To confirm the validity of our results, we computed the Spearman correlation coefficient, which does not assume linearity and is more robust to outliers. The HiS metric still correlated more strongly than the other score-based metrics.

The STD showed strong correlations with the MSD, but only for the cervical cancer GTV segmentation task. For rectal cancer, the correlation was much lower and importantly also changed sign. A similar behavior was observed for the mean over the entropy map, commonly used in literature. This metric showed strong correlations for the rectal cancer segmentation task, but for the cervical cancer GTV the correlations were poor for the segmentation metrics and also changed sign. Therefore, these metrics appear to be less robust for QA. Tumors (like the cervical cancer GTV) are more heterogeneous in size, shape and texture than anatomical structures (like the rectal cancer CTV, or mesorectum). Uncertainties in tumor auto-segmentation networks are likely more prominent than those of auto-segmentation networks of anatomical structures. This may explain the difference in behavior of the metrics across the two tasks. Previous works have mostly focused on segmentation tasks with arguably lower uncertainty, such as the segmentation of anatomical structures [36,100] or the segmentation of brain tumors [32,33].

Although most studies propose using the average of the entire entropy map, other works [32,102] have trained models to automatically predict the DSC coefficient directly from the entropy maps, thereby incorporating the metric definition into the learning task. Learning-based metrics can be more generic than the pre-specified average, but they are also less interpretable and therefore might be less desirable for QA purposes.

Recent literature has focused on other methods for computing the score maps, such as Monte Carlo dropout [32,33,107], which averages the scores resulting from multiple instances of the network. We expect our metric to also be applicable to Monte Carlo dropout estimates. However, using the softmax layer outputs eliminates the need for specific architectural or training scheme modifications. Furthermore, it does not require running inference multiple times which could hinder the clinical implementation of the method.

In clinical settings, the clinician could be provided with both the automatic segmentation and its associated HiS score that would serve as a quality metric. Prior to clinical implementation, a pilot study could be set up to assess the time savings achieved by using this tool in a clinical setting. The trade-off between the amount of cases that would not need to be reviewed manually and the missed faulty cases that would require reviewing, should also be assessed.

In conclusion, we identified a simple metric derived directly from the output of the segmentation network that correlated strongly with commonly used segmentation metrics, not only for the case of DSC but also for the more clinically relevant distance-based metrics. The proposed metric was able to flag segmentations that would require review. It is also easy to compute, as it does not require any architecture or training scheme modifications. The proposed metric has potential as a tool for QA of automatic target segmentations.

SUPPLEMENTAL MATERIAL

Table S1. Patient characteristics in the cervical cancer cohort.

	Training	**Evaluation**	**Total**
Total	156	39	195
Age (years)			
Mean	53	56	53
Standard deviation	15	17	15
FIGO stage			
FIGO I	18 (11.5 %)	6 (15.4 %)	24 (12.3 %)
FIGO II	92 (59.0 %)	22 (56.4 %)	114 (58.6 %)
FIGO III	29 (18.6 %)	8 (20.5 %)	37 (19.0 %)
FIGO IV	12 (7.7 %)	3 (7.7 %)	15 (7.7 %)
Unknown	5	0	5 (2.6 %)
Histopathological type			
Squamous cell carcinoma	128 (82.0 %)	33 (84.6 %)	161 (82.6 %)
Adenocarcinoma	22 (14.10 %)	4 (10.3 %)	26 (13.3 %)
Adeno-squamous cell carcinoma	2 (1.3 %)	1 (2.6 %)	3 (1.5%)
Non specified/unknown	4 (2.6 %)	1 (2.5 %)	5 (2.6 %)
External beam radiotherapy scheme (prior to brachytherapy and combined with cisplatin (40 mg/m2 weekly)			
23 x 2 Gy	124 (79.5 %)	32 (82.0 %)	156 (80 %)
25 × 1.8 Gy	32 (20.5 %)	7 (18.0 %)	39 (20 %)

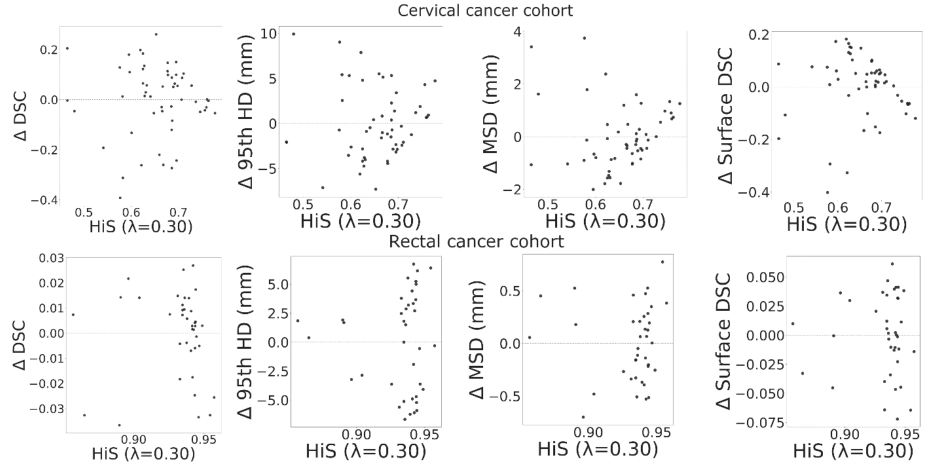
Table S2. Patient characteristics in the rectal cancer cohort.

	Training	Evaluation	Total
Total	25	5	30
Sex			
Male	15 (60 %)	5 (100%)	20 (66 %)
Female	10 (40 %)	0	10 (34 %)
Age (years)			
Mean	59	61	59
Standard deviation	11	16	12
T stage			
T2	8 (32 %)	0	8 (26.7 %)
T3	16 (64 %)	5 (20 %)	21 (70 %)
T4	1 (4 %)	0	1 (3.3 %)
N stage			
N0	14 (56 %)	1 (20 %)	15 (50 %)
N1	9 (36 %)	2 (40 %)	11 (36.7 %)
N2	2 (0.08 %)	2 (40 %)	4 (13.3 %)
External beam radiotherapy scheme			
5 x 5 Gy	20 (80 %)	2 (40 %)	22 (73.3 %)
25 x 2 Gy	5 (20 %)	3 (60 %)	8 (26.7 %)

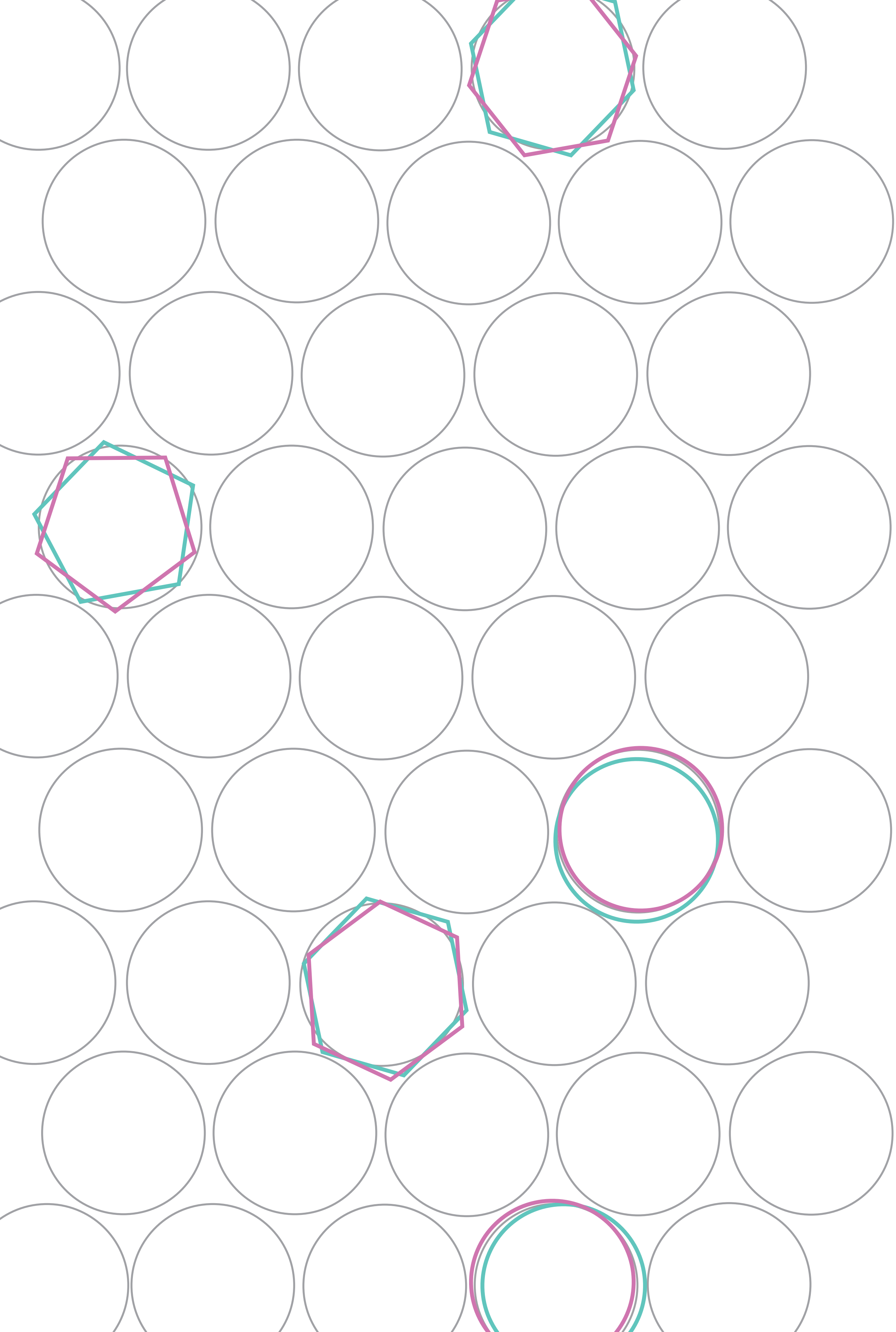
Table S3. Spearman correlation coefficients (mean ± standard deviation among folds) between the metrics and the segmentation performance metrics. Bold letters show the highest correlation among the different metrics.

	DSC	95th HD	MSD	Surface DSC
Cervical cancer cohort				
Mean	0.73 ± 0.09	-0.59 ± 0.09	-0.62 ± 0.09	0.61 ± 0.09
STD	0.52 ± 0.10	-0.48 ± 0.04	-0.49 ± 0.07	0.51 ± 0.07
Mean (over entropy map)	0.30 ± 0.13	-0.29 ± 0.10	-0.31 ± 0.11	0.30 ± 0.11
HiS ($\lambda = 0.30$)	**0.80 ± 0.04**	**-0.64 ± 0.04**	**-0.67 ± 0.04**	**0.65 ± 0.06**
Rectal cancer cohort				
Mean	0.41 ± 0.11	-0.25 ± 0.16	-0.38 ± 0.16	0.34 ± 0.08
STD	-0.26 ± 0.16	0.17 ± 0.21	0.27 ± 0.23	-0.22 ± 0.19
Mean (over entropy map)	-0.47 ± 0.09	0.22 ± 0.04	0.37 ± 0.08	-0.35 ± 0.07
HiS ($\lambda = 0.30$)	**0.51 ± 0.07**	**-0.30 ± 0.06**	**-0.43 ± 0.06**	**0.40 ± 0.05**

Supplemental Figures



Supplemental Figure 1. Residual plots of the HIS metric with the segmentation performance. The results are shown for one of the folds.



Chapter 6

General discussion

In this thesis, we explored the topic of automated segmentation of tumors on MRI images. Deep learning (DL) techniques are already employed clinically in radiotherapy (RT) departments for organ-at-risk segmentation. However, tumor segmentation so far remains limited to a research setting. Furthermore, most existing studies about automatic segmentation of tumors use CT or FDG-PET images as the primary modality, even though MRI is often preferred to visualize cancer tissue in several tumor sites. We aimed to implement DL techniques to deliver clinically acceptable tumor segmentations in MRI cohorts. Two different MRI cohorts acquired in a clinical setting were used throughout this thesis: a cohort of oropharyngeal primary tumors in multiparametric diagnostic MRIs (in chapters 2 and 3) and a cohort of cervical cancer gross tumor volume in MRI images of brachytherapy treatment (in chapters 4 and 5).

INTRODUCING MULTIPLE MRI SEQUENCES AS INPUT FOR SEGMENTATION

When physicians manually segment tumors, they often rely on information from various sources, such as different imaging modalities or MRI sequences. Each of these images can provide distinct insights of the anatomy of the patient. The physicians can define the boundaries of the tumor by combining these different insights in their minds. Therefore, we also expect DL methods to benefit from combining different images as input.

In chapter 2, we investigated the effect on the segmentation performance of using different anatomical MRI sequences as input for the task of oropharyngeal cancer segmentation. The investigated MRI sequences were T1-weighted (T1w), T2-weighted (T2w) and T1 weighted after gadolinium injection (T1gd). We compared the segmentation performance of the networks trained with each of those sequences as input and with all the sequences together. Indeed, the network trained with all the available sequences outperformed the networks trained with one sequence only. This suggests that DL segmentation techniques also benefit from combining several input images.

Similar conclusions have been reached in other studies. Wahid et al. [42] investigated the use of both anatomical MRI sequences (T1w and T2w) and quantitative MRI sequences (ADC, Ktrans, and Ve) for automatic segmentation of the oropharyngeal tumor in an MRI-only workflow. Their study demonstrated that the best segmentations were achieved when combining both anatomical sequences. Ren et al. [80] explored the optimal combination of imaging modalities (MRI, CT, and PET) to improve the automatic segmentations. The model that incorporated the information from all the imaging modalities outperformed the rest of segmentation models. Thus, these works further support the claim that DL segmentation techniques benefit from being trained with multiple input images.

In all of the aforementioned studies, the quality of the automatic segmentations was assessed by comparing them to ground truth segmentations made by physicians. These ground truth segmentations were carried out on a single reference image. Although other images are consulted, the final voxel-level decision is based on just this reference image. This reference image may differ across different studies. For instance, in our work the ground truth segmentations were made in the T1gd sequence. In the study of Wahid et al [42], the ground truth segmentations were made in the T2w sequence. This can create differences in the ground truth segmentations. Given that the ground truth is used for training and evaluation of the network, it may also make the comparison between studies more challenging.

An alternative approach to compute the ground truth segmentations without the bias of a reference image is desirable. The presence of tumor at a pixel level could be confirmed with histopathology data, considered the gold standard in cancer diagnostics. However, obtaining this type of data is a complex process. It involves the surgical removal of the tumor tissue, subsequent staining, and precise registration with the 3D radiological images. Hence, histopathological validation is limited in the most tumor auto-segmentation studies, as acknowledged by Jager et al. [108].

In some tumor sites, the physicians also use the information from physical examinations to manually segment the tumor. This is the case for the two tumor sites studied throughout this thesis (i.e. the head and neck cancer and the cervical cancer), for which the tumor is reachable by the physicians. This information is not taken into account by the DL methods described in this section, which only rely on imaging data. The physicians in charge of the treatment might need to edit the automatic segmentation after the physical examination of the patient to include this information.

IMPROVING THE SEGMENTATIONS BY INCORPORATING PRIOR KNOWLEDGE

A common promise in the machine learning field is that underperforming methods would improve their performance by being trained on more data. As already stated in the introduction, inclusion of new data is challenging in the field of medicine. In the context of automatic segmentation of medical structures Fang et al. [109] observed that the performance improved logarithmically with the dataset size. They showed quantitatively that for the structures that were more dependent on the size of the dataset (i.e. the optical nerves, in their work) an improvement of 0.04 on the Dice Similarity Coefficient (DSC) was achieved with a training set 10 times larger. This suggests that even by collecting more data, performance gains may be rather modest. Therefore, a different approach to improve the segmentation performance is preferable.

In chapters 2 and 3, we hypothesized that the segmentation performance would improve by reducing the amount of context present in the image given as input. The rationale behind it is that by reducing the context around the tumor, the network does not need to spend any resources in localizing the tumor in the whole image. This simplifies the segmentation task, leaving more resources for the network to accurately segment the tumor. Our results confirmed that indeed, reducing the context led to improved performance.

Other strategies of simplifying the segmentation task have been investigated to improve the automatic segmentations in medical imaging. One approach is to use shape or anatomical constraints during training, thereby regularizing the solution space to only anatomically reasonable segmentations. This has been shown to result in more accurate segmentations than when training the segmentation network from scratch [110,111].

Another approach is to combine the segmentation task together with other relevant tasks, such as registration. In the context of adaptive image-guided radiotherapy for prostate cancer, S. Elmahdy et al. [112] framed the registration and segmentation as a joint problem within a multi-task learning setting. This yielded improved segmentation performance compared to the single-task setting.

Self-supervised learning (SSL) has been often posed as a promising strategy to improve the performance of machine learning techniques, particularly in low-data regimes [113,114]. SSL generally consists of leveraging information from unlabeled data during a pretraining phase. This approach reduces the need for extensive datasets for subsequent trainings. The work by Chaitanya et al [115]. serves as an example of the application of SSL to the field of automatic segmentation of medical images. In their study, they demonstrated improved segmentation performance for the networks trained with SSL compared to the networks initialized from scratch for three different medical segmentation tasks.

Leveraging the anatomical information of an individual patient by utilizing previous images and contours of that same patient can also improve the segmentations. This approach is sometimes referred to as "patient-specific fine-tuning" and is especially promising in scenarios that require segmentations for the same patient across different time points, such as the different fractions in RT. Li et al. [116] proposed such an approach for online contouring for MR-guided adaptive radiotherapy. Their results demonstrated improved segmentation performance compared to the network trained without patient-specific fine-tuning. Their segmentations were also more accurate than the segmentations generated by existing deformable registration algorithms commonly employed in clinical settings.

Though employing fairly different methodologies, these techniques share a common thread with the work presented in chapters 2 and 3: the integration of clinically relevant prior knowledge during the training of segmentation networks is a promising strategy to improve the quality of the resulting automatic segmentations.

EVALUATION OF AUTOMATIC SEGMENTATIONS: WHAT IS A GOOD (AUTOMATIC) SEGMENTATION?

Determining the quality of the automatics segmentations was done in two ways throughout this thesis. In chapters 2 and 3, we geometrically compared the automatic segmentations to the manual segmentations made by expert radiologists. The aim of the radiologists was to accurately draw the extent of the tumor visible in the available imaging modalities. Therefore, a geometric comparison between the manual and automatic segmentations was adequate. This comparison was assessed with commonly used metrics: DSC, 95th Hausdorff distance (95th HD) and the mean surface distance (MSD).

In chapters 4 and 5, the manual segmentations used as ground truth for training and evaluation were performed by radiation oncologists. These segmentations were used in clinical practice to derive a treatment plan. A dosimetric evaluation was in this case also pertinent to evaluate the quality of the automatic segmentations. More specifically, we determined dose-volume parameters D90 and D98 for the automatic and expert delineations using the clinical dose distribution. Dosimetric evaluation of the automatic segmentation is important, because an error in the segmentation may have a different clinical impact depending on the dose that will be given to that point. This effect cannot be represented with geometric evaluation metrics, given that they do not take into account the dose distribution delivered to the patient.

In chapter 4, the segmentation performance of automatic segmentation of tumors in the cervix was further stratified in subgroups based on the tumor volume and the FIGO stage. Tumors with different FIGO stage or volume may appear differently on the MRI. Specifically, tumors with FIGO stage I will be limited to the cervix, while tumors with FIGO stage II and higher may extend to other anatomical structures, such as the vagina or the pelvic wall. The segmentation network potentially needs to look at different anatomical areas in the image to segment tumors depending on these clinical parameters. Therefore, differences in segmentation performance between the different subgroups of patients can arise. Another potential reason for performance differences is that some of these subgroups were under-represented in the training set. Regardless of the source of these differences, analyzing the segmentation performance separately for FIGO and volume can reveal a bias of the trained network towards certain subgroups of patients. Physicians could take this information into account to only use the segmentation network in the subgroups it performs best.

We used different evaluation metrics compared to other auto-segmentation works in literature. Firstly, it is not uncommon to find articles that report the DSC only. However, the DSC is volume-dependent by construction. Therefore, larger and more rounded structures typically result in higher values than smaller or more eccentric structures. This is particularly critical for structures with variable sizes, such as the tumors. Therefore we opted to always provide distance-based metrics together with the DSC. Secondly, the normalized surface distance (NSD) and the added path length (APL) are two quality metrics that are used in other works but not in this thesis. Although interesting metrics, they present shortcomings. The NSD is defined with a certain degree of tolerance, making them dependent on this hyper-parameter. The APL is not normalized or expressed with known units, making it less interpretable. Arguably, the distance-based metrics used in this work (95th HD, MSD and the surface DSC), also present shortcomings. The 95th HD primarily reflects the most significant error in the segmentation, often ignoring other relevant errors. MSD considers the entire contour but averages all errors together, which can lead to bias in the results, especially when gross errors are present. The surface DSC also depends on a tolerance parameter. To provide a more comprehensive assessment of contour quality, we recommend reporting a combination of these metrics.

Specifically for RT purposes, some geometric metrics may be more relevant than others. DSC has been shown not to correlate strongly with editing time [104] nor with dose/volume parameters [103]. In contrast, distance-based metrics, such as the 95th HD, the MSD, the surface DSC and the APL, have been shown to correlate more strongly with the editing time. Furthermore, a certain degree of geometric variability is expected in the manual tumor delineations due to interobserver variability. These geometric differences are taken into account by treatment margins for some tumor areas. If the error of the automatic segmentations is within the interobserver variability, the tumor segmentation might be clinically valid. Given that the treatment margins are defined as a certain distance around the tumor delineation, distance based metrics may also be more appropriate for this type of assessment.

Other works [42,117–119] have used the Turing test (or “Imitation game”) to assess the quality of the automatic segmentations. The rationale behind it is that if a human observer cannot distinguish whether the contour was automatically generated or not, the contour closely resembles a manually created contour. Therefore, it is likely clinically acceptable. However, the source of the contour (automatic or human) does not necessarily indicate that the contour is clinically acceptable. Gooding et al. [117] proposed to complement the Turing test with additional questions that directly refer to clinical acceptability of the contour. These questions related to whether the observer would perform changes to the contour or how large those changes would be. In any case, these tests are strongly tied to the expertise and environment of the observer. Consequently, we would recommend using them in complement with other evaluation techniques.

Overall, there are many approaches to evaluate the quality of a contour, each of them with their own advantages and limitations. Our recommendations would be two-fold: 1) to consider the clinical end goal of the contour, and use the evaluation framework that more closely assesses if that contour is acceptable, and 2) to combine the different evaluation metrics to describe the quality of the contour.

OPTIMIZING FOR THE RELEVANT LOSS FUNCTION

The loss function is a crucial part of the training of any neural network, including segmentation networks. Because of its importance, a multitude of different loss functions have been proposed in literature [30,70]. Ma et al. [30] grouped the available loss functions in three main categories: overlap-based loss functions, such as the DSC loss; distribution-based loss functions, such as the cross entropy loss or the focal loss; and boundary-based loss functions, such as the Hausdorff distance loss. Despite the large amount of loss functions available, the DSC loss is still selected in most works. A comprehensive evaluation of which loss function renders best segmentation performance for each specific task is rarely investigated.

In chapter 3 of this thesis, we trained the segmentation networks with different loss functions for the task of oropharyngeal cancer segmentation. The investigated loss functions were two overlap-based loss functions (DSC loss and Generalized DSC loss) and two distribution-based loss function (Focal Tversky loss and Unified Focal loss). No significant differences in DSC, 95th HD or MSD were found when training any of the loss functions. This suggests that the DSC loss was sufficient for our specific task. Similar conclusions were reached by Ma et al [30]. In their work, they compared 20 different loss functions on four different segmentation tasks. The showed that loss functions derived from the DSC were overall performing best.

All the loss functions compared in both our work and the work by Ma et al. reflect geometric differences between the automatic segmentation and the ground truth. This means that the segmentation networks are optimized during training to resemble the shape of the manual segmentations as closely as possible. However, the desired automatic segmentations for RT purposes are not necessarily those with the exact same shape than the manually acquired segmentations. In reality, the desired segmentations are those that reach the same dosimetric impact as the manual segmentations. Consequently, an interesting future line of research is to build loss functions that consider the dosimetric impact in its definition. Further in the future, the loss function could even consider the RT patient outcomes in their definition, such as tumor control and toxicity.

QUALITY ASSURANCE OF AUTOMATIC SEGMENTATION OF TARGETS

Current automatic segmentation techniques do not render good quality segmentations for tumors in all the cases. This hinders their use in clinical practice, because physicians would still need to check, and potentially correct, the automatic segmentations. A possible solution is the implementation of quality assurance (QA) algorithms for the automatic segmentations. These QA algorithms could help the physicians by flagging the segmentations that would require review.

In chapter 5, we identified a metric that could be used for QA of the automatic segmentations. The proposed quality metric showed good capability to distinguish between cases that would require review as compared to adequate automatic segmentations that could be used without further check. Furthermore, the proposed metric correlated strongly with the DSC but also with clinically relevant distance-based metrics. These results indicate the potential of the metric as a QA tool.

Our work differed from the current literature of QA for automatic segmentations in two ways. Firstly, most works only correlate QA metrics to the DSC [32–34,102]. As discussed in previous sections, DSC does not fully describe the quality of the contour. Therefore, we considered it important to correlate our metric to distance-based metrics as well. Secondly, our metric can be derived directly from the output of the segmentation network. Instead, other works often extract their QA metrics from uncertainty maps [32,33,98], computed by applying the entropy operator on the output of the network. However, the entropy operator is not injective and can therefore destroy relevant information of the score maps.

Even if we can detect the segmentations that would require review, physicians would still need to spend time in the correction. A potential solution would consist on signaling the areas where the automatic segmentation is likely wrong. This information would assist the physicians in adjusting the contour in a semi-automatic manner. Even further in the future, the QA approaches could provide insights on how each potential editing would affect the final RT treatment outcome. With this relevant clinical information at hand, physicians could then make the final decision regarding the contour adjustments.

CLINICAL IMPLEMENTATION OF TUMOR AUTO-SEGMENTATION.

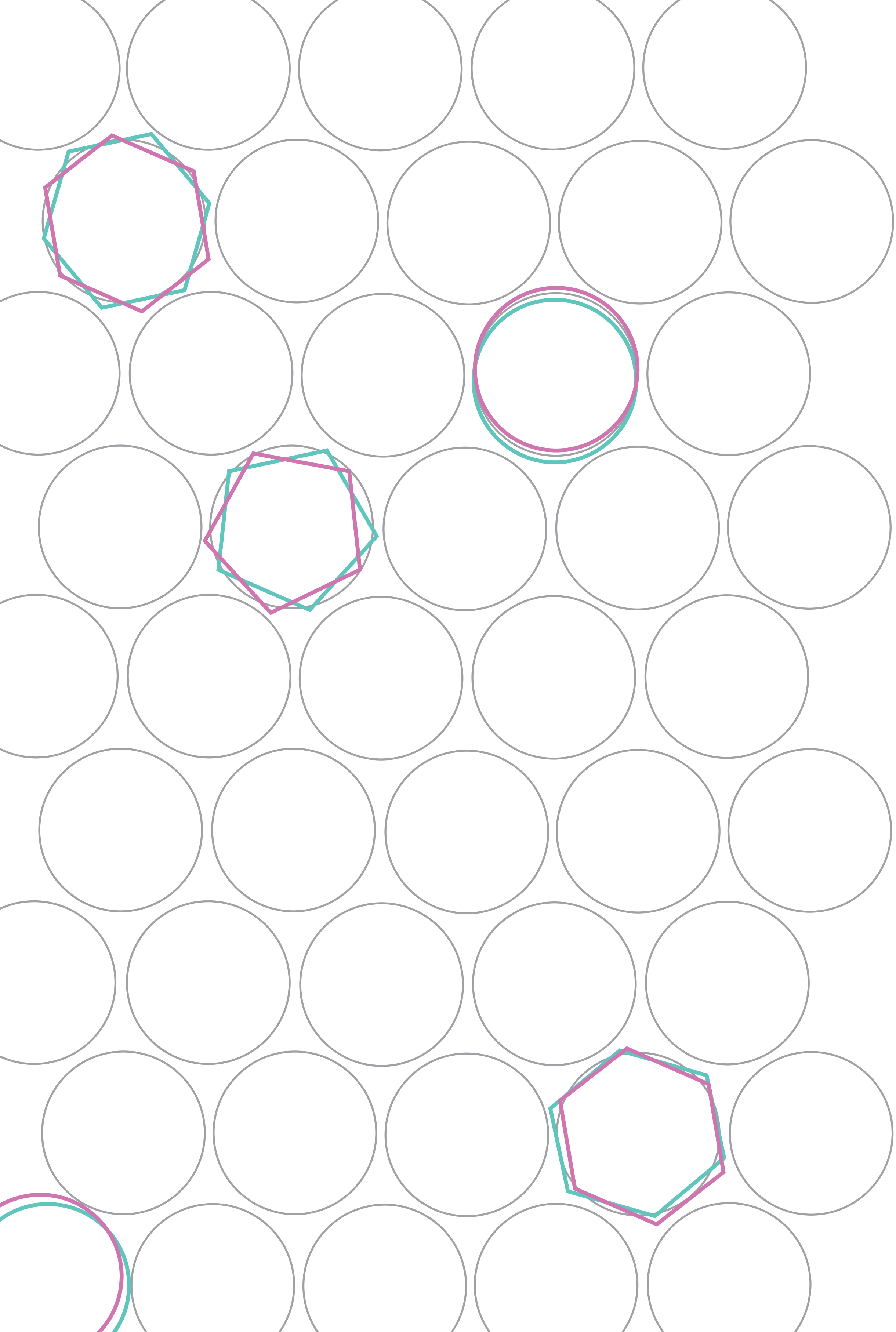
Ultimately, the aim is to implement automatic segmentation techniques for the tumors in clinical practice, as it is already the case for the OARs. As stated in the previous section, these techniques do not render acceptable segmentations in all the cases yet. Any errors in the tumor segmentation can have a large impact in the outcome of the patient, because tumors receive the highest radiation dose during the RT treatment. The responsibility of accepting the tumor segmentations lies with the radiation oncologists. However, they will not accept a segmentation unless they are certain of its quality.

The studies presented in this thesis can approach us towards clinical implementation in two ways. In chapters 2, 3 and 4, we focused on improving the quality of the automatic segmentations. By continuing to improve the auto-segmentations, we could eventually provide segmentations that radiation oncologists find acceptable in all cases. Furthermore, in chapter 5, we proposed a metric that can be used for QA purposes. With this tool in hand, the radiation oncologist could potentially distinguish which segmentations are clinically acceptable without the need of checking each auto-segmentation. However, further efforts are likely still needed to reach the clinical implementation of these techniques.

Vinod et al. [120] reviewed different applications to reduce the interobserver variability in target and OARs contouring. They observed that providing radiation oncologists with an automatic segmentation as a starting point that can be subsequently edited is an effective method to reduce contouring variability. One example of this was shown by Ferreira Silvério et al. [106] where they automatically segmented the mesorectum (CTV in MRI-guided rectal cancer treatment) and then asked an expert to manually correct the automatic segmentations. These corrections were not only comparable in terms of quality to the current clinical standard but also completed faster than the delineations made from scratch. Therefore, the clinical implementation of automatic segmentation techniques could already provide clinical value as an aid to the radiation oncologist to edit in a semi-automatic setting. Arguably, such an evaluation framework is a necessary step towards the clinical implementation of auto-segmentation tools. This approach can provide clinically acceptable segmentations more quickly than the current clinical approaches, while also increasing the trust in the auto-segmentations.

CONCLUSIONS

In this thesis, we explored the topic of automatic segmentation of tumors on MRI. Deep learning methods are already employed in clinical settings to segment anatomical structures. However, automatic segmentation of tumors remains in a research capacity, showcasing the complexity of the problem. A promising approach to improve the quality of the segmentations consists of utilizing prior information to guide the training of the segmentation network. Two examples have been demonstrated in this thesis: the reduction of context around the tumor and the incorporation of different MRI sequences. Furthermore, clinical relevant information should be considered both during the training and evaluation of these methods. During training, this can be achieved by defining clinically relevant loss functions. During the evaluation, this is possible by defining clear clinical end points. Besides potential improvements in the quality of the automatic segmentations, automatic QA can also play a crucial role in advancing towards clinical applicability. QA methods are not only important for safety reasons but also to increase the trust of clinicians in these techniques. Finally, a currently viable strategy involves an interactive approach in which a candidate auto-segmentation is provided as a starting point. This could presently reduce the time spent by the clinical staff in the segmentations, thereby enhancing the efficiency of the RT workflow.



Chapter 7

References

REFERENCES

1. World Health Organization (WHO). Cancer overview. 2021.
2. Abdel-Wahab M, Gondhowiardjo SS, Arthur ;, Rosa A, Lievens Y, Noura El-Haj ;, et al. Global Radiotherapy: Current Status and Future Directions-White Paper [Internet]. Vol. 7, JCO Global Oncol. 2021. Available from: https://ascopubs.org/go/authors/open-access
3. Borras JM, Lievens Y, Dunscombe P, Coffey M, Malicki J, Corral J, et al. The optimal utilization proportion of external beam radiotherapy in European countries: An ESTRO-HERO analysis. Radiotherapy and Oncology. 2015 Jul 1;116(1):38–44.
4. Barton MB, Jacob S, Shafiq J, Wong K, Thompson SR, Hanna TP, et al. Estimating the demand for radiotherapy from the evidence: A review of changes from 2003 to 2012. Radiotherapy and Oncology. 2014;112(1):140–4.
5. Njeh CF. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. Vol. 33, Journal of Medical Physics. 2008.
6. Vorwerk H, Zink K, Schiller R, Budach V, Böhmer D, Kampfer S, et al. Protection of quality and innovation in radiation oncology: The prospective multicenter trial the German Society of Radiation Oncology (DEGRO-QUIRO study): Evaluation of time, attendance of medical staff, and resources during radiotherapy with IMRT. Strahlentherapie und Onkologie. 2014;190(5):433–43.
7. Lindberg J, Holmström P, Hallberg S, Björk-Eriksson T, Olsson CE. A national perspective about the current work situation at modern radiotherapy departments. Clin Transl Radiat Oncol. 2020 Sep 1;24:127–34.
8. Sonke JJ, Aznar M, Rasch C. Adaptive Radiotherapy for Anatomical Changes. Vol. 29, Seminars in Radiation Oncology. W.B. Saunders; 2019. p. 245–57.
9. Lim-Reinders S, Keller BM, Al-Ward S, Sahgal A, Kim A. Online Adaptive Radiation Therapy. Vol. 99, International Journal of Radiation Oncology Biology Physics. Elsevier Inc.; 2017. p. 994–1003.
10. Jager EA, Kasperts N, Caldas-Magalhaes J, Philippens MEP, Pameijer FA, Terhaard CHJ, et al. GTV delineation in supraglottic laryngeal carcinoma: interobserver agreement of CT versus CT-MR delineation. Radiation Oncology [Internet]. 2015;10(1):26. Available from: https://doi.org/10.1186/s13014-014-0321-4
11. Blinde S, Mohamed ASR, Al-Mamgani A, Newbold K, Karam I, Robbins JR, et al. Large Interobserver Variation in the International MR-LINAC Oropharyngeal Carcinoma Delineation Study. International Journal of Radiation Oncology*Biology*Physics. 2017;99(2):e639–40.
12. Nowee ME, Voncken FEM, Kotte ANTJ, Goense L, van Rossum PSN, van Lier ALHMW, et al. Gross tumour delineation on computed tomography and positron emission tomography-computed tomography in oesophageal cancer: A nationwide study. Clin Transl Radiat Oncol. 2019 Jan 1;14:33–9.
13. Bowden P, Fisher R, Manus M Mac, Wirth A, Duchesne G, Millward M, et al. MEASUREMENT OF LUNG TUMOR VOLUMES USING THREE-DIMENSIONAL COMPUTER PLANNING SOFTWARE. 2002.
14. Cheon W, Jeong S, Jeong JH, Lim YK, Shin D, Lee SB, et al. Interobserver Variability Prediction of Primary Gross Tumor in a Patient with Non-Small Cell Lung Cancer. Cancers (Basel). 2022 Dec 1;14(23).

15. Burbach JPM, Kleijnen JPJ, Reerink O, Seravalli E, Philippens MEP, Schakel T, et al. Interobserver agreement of MRI-based tumor delineation for preoperative radiotherapy boost in locally advanced rectal cancer. Radiotherapy and Oncology. 2016 Feb 1;118(2):399–407.

16. Petrič P, Hudej R, Rogelj P, Blas M, Tanderup K, Fidarova E, et al. Uncertainties of target volume delineation in MRI guided adaptive brachytherapy of cervix cancer: A multi-institutional study. Radiotherapy and Oncology. 2013 Apr;107(1):6–12.

17. Cox S, Cleves A, Clementel E, Miles E, Staffurth J, Gwynne S. Impact of deviations in target volume delineation – Time for a new RTQA approach? Vol. 137, Radiotherapy and Oncology. Elsevier Ireland Ltd; 2019. p. 1–8.

18. Brunenberg EJL, Steinseifer IK, van den Bosch S, Kaanders JHAM, Brouwer CL, Gooding MJ, et al. External validation of deep learning-based contouring of head and neck organs at risk. Phys Imaging Radiat Oncol [Internet]. 2020;15:8–15. Available from: https://doi.org/10.1016/j.phro.2020.06.006

19. Chen X, Sun S, Bai N, Han K, Liu Q, Yao S, et al. A deep learning-based auto-segmentation system for organs-at-risk on whole-body computed tomography images for radiation therapy. Radiotherapy and Oncology. 2021 Jul 1;160:175–84.

20. Mohammadi R, Shokatian I, Salehi M, Arabi H, Shiri I, Zaidi H. Deep learning-based auto-segmentation of organs at risk in high-dose rate brachytherapy of cervical cancer. Radiotherapy and Oncology [Internet]. 2021;159:231–40. Available from: https://doi.org/10.1016/j.radonc.2021.03.030

21. Savenije MHF, Maspero M, Sikkes GG, Van Der Voort Van Zyp JRN, Alexis AN, Bol GH, et al. Clinical implementation of MRI-based organs-at-risk auto-segmentation with convolutional networks for prostate radiotherapy. Radiation Oncology. 2020 May 11;15(1).

22. Mirada Medical. Mirada Autocontouring DLCexpert.

23. Raysearch Labs. Machine learning in Raystation.

24. Therapanacea Annotate.

25. Limbus AI.

26. Lin H, Xiao H, Dong L, Teo KBK, Zou W, Cai J, et al. Deep learning for automatic target volume segmentation in radiation therapy: A review. Vol. 11, Quantitative Imaging in Medicine and Surgery. AME Publishing Company; 2021. p. 4847–58.

27. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated Tumor Segmentation in Radiotherapy. Vol. 32, Seminars in Radiation Oncology. W.B. Saunders; 2022. p. 319–29.

28. Savjani RR, Lauria M, Bose S, Deng J, Yuan Y, Andrearczyk V. Automated Tumor Segmentation in Radiotherapy. Vol. 32, Seminars in Radiation Oncology. W.B. Saunders; 2022. p. 319–29.

29. Zhou T, Ruan S, Canu S. A review: Deep learning for medical image segmentation using multi-modality fusion. Array. 2019 Sep 1;3–4.

30. Ma J, Chen J, Ng M, Huang R, Li Y, Li C, et al. Loss odyssey in medical image segmentation. Med Image Anal. 2021;71.

31. Meyer P, Noblet V, Mazzara C, Lallement A. Survey on deep learning for radiotherapy. Comput Biol Med. 2018;98:126–46.

32. Jungo A, Balsiger F, Reyes M. Analyzing the Quality and Challenges of Uncertainty Estimations for Brain Tumor Segmentation. Front Neurosci. 2020 Apr 8;14.

33. Roy AG, Conjeti S, Navab N, Wachinger C. Bayesian QuickNAT: Model uncertainty in deep whole-brain segmentation for structure-wise quality control. Neuroimage. 2019 Jul 15;195:11–22.

34. Judge T, Bernard O, Porumb M, Chartsias A, Beqiri A, Jodoin PM. CRISP - Reliable Uncertainty Estimation for Medical Image Segmentation. Medical Image Computing and Computer Assisted Intervention – MICCAI 2022 . 2022;492–502.

35. Abdar M, Pourpanah F, Hussain S, Rezazadegan D, Liu L, Ghavamzadeh M, et al. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. Vol. 76, Information Fusion. Elsevier B.V.; 2021. p. 243–97.

36. Sander J, de Vos BD, Wolterink JM, Išgum I. Towards increased trustworthiness of deep learning segmentation methods on cardiac MRI. 2018 Sep 27; Available from: http://arxiv.org/abs/1809.10430

37. Carannante G, Dera D, Bouaynaya NC, Fathallah-Shaykh HM, Rasool G. SUPER-Net: Trustworthy Medical Image Segmentation with Uncertainty Propagation in Encoder-Decoder Networks. 2021 Nov 10; Available from: http://arxiv.org/abs/2111.05978

38. Ng J, Gregucci F, Pennell RT, Nagar H, Golden EB, Knisely JPS, et al. MRI-LINAC: A transformative technology in radiation oncology. Vol. 13, Frontiers in Oncology. Frontiers Media S.A.; 2023.

39. Bhalodiya JM, Lim Choi Keung SN, Arvanitis TN. Magnetic resonance image-based brain tumour segmentation methods: A systematic review. Vol. 8, Digital Health. SAGE Publications Inc.; 2022.

40. Tanderup K, Fokdal LU, Lindegaard DMSc JC, Jürgenliemk-Schulz I, C De Leeuw AA, Segedin B, et al. MRI-guided adaptive brachytherapy in locally advanced cervical cancer (EMBRACE-I): a multicentre prospective cohort study [Internet]. Vol. 22, Articles Lancet Oncol. 2021. Available from: https://www.embracestudy.dk/

41. Delaney G, Jacob S, Barton M. Estimation of an optimal external beam radiotherapy utilization rate for head and neck carcinoma. Cancer. 2005;103:2218.

42. Wahid KA, Ahmed S, He R, van Dijk L V., Teuwen J, McDonald BA, et al. Evaluation of deep learning-based multiparametric MRI oropharyngeal primary tumor auto-segmentation and investigation of input channel effects: Results from a prospective imaging registry. Clin Transl Radiat Oncol [Internet]. 2022;32(July 2021):6–14. Available from: https://doi.org/10.1016/j.ctro.2021.10.003

43. Pötter R, Tanderup K, Kirisits C, de Leeuw A, Kirchheiner K, Nout R, et al. The EMBRACE II study: The outcome and prospect of two decades of evolution within the GEC-ESTRO GYN working group and the EMBRACE studies. Clin Transl Radiat Oncol. 2018;9(March):48–60.

44. Fitzmaurice C, Allen C, Barber RM, Barregard L, Bhutta ZA, Brenner H, et al. Global, Regional, and National Cancer Incidence, Mortality, Years of Life Lost, Years Lived With Disability, and Disability-Adjusted Life-years for 32 Cancer Groups, 1990 to 2015. JAMA Oncol. 2017 Apr 1;3(4):524.

45. DSouza AM, Chen L, Wu Y, Abidin AZ, Xu C, Wismüller A. MRI tumor segmentation with densely connected 3D CNN. In 2018. p. SPIE Medical Imaging Proceedings VOL 10574.

46. Li J, Chen H, Li Y, Peng Y. A novel network based on densely connected fully convolutional networks for segmentation of lung tumors on multi-modal MR images. In: ACM International Conference Proceeding Series. 2019. p. 1–5.

47. Li X, Chen H, Qi X, Dou Q, Fu CW, Heng PA. H-DenseUNet: Hybrid Densely Connected UNet for Liver and Tumor Segmentation from CT Volumes. IEEE Trans Med Imaging. 2018;37(12):2663–74.

48. Trebeschi S, Van Griethuysen JJM, Lambregts DMJ, Lahaye MJ, Parmer C, Bakers FCH, et al. Deep Learning for Fully-Automated Localization and Segmentation of Rectal Cancer on Multiparametric MR. Sci Rep. 2017;8(1):2589.

49. Guo Z, Guo N, Gong K, Zhong S, Li Q. Gross tumor volume segmentation for head and neck cancer radiotherapy using deep dense multi-modality network. Phys Med Biol. 2019;64(20).

50. Cardenas CE, McCarroll RE, Court LE, Elgohari BA, Elhalawani H, Fuller CD, et al. Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function. Int J Radiat Oncol Biol Phys. 2018;101(2):468–78.

51. Men K, Chen X, Zhang Y, Zhang T, Dai J, Yi J, et al. Deep deconvolutional neural network for target segmentation of nasopharyngeal cancer in planning computed tomography images. Front Oncol. 2017;7(315).

52. Anderson CM, Sun W, Buatti JM, Maley JE, Policeni B, Mott SL, et al. Interobserver and intermodality variability in GTV delineation on simulation CT, FDG-PET, and MR Images of Head and Neck Cancer. Jacobs J Radiat Oncol. 2014;1(1):006.

53. Çiçek Ö, Abdulkadir A, Lienkamp SS, Brox T, Ronneberger O. 3D U-Net: Learning Dense Volumetric Segmentation from Sparse Annotation BT - Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. 2016. p. 424–32.

54. Zeng G, Yang X, Li J, Yu L, Heng PA, Zheng G. 3D U-net with multi-level deep supervision: Fully automatic segmentation of proximal femur in 3D MR images. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. p. 274–82.

55. Gordienko Y, Gang P, Hui J, Zeng W, Kochura Y, Alienin O, et al. Deep learning with lung segmentation and bone shadow exclusion techniques for chest X-ray analysis of lung cancer. Advances in Intelligent Systems and Computing. 2019;754:638–47.

56. Norman B, Pedoia V, Majumdar S. Use of 2D U-net convolutional neural networks for automated cartilage and meniscus segmentation of knee MR imaging data to determine relaxometry and morphometry. Radiology. 2018;288(1):177–85.

57. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2015. p. 234–41.

58. Sudre CH, Li W, Vercauteren T, Ourselin S, Jorge Cardoso M. Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017. p. Vol. 10553.

59. Kingma DP, Ba JL. Adam: A method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015.

60. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Med Phys. 2019;46(1):e1–36.

61. Bos P, van den Brekel MWM, Gouw ZAR, Al-Mamgani A, Waktola S, Aerts HJWL, et al. Clinical variables and magnetic resonance imaging-based radiomics predict human papillomavirus status of oropharyngeal cancer. Head Neck. 2021;43(2):485–95.

62. Brouwer CL, Dinkla AM, Vandewinckele L, Crijns W, Claessens M, Verellen D, et al. Machine learning applications in radiation oncology: Current use and needs to support clinical implementation. Phys Imaging Radiat Oncol. 2020;16:144–8.

63. Boldrini L, Bibault JE, Masciocchi C, Shen Y, Bittner MI. Deep Learning: A Review for the Radiation Oncologist. Front Oncol. 2019;9:977.

64. Vrtovec T, Močnik D, Strojan P, Pernuš F, Ibragimov B. Auto-segmentation of organs at risk for head and neck radiotherapy planning: From atlas-based to deep learning methods. Med Phys. 2020;47(9):e929–50.

65. Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning. Phys Imaging Radiat Oncol. 2021;19:39–44.

66. Small H, Ventura J. Handling Unbalanced Data in Deep Image Segmentation. 2017;

67. Kochkarev A, Khvostikov A, Korshunov D, Krylov A, Boguslavskiy M. Data balancing method for training segmentation neural networks. CEUR Workshop Proc. 2020;2744:1–9.

68. Tsung-Yi Lin. Focal Loss for Dense Object Detection (RetinaNet). 13C-NMR of Natural Products. 2017;30–3.

69. Salehi SSM, Erdogmus D, Gholipour A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2017;10541 LNCS:379–87.

70. Yeung M, Sala E, Schönlieb CB, Rundo L. Unified Focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. Vol. 95, Computerized Medical Imaging and Graphics. 2022.

71. Abraham N, Khan NM. A novel focal tversky loss function with improved attention u-net for lesion segmentation. Proceedings - International Symposium on Biomedical Imaging. 2019;2019-April:683–7.

72. Feng X, Qing K, Tustison NJ, Meyer CH, Chen Q. Deep convolutional neural network for segmentation of thoracic organs-at-risk using cropped 3D images. Med Phys. 2019;46(5):2169–80.

73. Balagopal A, Kazemifar S, Nguyen D, Lin MH, Hannan R, Owrangi A, et al. Fully automated organ segmentation in male pelvic CT images. Phys Med Biol. 2018;63(24).

74. Wang Y, Zhao L, Wang M, Song Z. Organ at Risk Segmentation in Head and Neck CT Images Using a Two-Stage Segmentation Framework Based on 3D U-Net. IEEE Access. 2019;7:144591–602.

75. Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. In: 2016 Fourth International Conference on 3D Vision (3DV) [Internet]. IEEE; 2016. p. 565–71. Available from: https://iopscience.iop.org/article/10.1088/0022-3735/6/9/035

76. 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;18:203–11.

77. Bielak L, Wiedenmann N, Nicolay NH, Lottner T, Fischer J, Bunea H, et al. Automatic tumor segmentation with a convolutional neural network in multiparametric mri: Influence of distortion correction. Tomography. 2019;5(3):292–9.

78. Schouten JPE, Noteboom S, Martens RM, Mes SW, Leemans CR, de Graaf P, et al. Automatic segmentation of head and neck primary tumors on MRI using a multi-view CNN. Cancer Imaging. 2022;22(1):1–9.

79. Andrearczyk V, Oreiller V, Jreige M, Vallières M, Castelli J, Elhalawani H, et al. Overview of the HECKTOR Challenge at MICCAI 2020: Automatic Head and Neck Tumor Segmentation in PET/CT. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2021;12603 LNCS(April):1–21.

80. Ren J, Eriksen JG, Nijkamp J, Korreman SS. Comparing different CT, PET and MRI multi-modality image combinations for deep learning-based head and neck tumor segmentation. Acta Oncol (Madr) [Internet]. 2021;60(11):1399–406. Available from: https://doi.org/10.1080/0284186X.2021.1949034

81. Haie-Meder C, Pötter R, Van Limbergen E, Briot E, De Brabandere M, Dimopoulos J, et al. Recommendations from Gynaecological (GYN) GEC-ESTRO Working Group (I): Concepts and terms in 3D image based 3D treatment planning in cervix cancer brachytherapy with emphasis on MRI assessment of GTV and CTV. Radiotherapy and Oncology. 2005;74(3):235–45.

82. Yoganathan SA, Paul SN, Paloor S, Torfeh T, Chandramouli SH, Hammoud R, et al. Automatic segmentation of magnetic resonance images for high-dose-rate cervical cancer brachytherapy using deep learning. Med Phys. 2022 Mar 1;49(3):1571–84.

83. Wong J, Kolbeck C, Giambattista J, Giambattista JA, Huang V, Jaswal JK. Deep Learning-based Auto-Segmentation for Pelvic Organs at Risk and Clinical Target Volumes in Intracavitary High Dose Rate Brachytherapy. International Journal of Radiation Oncology*Biology*Physics [Internet]. 2020 Nov;108(3):e284. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0360301620321003

84. Zhang D, Yang Z, Jiang S, Zhou Z, Meng M, Wang W. Automatic segmentation and applicator reconstruction for CT-based brachytherapy of cervical cancer using 3D convolutional neural networks. J Appl Clin Med Phys. 2020 Oct 1;21(10):158–69.

85. He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 770–8.

86. Tomizawa K, Kaminuma T, Murata K, Noda SE, Irie D, Kumazawa T, et al. Figo 2018 staging for cervical cancer: Influence on stage distribution and outcomes in the 3d-image-guided brachytherapy era. Cancers (Basel). 2020 Jul 1;12(7):1–10.

87. Pötter R, Haie-Meder C, Van Limbergen E, Barillot I, De Brabandere M, Dimopoulos J, et al. Recommendations from gynaecological (GYN) GEC ESTRO working group (II): Concepts and terms in 3D image-based treatment planning in cervix cancer brachytherapy - 3D dose volume parameters and aspects of 3D image-based anatomy, radiation physics, radiobiology. Radiotherapy and Oncology. 2006 Jan;78(1):67–77.

88. Hellebust TP, Tanderup K, Lervåg C, Fidarova E, Berger D, Malinen E, et al. Dosimetric impact of interobserver variability in MRI-based delineation for cervical cancer brachytherapy. Radiotherapy and Oncology. 2013 Apr;107(1):13–9.

89. Liu Z, Tong L, Chen L, Jiang Z, Zhou F, Zhang Q, et al. Deep learning based brain tumor segmentation: a survey. Complex and Intelligent Systems. 2022 Feb 1;

90. Biratu ES, Schwenker F, Ayano YM, Debelee TG. A survey of brain tumor segmentation and classification algorithms. J Imaging. 2021 Sep 1;7(9).

91. Rodríguez Outeiral R, González PJ, Schaake EE, van der Heide UA, Simões R. Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy. Radiation Oncology. 2023 Dec 1;18(1).

92. Zabihollahy F, Viswanathan AN, Schmidt EJ, Lee J. Fully automated segmentation of clinical target volume in cervical cancer from magnetic resonance imaging with convolutional neural network. J Appl Clin Med Phys. 2022 Sep 1;23(9).

93. Fransson S, Tilly D, Strand R. Patient specific deep learning based segmentation for magnetic resonance guided prostate radiotherapy. Phys Imaging Radiat Oncol. 2022 Jul 1;23:38–42.

94. Cha E, Elguindi S, Onochie I, Gorovets D, Deasy JO, Zelefsky M, et al. Clinical implementation of deep learning contour autosegmentation for prostate radiotherapy. Radiotherapy and Oncology. 2021 Jun 1;159:1–7.

95. Sahiner B, Pezeshk A, Hadjiiski LM, Wang X, Drukker K, Cha KH, et al. Deep learning in medical imaging and radiation therapy. Vol. 46, Medical Physics. John Wiley and Sons Ltd; 2019. p. e1–36.

96. van den Berg CAT, Meliadò EF. Uncertainty Assessment for Deep Learning Radiotherapy Applications. Vol. 32, Seminars in Radiation Oncology. W.B. Saunders; 2022. p. 304–18.

97. Claessens M, Oria CS, Brouwer CL, Ziemer BP, Scholey JE, Lin H, et al. Quality Assurance for AI-Based Applications in Radiation Therapy. Vol. 32, Seminars in Radiation Oncology. W.B. Saunders; 2022. p. 421–31.

98. McClure P, Rho N, Lee JA, Kaczmarzyk JR, Zheng CY, Ghosh SS, et al. Knowing What You Know in Brain Segmentation Using Bayesian Deep Neural Networks. Front Neuroinform. 2019 Oct 17;13.

99. van Rooij W, Verbakel WF, Slotman BJ, Dahele M. Using Spatial Probability Maps to Highlight Potential Inaccuracies in Deep Learning-Based Contours: Facilitating Online Adaptive Radiation Therapy. Adv Radiat Oncol. 2021 Mar 1;6(2).

100. Isaksson LJ, Summers P, Bhalerao A, Gandini S, Raimondi S, Pepa M, et al. Quality assurance for automatically generated contours with additional deep learning. Insights Imaging. 2022 Dec 1;13(1).

101. Chen X, Men K, Chen B, Tang Y, Zhang T, Wang S, et al. CNN-Based Quality Assurance for Automatic Segmentation of Breast Cancer in Radiotherapy. Front Oncol. 2020 Apr 28;10.

102. DeVries T, Taylor GW. Leveraging Uncertainty Estimates for Predicting Segmentation Quality. 2018 Jul 2; Available from: http://arxiv.org/abs/1807.00502

103. Kaderka R, Gillespie EF, Mundt RC, Bryant AK, Sanudo-Thomas CB, Harrison AL, et al. Geometric and dosimetric evaluation of atlas based auto-segmentation of cardiac structures in breast cancer patients. Radiotherapy and Oncology. 2019 Feb 1;131:215–20.

104. Vaassen F, Hazelaar C, Vaniqui A, Gooding M, van der Heyden B, Canters R, et al. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. Phys Imaging Radiat Oncol. 2020 Jan 1;13:1–6.

105. Valentini V, Boldrini L, Damiani A, Muren LP. Recommendations on how to establish evidence from auto-segmentation software in radiotherapy. Vol. 112, Radiotherapy and Oncology. Elsevier Ireland Ltd; 2014. p. 317–20.

106. Ferreira Silvério N, van den Wollenberg W, Betgen A, Wiersema L, Marijnen C, Peters F, et al. Evaluation of Deep Learning target auto-contouring for MRI-guided online adaptive treatment of rectal cancer. Under review.

107. Gal Y, Ghahramani Z. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. 2015 Jun 6; Available from: http://arxiv.org/abs/1506.02142

108. Jager L. Towards accurate target delineation for head and neck cancer. 2017.

109. Fang Y, Wang J, Ou X, Ying H, Hu C, Zhang Z, et al. The impact of training sample size on deep learning-based organ auto-segmentation for head-and-neck patients. Phys Med Biol. 2021 Sep 21;66(18).

110. Ravishankar H. and Venkataramani R and TS and SP and V V. Learning and Incorporating Shape Models for Semantic Segmentation. In: Descoteaux Maxime and Maier-Hein L and FA and JP and CDL and DS, editor. Medical Image Computing and Computer Assisted Intervention – MICCAI 2017. Cham: Springer International Publishing; 2017. p. 203–11.

111. Tong N, Gou S, Yang S, Cao M, Sheng K. Shape constrained fully convolutional DenseNet with adversarial training for multiorgan segmentation on head and neck CT and low-field MR images. Med Phys. 2019 Jun 1;46(6):2669–82.

112. Elmahdy MS, Beljaards L, Yousefi S, Sokooti H, Verbeek F, Van Der Heide UA, et al. Digital Object Identifier Joint Registration and Segmentation via Multi-Task Learning for Adaptive Radiotherapy of Prostate Cancer [Internet]. Available from: https://github.com/moelmahdy/JRS-MTL.

113. Huang SC, Pareek A, Jensen M, Lungren MP, Yeung S, Chaudhari AS. Self-supervised learning for medical image classification: a systematic review and implementation guidelines. Vol. 6, npj Digital Medicine. Nature Research; 2023.

114. Wang WC, Ahn E, Feng D, Kim J. A Review of Predictive and Contrastive Self-supervised Learning for Medical Images. Machine Intelligence Research. 2023 Aug 3;

115. Chaitanya K, Erdil E, Karani N, Konukoglu E. Contrastive learning of global and local features for medical image segmentation with limited annotations [Internet]. Available from: https://github.com/krishnabits001/domain_specific_cl.

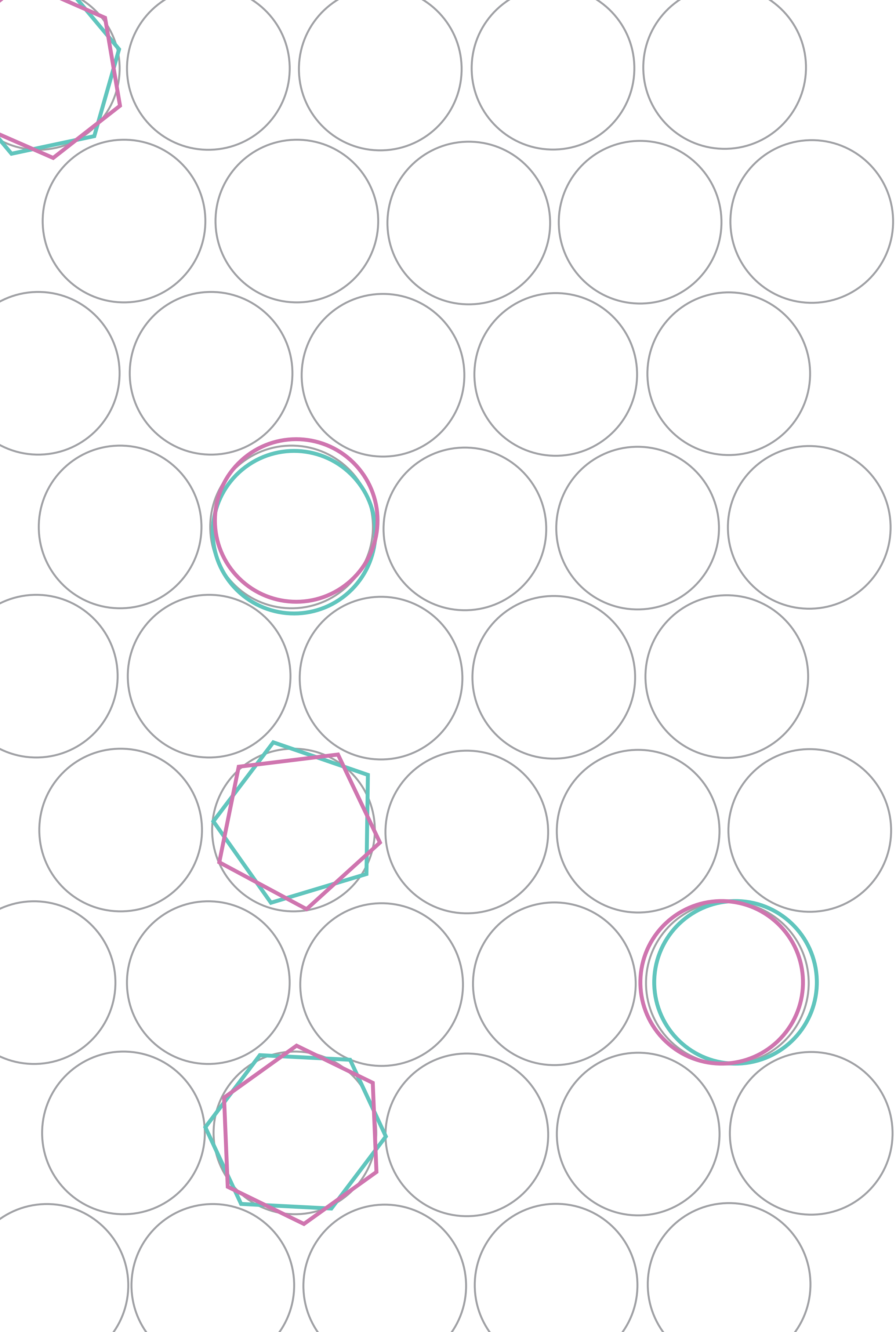
116. Li Z, Zhang W, Li B, Zhu J, Peng Y, Li C, et al. Patient-specific daily updated deep learning auto-segmentation for MRI-guided adaptive radiotherapy. Radiotherapy and Oncology. 2022 Dec 1;177:222–30.

117. Gooding MJ, Smith AJ, Tariq M, Aljabar P, Peressutti D, van der Stoep J, et al. Comparative evaluation of autocontouring in clinical practice: A practical method using the Turing test. Med Phys. 2018 Nov 1;45(11):5105–15.

118. Chung SY, Chang JS, Kim YB. Comprehensive clinical evaluation of deep learning-based auto-segmentation for radiotherapy in patients with cervical cancer. Front Oncol. 2023;13.

119. Wu Y, Kang K, Han C, Wang S, Chen Q, Chen Y, et al. A blind randomized validated convolutional neural network for auto-segmentation of clinical target volume in rectal cancer patients receiving neoadjuvant radiotherapy. Cancer Med. 2022 Jan 1;11(1):166–75.

120. Vinod SK, Min M, Jameson MG, Holloway LC. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. Vol. 60, Journal of Medical Imaging and Radiation Oncology. Blackwell Publishing; 2016. p. 393–406.



Appendices

Summary
Samenvatting
Acknowledgements
List of publications
Curriculum vitae

SUMMARY

Tumor segmentation is a crucial part of the radiotherapy treatment workflow. In the current clinical practice, this task is done manually by expert clinicians. This is time-consuming. A promising alternative is to automate this task with deep learning (DL) techniques. These methods have already been successfully applied clinically to segment other relevant structures for the radiotherapy treatment planning, such as the organs at risk. However, the automatic segmentation of tumors is not yet part of the clinical workflow. The aim of this thesis was to implement DL techniques to deliver clinically acceptable tumor segmentations. These algorithms were applied in two different MRI-based cohorts: a cohort of oropharyngeal primary tumors in multiparametric diagnostic MRIs and a cohort of cervical cancer gross tumor volume in MRI images of brachytherapy treatment.

When clinicians delineate the oropharyngeal cancer, they often rely on the information from several MRI sequences, such as the T1 weighted after gadolinium injection, T2 weighted or T1 weighted sequences. We hypothesised that DL methods may also benefit from combining the information from these different MRI sequences. Therefore, in **Chapter 2**, we studied the effect of combining different MRI sequences as input for the segmentation networks. Indeed, the performance of the DL networks improved when combining different MRI sequences as inputs compared to the use of single sequences.

Given the complexity of the task of oropharyngeal cancer segmentation, we hypothesised that we could achieve better auto-segmentations by simplifying the task. To that end, we split the task in two stages: first a coarse localization of the tumor and then its fine segmentation. In **Chapter 2**, we implemented this two stage approach in a semi-automatic manner. The first stage consisted of a localization step manually performed by clinicians, who were asked to draw a box around the tumor. Then, the final segmentation was performed by a DL network within the drawn box. As expected, the auto-segmentations rendered by this semi-automatic two stage approach outperformed the auto-segmentations achieved when directly providing the whole image as input.

The oropharyngeal cancer is present in a substantially smaller amount of voxels compared to the rest of structures in the MRI image. In the field of DL for automatic segmentation, this is known as the class imbalance problem, and can result in poor segmentation performance. In **Chapter 3**, we studied two different strategies to tackle this problem. One of these strategies was the use of different loss functions to train the segmentation networks. The most commonly used loss function when training segmentation networks is the Dice loss function, even though it is known to be suboptimal for smaller structures. In this chapter, we trained the segmentation networks with different loss functions designed to tackle class imbalance. Our results showed that all the proposed loss functions performed comparably to each other for the case of oropharyngeal cancer segmentation.

Another strategy to tackle the class imbalance problem consisted of the implementation of a two stage approach. Similarly to the previous chapter, we split the task in a localization step and a segmentation step. However, in **Chapter 3**, both stages were performed by neural networks thereby fully automating the semi-automatic approach proposed in the previous chapter. We demonstrated that our proposed two stage approach was an effective strategy to mitigate the class imbalance problem.

Brachytherapy is part of the standard of care for locally advanced cervical cancer. In this type of therapy, an applicator is placed inside the patient for treatment delivery. The tumor segmentations are made with this applicator inside, which can be uncomfortable for the patient. Therefore, it is desirable to acquire the tumor segmentations as promptly as possible. Due to these time constraints, the need for automatic segmentation is even more critical in the case of brachytherapy. In **Chapter 4**, we assessed the quality of the automatic segmentations of the cervical cancer gross tumor volume on brachytherapy MRI images. This assessment was performed both geometrically and dosimetrically. Our results showed similar dose-volume parameters as the manual segmentations used clinically, indicating that current DL methods can already render close to clinically valid tumor auto-segmentation in some cases.

In other cases, DL methods still produce tumor auto-segmentations that would not be clinically valid. Consequently, clinicians would still need to verify whether these auto-segmentations are acceptable for clinical use, limiting the time-gains of automatic segmentation methods. Therefore, there is a need for metrics that describe the quality of the auto-segmentations. However, common metrics to assess the quality of auto-segmentations also rely on comparing them to manually drawn segmentations, making them unsuitable for quality assurance. In **Chapter 5**, we identified a quality metric that can be generated directly from the output of the network. This quality metric had a high capability to distinguish between well and poorly performing auto-segmentations, showcasing its potential for quality assurance.

In conclusion, in this thesis we implemented different DL based approaches for automatic segmentation of tumors on MRI images. Strategies that provided relevant prior information to the segmentation networks were proved effective to increase the quality of the auto-segmentations, such as combining different MRI sequences as input or restricting the context around the tumor. Furthermore, we illustrated that current auto-segmentation frameworks can already render auto-segmentations that are comparable to clinically valid segmentations. Finally, besides the demonstrated improvements in the quality of the tumor auto-segmentations, we proposed a quality assurance metric that can distinguish between well and poorly performing cases. This type of metrics will potentially play a crucial role in advancing auto-segmentation methods for tumors towards clinical applicability.

SAMENVATTING

Tumorsegmentatie is een cruciaal onderdeel van de workflow voor een radiotherapeutische behandeling. In de huidige klinische praktijk wordt deze taak handmatig uitgevoerd door deskundige artsen. Dit is een tijdrovende activiteit. Een veelbelovend alternatief is om deze taak te automatiseren met behulp van deep learning (DL) technieken. Deze methoden zijn al succesvol toegepast in de kliniek om andere relevante structuren voor de radiotherapie behandeling te segmenteren, zoals de risico-organen. Echter, de automatische segmentatie van tumoren maakt nog geen deel uit van de klinische workflow. Het doel van dit proefschrift was om DL-technieken te implementeren waarmee klinisch acceptabele tumorsegmentaties worden geleverd. Deze algoritmen werden toegepast in twee verschillende MRI cohorten: een cohort van patiënten met primaire tumoren in de mond-keelholte (orofarynx) met multiparametrische diagnostische MRI scans en een cohort van patiënten met baarmoederhalskanker met MRI-beelden ter voorbereiding van de behandeling met brachytherapie.

Bij het segmenteren van de orofaryngeale tumor maken artsen vaak gebruik van informatie uit verschillende MRI-sequenties, zoals de T1-gewogen sequenties voor en na gadoliniuminjectie, en T2-gewogen sequenties. Onze hypothese was dat DL-methoden ook baat zouden kunnen hebben bij het combineren van de informatie uit deze verschillende MRI-sequenties. Daarom hebben we in **Hoofdstuk 2** het effect bestudeerd van het combineren van verschillende MRI-sequenties als input voor de segmentatie netwerken. De prestaties van de DL-netwerken verbeterden inderdaad toen verschillende MRI-sequenties als input werden gecombineerd in vergelijking met het gebruik van een enkele sequentie.

Gezien de complexiteit van de taak van het segmenteren van orofaryngeale tumoren, verwachtten we dat we betere autosegmentaties konden bereiken door de taak te vereenvoudigen. Met dat doel hebben wij de taak opgesplitst in twee fasen: Eerst een grove lokalisatie van de tumor, en vervolgens een gedetailleerde segmentatie. In **Hoofdstuk 2** hebben we deze methode op een semi-automatische manier geïmplementeerd. De eerste fase bestond uit handmatige lokalisatie van de tumor door artsen, die werden gevraagd om een kubus rond de tumor te tekenen. Vervolgens werd de definitieve segmentatie uitgevoerd door een DL-netwerk binnen het getekende vak. Zoals verwacht, presteerden de auto-segmentaties die door de twee-fasen methode werden verkregen beter dan de auto-segmentaties die werden verkregen door direct de hele afbeelding als input te geven.

Orofaryngeale kanker is aanwezig in aanzienlijk minder voxels vergeleken met de rest van de structuren in de MRI-afbeelding. In het veld van automatische segmentatie met DL staat dit bekend als het ‘class imbalance’ probleem en kan leiden tot segmentaties van slechte kwaliteit. In **Hoofdstuk 3** hebben we twee verschillende strategieën bestudeerd om dit probleem aan te pakken. Een van deze strategieën was het gebruik van verschillende loss functies om de segmentatie netwerken te trainen. De meest gebruikte loss functie bij het trainen van segmentatie netwerken is de Dice loss functie, ook al is bekend dat deze suboptimaal is voor kleinere structuren. In dit hoofdstuk hebben we de segmentatie netwerken getraind met verschillende loss functies die zijn ontworpen om het 'class imbalance' probleem aan te pakken. Onze resultaten toonden aan dat alle geïncludeerde loss functies vergelijkbaar presteerden voor orofaryngeale kankersegmentatie.

Een andere strategie om het 'class imbalance' probleem aan te pakken, bestond uit de implementatie van een twee-fasen methode. Net als in het vorige hoofdstuk hebben we de taak opgesplitst in een lokalisatiestap en een segmentatiestap. Echter, in **Hoofdstuk 3** werden beide fasen uitgevoerd door neurale netwerken, waardoor de semi-automatische methode volledig werd geautomatiseerd. We hebben aangetoond dat onze voorgestelde twee-fasen methode een effectieve strategie was om het 'class imbalance' probleem te verminderen.

Brachytherapie maakt deel uit van de behandeling voor lokaal gevorderde baarmoederhalskanker. Bij deze soort therapie wordt een applicator bij de patiënt ingebracht waarbinnen radioactieve bronnen worden geleid voor bestraling. De tumorsegmentaties worden gemaakt wanneer deze applicator is ingebracht, Omdat de applicator ongemakkelijk kan zijn voor de patiënt, is het wenselijk om de tumorsegmentaties zo snel mogelijk te verkrijgen. In **Hoofdstuk 4** hebben we de kwaliteit beoordeeld van de automatische segmentaties van de primaire tumor op MRI-beelden verkregen bij brachytherapie. Deze beoordeling werd zowel gebasserd op geometrische als dosimetrische criteria. Onze resultaten lieten dosis-volume parameters zien die vergelijkbaar waren met die van de klinisch gebruikte handmatige segmentaties. Dit toont aan dat huidige DL-methoden in een aantal gevallen auto-segmentaties kunnen genereren die vergelijkbaar zijn met handmatige tumorsegmentaties. In andere gevallen produceren DL-methoden nog steeds tumorauto-segmentaties die niet klinisch acceptabel zouden zijn. Om die reden zouden artsen alle auto-segmentaties nog steeds moeten controleren, waardoor de tijdwinst van automatische segmentatiemethoden deels wordt verloren. Daarom is er behoefte aan maten die de kwaliteit van de auto-segmentaties beschrijven. Echter, gangbare maten om de kwaliteit van auto-segmentaties te beschrijven, zijn meestal gebaseerd op een vergelijkingen met handmatig getekende segmentaties, waardoor ze ongeschikt zijn voor kwaliteitsborging in een klinische setting. In **Hoofdstuk 5** hebben we een kwaliteitsmaat geïdentificeerd die rechtstreeks kan worden gegenereerd uit de output van het netwerk. Deze kwaliteitsmaat was succesvol in het onderscheiden van wel en niet acceptabele segmentaties.

We hebben in dit proefschrift verschillende op DL gebaseerde technieken geïmplementeerd voor automatische segmentatie van tumoren op MRI-beelden. Strategieën die relevante vooraf beschikbare informatie aan de segmentatie netwerken leverden, bleken effectief om de kwaliteit van de auto-segmentaties te verbeteren, zoals het combineren van verschillende MRI-sequenties als invoer of het beperken van de context rond de tumor. Verder hebben we aangetoond dat huidige auto-segmentatie frameworks al auto-segmentaties kunnen produceren die vergelijkbaar zijn met klinische segmentaties. Ten slotte hebben we naast de aangetoonde verbeteringen in de kwaliteit van de auto-segmentaties van tumoren een kwaliteitsmaat voorgesteld die onderscheid kan maken tussen goed en slecht gesegmenteerde gevallen. Dit type maten zal naar verwachting een cruciale rol spelen bij het toepasbaar maken van auto-segmentatiemethoden voor tumoren in de kliniek.

ACKNOWLEDGEMENTS

Uulke, a great deal of the things I have learnt in the last years were thanks to the many hours you spent helping me, and for that, I am very thankful. From all those things I learnt, I would especially like to thank you for teaching me how to worry (a little bit) less.

Dear Rita, I would like to thank you not only for your broad technical expertise that was instrumental for this thesis to be but for your kindness, patience and understanding nature that were in countless moments what really kept me going. I feel deeply grateful for having you as a supervisor but even more grateful to be your friend.

Tomas, I highly appreciate all the relevant input you provided for my work. Thank you as well for welcoming me to the data science meetings, where I was able to enjoy not only many interesting technical discussions but also an outstandingly friendly and warm environment.

A very special ‘thank you’ goes to my paranymphs, not just for all the effort helping me organise the defence, but for the very special roles you played during my PhD. Ernst, my original officemate/workplace proximity acquaintance who was there since the beginning of this PhD journey. Thank you for all the fun moments in and outside the office, such as exchanging our favourite sitcom one-liners or the good games of chess. Your point of view has helped me on many occasions and I am very glad that you are by my side on this special day. Mar, el día en el que nos presentaron nos tuvieron que interrumpir, porque vieron que si no, no íbamos a parar de hablar. Muchas gracias por eso, por nunca dejar de hablar, y por hacer de mis últimos años en ese despacho una temporada mucho más feliz y de menos estrés. Karolina, for not only understanding my type of humour but for always having the perfect comeback. Your friendship is very special to me and I am honoured you are one of my paranymphs.

Muchas gracias también a Celia, que durante los últimos cinco años hizo a la vez de compañera de sesiones de escritura, de confidente y de amiga. Me alegra que, aun después de salir del NKI, sigamos presentes en la vida del otro.

Another big ´thank you´ is for George. It can get rather repetitive to hear me complain about the papers, deadlines or the general stress associated with the PhD. However, you were there through the whole process listening with compassion and kindness. Thank you.

I would also like to express my gratitude to all the people in the department, both past and present. Special thank you to the people of my group: Anke, Bart, Bas, Chavelli, Edward, Ghazaleh, Koen, Marcel, Paula, Petra, Rick, Robin, Roelant, Nikita, Nicole, Stefan, Thijs. Thank you, not only for the relevant scientific discussions, but also for the fun lunches, borrels, group activities, board game nights during these years, which made the whole experience so much more enjoyable. I would also like to thank Vineet, Federica and Carmen, for our fun padel games and for the relaxing drinks afterwards.

Unas palabras de agradecimiento a todos mis amigos de España, que incluso desde la distancia supieron apoyarme durante este periodo. En especial, me gustaría nombrar a los siguientes. A Claucis, que sei ben que sempre está ai cando o preciso. A Ele, por siempre acertar a entenderme en nuestras intermitentes pero maravillosas conversaciones. A Guerris, quien no sólo comprende cuando necesito tomarme una cerveza o irme a la playa, si no que siempre me hace feliz al apuntarse. E a Javi, polas moitas chamadas pero sobre todo por ser un dos meus mais vellos e mellores amigos.

Muchas gracias también a mi hermano Lino. En varias ocasiones durante este doctorado, fue gracias a tu firme convicción de que todo iba a salir bien, que fui capaz de salir adelante. Por eso, y por otras muchas cosas, gracias. Un abrazo también para Álvaro, con quién me alegro de ahora poder discutir los altibajos de los proyectos de *deep learning*.

To Dries, who not only supported me during the most stressful part of the PhD but with whom I lived some of the most beautiful experiences in The Netherlands.

Y finalmente, un agradecimiento muy especial a papá y mamá. Vuestro constante apoyo y amor incondicional fue clave para llegar hasta aquí. Os quiero.

LIST OF PUBLICATIONS

This thesis

Oropharyngeal primary tumor segmentation for radiotherapy planning on magnetic resonance imaging using deep learning

Rodríguez Outeiral R, Bos P, Al-Mamgani A, Jasperse B, Simões R, van der Heide UA. Phys Imaging Radiat Oncol. 19:39–44. (2021)

doi: 10.1016/j.phro.2021.06.005

Strategies for tackling the class imbalance problem of oropharyngeal primary tumor segmentation on magnetic resonance imaging

Rodríguez Outeiral R, Bos P, van der Hulst HJ, Al-Mamgani A, Jasperse B, Simões R, et al. Phys Imaging Radiat Oncol. 23;144–9 (2022)

doi: 10.1016/j.phro.2022.08.005

Deep learning for segmentation of the cervical cancer gross tumor volume on magnetic resonance imaging for brachytherapy.

Rodríguez Outeiral R, González PJ, Schaake EE, van der Heide UA, Simões R. Radiat Oncol. 1:18. (2023)

doi: 10.1186/s13014-023-02283-8

A network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations

Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, Simões R. Phys Imaging Radiat Oncol. 28. (2023)

doi: 10.1016/j.phro.2023.100500

Other publications

Response letter to Wahid et al. Regarding our publication "a network score-based metric to optimize the quality assurance of automatic radiotherapy target segmentations"

Rodríguez Outeiral R, Ferreira Silvério N, González PJ, Schaake EE, Janssen T, van der Heide UA, Simões R. Phys Imaging Radiat Oncol. 28. (2023)

doi: 10.1016/j.phro.2023.100528

Realce de imágenes mamográficas para su análisis y clasificación mediante un sistema CAD basado en redes neuronales convolucionales.

Rodríguez R, Planchuelo A, Yébenes B, Ríos B, Sánchez C

XXXIV Congreso anual de la Sociedad Española de Ingeniería Biomédica. (2016)

ISBN: 978-84-9048-531-6.

CURRICULUM VITAE

I am Roque Rodríguez Outeiral and I was born in Vigo on the 29th of April 1994. I was raised in the same city, where I also attended high school. When I was 18 years old, I moved to Madrid to study biomedical engineering at the Polytechnical University of Madrid (UPM). In these studies, I specialized in the track of medical imaging. Furthermore, as my bachelor's final thesis I worked on my first project on the topic of deep learning applied to medical image analysis, which would later be part of my first publication. In 2016, I relocated to Barcelona, where I enrolled in the masters of computer vision at the Autonomous University of Barcelona (UAB). During these studies, I delved in both the theory and application of deep learning for different computer vision tasks. After completing the masters, I moved to Aachen (Germany) to work as a research intern in Nuance communications. In this role, I was responsible for implementing deep learning techniques for the automatic classification of lesions on X-ray images. In 2018, I started my PhD at the Radiation Oncology department of the Netherlands Cancer Institute (NKI-AvL) in Amsterdam. In this thesis, I describe the research conducted during this PhD on the implementation of deep learning techniques to automatically segment tumors on MRI images. Currently, I am working in Agendia as a deep learning engineer. In this position, I am implementing different deep learning techniques for the analysis of pathology images, with the aim of improving the treatment of breast cancer patients.