Master's thesis

NTNU
Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

Jeanette Bonden Isachsen

Deep learning image segmentation and automatic treatment planning in breast cancer radiotherapy

Master's thesis in Applied Physics and Mathematics
Supervisor: Sigrun Saur Almberg
Co-supervisor: Kathrine Røe Redalen
June 2021

NTNU
Norwegian University of Science and Technology

ST. OLAVS HOSPITAL
TRONDHEIM UNIVERSITY HOSPITAL

Jeanette Bonden Isachsen

Deep learning image segmentation and automatic treatment planning in breast cancer radiotherapy

Master's thesis in Applied Physics and Mathematics
Supervisor: Sigrun Saur Almberg
Co-supervisor: Kathrine Røe Redalen
June 2021

Norwegian University of Science and Technology
Faculty of Natural Sciences
Department of Physics

NTNU
Norwegian University of
Science and Technology

Abstract

Background and purpose: For radiotherapy, organ at risk (OAR) and target volume segmentation and the following treatment planning are today, more or less, done manually and therefore very time-consuming and prone to inter-observer variability. The use of deep learning (DL) models for automatic segmentation has the potential to both save time and lead to a more standardized process. The performance of a DL segmentation model trained on local patient data and a pre-trained DL segmentation model available from Siemens Healthineers has been evaluated for use at St. Olavs hospital. Additionally, a protocol-based script for automatic plan optimization was evaluated.

Materials and methods: The local model was trained by RaySearch Laboratories AB (Stockholm, Sweden) on CT images of 168 left-sided breast cancer patients treated with radiotherapy. Geometric and dosimetric evaluations were done for 15 patients where manual delineations were used as ground truth. Additionally, clinical evaluations were done for the pre-trained Siemens model. The protocol-based script for automatic plan optimization was evaluated dosimetrically by comparing automatic volumetric-modulated arc therapy (VMAT) plans to clinical hybrid and clinical VMAT plans for 16 patients in total.

Results: The heart, left lung, right lung, spinal canal, esophagus, sternum, right breast, and left breast (primary target volume) were evaluated for both segmentation models. Additionally, locoregional lymph node areas (nodal target volume) were evaluated for the local model. The local model was significantly better than the Siemens model based on the geometric evaluation. The dosimetric differences were statistically significant for 9 of the 12 main metrics for the Siemens model and for 4 of the same metrics for the local model. Larger dosimetric differences were found for the lymph node areas. Clinical scoring of five structures segmented by the Siemens model was promising for breast radiotherapy. The evaluation of the automatic plan optimization indicates that the target volume coverage and treatment quality are preserved when using automatic planning. OAR doses were generally reduced with the automatic plans. Compared to the hybrid plans, large dose reductions were found for the heart and left lung.

Conclusion: The evaluation of the DL models indicates that the quality of both models is adequate to segment OARs for breast radiotherapy. However, in some cases, manual adjustments might be required, especially when using the Siemens model. The local model is preferable for target volumes and will likely be good enough for clinical use when some adjustments have been done to the lymph node areas. The final version is now being trained. The script for automatic plan optimization has been validated and is now being implemented in the clinic.

Sammendrag

Bakgrunn og formål: Behandlingsplanlegging og tilhørende segmentering av risikoorgan og målvolum for stråleterapi blir i dag mer eller mindre gjort manuelt. Prosessene er derfor tidkrevende og utsatt for variasjon mellom observatørene. Bruken av modeller basert på dyp læring (DL) til segmentering har potensialet til å både spare tid og føre til en mer standardisert prosess. En modell trent opp på lokal pasientdata og en ferdig trent modell fra Siemens Healthineers har blitt evaluert til bruk på St. Olavs Hospital. I tillegg har et protokoll-basert skript for automatisk optimalisering av behandlingsplan blitt evaluert.

Materiale og metode: Den lokale modellen ble trent opp av RaySearch Laboratories AB (Stockholm, Sverige) på CT bilder av 168 pasienter som ble behandlet for venstresidig brystkreft med stråleterapi. Begge modellene ble testet på 15 pasienter og sammenlignet med manuelle inntegninger med geometriske og dosimetriske parametere. Klinisk evaluering ble også gjort for Siemens modellen. Skriptet for automatisk planlegging ble evaluert dosimetrisk ved å sammenligne automatiske VMAT-planer med kliniske hybridplaner og kliniske VMAT-planer.

Resultater: Hjertet, venstre lunge, høyre lunge, spinal kanalen, øsofagus, sternum, høyre bryst og venstre bryst (primært målvolum) ble evaluert for begge segmenteringsmodellene. I tillegg ble regionale lymfeknuteområder (nodalt målvolum) evaluert for den lokale modellen. Den lokale modellen var betydelig bedre enn Siemens modellen, basert på den geometriske evaluering. De dosimetriske forskjellene var statistisk signifikante for 9 av 12 hovedparametere for Siemens modellen, og for fire av de samme parameterne for den lokale modellen. Større dosimetriske forskjeller ble funnet for lymfeknuteområdene. Klinisk scoring for fem strukturer segmentert av Siemens modellen ga lovende resultater til bruk for brystbestråling. Evalueringen av automatisk planoptimalisering indikerer at dekning til målvolumene og behandlingskvaliteten er bevart ved automatisk planlegging. Dosene til risikoorganene var generelt redusert for de automatiske planene og sammenlignet med hybridplanene var det større reduksjoner for hjertet og venstre lunge.

Konklusjon: Evalueringen av DL modellene indikerer at kvaliteten på begge modellene er tilstrekkelig ved segmentering av risikoorganer for brystbestråling. Det vil imidlertid være nødvendig å gjøre manuelle justeringer i noen tilfeller, spesielt ved bruk av Siemens modellen. Den lokale modellen er foretrukken for segmentering av målvolum og modellen vil trolig være tilstrekkelig til klinisk bruk når noen justeringer har blitt gjort ved lymfeknuteområdene. Den endelige modellen blir nå trent opp. Skriptet for automatisk planoptimalisering har blitt validert og blir nå implementert i klinikken.

Preface

With this master thesis, I am finishing my five-year study of Applied Physics and Mathematics at the Norwegian University of Science and Technology. The work for this master thesis is done in the spring of 2021. In writing this master thesis and my previous project thesis, there are several people that need acknowledgments for their support and assistance.

First and foremost, I am extremely grateful to both my supervisors, Sigrun Saur Almberg and Kathrine Røe Redalen. Sigrun for always answering all my questions and assisting me quickly without losing patience. Her help and guidance in both the work and writing process for my thesis has been greatly appreciated. Kathrine has let me be a part of her research group. Thereby provided me with a broader perspective in the field of medical physics, motivation, and helpful advice from the group. Kathrine has also provided me with encouragement and guidance throughout the past year. I would also like to thank the three oncologists and the radiotherapist who performed the clinical evaluations of the automatic segmentations, as well as Jomar and Marit for letting me use their script for automatic plan optimization.

I am very thankful to my friends and fellow students who have made both my time of studying and spare time a joy. Especially, I am grateful to my collective with whom much of my time has been spent during this past year with varying degrees of lockdown, due to the coronavirus. I would also like to thank my family, who believe in me and provide me with support in everything I set my mind to.

Trondheim, June 2021

Jeanette Bonden Isachsen

Table of Contents

List of Figures

List of Tables

List of Abbreviations

3D-CRT	three-dimensional conformal radiation therapy 7, 8, 18, 21
AI	artificial intelligence xiii, 10, 11, 15, 16, 28, 53
ALARA	as low as reasonably achievable 19, 27, 49
CI	conformity index 26
CNN	convolutional neural network 14, 15, 43
CT	computed tomography xiii, 3, 4, 6, 7, 23, 27, 28
CTV	clinical target volume xiii, 6, 7, 17, 18, 24, 28, 31, 34, 36, 39, 40, 44, 49
DL	deep learning v, xiii, 1, 9–11, 13–16, 22, 23, 25, 55
DSC	Dice similarity coefficient xiii, xv, 25, 31, 32, 43, 50, 81
DVH	dose-volume histogram xiii, xiv, 26, 27, 40–42, 73, 89
GTV	gross target volume 6, 17, 48
HD	Hausdorff distance xiii, xv, 25, 26, 31, 32, 43, 50, 81
HI	homogeneity index 26
IMRT	intensity-modulated radiotherapy 7, 8, 18
ML	machine learning xiii, 10–12, 15
MLC	multi leaf collimator 5–8
MSE	mean squared error 13
OAR	organ at risk v, xiii, xiv, 1, 3, 6–9, 18, 22–24, 26, 27, 31, 35, 40, 41, 43, 48–50, 52, 53, 55
PTV	planning target volume xiii, 6–8, 18, 23, 24, 26, 27, 31, 35, 49
ReLU	rectified linear unit 15
RNN	recurrent neural network 14
SD	standard deviation 31, 34, 39, 40, 43, 49, 81, 83, 85–87, 89, 90
VMAT	volumetric-modulated arc therapy v, vii, xv, 7, 8, 18, 21, 24, 38–42, 49, 50, 55, 73

1 Introduction

The field of radiotherapy has come a long way since the first breast cancer patient was treated with radiation in 1896 [1]. This was only a year after Wilhelm Conrad Röntgen first discovered x-rays (photon radiation) and they were not familiar with the physical properties of radiation at the time [1]. Today, we know much more about the properties of radiation and how to use it with care to achieve the best possible result for the patient.

Modern radiotherapy techniques allow precise treatment delivery. However, the accuracy of the whole process is limited by the weakest link, which in radiotherapy is considered to be segmentation [2]. This is the procedure of delineating target volumes and organs at risk, which is then used to make a personalized treatment plan for each patient. Segmentation is today mostly done manually, which is time-consuming, highly affected by the competence of the observer, and generally affected by inter-observer variability [3]. Following segmentation is plan optimization, also a time-consuming procedure that is associated with large inter-observer variability and dependent on the competence of the observer [3].

Increased automation is expected to have a major impact on further development within radiotherapy. Like in many other technological fields, artificial intelligence will likely have a major role in this development. The use of artificial intelligence, and more specifically deep learning (DL) and knowledge-based algorithms, should lead to increased efficiency, standardization, and quality. Introducing these methods to automatic segmentation and automatic plan optimization is looking promising in terms of efficiency and consistency [4–6].

Before automatic models can be implemented clinically they need to be properly tested and evaluated [3]. The goal of this project was to test and evaluate two DL segmentation models and a plan optimization script, all related to left-sided breast cancer.

Different hospitals use different guidelines, both for segmentation and plan optimization. To ensure that the automatic method follows the applicable guidelines, it might be necessary for each hospital to develop automatic methods locally [3]. One locally trained and one pre-trained segmentation model was evaluated for use at St. Olavs hospital.

Specifically, the three main aims of this project were to:

1. Evaluate a locally trained DL segmentation model for left-sided breast cancer patients.
2. Evaluate a pre-trained DL segmentation model for organs in the thorax-area made by Siemens Healthineers.
3. Evaluate a locally made protocol-based script for an optimization of left-sided breast cancer.

Both segmentation models and the plan optimization script were evaluated on left-sided breast cancer patients treated with locoregional radiotherapy, including both target volumes and organs at risk (OARs). The treatment planning system RayStation 9B was used to test and evaluate all three automatic methods. Both segmentation models were evaluated geometrically and dosimetrically. Additionally, the pre-trained Siemens model was evaluated qualitatively. The plan optimization script was dosimetrically evaluated.

The locally trained model is a preliminary version, and the final model is intended to be used clinically at St. Olavs Hospital. Only minor changes are expected to be done for the final model compared to this preliminary model.

2 Theory

This section of theory is highly inspired by the theory written for the project thesis written prior to this master thesis [7]. Specifically, the introduction to section 2.1 about radiotherapy process, subsection 2.1.2 about the linear accelerator, subsection 2.1.3 about target volume and organ delineation, subsection 2.1.4 about treatment planning and treatment techniques for photon radiation and section 2.4 about breast cancer are copied from Isachsen [7], with some changes and additions. Section 2.2 about automatic segmentation and section 2.3 about artificial intelligence have been partially copied from Isachsen [7], but larger changes and additions have been made.

2.1 The radiotherapy process

In radiotherapy, ionizing radiation is used to kill cancer cells by damaging their DNA. The aim is tumor control and minimization of normal tissue damage. The treatment can be for a curative or palliative purpose. Radiotherapy is generally a non-invasive treatment of cancer using a linear accelerator and delivering radiation through fractions, over the course of several weeks.

The typical workflow for radiotherapy is shown in figure 1. The process begins with a consultation where the physician and patient decide to proceed with radiotherapy and ends with follow-up after the treatment delivery [8].

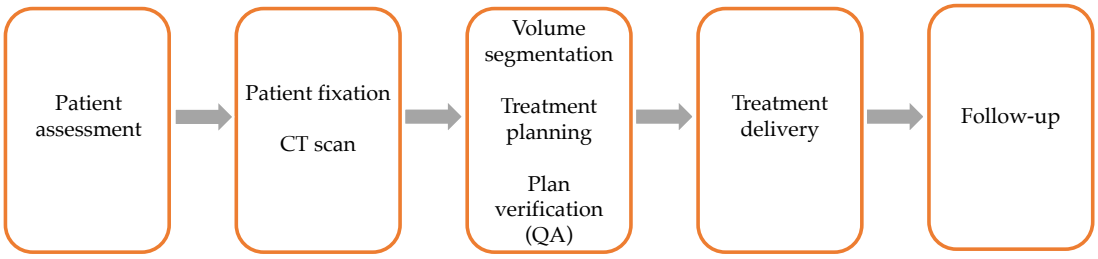


Figure 1: Typical workflow of radiotherapy.

Not only cancer cells are killed during radiotherapy. Killing too many normal tissue cells in an organ can lead to loss of vital functions and it is thus imperative to limit radiation dose to OARs and thereby minimize the normal tissue complication probability. OARs are organs that are especially close to the target volume and should be taken into consideration when planning the treatment. A computed tomography (CT) scan is taken of the patient and used as a 3D model of the patient for treatment planning. The CT-scan should be representative of every treatment fraction and should therefore be taken with the necessary preparations (i.e., fixation, bladder filling, breath-hold, etc.). The OARs and target volume(s) are delineated in the CT-images by a physician, so-called image segmentation. The segmented images are then used for treatment planning [8].

A treatment plan results in a radiation dose distribution that aims to maximize the therapeutic ratio. The therapeutic ratio is the relationship between the probability of tumor control and the probability of normal tissue damage. A verified plan can be delivered to the patient, while carefully monitoring the process.

2.1.1 Computed tomography

Today, CT is a necessary part of the process for anatomic imaging and for mapping the electron density which is used for radiation dose calculations [9]. This modality is therefore used as the basis of the treatment planning [9]. CT is one of the oldest medical imaging techniques, according to Kalender [10]. CT uses x-rays and sends these into the patient/object, from different directions, and measure the intensity, I, of the radiation that leaves the patient/object. Knowing the initial intensity, I_0, one can then reconstruct an image based on the attenuated value from each ray [10].

The linear attenuation coefficient, μ, is tissue-dependent and can in a simplified case be found from $I = I_0 \exp(\mu d)$, where d is the absorber thickness [10]. If the absorber thickness is known, one can simply solve for μ. In this simplified case the distribution of μ along the beam path would be unknown, making this a 2D projection of the patient instead of a 3D model, but this is the general idea behind x-ray imaging techniques. For a CT-scan the beam and detector is rotated around the patient, in a spiral, from head to toe, while the patient is lying still. This is shown in figure 2. After this, the slices are reconstructed from the scan by an algorithm and can be viewed either as slice by slice in 2D or as a 3D model. The linear attenuation coefficient is related to the electron density [11]. Therefore, a CT-scan is ideal to use for dose calculations in treatment planning for radiotherapy [9].

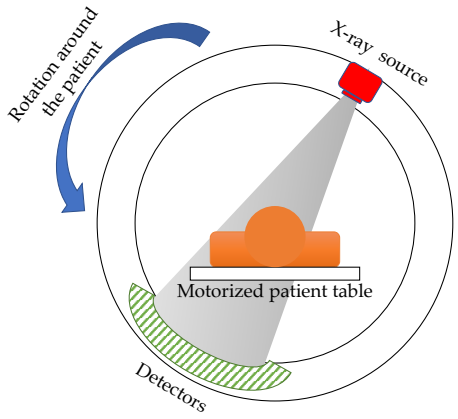


Figure 2: During a CT-scan the x-ray source and detector rotates around the patient while the patient table moves in the horizontal direction through the scanner.

2.1.2 Linear accelerator

This subsection about the linear accelerator is based on [12], unless otherwise stated.

The linear accelerator or linac is the workhorse of radiotherapy worldwide [13]. It is a particle accelerator that accelerates electrons to almost the speed of light. These electrons are then used directly or converted into photons and directed towards the patient. An overview of the main components in a typical linac is given in figure 3 and a photo is shown in figure 4.

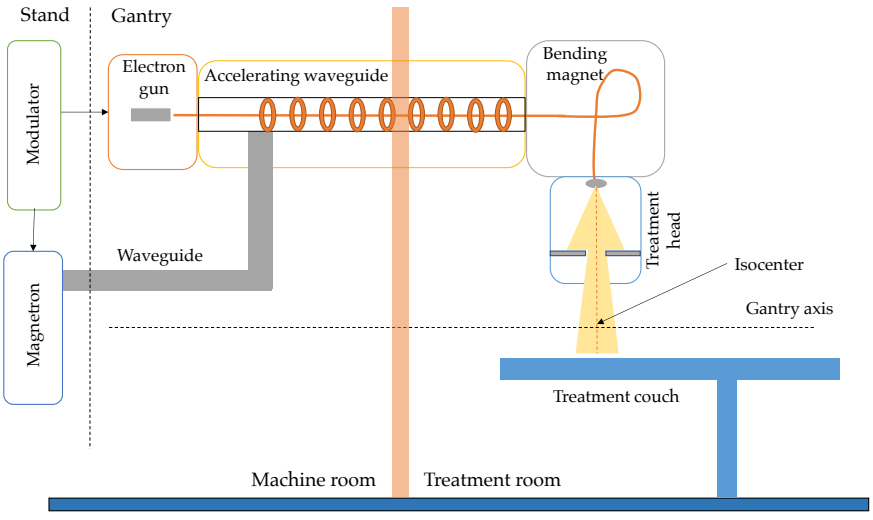


Figure 3: A schematic overview of a typical linear accelerator with the main components.

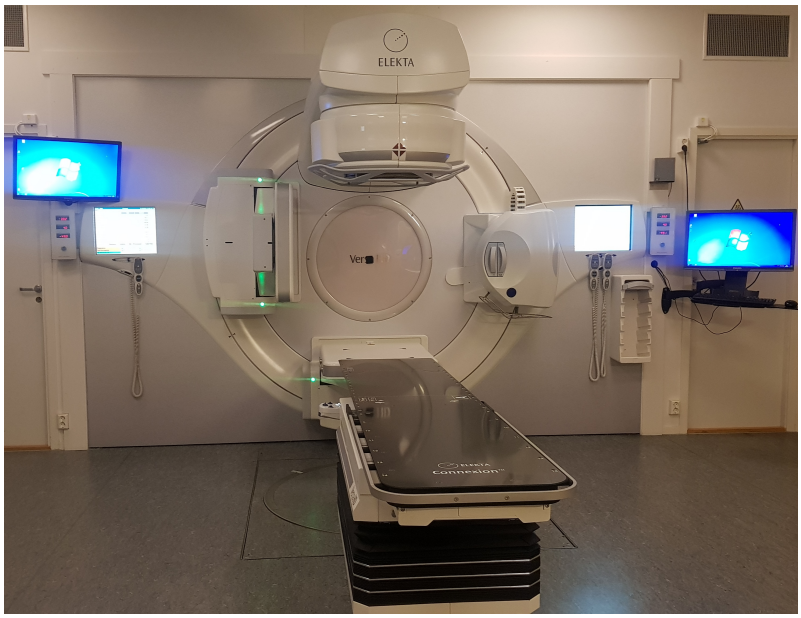


Figure 4: An Elekta Versa HD linear accelerator from St. Olavs Hospital.

The linac can be divided into the stand and the gantry. The stand is the stationary part. The gantry is the rotating part and rotates 360° around the patient delivering radiation at different gantry angles. The rotation is around the gantry axis, which goes through the isocenter. The isocenter is the crossing point between the gantry axis and the center of the beam that exits the treatment head. The gantry includes the electron gun, the accelerating waveguide, and the treatment head.

The magnetron produces radio frequency oscillations that are sent through the waveguide. The electron gun sends electrons into the waveguide synchronized with the radio frequency pulses. In the accelerating waveguide, the radio frequency field accelerates the electrons. For clinical use, the linear accelerator often has a horizontal accelerating waveguide because otherwise, the height needed for the linac would not be practical. Therefore, the electron beam needs to be bent 90° or 270°. This happens in the bending magnet.

At this point, the electron beam is narrow, focused, and directed towards the target area, but before the beam is ready to reach the patient it needs to be converted into photons. This happens in the treatment head. The electron beam is directed towards a tungsten target where the electrons are converted into photon radiation (Bremsstrahlung). Electrons are sometimes used directly for treatment, and then the narrow beam needs to be appropriately scattered.

One can see a simplified diagram of a treatment head in figure 5. First, the beam hits the tungsten target and then goes through the primary collimator, which limits the beam to the appropriate size. The photon beam is more intense in the center, so to achieve a more homogeneous field a flattening filter is inserted. The beam then passes through an ion chamber, where the dose and uniformity are monitored.

A multi leaf collimator (MLC) shapes the beam so that it fits the target volume shape according to the treatment plan. The MLC typically consists of 80 tungsten "leaves" that move independently of each other and allow a flexible beam shape. The leaves of the MLC usually have a width of 1 cm. Since this type of field shaping is dynamic and the leaves need to move swiftly, some spacing between the leaves is necessary. This spacing leads to leakage of radiation between the leaves and the leakage needs to be reduced before the beam reaches the patient. Partially, this problem is solved by having stepped or overlapping leaves, but a backup collimator is used to get the leakage level down to an acceptable level.

Until MLCs started to appear in the 1980s, the beam was shaped into a rectangular field. MLC was introduced to reduce the amount of radiation given to healthy tissue and thereby also allowing

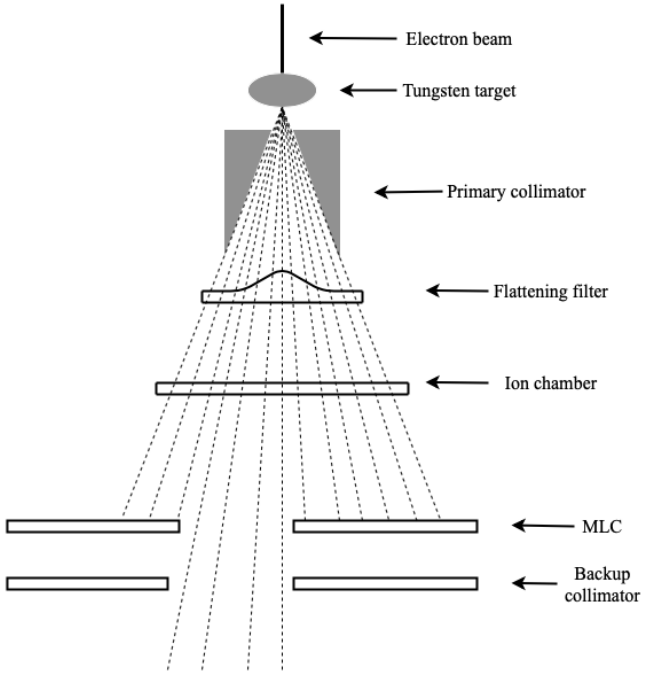


Figure 5: Diagram of a linear accelerator treatment head with MLC as beam shaping collimator.

an increased dose to the target volume, leading to a more conform treatment. When the beam has been shaped to fit the target, it is ready to reach the patient. The planned dose is then delivered as the gantry rotates and the beam is shaped for each new angle.

2.1.3 Target volume and organ delineation

Segmentation of medical images is the process of delineating structures in the images. In radiotherapy, these structures can be OARs, clinical target volumes, or other regions of interest. The delineated images are then used for treatment planning. The dose distribution aims to spare OARs and secure coverage of the target volume.

The most common method used for segmentation today is manual delineation. This method consists of using different tools to draw the contours around the organs and target volume(s) in the CT-slices. The slices are usually 3 mm thick and there are around 200 slices from one scan [14]. Manual segmentation is a time-consuming process. This method is also highly dependent on the anatomical knowledge and therefore experience of the physician [15]. Typically, the physician uses more than one imaging modality to examine the extent of the disease for target volume delineation [9, 16].

Target volumes

Gross target volume (GTV) is an anatomical volume and is the visible part of the tumor that the physician can see from the images. The clinical target volume (CTV) includes GTV, but also areas around the GTV where microscopic disease is suspected. The CTV may include lymph nodes that are suspected cancerous [14].

Movements and differences in patient set-up can affect the position of the target volume during treatment [14]. Planning target volume (PTV) is a geometric volume defined to take these effects into account. The PTV is in the end the area that is treated during radiotherapy. An overview of the relationship between GTV, CTV and PTV can be seen in figure 6. To secure dose coverage of

the CTV in practice, the treatment plan is made based on the PTV [14].

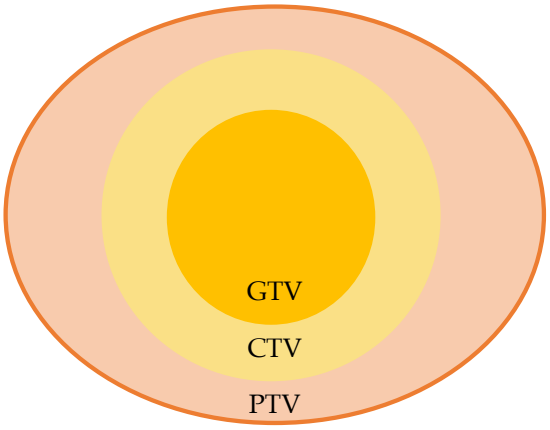


Figure 6: Overview of the volume definitions in radiotherapy planning and the relationship between them.

2.1.4 Treatment planning and treatment techniques for photon radiation

This subsection about treatment planning and treatment techniques is based on Khan et al. [14], unless otherwise stated.

When the 3D model of the patient has been acquired through CT, as well as any additional imaging modalities, and the OARs and target volumes have been delineated, the treatment planning can begin. The treatment planning consists of setting up beams from different angles around the patient and then calculating the predicted dose to each voxel in the CT-scan. The plan must be evaluated to see if it meets all the clinical goals, and is adjusted until it does. Once it meets all the clinical goals and is optimized to spare as much normal tissue as possible the plan can be approved, and used for treatment.

Dose calculations are done by algorithms. Dose is measured in Gy, which is J/kg and is hence defined as absorption of one joule of radiation energy per kilogram of matter [17]. Dose calculations are the planned amount of dose, in Gy, given to each voxel in the CT-scan. The dose calculations require a description of the anatomy of the patient, i.e. CT-scan, and a description of the radiation source. Description of the radiation source includes beam energy, source size, and the behavior of the photons through the head of the linear accelerator [12].

There is a wide variety of different computer algorithms for photon dose calculation. Generally, choosing the best method is a compromise between speed and accuracy [12]. The most used method, according to Mayles et al. [12], is three-dimensional convolution of the point-spread function. The point-spread functions are derived from Monte Carlo simulations in water [12]. Although full Monte Carlo dose calculations have been implemented in commercial treatment planning systems recently, it is not yet widely used.

Three main treatment techniques for photon radiotherapy are three-dimensional conformal radiation therapy (3D-CRT), intensity-modulated radiotherapy (IMRT) and volumetric-modulated arc therapy (VMAT). All three are CT-based and the OARs and PTV are drawn into the images by a physician. A treatment plan is then made using a MLC to avoid and spare the OARs and hit the PTV more conformly from different directions, compared to using simpler box-techniques.

When using 3D-CRT, each field is added manually to the plan by choosing the direction and intensity. While 3D-CRT uses a so-called, forward planning technique, the other techniques use inverse planning. Forward planning is more time-consuming and it is not possible to explore all options. Inverse treatment planning, on the other hand, lets the computer find the optimized treatment plan automatically given objectives for OARs and target volume(s).

The optimization aims to minimize the difference between the calculated dose distribution and the prescription dose distribution. It can be practically implemented by minimizing the quadratic cost function

$$C = \sum_{i}^{N} I_i (D_i - D_i^P)^2,$$

where D_i is the dose delivered to the ith voxel, D_i^P is the prescription dose to the ith voxel, I_i is the importance factor to the ith voxel and N is the total number of voxels. The prescription dose is made based on getting full coverage of the target volume and meeting clinical goals for the OARs. The importance factor is a weight that allows the user to determine which objectives are most important during the optimization [12].

IMRT can reduce the fluence in some areas and increase it in other areas, hence it is intensity-modulated. Fluence is the energy delivered per unit area. In IMRT, 5-9 gantry angles are usually used where around 10 segments are delivered at each angle.

VMAT is a more advanced version of IMRT. VMAT also uses inverse treatment planning but gives the radiation continuously through the whole gantry rotation. The MLC positions also change continuously during the rotation. The entire treatment is then delivered in the one gantry rotation while the MLC changes, the dose rate changes, and the speed of the rotation changes. This technique is faster than IMRT.

Both IMRT and VMAT have become routine for most modern treatment planning. They are superior to the standard treatment using 3D-CRT because these methods allow shaping the dose distribution so that one achieves conform delivery of the dose to the PTV while sparing the OARs. However, these techniques may irradiate larger volumes with small doses, which can be a concern in some situations.

2.2 Automatic segmentation

Manual delineation of target volume(s) and OARs is a time-consuming process [14]. Target volume delineation is also known to represent the largest uncertainty in the radiotherapy process [2, 18]. Errors at this stage of the treatment generate systematic errors during the treatment [18]. With conformal techniques, accurate delivery becomes very important. If the target volume delineation is slightly wrong, it could lead to insufficient tumor control and unnecessary damage to the critical organs. These uncertainties can also make it harder to find correlations in clinical studies and cause confusion when comparing different techniques [2]. In figures 7 and 8, examples of inter-observer variability between manual delineations is shown.

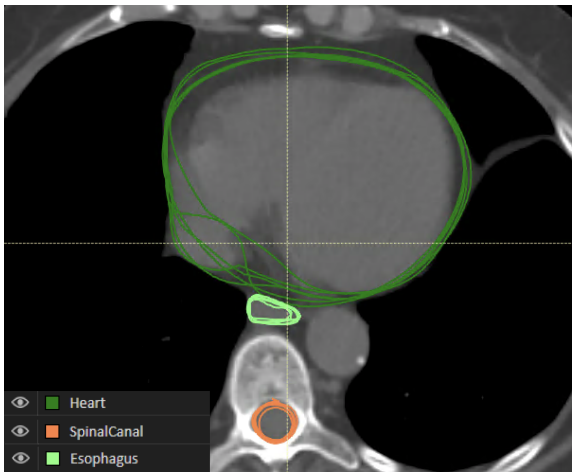
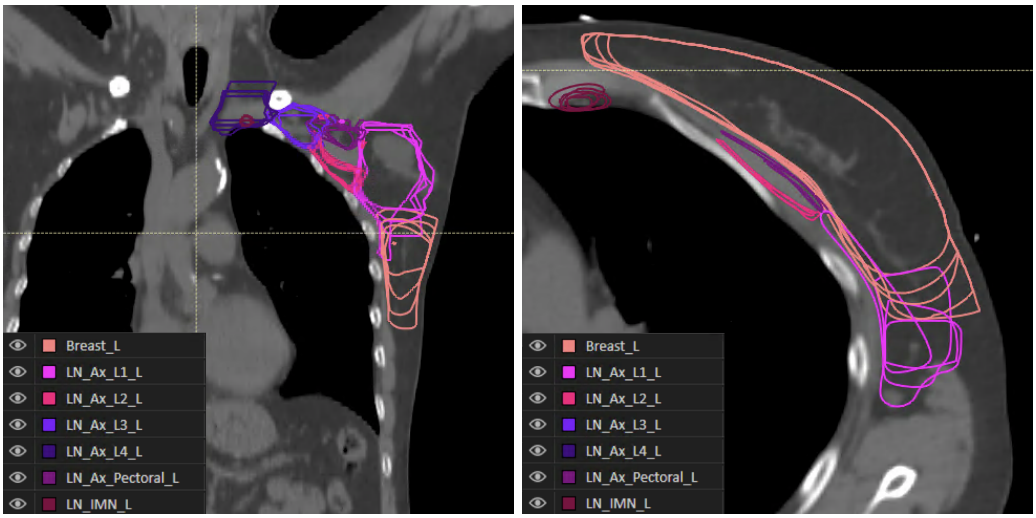


Figure 7: Example of inter-observer variability between manual delineations from six different observers at St. Olavs Hospital of organs at risk, transversal plane.



(a) Coronal plane **(b)** Transversal plane

Figure 8: Example of inter-observer variability between manual delineations from five different observers at St. Olavs Hospital for target volumes for left-sided breast cancer.

The uncertainties in target volume delineation originate from a lack of ground truth and inter-observer variability among the physicians. Although these uncertainties are the largest for target volume delineation, this also applies to OAR delineation. Especially, the inter-observer variability is said to affect the accuracy of OAR delineation [15, 19]. Nelms et al. [20] studied the variations between different clinics in the delineation of OARs. They found that there was significant inter-clinician variability and stresses the importance of accuracy not only in target volume delineation but in OAR delineation as well [20].

It has become of great interest to reduce these uncertainties and also find a less time-consuming method. Through automatic segmentation, the element of the eye of the beholder can be partially or completely removed and there is potential to save time and valuable resources. The standardization will also be increased, making it easier to perform clinical studies to improve radiotherapy further. Overall, automatic segmentation may lead to increased quality of treatment.

There are several approaches to the automation of image segmentation in radiotherapy. Below, the three most commons methods are presented, i.e, atlas-based, model-based, and DL based. Automatic segmentation methods can also be a hybrid of two or more methods to compensate for their weaknesses. Delineation guidelines and practices vary from clinic to clinic as this is not standardized. Therefore, a model may be suitable for one clinic and not for another. According to Liesbeth et al. [3], each clinic needs to perform an evaluation of the model, used on data similar to which the model will be used clinically [3].

2.2.1 Atlas-based segmentation

Atlas-based image segmentation uses a reference image, an atlas, to segment the new image. In the atlas, the structures of interest are already segmented. The image that is to be segmented is mapped or paired with a suitable atlas from the library. A transformation is done between the atlas and the new image to transfer and fit the segmentations in the atlas to the new image [21].

The similarity of the atlas to the image is important for the quality of the segmentations. Using an average of several suitable atlases can therefore reduce errors due to lack of correspondence between the atlas and the images. This is called multi-atlas-based segmentation and improves the robustness of the segmentation. Voxels are decided as part of a structure or not by a voting system from all the suitable atlases. A downside of this method is that it can lead to topological errors where the structures are not closed, and this demands manual editing which is time-consuming [21].

2.2.2 Model-based segmentation

Model-based image segmentation generates closed and anatomically correct surfaces by using statistical shape models or statistical appearance models. These models can restrict the final segmentation to something that is an anatomically correct shape. The shapes are attained in compact form, together with voxel intensities, based on the training data. The best-fitted model is chosen and used for each structure of interest and the segmentations are made. The models are trained on delineated structures in a training set, delineated by experts and the flexibility of the models are highly dependent on the training data size and content [21].

2.2.3 Deep learning segmentation

Image segmentation methods using artificial intelligence (AI), specifically DL methods, have shown promising results in the last years and have the potential to outperform other automatic segmentation methods [22, 23]. DL methods can be explained as algorithms that mimic the way the human brain works to segment regions of interest in medical images. U-net is the most promising algorithm for medical image segmentation. See section 2.3 for more details about AI, DL and U-net.

Just as for physicians, the more images the model is trained on, the better it will perform [24]. So, with big enough and good enough training data, one should be able to train a model that performs satisfactorily, i.e. follows guidelines, and is more consistent than today's clinical practice. DL segmentation is documented to be both faster and better than atlas-based segmentation [25, 26].

2.3 Artificial intelligence

AI is defined as the simulation of intelligent human thinking and acting [27]. This is not a new concept, the term was introduced already in 1956, but it has become more popular in the last 20 years due to the availability of massive amounts of data power [28]. The concept is to train a model to make decisions based on inductive reasoning. A large training set for the model makes it possible to improve the model significantly more than trying to improve the algorithm that the model uses [28].

AI can be divided into subdomains that focus on different fields. These are, among others, natural language processing, vision, robotic processes, and machine learning (ML). ML will be covered in more detail below, as well as DL, which is the subdomain of ML of interest in this thesis. Their relationship to AI can be seen illustrated in figure 9. Some examples of practical use of AI today are Apple's Siri or Amazon's Alexa, spam filters, Google Translate, and self-driving cars.

2.3.1 Machine learning

Subsection 2.3.1 about ML is based on Theodoridis [29], unless otherwise stated.

Arthur Samuel defined the term ML in 1959 as "the ability to learn without being explicitly programmed."[30]. ML is a science that uses AI and algorithms borrowed from statistics to make computers learn from data, find hidden structures in data, and then make rational decisions. Traditionally, computers need explicit instructions and rules for their data processing and decision making, but with ML, computers can learn more advanced decision making from examples instead [31]. This allows computers to solve more complex problems [31].

Models are built through a process called training. This is the process of letting the model look at the observations/examples that it should learn from. The model performance increases with the amount of training data. This is comparable to humans, but computers lack common sense and therefore need to see a lot more examples than humans [31]. Good quality training data is crucial for the performance of the model and obtaining good enough training data is often the challenge

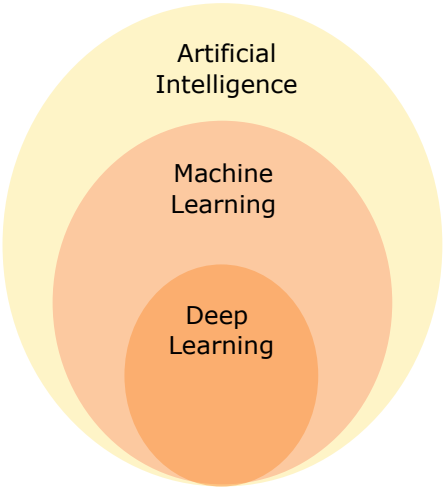


Figure 9: Relation between AI, ML and DL

when making a ML model [31]. Once the model has been trained it should be able to take in a new case and give a prediction as output.

Before the model can be used it needs to be tested. A part of the dataset should be set aside for this. A compromise needs to be made so that the training dataset is large enough to make a good model, as well as that the test dataset should be large enough to represent the relevant variations. A larger dataset allows a larger fraction to be used for training because the size of the test dataset only needs to be big enough and increasing above this level is probably unnecessary.

The dataset is actually split into three parts, training, validation, and test datasets. The validation dataset is used during the training of the model to fine-tune parameters or decide when the model is good enough and the training can stop. Training error is the error found using the validation dataset while training the model, and the test error is the error found using the test dataset on the final model and is the same as a generalization error. It is important that the test dataset does not include data that is used for the training. Otherwise, the model will likely be too well fitted to the training data and the training error could go towards zero, while the generalized performance will be poor, leading to a high test error. This model would be overfitted to the training data and therefore perform badly on new data. This can also happen if there are too many parameters in the model, compared to how much training data is available. On the other hand, a model can also become underfitted if the number of parameters does not fit the actual complexity of the situation. Examples of overfitted and underfitted models can be seen in figure 10.

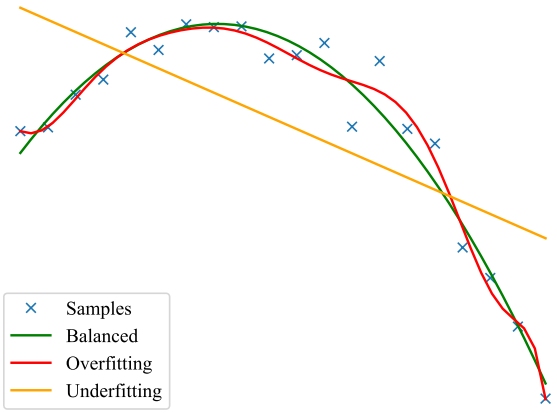


Figure 10: A curve fitted to data in three different manners, underfitted, balanced and overfitted.

Different algorithms

The complexity of the algorithm used for the model should reflect the complexity of the data and the problem to be solved. For simple data, as for example text and numbers, the classical ML approaches will give a simple model that works fast. For images and videos, it is recommended to approach the problem with a neural network algorithm. Actually, neural networks can be used for most problems and are being used increasingly, but they have the drawback of being less intuitive. Neural networks will be explained in more detail in subsection 2.3.2.

ML can be used for regression or classification problems. Regression problems aim to predict continuous values, while classification aims to predict discrete values, for example, true or false. The problems can be trained by supervised learning or unsupervised learning. Figure 11 illustrates the difference between these two learning methods. Supervised learning refers to when the machine has labeled data or the ground truth during the training of the model. Unsupervised learning uses unlabeled data, that is data with no ground truth, and the goal is not necessary to predict anything specific but let the machine try to find patterns and similarities in the data [32].

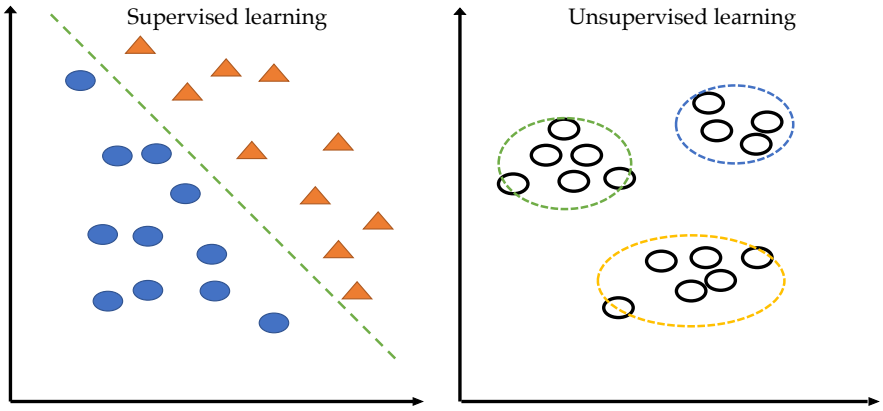


Figure 11: Supervised and unsupervised training for a classification problem.

The algorithms referred to here as classical ML approaches are the ones that have descended, more or less, directly from statistics. On the supervised part, we have regression and classification algorithms. Regression can be, for example, linear or polynomial, which can be used for predicting any kind of continuous values that vary over time. Figure 10 is an example of a regression problem, while figure 11 shows a classification problems. Clustering is an example of an unsupervised learning algorithm, where the goal is to separate the data into different clusters that have more in common with each other than the data from other clusters [32]. Practical applications of classical ML can be housing market predictions for regression and spam filters for classification. Clustering is used as an analysis tool to find patterns make data more easily understandable.

Examples of classification algorithms are decision trees, k-nearest neighbor, and support vector machines. Decision trees or classification trees are based on the computer choosing suitable yes/no questions to separate the data into the correct classes. These are easy to explain and are popular for this reason. One can increase the performance of decision trees by making an ensemble of trees called a random forest and then use the average of several decision trees as a result [32]. K-nearest neighbor uses the k number of nearest neighboring samples to predict the classification of a new sample. K-nearest neighbor is also easy to understand and simple, but one major drawback is that the distance to the other samples needs to be calculated for each new sample to decide which are the k closest and is therefore dependent on a good searching technique. Support vector machines is one of the most popular ML algorithms. It is also called optimal margin classifier and, simplified, it tries to find the optimal separation line or plane to differentiate two or more different classes. This is a robust algorithm that also performs well when modeling non-linear relations. It is used in many different areas including image analysis [33].

According to Seo et al. [34], the use of random forests, k-nearest neighbor, and support vector

machines for image segmentation have been studied a lot in the last decade, but the success is limited. Support vector machines can be trained on small amounts of training data and are easy to train because they are less complex than neural networks, but this also makes them less flexible for more complex features [34]. Random forests are also simple to train and have high accuracy, but because this is an ensemble of many trees, the internal process becomes difficult to follow [34].

2.3.2 Deep learning and neural networks

This subsection about DL and neural networks is based on LeCun et al. [35], unless otherwise stated.

DL is a method that mimics the human brain, using artificial neural networks. Training can be supervised or unsupervised. Neural networks consist of nodes that represent neurons and are connected to each other and organized into different layers. The typical layers are the input layer, the output layer, and hidden layers. In figure 12 a simplified neural network is illustrated. DL is all neural networks with more than three hidden layers between the input and output layer.

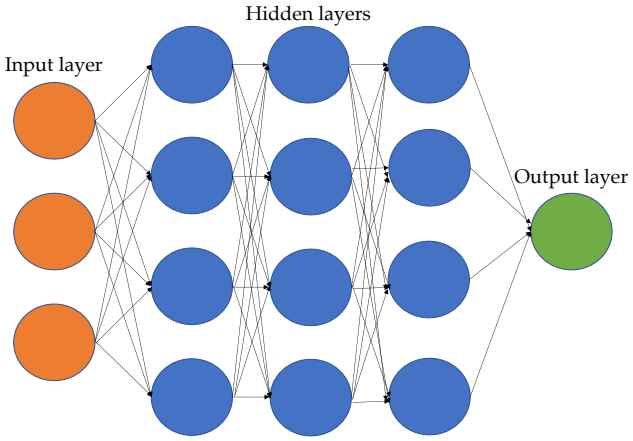


Figure 12: A simplified illustration of how a neural network is connected.

The input layer could be an image that is to be classified by the neural network model and could include the same number of nodes as pixels in the image. The output layer is the output of the model and can include several nodes depending on the model. The model chooses the output that is most probable to be correct. The hidden layers are where the computations take place. The number of hidden layers varies from model to model, and the number of neuron nodes in each layer can vary from layer to layer depending on the specific task of the layer. Each neuron has a set of weights, inputs, and an activation function that determines the output value of the neuron. The output value determines whether or not the specific feature is important or not. The weights of the nodes are the learnable characteristics of the neural network and are optimized during the training of the model.

The loss function is used to optimize the model during training and is calculated once the model has made a prediction for a single training case. The prediction can then be compared to the labels of the data. The training aims to minimize the loss function and thereby get a model that has outputs as close to the correct labels as possible. A loss function can be many different things, but one simple example is the mean squared error (MSE) [36].

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(y_i - y_p(x_i))^2,$$

where N is the number of inputs, y_i is the actual output, y_p is the predicted output based on the input x_i. Minimizing this function will lead to the smallest possible differences between actual output and predicted output.

Backpropagation is a procedure used for supervised training of neural networks to calculate the gradient descent of the loss function. The gradient descent can be used to adjust the learnable parameters of the model in order to minimize the loss function. After a single training case moves forward through the network, the backpropagation moves from the output layer to the input layer of the same case and calculates the gradient of the loss function. This gradient is then used to adjust the parameters of the model in the direction that minimizes the loss function. For multilayer neural networks, having local minima does not seem to be an issue for achieving good results, neither does saddle points. Backpropagation allows adjusting all the model parameters efficiently and is therefore widely used for training neural networks [36].

Overfitting of the model is generally a problem for neural networks, and as the models are already large, i.e. many layers and nodes, it is not efficient to make an ensemble of several networks to deal with this [37]. Dropout is a technique that significantly reduces overfitting. The way dropout works is by giving each node in the hidden layers a probability of being removed, often set to 0.5, and then removing the node with its connections to other nodes [37]. This thins out the network and is done for each training case that is presented [37]. The final network is then made by combining the thinned networks from the training but scaling down the weights of the nodes with the probability that the node was removed during training [37]. The purpose of this process is to not let the network become overfitted to the training data by combining networks that are built slightly differently.

Neural networks have shown segmentation results similar to the performance of manual segmentation and have become more popular these recent years due to significant improvements in computational power and the ability of the network to automatically choose the best features to learn, from large amounts of training data [34]. Insufficient training data is the largest issue for neural networks, although it is also a problem that these models are less intuitive and can feel like a black box [34]. According to Litjens et al. [38] the black box problem is especially important in the field of medicine, and several approaches have been made to make the models understandable for the user. Another way to improve the trust in the model is to access uncertainty estimates from the network [38].

Two common types of neural networks and DL are recurrent neural network (RNN) and convolutional neural network (CNN). RNN are often used for speech and language tasks because the network has access to a vector with information about the history of the past elements in the sequence. This makes RNN good at predicting the next letter in a sentence, for example, but they are used for much more complex tasks than that, as well. Generally, these networks are used for tasks that include time steps. CNNs are designed for processing multiple arrays, for example, tasks involving images or videos. CNNs are good at this because convolutional layers are good at extracting features in images, while the computational cost is kept low [15]. As the input moves further into the layers, the features that are detected become increasingly complex, from edges to shapes.

U-Net

U-net was introduced by Ronneberger et al. [39] in 2015 and these next paragraphs about U-net are based on his paper, unless otherwise stated. A U-Net is a learning algorithm within DL, a type of CNN that uses supervised learning and is specialized for biomedical image segmentation. U-Net is a fully CNN, as this network only has convolutional layers in the hidden layers. What separates this network from a regular CNN is that U-Net needs fewer training images and gives more precise segmentations. These improvements are due to data augmentation with elastic deformations and a supplementary path with up-sampling, respectively. Regular CNNs are usually used just for classification, while the U-Net can make a classification for each pixel and thereby achieve both classification and localization. A 3D version of the U-net was presented in 2016 by Çiçek et al. [40], where all the 2D operations are replaced with their 3D counterparts and the output is volumetric segmentations.

When U-net is used to train a model for organ segmentation, it is presented with patient cases, already segmented, and learns features from these cases. The number of features a model should

learn is specified by the input type before the training begins. This means that a model could be trained on any number of cases and the run-time for the model would be the same because the number of features is the same. A trained model classifies each voxel in the images to either be part of a specific organ or unspecified tissue.

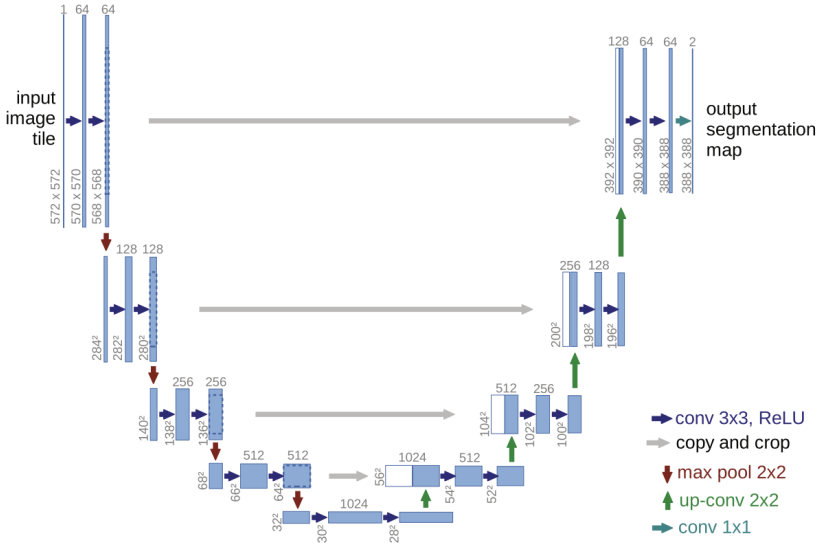


Figure 13: The U-Net architecture (Ronneberger et al. [39]). The left part corresponds to the contracting path, and the right part corresponds to the expansive path. The number above the box is the number of feature channels and the number to the left of the box is the image dimension.

The architecture of the U-Net model is presented in figure 13. The U-shape of the model is made by the contracting path to the left and the expansive path to the right. The contracting path captures context, and the expansive path enables precise localization. The resolution of the input is decreased during the contracting path and without the expansive path, the network could not localize the classification because of low resolution in the output. The contracting path is typical for CNNs and consists of two repeated convolutional layers, both using a rectified linear unit (ReLU) as activation function, and a max-pooling unit for downsampling. The expansive path up-samples by convolution and then combines this with the corresponding part of the contracting path and performs two convolutions with ReLU activation functions, assembling a more precise output than would be possible at the bottom of the "U"-shape. Dropout layers are added to the end of the contracting path to avoid overfitting. In the end, the output is a segmentation map.

2.3.3 Artificial intelligence in radiation oncology

AI can be used for most of the processes in radiation oncology. The use of AI should lead to increased quality, standardization, and acceleration of many of the processes involved [3]. The most popular AI applications in radiotherapy are automatic segmentation, treatment planning, and synthetic CT generation [3]. AI can also be used for quality assurance in radiotherapy [3]. In figure 14 an overview of AI in the radiotherapy workflow is shown. Before clinical use, an automatic method needs to be properly tested and validated.

According to Lin et al. [41] the use of DL models for medical image segmentation is promising. The use of multi-atlas registration combined with more traditional ML has been tested out, but registration-based methods are not stable enough for non-rigid organs, like abdominal organs for example [41]. Today, pure DL models are being implemented for image segmentation in radiotherapy [41], and according to Liesbeth et al. [3] these DL models are already outperforming the traditional automatic segmentation methods and are reaching the same accuracy as manual segmentation. The performance of these DL models depends on the quality and quantity of the training data. As patients have different builds, it is important that the model has been trained

on cases that represent the variability in the clinical data. Having high-quality training data can decrease the amount of data needed for good performance.

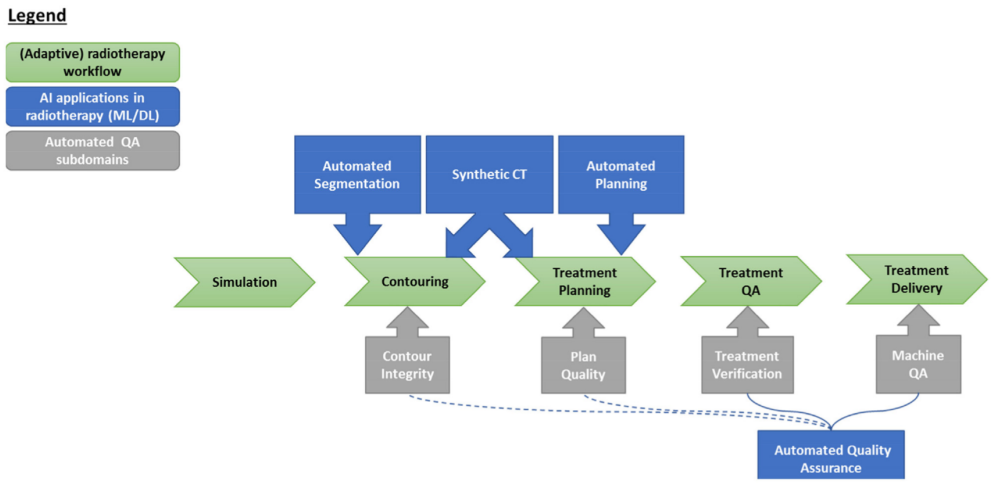


Figure 14: Overview of the radiotherapy workflow and where AI is being introduced (Liesbeth et al. [3]).

The time-saving aspects of using AI for automatic segmentation are highly dependent on the quality of the model and how much editing must be done to achieve acceptable delineations. However, several studies already show results indicating an increase in efficiency. The time-saving ranges from 12 % to 77 %, with a median of about 30 % [4, 42, 43]. Van der Veen et al. [4] also found that the inter-observer variability was reduced with the use of DL.

Treatment planning is, like image segmentation, a time-consuming process, requires a high skill level, and is associated with large inter-observer variability [3]. The use of AI has the potential to increase efficiency and lead to a more standardized process [3]. A patient dataset must be acquired with treatment plans that follow the applicable guidelines and have consistency in treatment technique and setup [3]. Some ways to automate treatment planning are knowledge-based algorithms, scripting, and protocol-based iterative planning [3]. Although scripting and protocol-based iterative planning do not necessarily use AI, they have the same intention as treatment planning based on AI. Chang et al. [44] compared manual treatment planning with knowledge-based algorithms. They concluded that this method could significantly improve planning efficiency and produce quality plans [44]. The total time saved was 78 % with the automatic method compared to the conventional method [44].

2.4 Breast cancer

Section 2.4 about breast cancer is based on Norwegian Breast Cancer Group [45], unless otherwise stated.

The most common type of cancer for women is breast cancer. In 2018, 3568 women were diagnosed with breast cancer in Norway. Breast cancer for men is rare but possible. Like any type of cancer, it starts as a mutation of a healthy cell and this mutation can either lead to increased cell division or reduced cell death. This mutated cell can over time become a tumor if the immune system does not detect it and take care of it. The tumor can either be benign (non-cancerous) or malignant (cancerous).

According to Norwegian Breast Cancer Group, the cumulative probability of females getting breast cancer before the age of 75 is 8.9 %. Survival is very highly dependent on the stage of cancer; therefore, early detection can increase the chances of successful curative treatment. For this reason, the Norwegian government offers mammography screening for women over the age of 50, with a new screening every 2 years.

2.4.1 Breast anatomy

In figure 15, the anatomy of the female breast is shown. The breast consists of 15-20 lobes that are made up of several lobules [46]. The lobe and lobules are where milk is produced, and they are connected by ducts that transport the milk to the nipple [46]. From the figure, one can see that there is fatty tissue surrounding the lobes and ducts, and lymph nodes outside the breast. The chest/thoracic wall is also visible in this figure.

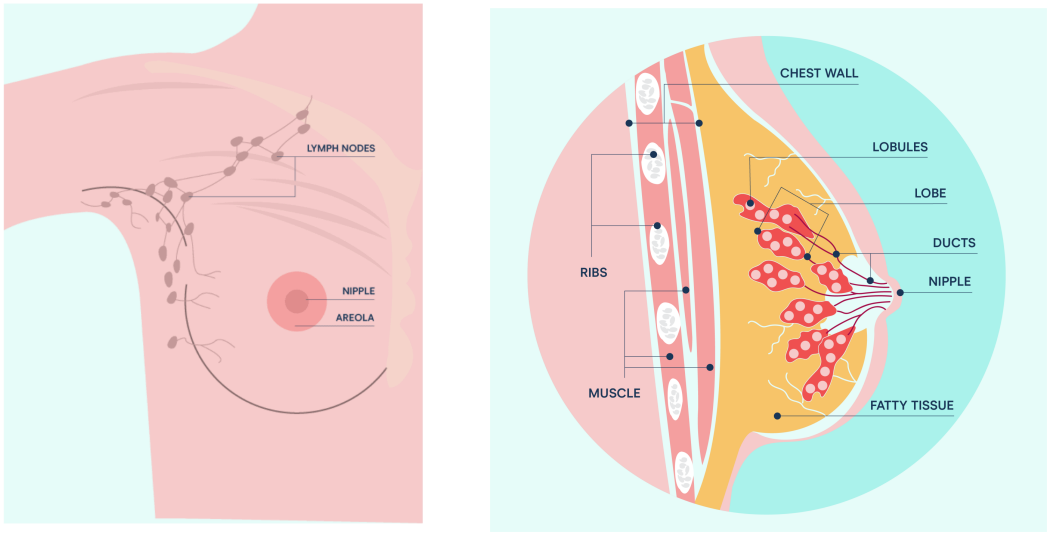


Figure 15: Anatomy of the female breast (National Breast Cancer Foundation [47]).

2.4.2 Treatment modalities

Breast cancer can either be invasive or non-invasive (in situ). Invasive cancer has spread from the lobe or duct where it originated, while non-invasive has not [48]. Invasive cancer can then spread through the bloodstream or lymph nodes to other parts of the body [48].

Treatment modalities for breast cancer usually include surgery followed by radiotherapy and chemotherapy. The surgery can be either mastectomy or lumpectomy, i.e., removal of all breast tissue or breast-conserving surgery. Of the Norwegian breast cancer patients in 2019, 81.2 % received breast-conserving surgery and the goal is to increase this to 85 % [49]. In breast-conserving surgery, only the tumor and some of the surrounding healthy tissue are removed and most of the breast is conserved. Depending on how much cancer has spread, removal of lymph nodes is also considered during surgery. After surgery follows radiotherapy to reduce the risk of relapse and to increase the chance of survival by removing any microscopic residues in the breast or areas around the breast. Chemotherapy can be given before or after operation or radiotherapy to decrease the size of the tumor or remove any leftover cancer cells. The chosen treatment combination of surgery, radiotherapy, and chemotherapy depends on the size, spread, and location of the tumor and how aggressive it is.

According to Norwegian Breast Cancer Group, five-year survival from 2014-2018 was 100 % for patients diagnosed with breast cancer without spread to lymph nodes and a tumor smaller than 2 cm (stage 1). On the other hand, patients with distant metastases had a five-year survival of 29.2 % in the same period. The prognosis is therefore highly dependent on the stage at diagnosis and can be increased with early diagnosis.

2.4.3 Radiotherapy target volumes and organs at risk

For breast cancer, the tumor has usually already been removed when the radiotherapy planning begins and therefore there will only be a CTV and no GTV. Typically, the whole breast is a CTV

together with cancerous lymph nodes. If a mastectomy has been performed, the primary CTV will be the thoracic wall instead of the breast. The primary CTV is generally limited 5 mm below the skin surface and by the major breast muscle.

The regional lymph nodes that may be treated together with the primary CTV are the axillary lymph nodes levels 1, 2, and 3, supraclavicular lymph nodes (level 4), pectoral axillary lymph nodes, and internal mammary lymph nodes. The lymph nodes included as target volumes are a part of the nodal CTV. These can be seen delineated in figure 16, together with the breast delineated as primary CTV.

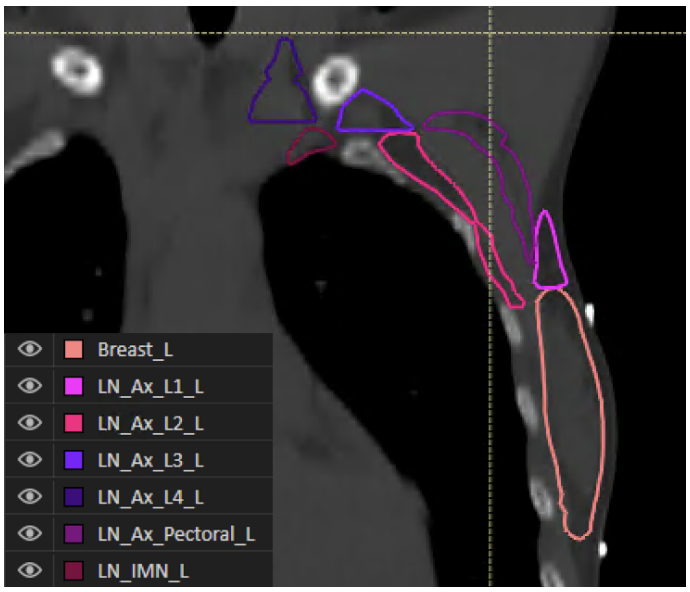


Figure 16: Left breast delineated together with axillary lymph nodes levels 1-4, pectoral axillary lymph nodes and internal mammary lymph nodes (IMN), coronal view.

It is desirable to achieve 95-107 % of the prescribed dose to the CTV. The PTV is delineated as CTV plus a 5-7 mm margin. According to Norwegian Breast Cancer Group, the PTV should be covered by at least 90 % of the prescribed dose, but St. Olavs Hospital uses 95 % coverage to the PTV.

When it comes to radiotherapy techniques, 3D-CRT is still considered the standard technique. However, there is an increased interest in using more advanced techniques, such as IMRT or VMAT, also for treating breast cancer. These techniques may give a higher amount of low-dose radiation to the contralateral breast and lungs. There is also a lack of research on late effects when using these techniques.

Different OAR have different structures and tolerances and may react differently to the same dose. The subunits of an organ can be structured more parallel or more serial. Parallel structured organs can withstand a higher maximum dose and it is the average dose that needs to be limited. These organs can keep functioning even though some subunits are damaged, but with increasing average dose the probability of normal tissue complication increases, as more subunits will be damaged. For serial organs, one needs to monitor the maximum dose given to the organ as the probability of normal tissue complication increases as a function of this. One subunit failing will affect several other subunits. The spinal canal is an example of a serial organ, while lungs have more of a parallel structure [12].

The most important OARs for breast radiotherapy are the heart and lungs. The chance of getting heart diseases and lung cancer increase with increased dose given to these organs, it is therefore the intention to radiate these organs as little as possible. For left-sided breast radiotherapy, the treatment should be done by using deep inspiration breath-hold, to increase the distance between the heart and breast and thereby minimize the radiation to the heart. According to Norwegian Breast Cancer Group [45], there is a relative heart disease risk increase of 7.4 % for each Gy the average heart dose increases with and the dose should therefore be kept below 2 Gy. For the lungs,

with a fractionation scheme of 15 x 2.67 Gy and radiation of breast and regional lymph nodes, the volume receiving 18 Gy or above should be less than 35 %. However, the dose should always be kept as low as reasonably achievable (ALARA), even below these limits [50].

3 Materials and methods

In this section, the overall procedure, patient data, segmentation models, plan optimization script, and different evaluation methods are presented. Statistical analyses used are also presented. Some parts are borrowed from the project thesis written prior to this master thesis [7]. Specifically, subsections 3.4.1 and 3.5.1 about geometric evaluation and boxplots, respectively, is copied from Isachsen [7] with minor changes.

3.1 Patient data

Two different patient datasets, both anonymized, were used for this study.

- COBRA. These patients are taken from the "Acute and Long-term Cardiovascular Toxicity After Modern Radiotherapy for Breast Cancer" (COBRA) study ongoing at St. Olavs Hospital in Trondheim and Ålesund Hospital [51]. The manual segmentations have been redone for the training of the breast model and are therefore regarded as high-quality segmentations. The patients included in this study were CT-scanned during deep inspiration breath-hold, a technique used to minimize radiation to the heart and lungs during treatment for left-sided breast cancer. This dataset was split into a training dataset and a test dataset. The training dataset includes 168 patients, and the test dataset includes 15 patients.
- CLINICAL. These are patients that were treated for left-sided breast cancer in 2020 at St. Olavs Hospital. This dataset includes 16 patients, 15 who were CT-scanned during deep inspiration breath-hold and one that was not. All the patients have a clinical plan, either VMAT or hybrid plan. A hybrid plan is a combination of conventional tangential arcs (3D-CRT) and VMAT.

Table 1 gives an overview of how the patient data was used in this thesis. Both patient datasets have separate lymph node-areas delineated. These are the left pectoral axillary lymph nodes, left axillary lymph nodes levels 1-4, and left internal mammary lymph nodes. The internal mammary lymph nodes were not used as target volume for any of the patients in this thesis. Also, one patient in the CLINICAL dataset does not have axillary lymph nodes level 1.

Table 1: Overview of the patient data involved in this thesis, and how the data was used. Two different clinical evaluations are included, distinguished here by S (standard) and T (modified Turing test). "-" indicates not applicable.

	Local segmentation model	Siemens segmentation model	Automatic plan optimization
Training data	COBRA (n = 168)	unknown external data	-
Geometric eval.	COBRA (n = 15)	COBRA (n = 15)	-
Dosimetric eval.	COBRA (n = 15)	COBRA (n = 15)	CLINICAL (n = 16)
Clinical eval., S	-	COBRA (n = 15)	-
Clinical eval., T	-	CLINICAL (n = 16)	-

3.2 Automatic segmentation

In this subsection, the locally trained segmentation model and the pre-trained Siemens model are presented.

3.2.1 Local model

This segmentation model was trained with local data from St. Olavs Hospital and trained together with RaySearch Laboratories AB (Stockholm, Sweden). This model is the second last version and

not the final model that is planned to be used clinically at St. Olavs Hospital. This version was made available in the treatment planning system RayStation 9B at the hospital in March 2021. This model is a collaboration with Ålesund Hospital, as well.

The local model is a DL segmentation model for OARs and breast/lymph node segmentation. It was trained by supervised learning using the COBRA training dataset with manual segmentations as a DL network of type U-net. The model can segment 24 organs/regions of interest in the breast/thorax area. These segmentations take less than 2 minutes to generate for one patient. In figure 17, an example of the segmented structures for one patient can be seen.

All target volumes and a subset of the OARs, as shown in figure 18, were evaluated in this thesis. They were evaluated with geometric and dosimetric metrics using the manual delineations as "ground truth". For the dosimetric evaluation, a treatment plan was made using the automatic segmentations and the plan optimization script described in section 3.3.

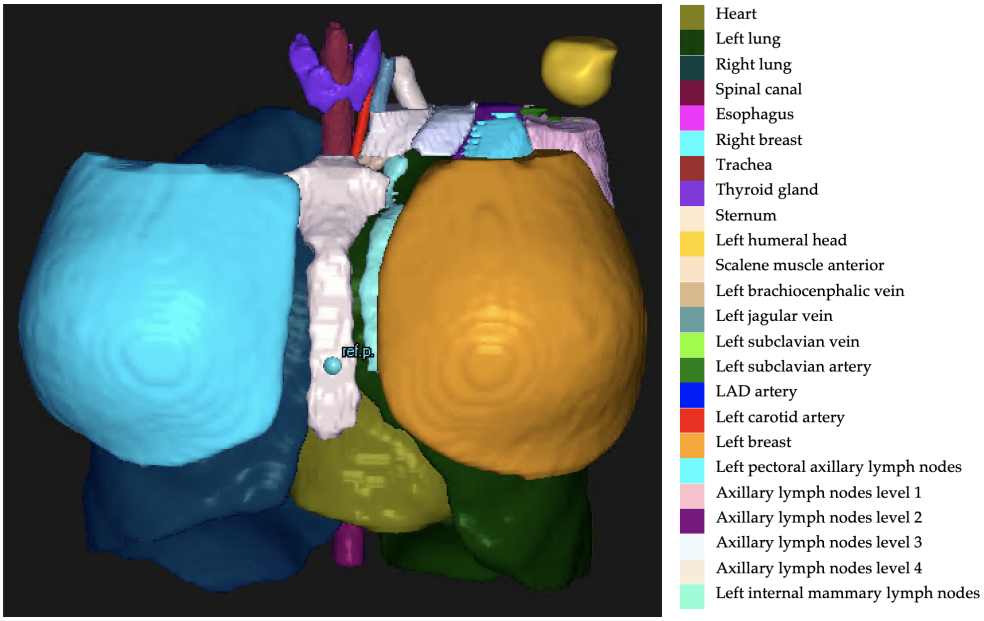


Figure 17: The local model segments 24 organs/regions of interest, as shown in this example.

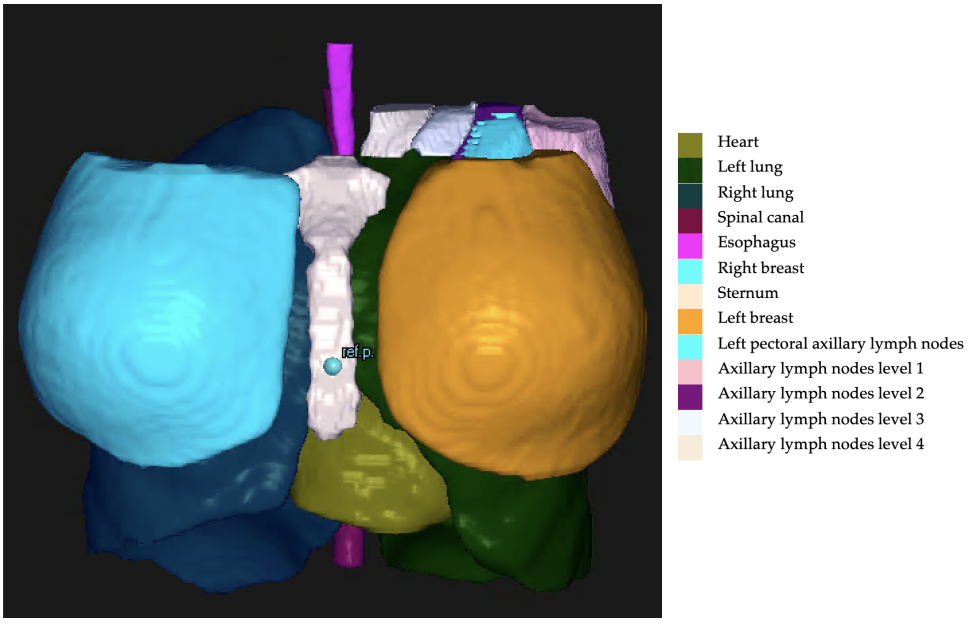


Figure 18: The structures evaluated in this thesis segmented by the local model.

For all evaluations of the model, the spinal canal and esophagus have been cut in the craniocaudal direction, so that these organs have the same length for the manual delineation and automatic segmentation. This has been done because the length is not as important for these organs, as long as they are segmented in the length corresponding to the PTV. The differences in length for the local model were minimal, about 2 slices different.

3.2.2 Siemens model

AI-Rad Companion Organs RT is a DL segmentation model trained by Siemens Healthineers. It is here referred to as the Siemens model which is available in Siemens' cloud-based solution Teamplay. An example of a patient where the organs/regions of interest in the thorax-area have been segmented is shown in figure 19. Details about the model, i.e., training data and algorithm details, are unknown.

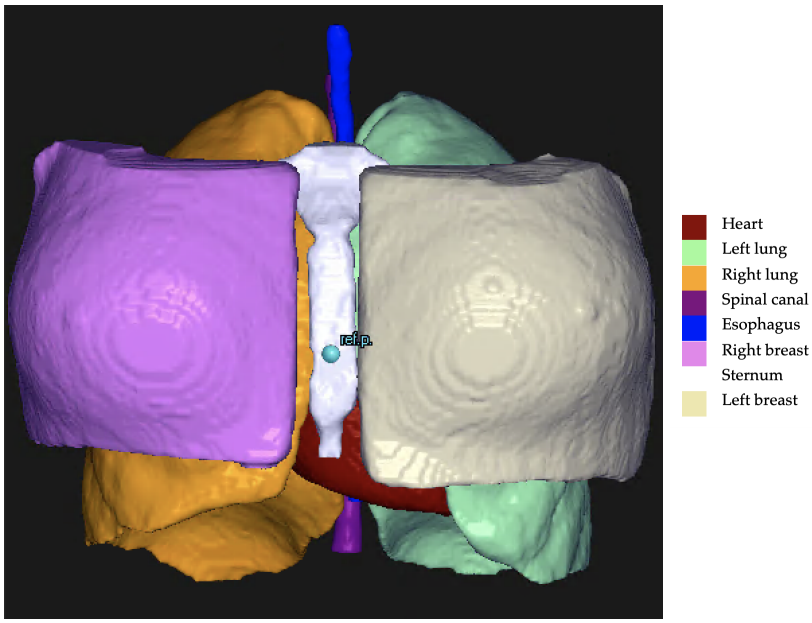


Figure 19: Shows a patient where the pre-trained Siemens model has been used and the organs/regions of interest have been segmented.

The pre-trained Siemens model is not available directly in RayStation, so the CT-scans need to be exported Teamplay, and after about 5 minutes the automatic segmentations can be imported. A geometric and dosimetric evaluation was done for this model, as well as clinical evaluations. Due to the limited amount of OARs and target volumes this model can segment, it was not possible to make treatment plans based on automatic segmentations. Instead, plans were made using the manual delineations and the plan optimization script described in section 3.3.

For the breasts, there were some obvious differences in guidelines used at St. Olavs Hospital and used to make the Siemens model. Therefore, some editing was also done manually to the breasts segmented by the Siemens model. The segmentations were cropped 5 mm below the body surface. The spinal canal and esophagus were also cut in the craniocaudal direction for the Siemens model and the difference in length were generally larger for this model than the local model. The editing of length was done separately for the two models and the manual volume of the spinal canal and esophagus, therefore, differ for the models. Examples of the manual editing can be seen in appendix D.

3.3 Automatic plan optimization

VMAT plans were automatically generated in RayStation based on python scripts developed by Jomar Frengen and Marit Funderud at St. Olavs Hospital. This script uses a protocol-based iterative method for planning. It is made for breast cancer patients and makes a plan with 15 fractions of 2.67 Gy to the breast as PTV. This equals a prescribed dose of 40.05 Gy.

The script requires the target volumes to be segmented, as well as certain OARs, i.e., the heart, lungs, and contralateral breast. The target volumes include the breast, pectoral axillary lymph nodes, and axillary lymph nodes levels 1, 2, 3, and 4, and internal mammary lymph nodes. Other OARs are also taken into account by the script if they are segmented, i.e., esophagus, humeral head, spinal canal, sternum, thyroid, and trachea. The script is made for different combinations of target volumes for either left-sided or right-sided breast radiotherapy. These specifics are chosen when running the script. For this thesis, the left breast was the primary target volume, with pectoral axillary lymph nodes and axillary lymph nodes levels 1, 2, 3, and 4 as nodal target volume.

The minimum dose objective for the nodal and primary PTV are 36.2 Gy and 38.2 Gy, respectively. The corresponding CTVs have objectives 0.3 Gy higher. The absolute goal for 98 % of the nodal and primary PTV was 90 % and 95 % of the prescribed dose, respectively. The script uses about 30 minutes for the whole process and ends up with a finished treatment plan that is ready to be evaluated by a physicist. To validate the script, plans were made and compared to manually made clinical plans with dosimetric evaluation. An example slice from a patient where the script has been used is shown in figure 20.

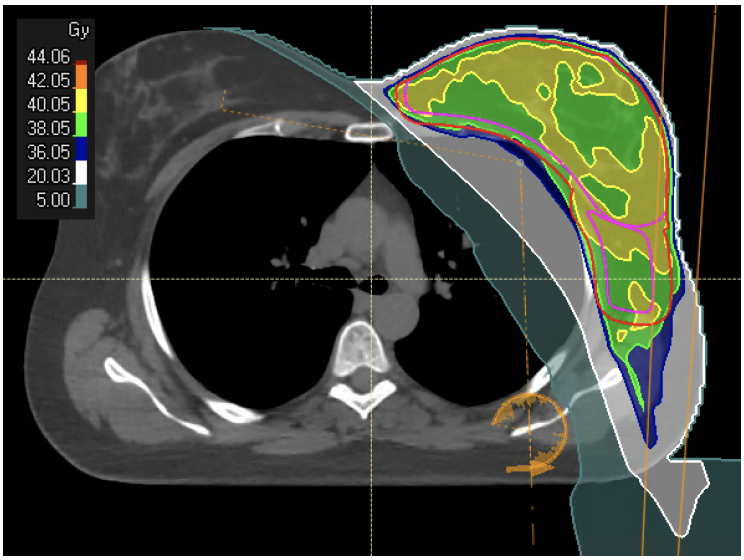


Figure 20: Transversal plane with dose distribution for a patient where the plan optimization script has been used. Pink segmentation is the CTV, and the red segmentation is the PTV, both delineated manually.

A hybrid plan has been the standard technique at St. Olavs Hospital for breast cancer patients for several years, but it can be favorable to make a VMAT plan instead, depending on the anatomy of the patient. If the heart dose or left lung dose is too high, a VMAT plan can often be better. For the patients used to evaluate the script, some had hybrid plans and some had VMAT plans. Since the patient population with a hybrid plan varies in anatomy from those with a VMAT plan, it was natural to divide the patients into two groups depending on which type of clinical plan was available. The script was therefore evaluated against the manually made VMAT plans and manually made hybrid plans separately.

3.4 Evaluation

In this thesis, geometric evaluation, dosimetric evaluation, and clinical evaluations were done.

3.4.1 Geometric evaluation

Three geometric metrics were used to evaluate the DL segmentations. The Dice similarity coefficient (DSC) was used as an overlap metric, Hausdorff distance (HD) was used as a distance metric and absolute volumes were compared to detect systematic differences. The HD mainly used was the 95^{th} percentile. For extracting these metrics, the scripting possibilities in RayStation were used. The script for extracting the geometric metrics for evaluation of segmentation models can be found in appendix A. The scripts for extracting dosimetric metrics can be found in appendices B and C for evaluation of segmentation models and automatic plan optimization, respectively.

Dice similarity coefficient

DSC is an overlap measure between two regions or volumes. This coefficient was developed by Lee R. Dice and published in 1945 [52]. It is defined as

$$\text{DSC} = \frac{2 \cdot |A \cap B|}{|A| + |B|},$$

where A and B are the regions of interest and $A \cap B$ is the intersecting area of A and B [52]. The terms involved in the DSC are illustrated in figure 21.

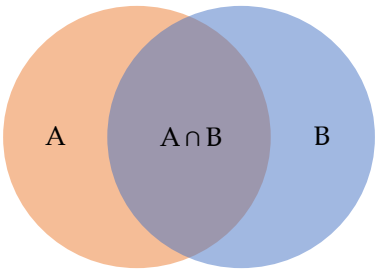


Figure 21: Illustrates the terms relevant for calculating the DSC of two regions or volumes A and B. DSC is the overlap, normalized by the area of the regions.

DSC describes the fraction of the overlapping region between two regions and is independent of which region is the reference. If the overlap is complete the coefficient will be 1 and if there is no overlap it will be 0.

Hausdorff distance

HD is a measure of the maximum distance between two sets of points in a metric space. A smaller value means a smaller geometric difference between the objects. Given two point sets $A = \{a_1, a_2, ..., a_p\}$ and $B = \{b_1, b_2, ..., b_q\}$ the HD is defined by Huttenlocher et al. [53] as

$$H(A, B) = \max(h(A, B), h(B, A)),$$

where

$$h(A, B) = \max_{a \in A} \min_{b \in B} ||a - b||.$$

$h(A, B)$ is called the directed HD distance using A as reference figure [53]. The directed HD can be seen as the maximum distance from a point on the reference figure directly to the other figure.

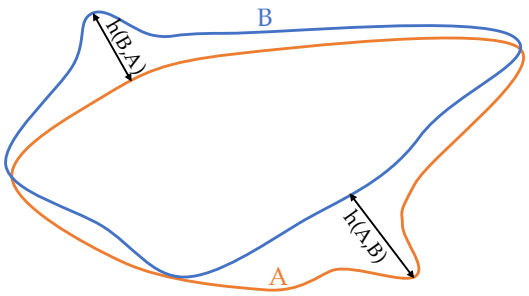


Figure 22: Illustrates the directed HD between two figures A and B. The HD would be h(A,B), since this is the largest of the two distances illustrated.

The HD is then the maximum of the directed HDs using both figures as a reference figure in turn. In figure 22 the directed HD has been drawn in for two shapes overlapping.

In practice, HD is sensitive to outliers [54]. Therefore, it is often useful to calculate different HD percentiles. For example, one can calculate the 95^{th} percentile and 99^{th} percentile HD. Calculating a HD percentile is selecting the distance that is in the quantile equivalent to the percentile one wishes to obtain. The 100^{th} percentile HD is what is calculated if no percentile is specified.

3.4.2 Dosimetric evaluation

Dosimetric metrics were used to evaluate both automatic segmentation and automatic plans.

Ideally, the target volume receives uniform coverage, while the OARs receive a dose as low as possible. Isodose curves are helpful to look at when evaluating whether the plan is reaching its goals. These curves connect the points that receive the same dose. The area enclosed by an isodose curve, therefore, receives the same or higher dose as the applicable dose. For target volumes, the isodoses are defined in percentage of the prescribed dose, while for the OARs it is more relevant to look at the absolute dose.

Different dose metrics can be used to evaluate how much dose is given to a specific volume or to quantify the relative volume covered by an isodose curve for a specific region of interest. It is also useful to use conformity index (CI) and homogeneity index (HI) to evaluate a plan. CI is the ratio between the PTV covered by the reference isodose and the total reference isodose volume. The reference isodose is 95 %, i.e., 95 % of the prescribed dose. CI is then defined as

$$CI = \frac{PTV \cap V95}{V95},$$

where V95 is the relative volume covered by 95 % of the prescribed dose. HI is the ratio between the maximum dose in the target volume and the reference isodose. HI is defined as

$$HI = \frac{D2}{D98},$$

where D2 and D98 are doses at 2 and 98 % volume, respectively. An overview of the different dose metrics used can be found in table 2. These dose metrics can also be used as clinical goals during the optimization of the treatment plan. For example, the 95 % isodose should be covered by 98 % of the PTV.

Dose-volume histogram (DVH) is a common analysis tool for evaluating treatment plans. They can be cumulative or differential. Cumulative graphs will be used in this thesis. This type of graph relates the relative volume of a region with the absolute dose and is therefore useful to evaluate coverage to the target volume and sparing of OARs. The target volume curve should be as far to

Table 2: Overview of the different dose metrics used. Relative sizes are given in percentage.

Metric	Unit	Explanation
Dx	[Gy]	Dose given to relative volume, x, given in %
Dxcm^3	[Gy]	Dose given to absolute volume in cm^3, x
Dmax	[Gy]	Maximum dose given to the region of interest, here defined as D0.03cm^3, i.e., the dose given to a volume of 0.03 cm^3
Dmean	[Gy]	Average dose given to the region of interest
Vx	[1]	Relative volume covered by the relative dose, x, given in %
VxGy	[1]	Relative volume covered by the absolute dose in Gy, x
CI	[1]	Conformity index
HI	[1]	Homogeneity index

the upper-right corner as possible and go straight down at the prescribed dose. For the OARs, the curve should be as far to the lower-left corner as possible, thereby following the ALARA-principle. Example curves for PTV and OAR is shown in figure 23.

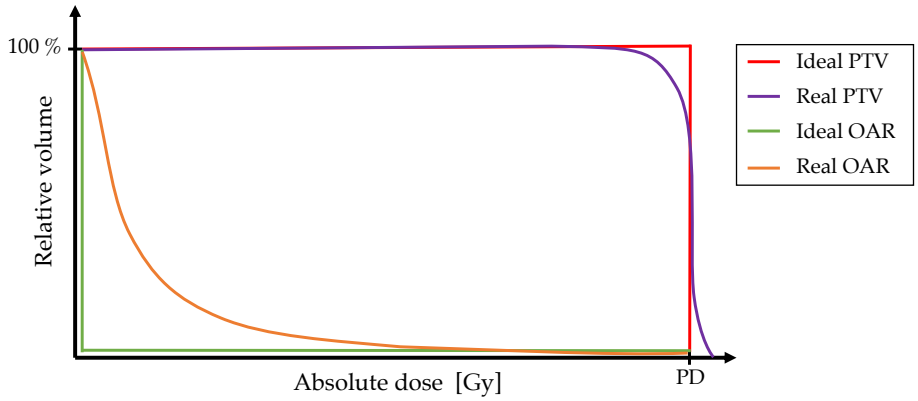


Figure 23: Cumulative DVH curves for PTV and OAR with an ideal case and a more clinically realistic case. PD is the prescribed dose.

3.4.3 Clinical evaluations

The Siemens model was evaluated clinically in two different ways, clinical scoring and modified Turing test.

Clinical scoring

Results from a clinical evaluation using a scoring system were included in this thesis. One experienced radiation therapist scored the segmentations using a scale as follows:

0 - Rejected

1 - Major corrections needed, but I would still use the model; small time gain

2 - Minor corrections needed; significant time gain

3 - Accepted without corrections

The score was made based on usage for adjuvant breast cancer radiotherapy. The radiation therapist had access to the manual segmentations while rating the automatic segmentations, as well as the whole CT-scan. This was done for both lungs, spinal canal, esophagus, and sternum using

the test dataset COBRA patients. The original segmentations were used for the spinal canal and esophagus and not the ones that had been edited to be the same length as the manual delineations.

Modified Turing test

A modified Turing test, very similar to that of Gooding [55], was executed as described below.

The clinical evaluation was done for the heart and left breast (CTVp). Two oncologists were presented with a set of questions put together in a free questionnaire website, `nettskjema.no`. The questionnaire was made together with a fellow student, Mari Rossvoll. The three main questions were:

1. For a single segmentation: "How was this segmentation drawn?", answer options:
   (a) By a human
   (b) By an AI model
2. For two segmentations (one AI model and one manual delineation): "Which segmentation do you prefer?", answer options are the colors of the two segmentations shown.
3. For a single segmentation: "You have been asked to quality assure this segmentation, would you...", answer options:
   (a) Require it to be corrected; there are large obvious errors
   (b) Require it to be corrected; there are minor errors
   (c) Accept it as it is; there are minor errors but correction is not necessary
   (d) Accept it as it is; this segmentation is very precise

   .

For each of the three main questions, 8 CT-slices were shown, in turn, for each of the organs. This made a total of 48 questions. For the first and third main question, manual delineations and AI model segmentations were shown equally many times but in random order. For the second main question, the color of the two segmentations was randomly set. The questions alternated between the organs so that the oncologists were less affected by what they have seen earlier for that organ. They were also asked not to go back to edit previous answers.

For each question, the slice was chosen by randomly choosing a patient and thereafter randomly choosing a slice according to the relevant range of slices. The relevant range of slices was chosen depending on what was to be shown in that specific question, i.e range of manual delineation slices, range of AI model slices, or range of slices that included both segmentations. The range did not include the most caudal or cranial slices as these are difficult for the oncologists to evaluate without more context presented.

The 16 patients from the CLINICAL dataset were used, but two patients with breast implants were excluded. The segmentation model is most likely not made for patients with implants.

3.5 Statistical analysis

In this subsection, boxplots, Wilcoxon signed-rank test, and Spearman's rank correlation are presented. Boxplots were used to visualize some of the results, Wilcoxon signed-rank test was used to find statistical significance, and Spearman's rank correlation was used to find the correlation coefficient between some metrics for the segmentation models.

3.5.1 Boxplot

In figure 24 an overview of how to interpret a boxplot is shown. The range between 25^{th} and 75^{th} percentile is called the interquartile range. Values that are more than 1.5 times the interquartile range larger than the upper quartile or less than 1.5 times the interquartile range smaller than the lower quartile are drawn in as an outlier point. The outliers are excluded from the minimum and maximum values. The lines from the interquartile range to the minimum and maximum values are called lower and upper whiskers, respectively.

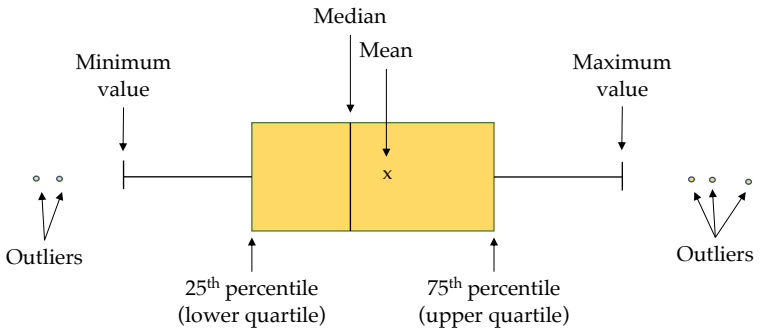


Figure 24: Shows how to read a boxplot. The box covers the mid 50 % of the samples, with the median indicated by the line. Mean value is the "x" and the whiskers illustrates the minimum and maximum values of the results, excluding the outliers.

3.5.2 Wilcoxon signed-rank test

Wilcoxon signed-rank test was presented by Frank Wilcoxon in 1945 as a method to test the statistical significance of paired differences [56]. This test is an alternative to the paired Student's t-test and can be used when the differences cannot be assumed normally distributed. This is because it is a non-parametric test and is therefore especially useful when the sample size is small.

There are small variations in how this test is implemented, especially when it comes to which test statistic is used. In this thesis, Wilcoxon signed-rank test was used as described below.

The null hypothesis, H_0, is that the differences follow a symmetric distribution around 0.

1. Pairwise differences between the two related samples, of size N, are calculated, as well as the absolute differences.
2. Differences that are equal to 0 are removed, also from total number of differences, N.
3. The absolute differences are ranked from closest to 0 (rank 1) and furthest away from 0 (rank N).
4. Sum the positive ranks and the negative ranks, separately. The smallest of these sums are set as test statistic, T.
5. $T_{crit}(\alpha, N, \text{two-sided})$ are found as table values (Wilcoxon signed-ranks table) but can also be calculated from a normal distribution if N is sufficiently large. α is the chosen significance level.
6. If $T > T_{crit}$ the null hypothesis cannot be rejected.
7. Calculate p-value $= 2(1 - \mathcal{N}(z))$, where $\mathcal{N}$ is the normal distribution and z is the z-score calculated as $(T - \mu)/\sigma$, where $\mu = N(N + 1)/4$ and $\sigma^2 = \mu(2N + 1)/6$.

3.5.3 Spearman's rank correlation

Spearman's rank correlation coefficient is a nonparametric version of the more known Pearson correlation coefficient. The correlation between the data in this thesis is not necessarily linear and the distribution is unknown, therefore, Spearman's rank correlation coefficient is used. It can be used for data where at least one of the variables is ordinal, i.e., one can rank the data. Additionally, Spearman's rank correlation coefficient is not sensitive to outliers because the calculations are done on the rank of the values and not the values in themselves.

The coefficient is found by ranking the samples from highest to lowest, individually for both sets of data. Then the Pearson correlation coefficient is calculated on the ranks of the data. The formula for the coefficient is

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}},$$

where n is the sample size, x_i and y_i are the individual samples, i.e., the rankings of the two datasets, and $\bar{x}$ and $\bar{y}$ are the mean values of the rankings. When calculating Spearman's coefficient the mean values will be the same for both datasets, unless there are tied ranks.

The direction of the coefficient is known according to the sign, and the size determines the strength. In this thesis, a coefficient lower than 0.4 indicates no correlation or weak correlation, from 0.4 to 0.7 indicates a moderate correlation, and a coefficient higher than 0.7 indicates a strong correlation.

4 Results

The results for both automatic segmentation models are presented, followed by the results from the evaluation of the automatic plan optimization. All statistics are calculated by the Wilcoxon signed-rank test using a p-value less than 0.05 as significant.

Primary and nodal CTV are denoted as CTVp and CTVn, respectively. Primary CTV is the left breast, while the nodal CTV is the union of the left pectoral axillary lymph nodes and left axillary lymph nodes levels 1-4. The corresponding PTVs are PTVpc and PTVnc, where "c" indicates "cropped".

4.1 Automatic segmentation

In this subsection, the results from the geometric, dosimetric, and clinical evaluations are presented for both segmentation models. Additional metrics, i.e. 99^{th} percentile HD (HD99), 100^{th} percentile HD (HD100), average distance (AVD) and some dosimetric metrics, can be found in appendix D. The geometric evaluation for the individual lymph node areas segmented by the local model can also be found in this appendix.

4.1.1 Geometric evaluation

Mean DSC and 95^{th} percentile HD (HD95) are presented in table 3. DSC and HD95 is also presented as boxplots in figures 25 and 26 for OAR and target volumes, respectively. Statistically significant differences in DSC and HD95 between the models was found for all structures except the lungs. The differences are all in favor of the local model, except for HD95 for the sternum, where the Siemens model has a lower distance than the local model.

Table 3: Mean DSC and HD95 for both models. Standard deviation (SD) is denoted as the ± value.

	DSC		HD95 [cm]	
	Local model	Siemens model	Local model	Siemens model
Heart	0.96 ± 0.01	0.91 ± 0.03	0.5 ± 0.1	0.9 ± 0.4
Left lung	0.973 ± 0.008	0.968 ± 0.005	0.35 ± 0.03	0.34 ± 0.02
Right lung	0.979 ± 0.005	0.975 ± 0.004	0.33 ± 0.03	0.32 ± 0.03
Spinal canal	0.94 ± 0.02	0.83 ± 0.03	0.17 ± 0.03	0.22 ± 0.03
Esophagus	0.87 ± 0.02	0.81 ± 0.04	0.25 ± 0.03	0.29 ± 0.07
Sternum	0.92 ± 0.02	0.87 ± 0.01	0.7 ± 0.5	0.4 ± 0.2
Right breast	0.94 ± 0.01	0.90 ± 0.03	0.6 ± 0.1	1.4 ± 0.6
CTVp	0.94 ± 0.02	0.89 ± 0.03	0.6 ± 0.2	1.4 ± 0.6
CTVn	0.76 ± 0.07		1.3 ± 0.7	

The absolute volumes are plotted in figure 27. Statistically significant differences were found for all structures except the spinal canal and esophagus for the local model and for all except the esophagus for the Siemens model. Points lying above the dotted line indicate that the model has segmented a larger volume than manual delineation, and points lying below the dotted line indicate that the model has segmented a smaller volume than the manual delineation.

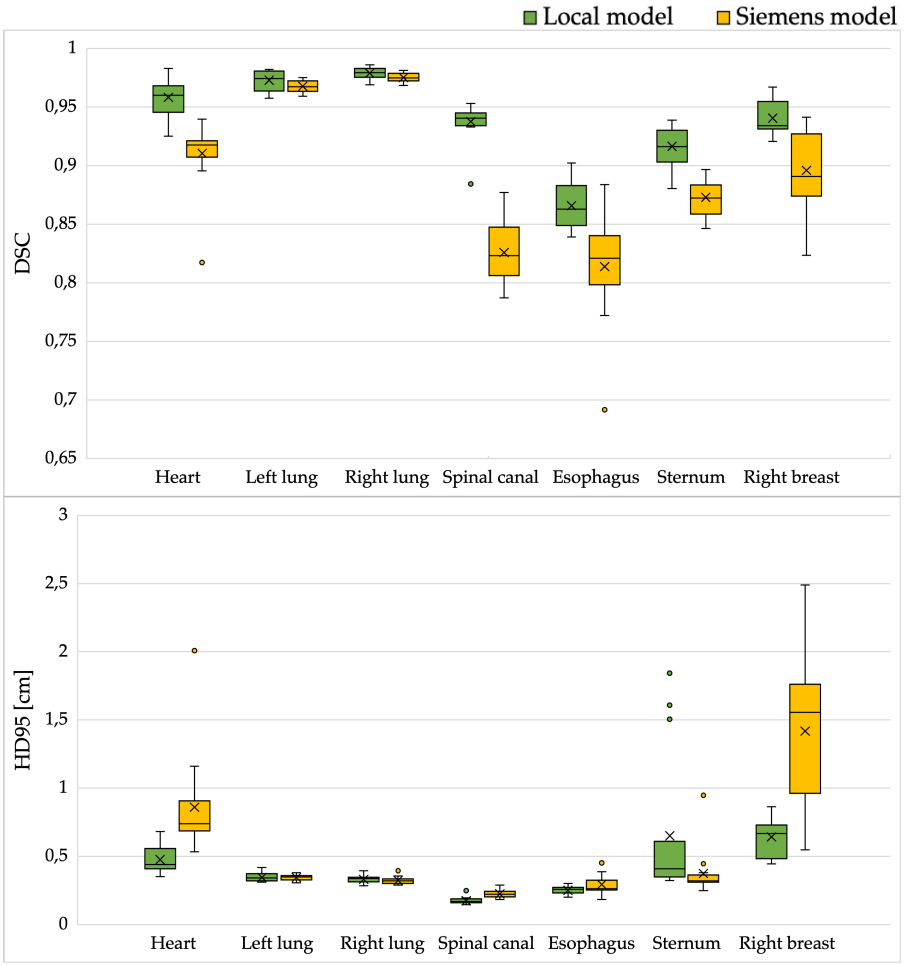


Figure 25: DSC and HD95 obtained by both segmentation models for the organs at risk.

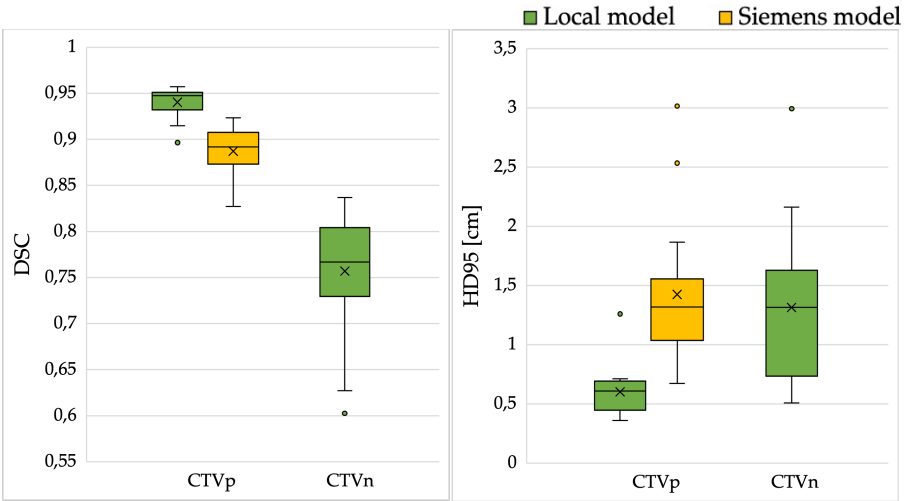


Figure 26: DSC and HD95 obtained by both segmentation models for the target volumes.

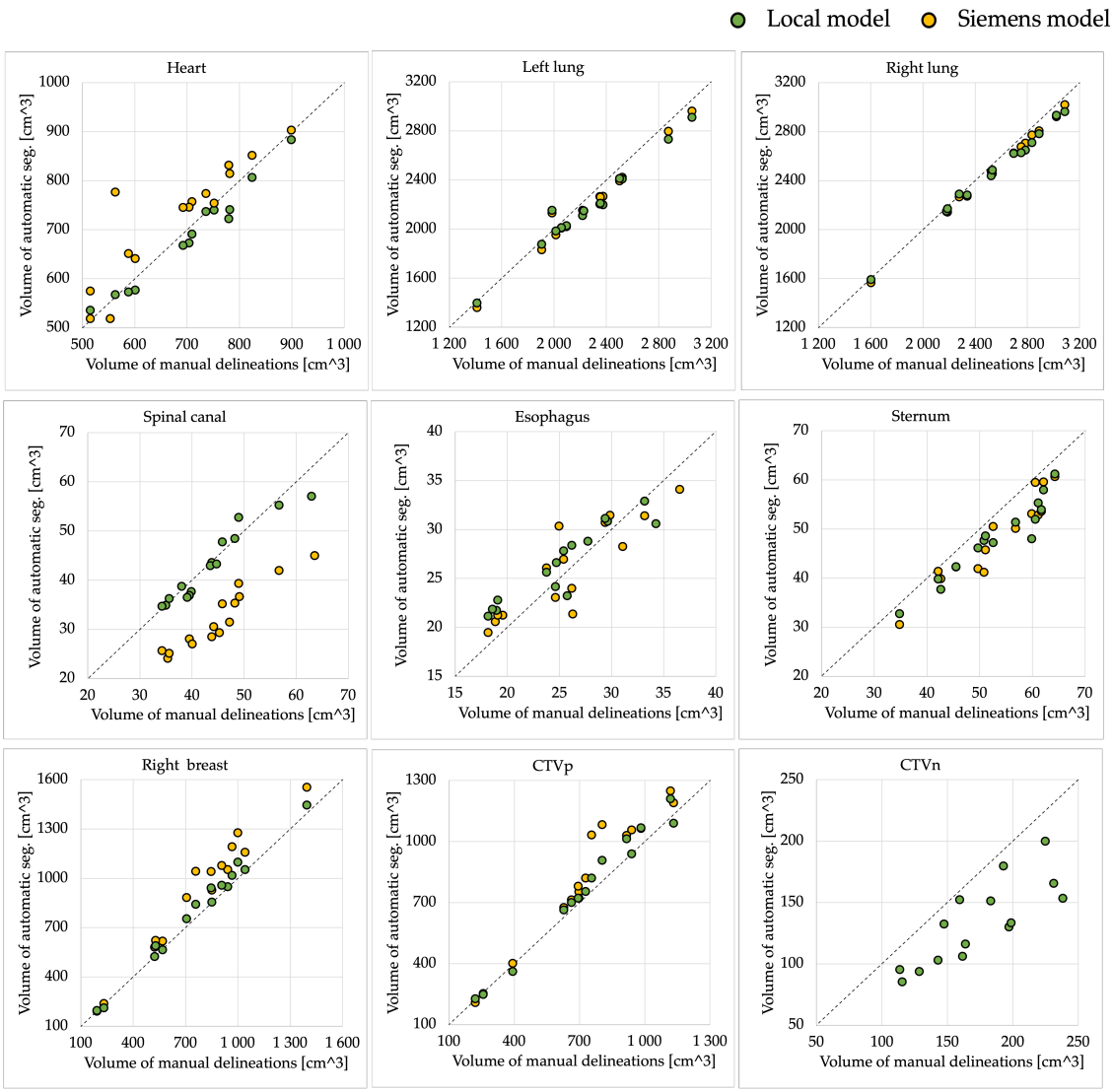


Figure 27: Volume of automatic segmentation plotted against the volume of the manual delineation for all patients. Dotted line represents equality.

4.1.2 Dosimetric evaluation

The mean values for the dosimetric metrics for the local model and Siemens model can be found in table 4 and 5, respectively. The mean pairwise difference is also presented together with a p-value indicating the level of statistical significance of the differences. The treatment plans used to evaluate the local model are made from segmentations based on local model, while the treatment plans used to evaluate the Siemens model are made based on manual delineations, i.e., clinical plans.

The dosimetric metrics are plotted for each individual patient in figure 28. The dosimetric effect of using automatic structures is evaluated based on the effect to the manual delineations, thought of as ground truth. Points below the dotted line, indicate that the model has underestimated the dose given to the structure, while points above the dotted line indicate that the model has overestimated the dose.

Table 4: Mean dosimetric metrics for the local model. The treatment plans are based on automatic segmentations by the local model. SD is denoted as the ± value. P-values in bold font are considered statistically significant.

Region	Metric	Manual del.	Local model	Pairwise difference (Manual - model)	P-value
Heart	Dmean [Gy]	1.4 ± 0.2	1.4 ± 0.2	-0.02 ± ± 0.02	**0.002**
Left lung	Dmean [Gy]	8.3 ± 0.8	8.3 ± 0.8	0.0± ± 0.1	0.363
	V18Gy [%]	17 ± 2	17 ± 2	0.1± ± 0.3	0.211
	V5Gy [%]	38 ± 4	38 ± 4	-0.2± ± 0.4	**0.041**
Right lung	Dmean [Gy]	0.82 ± 0.06	0.82 ± 0.06	0.001± ± 0.004	0.460
	V5Gy [%]	0.03 ± 0.04	0.02 ± 0.03	0.01± ± 0.01	**0.001**
Spinal canal	Dmax [Gy]	12 ± 3	12 ± 3	0.0± ± 0.4	0.570
Esophagus	D1cm3 [Gy]	8 ± 2	9 ± 2	-0.2± ± 0.5	0.088
Sternum	D1cm3 [Gy]	35 ± 4	34 ± 4	0.7± ± 0.7	**0.001**
Right breast	Dmean [Gy]	1.0 ± 0.2	1.0 ± 0.2	-0.03± ± 0.09	0.112
CTVp	D98 [Gy]	38.8 ± 0.1	38.8 ± 0.1	0.0± ± 0.1	0.460
	V95 [%]	99.8 ± 0.2	99.8 ± 0.1	0.0± ± 0.2	0.460
CTVn	D98 [Gy]	38 ± 1	39.08 ± 0.09	-1± ± 1	**0.001**
	V95 [%]	98 ± 2	99.998 ± 0.005	-2± ± 2	**0.001**
PTVpc	D98 [Gy]	37 ± 1	38.08 ± 0.04	-1± ± 1	**0.001**
	V95 [%]	97 ± 1	98.1 ± 0.2	-2± ± 1	**0.001**
PTVnc	D98 [Gy]	34 ± 2	38.09 ± 0.04	-4± ± 2	**0.001**
	V95 [%]	91 ± 3	98.2 ± 0.2	-7± ± 3	**0.001**

Table 5: Mean dosimetric metrics for the Siemens model. The treatment plans are based on manual delineations. SD is denoted as the ± value. P-values in bold font are considered statistically significant.

Region	Metric	Manual del.	Siemens model	Pairwise difference (Manual - model)	P-value
Heart	Dmean [Gy]	1.4 ± 0.2	1.5 ± 0.3	-0.1 ± 0.2	**0.001**
Left lung	Dmean [Gy]	8.6 ± 0.8	8.5 ± 0.8	0.01 ± 0.09	0.865
	V18Gy [%]	18 ± 2	18 ± 2	0.1 ± 0.3	0.496
	V5Gy [%]	38 ± 4	39 ± 4	-0.2 ± 0.3	**0.017**
Right lung	Dmean [Gy]	0.81 ± 0.06	0.81 ± 0.06	-0.001 ± 0.003	0.427
	V5Gy [%]	0.02 ± 0.04	0.01 ± 0.03	0.01 ± 0.01	**0.002**
Spinal canal	Dmax [Gy]	13 ± 3	12 ± 3	0.8 ± 0.4	**0.001**
Esophagus	D1cm3 [Gy]	10 ± 4	10 ± 4	-0.6 ± 0.8	**0.009**
Sternum	D1cm3 [Gy]	34 ± 4	35 ± 3	-2 ± 1	**0.001**
Right breast	Dmean [Gy]	1.0 ± 0.2	1.1 ± 0.3	-0.1 ± 0.2	**0.020**
CTVp	D98 [Gy]	38.8 ± 0.1	35 ± 5	4 ± 5	**0.001**
	V95 [%]	99.88 ± 0.07	96 ± 3	4 ± 3	**0.0007**

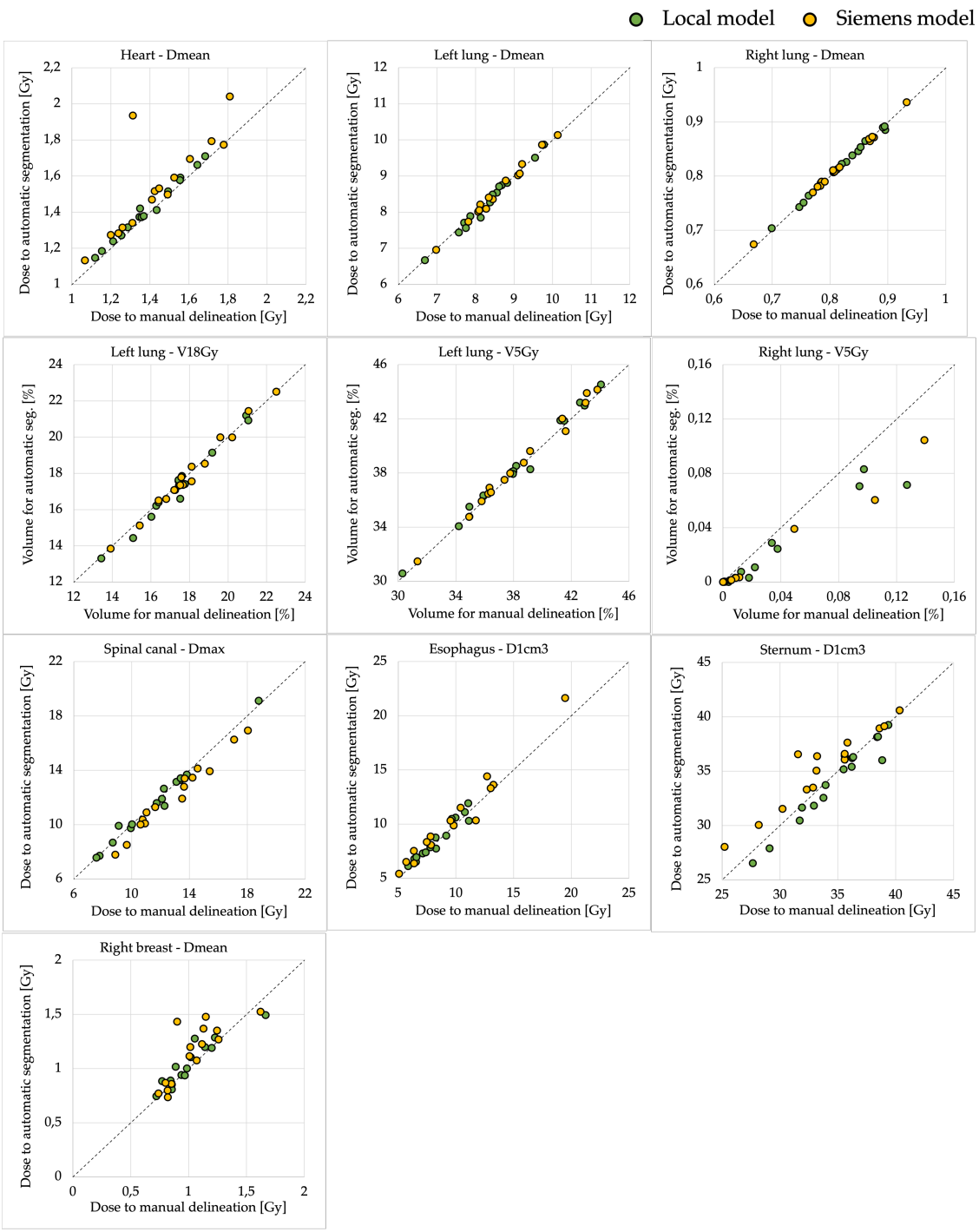


Figure 28: Dosimetric metrics plotted for all OARs. Dose metrics are plotted as dose to automatic segmentations against dose to manual delineations. Volume metrics are plotted as volume covered by the relevant isodose curve for the automatic segmentations against volume for the manual delineations. Dotted line represents equality.

The primary target volume coverage (V95) and near-minimum dose (D98) is plotted for the local model and Siemens model in figures 29 and 30, respectively. PTVpc was not made for the Siemens model, as no plans were made based on this model. Therefore the dose coverage to the PTVpc is not plotted for the Siemens model, as it is for the local model. Furthermore, the nodal target volume coverage and near-minimum dose is plotted for the local model in figure 31.

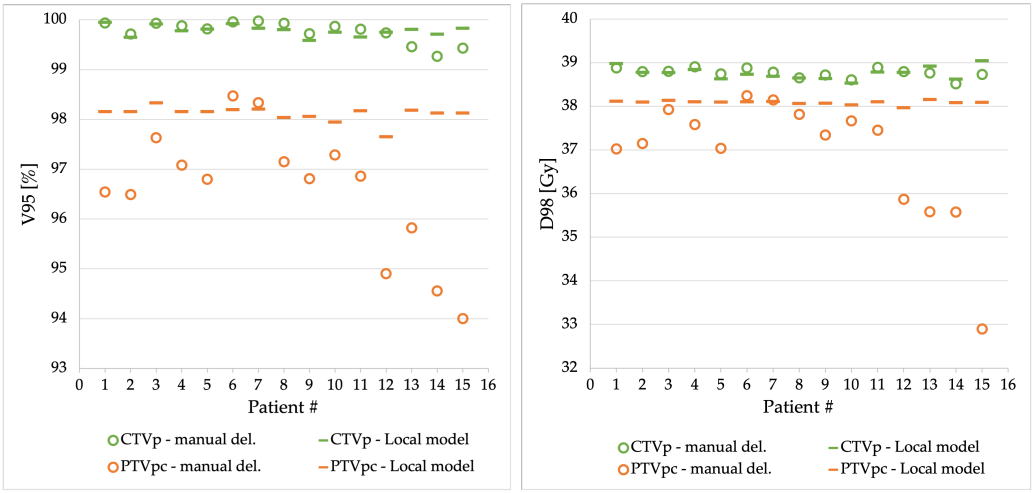


Figure 29: Primary target volume coverage (V95) and near-minimum dose (D98) to the manual delineations for plans based on automatic segmentations by the local model.

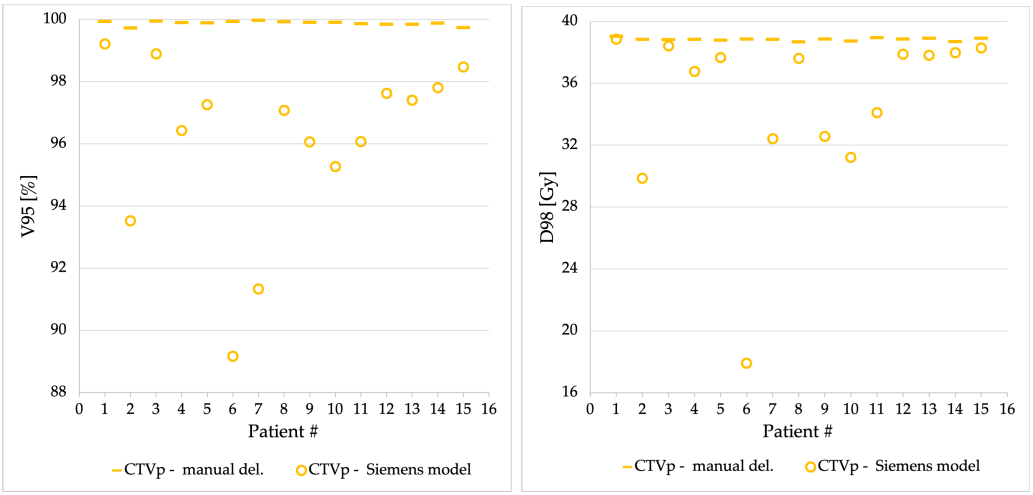


Figure 30: Primary CTV coverage (V95) and near-minimum dose (D98) to the automatic segmentations by the Siemens model for plans based on manual delineations.

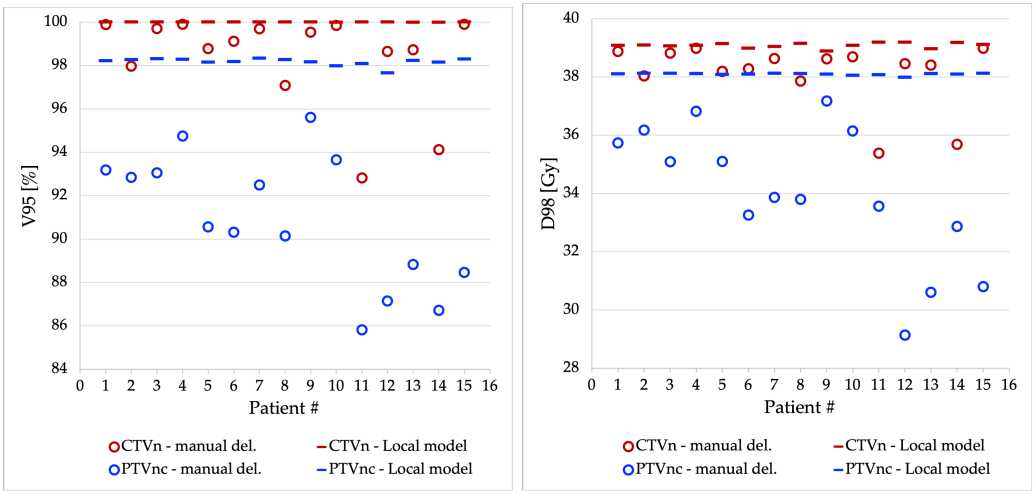


Figure 31: Nodal target volume coverage (V95) and near-minimum dose (D98) to the manual delineations for plans based on automatic segmentations by the local model.

4.1.3 Clinical evaluations

The clinical scoring of the Siemens model was done for the lungs, spinal canal, esophagus, and sternum, and the modified Turing test was done for the heart and left breast. The results are presented separately.

Clinical scoring

The results from the clinical scoring can be seen in figure 32. The scores are based on that the segmentations are to be used for radiation of breast cancer.

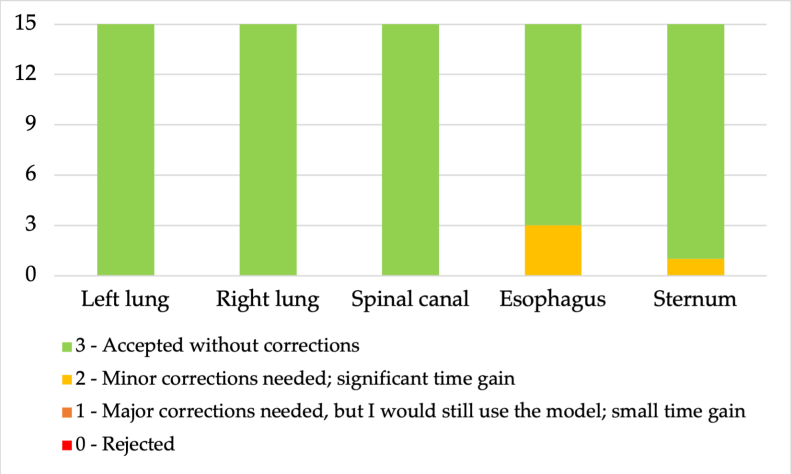


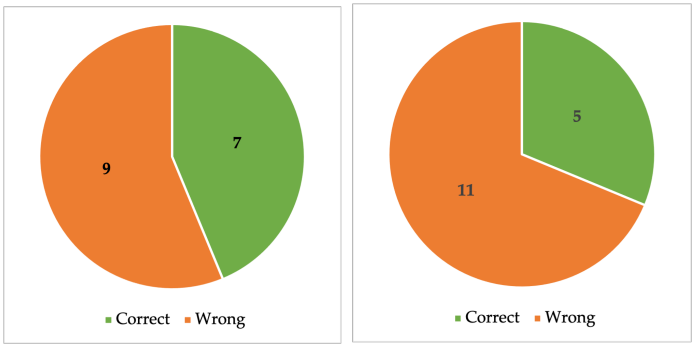
Figure 32: Results from the clinical scoring by the radiation therapist.

The evaluation came with some comments. These can be summarized as:

- Left lung - for six patients: "would do some minor editing around the bronchi, but not necessary for breast irradiation"
- Right lung: for all but one patient: "would do some minor editing around the bronchi, but not necessary for breast irradiation"
- Spinal canal - for two patients: "somewhat small in some areas" and for one patient: "a little strange caudal, but also strange anatomy of the patient"
- Esophagus - for 11 patients: "a little small in circumference", for 13 patients: "some lung and/or bone included", for five patients; "somewhat strange caudal or cranial", but still: "good enough for breast irradiation" for all patients
- Sternum - for all patients: "could have been segmented better, but good enough for breast irradiation."

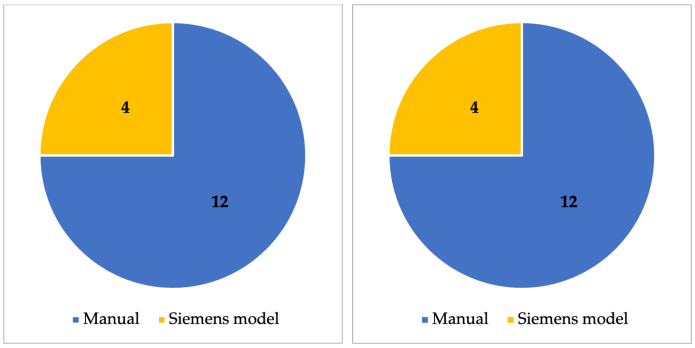
Modified Turing test

The results for the three main questions of the modified Turing test can be seen in figures 33, 34 and 35. The misclassification rate for the classical Turing question was 56 % for the heart and 69 % for the left breast. In 75 % of the slices shown, the manually delineated structure was preferable for both the heart and the left breast.



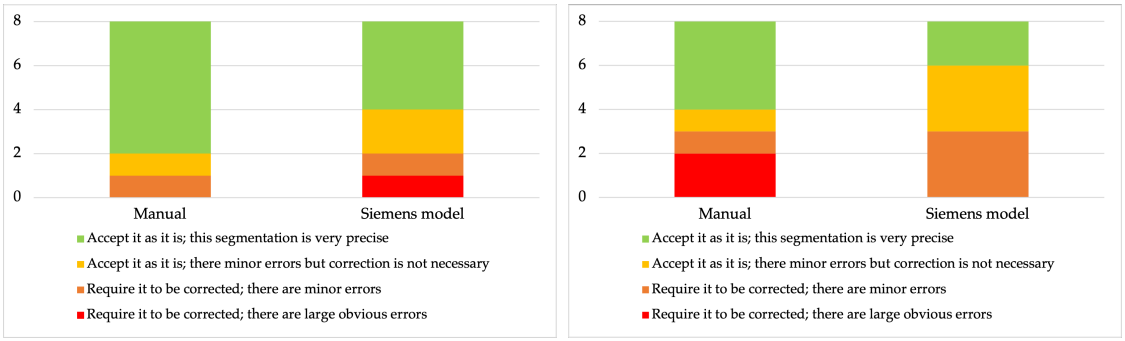
(a) Heart **(b)** Left breast

Figure 33: Results from the modified Turing test and the question "How was this segmentation drawn?".



(a) Heart **(b)** Left breast

Figure 34: Results from the modified Turing test and the question "Which segmentation do you prefer?".



(a) Heart **(b)** Left breast

Figure 35: Results from the modified Turing test and the question "You have been asked to quality assure this segmentation, would you...".

4.2 Automatic plan optimization

All mean dosimetric metrics, together with the mean pairwise difference and statistical significance can be found in tables 6 and 7, for comparison of automatic VMAT plans to clinical hybrid plans and clinical VMAT plans, respectively. For some of the pairwise differences and corresponding p-values, the difference was 0. These were left out of the Wilcoxon signed-rank test. This lead to some of the metrics having too few data points to calculate significance and p-value. This is indicated in the tables with a "-". Results for additional dosimetric parameters can be found in appendix E.

Table 6: Mean dosimetric metrics for the automatic VMAT plans and the clinical hybrid plans. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

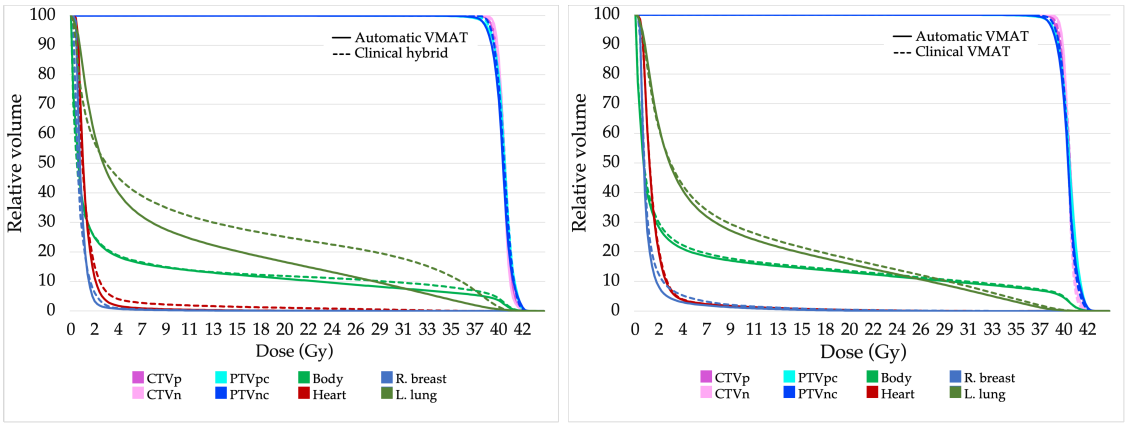
Region	Metric	Automatic VMAT	Clinical hybrid	Pairwise difference (Clinical - Automatic)	P-value
CTVp	D98 [Gy]	38.8 ± 0.1	38.7 ± 0.2	-0.1 ± 0.2	0.237
	V95 [%]	99.7 ± 0.1	99.5 ± 0.4	-0.2 ± 0.3	0.091
	V105 [%]	0.02 ± 0.04	0.1 ± 0.2	0.1 ± 0.2	-
CTVn	D98 [Gy]	39.1 ± 0.1	39.0 ± 0.3	-0.1 ± 0.3	0.499
	V95 [%]	99.998 ± 0.005	99.95 ± 0.09	-0.05 ± 0.08	-
	V105 [%]	0.1 ± 0.1	0.1 ± 0.3	0.0 ± 0.2	0.612
PTVpc	D98 [Gy]	38.10 ± 0.01	38.5 ± 0.3	0.4 ± 0.3	**0.028**
	V95 [%]	98.16 ± 0.06	99.1 ± 0.7	0.9 ± 0.7	**0.028**
	V105 [%]	0.1 ± 0.3	0.1 ± 0.1	-0.1 ± 0.4	1.000
PTVnc	D98 [Gy]	38.09 ± 0.01	38.7 ± 0.2	0.6 ± 0.2	**0.018**
	V95 [%]	98.17 ± 0.04	99.8 ± 0.2	1.6 ± 0.2	**0.018**
	V105 [%]	0.3 ± 0.2	0.1 ± 0.2	-0.1 ± 0.3	0.398
Body	Dmean [Gy]	5.1 ± 0.7	5.5 ± 0.8	0.3 ± 0.2	**0.018**
	V32Gy [%]	7 ± 1	9 ± 2	2.0 ± 0.7	**0.018**
	D2cm3 [Gy]	42.0 ± 0.2	42.0 ± 0.1	0.0 ± 0.2	0.499
Heart	Dmean [Gy]	1.4 ± 0.3	1.8 ± 0.4	0.3 ± 0.3	**0.018**
Right Breast	Dmean [Gy]	1.0 ± 0.3	0.9 ± 0.4	-0.1 ± 0.2	0.128
Left Lung	Dmean [Gy]	8 ± 1	11 ± 1	2.7 ± 0.9	**0.018**
	V18Gy [%]	18 ± 5	26 ± 4	8 ± 3	**0.018**
	V5Gy [%]	37 ± 4	43 ± 5	6 ± 3	**0.018**
Right Lung	Dmean [Gy]	0.74 ± 0.06	0.7 ± 0.1	-0.01 ± 0.06	0.735
	V5Gy [%]	0.02 ± 0.05	0.7 ± 0.5	0.7 ± 0.5	**0.018**
Esophagus	D1cm3 [Gy]	6 ± 2	25 ± 6	19 ± 6	**0.018**
Thyroid	D1cm3 [Gy]	27 ± 15	37 ± 4	10 ± 12	**0.018**
Trachea	D1cm3 [Gy]	19 ± 6	33 ± 4	14 ± 5	**0.018**
L. humeral head	D1cm3 [Gy]	33 ± 6	39 ± 2	6 ± 5	**0.018**
Spinal canal	Dmax [Gy]	14 ± 4	17 ± 5	3 ± 5	0.176

Table 7: Mean dosimetric metrics for the automatic VMAT plans and the clinical VMAT plans. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Automatic VMAT	Clinical VMAT	Pairwise difference (Clinical - Automatic)	P-value
CTVp	D98 [Gy]	38.7 ± 0.2	38.8 ± 0.2	0.1 ± 0.2	0.173
	V95 [%]	99.78 ± 0.08	99.7 ± 0.3	0.0 ± 0.3	0.953
	V105 [%]	0.03 ± 0.02	0.1 ± 0.1	0.1 ± 0.1	0.594
CTVn	D98 [Gy]	39.1 ± 0.2	39.2 ± 0.3	0.1 ± 0.3	0.314
	V95 [%]	100.0 ± 0.1	100.0 ± 0.1	-0.01 ± 0.02	-
	V105 [%]	0.08 ± 0.07	0.01 ± 0.02	-0.07 ± 0.06	**0.011**
PTVpc	D98 [Gy]	38.110 ± 0.009	38.1 ± 0.5	0.0 ± 0.5	0.678
	V95 [%]	98.19 ± 0.06	98 ± 1	0 ± 1	0.515
	V105 [%]	0.09 ± 0.08	0.1 ± 0.2	0.0 ± 0.1	0.767
PTVnc	D98 [Gy]	38.09 ± 0.01	38.5 ± 0.2	0.4 ± 0.2	**0.008**
	V95 [%]	98.17 ± 0.04	99 ± 1	1 ± 1	**0.008**
	V105 [%]	0.26 ± 0.17	0.1 ± 0.1	-0.2 ± 0.2	**0.028**
Body	Dmean [Gy]	6.0 ± 0.5	6.2 ± 0.5	0.3 ± 0.2	**0.015**
	V32Gy [%]	9 ± 1	9.5 ± 0.9	0.5 ± 0.3	**0.008**
	D2cm3 [Gy]	42.1 ± 0.1	42.0 ± 0.1	0.0 ± 0.1	0.515
Heart	Dmean [Gy]	1.8 ± 0.5	1.8 ± 0.6	0.0 ± 0.2	0.594
Right Breast	Dmean [Gy]	1.2 ± 0.4	1.5 ± 0.9	0.3 ± 0.5	0.086
Left Lung	Dmean [Gy]	8 ± 1	9 ± 1	0.5 ± 0.9	0.139
	V18Gy [%]	17 ± 4	19 ± 4	2 ± 3	0.086
	V5Gy [%]	38 ± 4	39 ± 4	2 ± 4	0.214
Right Lung	Dmean [Gy]	0.74 ± 0.09	0.9 ± 0.2	0.1 ± 0.1	**0.028**
	V5Gy [%]	0.001 ± 0.002	0.4 ± 0.4	0.4 ± 0.4	**0.018**
Esophagus	D1cm3 [Gy]	8 ± 6	19 ± 9	11 ± 4	**0.008**
Thyroid	D1cm3 [Gy]	28 ± 12	33 ± 10	6 ± 6	**0.008**
Trachea	D1cm3 [Gy]	19 ± 6	26 ± 5	7 ± 3	**0.008**
L. humeral head	D1cm3 [Gy]	32 ± 10	34 ± 8	2 ± 2	**0.011**
Spinal canal	Dmax [Gy]	13 ± 3	12 ± 3	0 ± 3	0.441

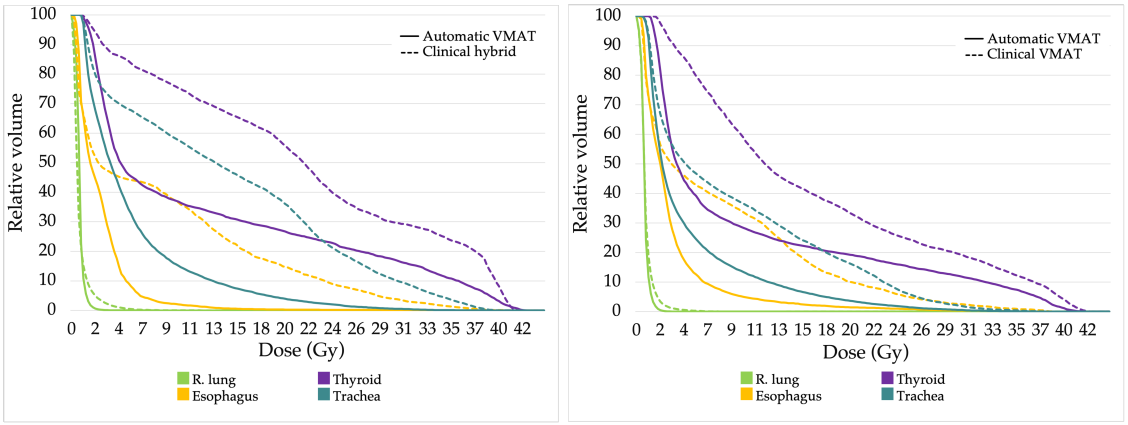
There were 15 statistically significant different metrics for the hybrid plan patients, where four of these were for the target volumes. For the VMAT plan patients, there were 12 statistically significant different metrics, where also four of them were for the target volumes. The statistically significant differences do not indicate the size and importance of the differences. The most important significant differences were found for the hybrid plan patients, specifically, for the heart and left lung.

DVH graphs for target volumes and the most critical OAR can be seen in figure 36. The less critical OAR can be seen in figure 37 and in appendix E. The target volumes can also be seen in figures 38 and 39 for primary and nodal structures, respectively. A solid line more towards the lower left corner for the OARs means that they are spared more using the automatic VMAT planning. A sharper descending from 100 % volume directly down to the prescribed dose of 40.05 Gy indicates better coverage and more homogeneous coverage of the target volume.



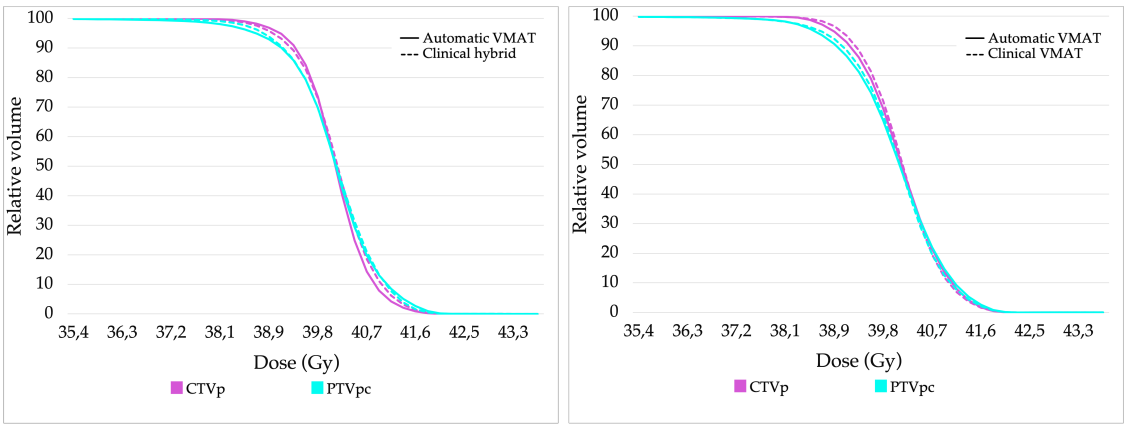
(a) Compared to hybrid plans **(b)** Compared to VMAT plans

Figure 36: Cumulative DVH for the target volumes and the most critical OAR.



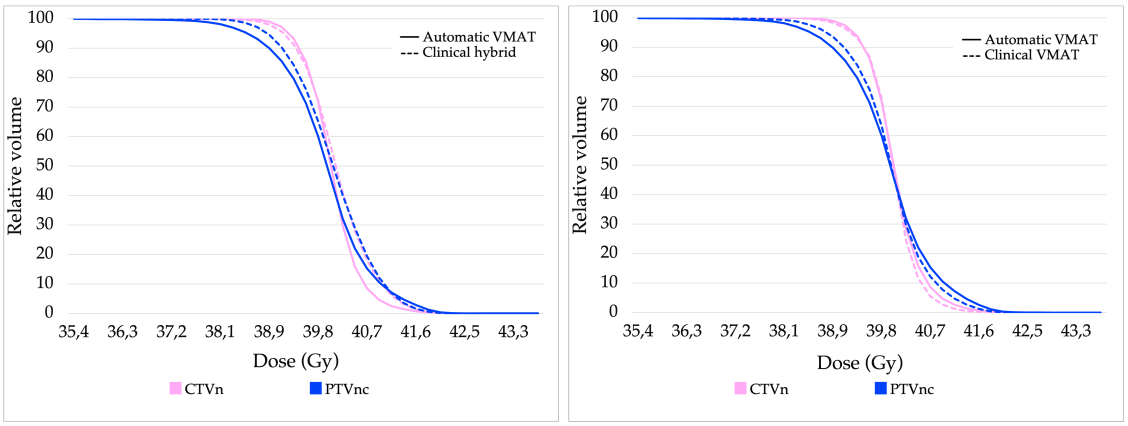
(a) Compared to hybrid plans **(b)** Compared to VMAT plans

Figure 37: Cumulative DVH for the less critical OAR



(a) Compared to hybrid plans **(b)** Compared to VMAT plans

Figure 38: Cumulative DVH for the primary target volumes.



(a) Compared to hybrid plans

(b) Compared to VMAT plans

Figure 39: Cumulative DVH for the nodal target volumes.

5 Discussion

In this section, the evaluations of the automatic segmentation models are discussed for each structure. This is followed by a discussion of the evaluation of the automatic plan optimization script. Methods used for quantitative evaluations are discussed, as well as methods used for qualitative evaluations. Finally, further work is discussed.

5.1 Automatic segmentation

To evaluate the performance of an automatic segmentation model, both quantitative and qualitative methods are recommended. However, there is no standard procedure recommended. For the quantitative analysis, different similarity metrics are usually used and may be combined with dosimetric metrics. The qualitative analysis is typically a clinical evaluation where one or more radiation oncologists compare the segmentations, and rate them as "pass" or "fail" or give them a score [3].

Manual delineation is subject to inter-observer variability, and therefore, there is no existing gold standard. Inter-observer variability was evaluated for the same patient group in the project thesis written prior to this master thesis [7]. The main results from the project thesis is included in appendix F. The calculated DSC and HD95 for the inter-observer variability of the heart, spinal canal, and esophagus can additionally be found in table 8. One should require a segmentation model to perform equal to or better than the inter-observer variability. Otherwise, the use of the automatic method will lead to reduced accuracy.

It is also of interest to compare with other similar studies. Dong et al. [57] compared their proposed method for automatic segmentation, using a type of U-Net, to the seven other methods that participated in the 2017 AAPM thoracic auto-segmentation challenges [58]. In comparison, they found that their model performed well based on three different geometric metrics. The DSCs from their model are presented in table 8. Zhu et al. [59] also evaluated a CNN model for the thorax-area and will be used for comparison for some OARs. Additionally, Simões et al. [60] evaluated an atlas-based automatic segmentation for the breast as a target volume with both DSC and HD95.

Table 8: Mean DSC and HD95 for the inter-observer variability (IOV) found in the project thesis [7] and the results presented by Dong et al. [57], together with the results from the local model and Siemens model. SD is denoted as the ± value. The best result for each organ is denoted by bold font.

Organ	Method	DSC	HD95 [cm]
Heart	IOV	**0.96 ± 0.01**	**0.39 ± 0.06**
	Dong et al.	0.87 ± 0.05	
	Local model	**0.96 ± 0.01**	0.5 ± 0.1
	Siemens model	0.91 ± 0.03	0.9 ± 0.4
Left lung	Dong et al.	0.97 ± 0.01	
	Local model	0.973 ± 0.008	0.35 ± 0.03
	Siemens model	**0.968 ± 0.005**	**0.34 ± 0.02**
Right lung	Dong et al.	0.97 ± 0.01	
	Local model	**0.979 ± 0.005**	0.33 ± 0.03
	Siemens model	0.975 ± 0.004	**0.32 ± 0.03**
Spinal canal	IOV	0.90 ± 0.02	0.21 ± 0.03
	Dong et al.	0.90 ± 0.04	
	Local model	**0.94 ± 0.02**	**0.17 ± 0.03**
	Siemens model	0.83 ± 0.03	0.22 ± 0.03
Esophagus	IOV	0.85 ± 0.03	0.25 ± 0.05
	Dong et al.	0.75 ± 0.08	
	Local model	**0.87 ± 0.02**	**0.25 ± 0.03**
	Siemens model	0.81 ± 0.04	0.29 ± 0.07

Dosimetric metrics are used to study the effect of geometric differences on the calculated dose and might therefore be considered as more clinically meaningful. Ideally, the dosimetric evaluation of the model should be done on plans based on the automatic segmentations. This was done for the local model. However, not all structures were available in the Siemens model, and making plans was not possible without the nodal CTV. Nevertheless, the goal for the dosimetric evaluation is still the same.

Dosimetric evaluations were done by Dong et al. [57] and Zhu et al. [25], but for lung cancer and esophageal cancer, respectively. As a dosimetric evaluation is highly affected by the diagnosis, these results are not used for comparison. Simões et al. [60] also did a dosimetric evaluation for breast cancer radiotherapy. However, the method was somewhat different, and this dosimetric evaluation is also not relevant.

5.1.1 Heart

The local model was significantly better than the Siemens model based on geometric metrics for the heart. The performance of the local model was similar to the inter-observer variability; same DSC, slightly worse HD95. Compared to the proposed model by Dong et al. [57], both the local model and the Siemens model perform better, when looking at DSC. However, for the heart, some of the models that competed in the auto-segmentation challenge in 2017 have a higher DSC than the model proposed by Dong et al.. The highest DSC in the challenge for the heart was 0.93. Zhu et al. [59] also found DSC for the heart to be 0.93 ± 0.04. The local model has a higher DSC than all models used for comparison.

The Siemens model has one outlier for the geometric metrics. For this patient, the Siemens model has segmented further in the anterior direction than both the manual delineation and the local model. This can be seen in figure 40. However, this patient has somewhat irregular anatomy, in that there is soft tissue between the heart and the thoracic wall. This patient was also used in the test dataset for the previous version of the local model [7] and then amounted to an outlier for the local model. It can be noticed that the present version of the local model did not have much trouble segmenting this patient. Generally, for the heart, the Siemens model over-segmented and the local model under-segmented based on absolute volume, both with statistical significance.

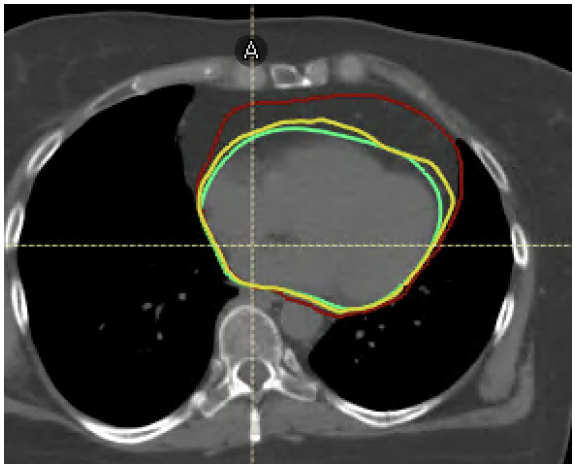


Figure 40: For the heart, the Siemens model has segmented too far in the anterior direction for one patient. Siemens model segmentation in red, local model segmentations in yellow, and manual delineation in green. Transversal plane.

Both models estimated statistically significant different mean heart dose than the manual delineations. Nevertheless, the differences are small as shown in figure 28. Only two of the differences might be considered clinically significant. Both these differences are found for the Siemens model. One of these patients is the same as for the geometric outlier. While the other is for a segmentation with average geometric metrics, but slightly over-segmented in the direction of the target volumes leading to an over-estimated dose. Even though these doses are not very high, one of them is

estimated above 2 Gy for the automatic segmentation. If the plan was based on the Siemens model segmentation of the heart, there might be trouble getting the dose below 2 Gy and one might have to compromise with target volume coverage.

In 2011, Feng et al. [61] developed a new delineation standard for the heart and reported the dosimetric differences to the heart before and after implementing the new standard. This can be seen as the inter-observer variability for dosimetric differences to the heart, as St. Olavs Hospital uses this standard to delineate the heart. They found that the mean difference in mean dose to the heart was (0.14 ± 0.14) Gy. Both models evaluated in this thesis have a lower mean dose difference than this.

Despite promising quantitative results, it should be mentioned that for some of the structures the manual delineations will be adjusted before the final local model is trained. This is because of some differences in guidelines between St. Olavs Hospital and Ålesund Hospital, which has also provided data for training the model. The heart is one of these structures and this is therefore not the final version for this organ.

The modified Turing test was performed for the Siemens model. The results have large uncertainties, due to few questions and few participants. Nevertheless, the results indicate that the oncologists found it difficult to determine which segmentation was made by model and which was made manually. These questions were differently answered by the two oncologists and therefore appear randomly answered. The misclassification rate larger than 50 % indicates that the model had succeeded in not being detected as a machine. However, there was agreement in that the oncologists preferred the manual delineation, but in 25 % of the cases, they preferred the model segmentations. There was generally large agreement on the quality assurance questions, but with some differences. For three of the four slices with an automatic segmentation presented, both oncologists agreed that it could be accepted. They also agreed that three of the manual delineations presented were acceptable, but disagreed on the fourth slice. Overall, these results are promising for the Siemens model.

5.1.2 Lungs

The models performed similarly on the geometric and dosimetric evaluation of the lungs. The local inter-observer variability is not available/quantified for the lungs. Compared to other studies, the local model and Siemens model performed similarly to the models proposed by Dong et al. [57] and Zhu et al. [59], based on similar DSC. The lungs are relatively large organs and will therefore generally get high DSC.

Despite statistically significant differences in V5Gy, the differences in all dosimetric parameters are small. According to Norwegian Breast Cancer Group [45], the volume receiving 18 Gy or more should be less than 35 % of the lung. None of the models are close to this threshold, as the largest V18Gy value is 22 % for the Siemens model. Therefore, these small differences in dose, will likely not affect treatment planning.

Clinical scoring was done for the lungs segmented by the Siemens model. The radiation therapist gave both lungs a top score of "3 - Accepted without corrections" for all 15 patients. There were, however, comments included that she would have done some minor editing around the bronchi, but that this is not necessary for breast irradiation. Overall, the dosimetric impact of replacing the current method for delineating the lungs with any of these two models is small. Also, considering the good score of the lungs segmented by the Siemens model, one can likely assume the lungs segmented by the local model are also clinically acceptable.

5.1.3 Spinal canal

The local model was statistically significantly better than the Siemens model on DSC and HD95 for the spinal canal. As seen from table 8, the local model was also better than the inter-observer variability and the model proposed by Dong et al. [57]. The Siemens model received a similar DSC

to the one found by Zhu et al. [59], which was 0.84 ± 0.04.

There was no statistically significant difference in volume for the local model, while the Siemens model systematically under-segmented compared to manual delineation. As both models were adjusted to the length of the manually delineated spinal canal in the cranial to caudal direction, these differences were in the transversal plane. This was also commented on by the radiation therapist doing the clinical scoring of the Siemens model.

Generally, for the Siemens model, lower doses are estimated to the automatic segmentations than to the manual delineations. This is probably because of under-segmentation. Even though the doses given to the spinal canal during breast irradiation are relatively low, using the Siemens model would probably compromise the accuracy of the manual method used today, at least compared to using the local model. The clinical scoring indicates that the Siemens model segmentations are good enough for clinical use for radiotherapy of breast cancer, despite somewhat small segmentation in some areas leading to a difference in maximum dose. However, the local model is probably preferable as these segmentations seem to be of the same quality as the manual delineations.

5.1.4 Esophagus

Similar to the results for the spinal canal, the geometric metrics were statistically significantly better for the local model than the Siemens model for the esophagus. The local model was also better than the inter-observer variability and the model proposed by Dong et al. [57]. Unlike for the spinal canal, the Siemens model also had a better DSC than the model proposed by Dong et al. and is more similar to the inter-observer variability.

For both models, the dose to the automatic segmentations was larger than to the manual delineations. No statistical significance was found for the local model, but the differences for the Siemens model segmentations, were larger and statistically significant. However, looking at figure 28, one can see that the differences are still small. Additionally, when making a treatment plan for breast cancer, the dose given to the esophagus is well below the dose limits of the esophagus. Therefore, these differences will likely not lead to compromise with target volume coverage.

Among the structures evaluated qualitatively, the esophagus received the lowest clinical score, but still got a top score for 12 of the patients. Three patients received a score of "2 - Minor corrections needed; significant time gain". Almost all the segmentations were with the comment "good enough for breast irradiation, but slightly small circumference and includes some bone and/or lung". Due to higher DSC and smaller HD95 for the breast model, there is reason to believe that the problems observed by the radiation therapist for the Siemens model are smaller for the local model.

5.1.5 Sternum

For the sternum, the local model was significantly better on DSC, while the Siemens model had a significantly lower HD95. It is therefore unclear which model performs better for this structure. They both perform similarly to an atlas-based automatic segmentation model for bone structures, evaluated by Fu et al. [62]. They found the mean DSC to be 0.89 ± 0.02 for the sternum. The sternum is one of the organs that will be retrained after adjustments are made for the final version of the local model. Hence, there is reason to believe that the local model will outperform the Siemens model.

Three outliers can be seen for the geometric metrics found for the local model, and one can be seen for the Siemens model. For all these outliers the model has not segmented far enough in the caudal direction, as seen in figure 41.

Both models have statistically significant differences for D1 cm^3, but the differences are relatively small compared to the dose estimated to the sternum. Additionally, this structure is mostly used as a matching structure during treatment, and according to the clinical scoring, the Siemens model has segmented good enough to use for breast cancer radiotherapy, even though it could have been done better. 14 of 15 segmentations received a top score, despite some errors in the caudal part, as

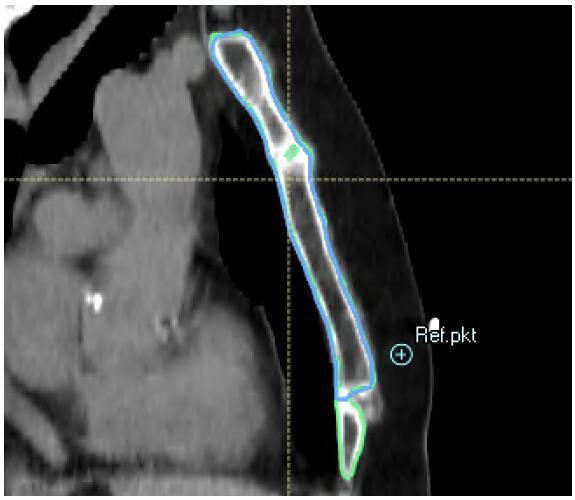


Figure 41: The local model has not segmented far enough in the caudal direction for this patients sternum. Local model segmentation in blue and manual delineation in green. Saggital plane.

seen for the geometric outlier. One segmentation received a score of "2 - Minor corrections needed; significant time gain" because the segmentation was not connected in the caudal part.

5.1.6 Right breast

The local model was significantly better than the Siemens model for the right breast. Both models over-segmented, but as one can see from figure 27, the Siemens model generally segmented larger volumes than the local model. The HD95 for the Siemens model is also the largest mean value found for any of the organs. This despite the segmentations from the Siemens model has been cropped 5 mm below the body surface.

Over-segmentation is probably the reason for the slightly higher mean dose estimated to the automatically segmented structures. For the local model, the mean dose difference is not statistically significant, while the differences are slightly larger and statistically significant for the Siemens model. The patient case with the largest relative differences can be seen in figure 42.

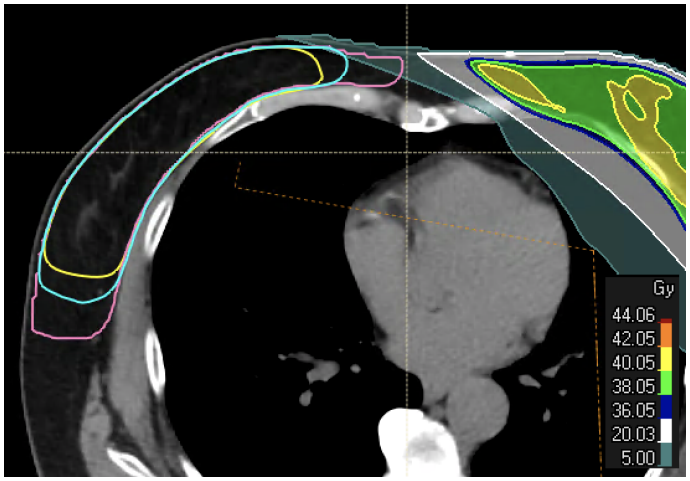


Figure 42: Both models over-segmented the right breast for this patient. Shown here together with the dose plan for the manual delineations. Manual delineation in yellow, local model segmentation in blue and Siemens model segmentation in pink. Transversal plane.

For the local model, the delineations used for the training were done for the right breast as an

OAR and the left breast as a target volume. Generally, target volumes require a delineation with higher precision than OARs. This is because the radiation beam is shaped conformally around the target volume. A small error will have a large effect on the actual coverage to the target.

5.1.7 Left breast (CTVp)

The local model was significantly better than the Siemens model for the left breast, as well. Local inter-observer variability is not available for the left breast. However, in 2013, the Danish Breast Cancer Cooperative Group developed national guidelines for delineation of target volumes for adjuvant radiotherapy of early breast cancer, including both primary target volume and nodal target volumes [63]. After the consensus, the DSC was found to be 0.95, ranging from 0.93 to 0.96, for the breast. Simões et al. [60] also found the median DSC to be 0.95. The local model has a similar DSC to these, while the Siemens model has a lower DSC. Simões et al. calculated the HD95, as well, and found the median value to be 0.97 cm. A larger HD95 than for the local model but smaller than for the Siemens model.

The dosimetric effect of the geometric differences for the left breast is seen by looking at the coverage to the CTVp and PTVpc. It should be mentioned that breast radiotherapy is in these cases adjuvant treatment. When irradiating the breast the GTV has already been removed and it might not always be necessary that the entire breast is irradiated.

For the local model, the coverage and near-minimum dose to the CTVp is approximately the same for the automatic segmentations and the manual delineations, as seen in figure 29. To the PTVpc, however, the differences are enhanced and also statistically significant. The differences might be significant for those patients that receive less than 96 % coverage to the manual delineations. For the local model, this is four of 15 patients. Probably due to larger differences in volume, the dosimetric metrics for the Siemens model are not as good as for the local model. For the Siemens model, the differences in the dosimetric metrics are statistically significant. Compared to no patients receiving lower CTVp coverage than 99 % for the local model, all but one patient receive lower coverage than 99 % for the Siemens model. If the plans were based on the segmentations by the Siemens model instead of manual delineations, these dosimetric differences would likely lead to over-treatment of the manual delineations, i.e., ground truth primary target volumes.

The results from the modified Turing test for the Siemens model indicate that the oncologists found it difficult to determine which segmentation was made by model and which was made manually. The misclassification rate was 69% and the oncologists did not generally answer the same. The preference was manual delineations in 75 % of the slices shown, but the oncologists did not agree for all slices. For the quality assurance question, manual and automatic segmentations were accepted equally. One of the manual delineations of the breast received the lowest score by both oncologists. The manual segmentations in this test were used clinically and do not have the same high quality as the COBRA dataset. Therefore, this shows that errors also occur in the "gold standard" manual delineation and sheds light on the issue of not having actual ground truth delineations when evaluating automatic segmentations.

Based on larger dosimetric differences and lower DSC than the one found after consensus by Nielsen et al. [63], the Siemens model does not segment the breast accurately enough to be used as a target volume. The local model seems more promising for left breast segmentation.

5.1.8 Lymph nodes (CTVn)

Lymph node segmentation is not available in the Siemens model, so the discussed results in this subsection are all for the local model. The geometric results for the union of the lymph node areas are the worst of all the structures evaluated. However, the manual delineations will also be adjusted and retrained for the final version of the model. Therefore, improvement is expected for the final version.

The inter-observer variability found by the Danish Breast Cancer Cooperative Group were a DSC

of 0.70 (0.60–0.77), 0.76 (0.67–0.84), 0.74 (0.66–0.82), 0.56 (0.43–0.73) and 0.66 (0.55–0.78) for axillary lymph nodes levels 1-4, and pectoral axillary lymph nodes, respectively [63]. The geometric metrics for the individual lymph node areas for the breast model can be found in appendix D. For the axillary lymph nodes levels 1, 3, and 4 the local model has higher or about the same DSC, while level 2 and pectoral axillary lymph nodes have lower DSC.

The dosimetric differences were statistically significant and might also be clinically significant. Due to under-segmentation by the model, the coverage to the manual delineations is most likely not good enough. None of the patients received a higher coverage to the PTVnc than 96 %, but hopefully, the adjustments to the model will give better results. A clinical evaluation of the segmentations would also be interesting and is required before approving the model.

5.2 Automatic plan optimization

The purpose of evaluating the script for automatic plan optimization was twofold: 1) it is used to make standardized plans for the dosimetric evaluation of the segmentation models and 2) to validate the script for clinical use.

For the hybrid plan patients, the coverage and near-minimum dose to the PTVs are somewhat better for the clinical plans than the automatic VMAT plans with a statistical significance. For the VMAT plan patients, there were statistically significant differences for the nodal PTV and in V105 for the nodal CTV and otherwise minor differences to the target volumes. Whether the statistically significant differences in target volume coverage are clinically significant has not been evaluated. However, a PTV coverage of 98 % is generally considered good enough and the script managed this for all the patients. The coverage can also be adjusted in the script but was chosen to 98 %. A larger coverage might lead to a larger dose to the heart and left lung than necessary. It should also be mentioned that the clinical plans were made with somewhat varying PTV margins and in some of the cases, the clinical plans had a 2 mm extra margin. The plans are evaluated on the PTVs made by the script and this is probably why the clinical plans seem to have a higher coverage in this evaluation. The maximum dose volume, i.e., V105, was slightly high for the automatic VMAT plans. Subsequently to this evaluation, the script has been adjusted so that the maximum dose has been reduced. It is also important to notice that the SD is generally lower for the automatic plans than the clinical plans. This indicates that the plans are more standardized and one gets the same quality plan for each patient.

A general concern for VMAT planning is that there is a larger amount of low-dose spill than for conventional planning. Looking at the dose effects on the body, however, one can see that the dose to the body is slightly lower for the automatic VMAT plans than for both the clinical hybrid and clinical VMAT plans. This shows that a carefully made VMAT plan does not need to have a higher mean dose to the whole body than a hybrid (more conventional) plan.

For the hybrid plan patients, larger dose differences were found for the heart and left lung. They are likely to be clinically significant as well as statistically significant. An average reduction of 0.3 Gy was found for the mean dose to the heart when using the script. For the left lung, this reduction was 2.7 Gy. Compared to the clinical VMAT plans, the dose difference to the heart and left lung did not reach statistical significance. There is more to gain when using the script for the hybrid plan patients. This is probably because VMAT planning is chosen for the patients where there is trouble with getting low enough heart dose and in some cases low enough left lung dose. If the heart and lung doses are considered low enough with the hybrid plan, this is chosen.

There was some dose reduction to the right lung, esophagus, thyroid, trachea, and left humeral head using the script compared to both the clinical plan types. Although these reductions are smaller, the ALARA principle is important to have in mind, and in that case, any reduction of dose to OARs is good. The remaining OARs, the right breast and spinal canal, did not have large differences and the differences were not statistically significant.

Automatic planning is a hot topic in radiotherapy, but there is still minimal relevant literature to compare with for script-based planning. RapidPlan, Varian's knowledge-based solution for automatic planning, was evaluated by Dumane et al. [64] for lung cancer, and Ling et al. [65]

evaluated a hybrid of RapidPlan and a script-based solution for esophageal cancer. Both found a reduction in treatment planning time, reduction of dose to critical OARs, and the same or increased treatment quality for target volumes, similar to what was found for the plan optimization script in this thesis. The script uses about 30 minutes to plan, half the time compared to the solution presented by Ling et al. [65].

Overall, the script performs well in making VMAT plans for left-sided breast cancer patients who are to be treated with locoregional radiotherapy. The target volume coverage is about the same as for hybrid plans and manually made VMAT plans. The dose to the OARs is reduced compared to both the hybrid plans and the VMAT plans, although a larger reduction is found compared to the hybrid plans. The script was good enough to make plans for dosimetric evaluation of the segmentation models, without any manual editing, and should also be good enough to be implemented in the clinic. In fact, it was implemented for clinical use at the end of May 2021, although this was a slightly modified version.

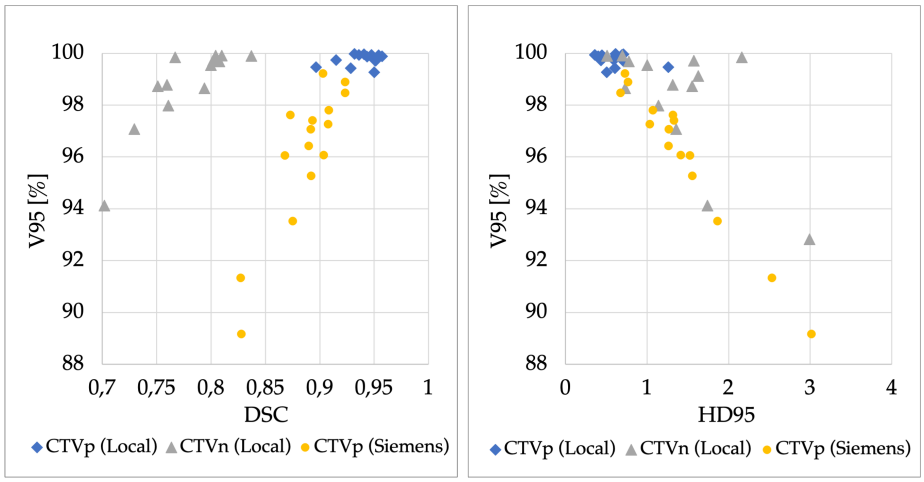
5.3 Metrics used for quantitative evaluation

DSC is one of the most used metrics for comparing segmentations, but it does have one well-known drawback. Larger volumes generally get higher DSC than smaller volumes. For this reason, HD is also included in the geometric analysis. Reinke et al. [66] illustrated important limitations of common image analysis metrics, with a focus on DSC and HD (HD100). The main limitations of DSC relevant to image segmentation are that DSC is sensitive to size, favors over-segmentation, and is unaware of shape. While, on the other hand, HD100 is sensitive to noise. Because of these limitations, one should include more than one metric when doing image analysis, but the metrics should complement each other [66]. For this thesis, DSC was chosen because it is widely used and the results can therefore more easily be compared to other research. HD95 was also calculated as a supplement to DSC, and chosen because it is more robust and therefore probably more clinically relevant than HD100. Additionally, the comparison of the basic metric volume allows knowledge about over- and under-segmentation. However, the volume says nothing about the spatial correlation between the segmentations and can not be used by itself either.

DSC and HD are good for geometric similarity but are not always correlated with clinical acceptability or time needed to adjust them [67]. Vaassen et al. [67] did an evaluation of measures for assessing time-saving of automatic OAR segmentation. They found that two new measures, surface DSC and added path length, are better at indicating time-saving and adjustments needed than the classical geometric measures. Surface DSC was introduced by Nikolov et al. [68] in 2018, and is a measure of overlap between the segmentation surfaces rather than the segmentation volumes, as in the classical volumetric DSC. As these metrics are new and not common yet, there is little to compare with, but it could still be interesting to test these metrics in combination with a clinical evaluation.

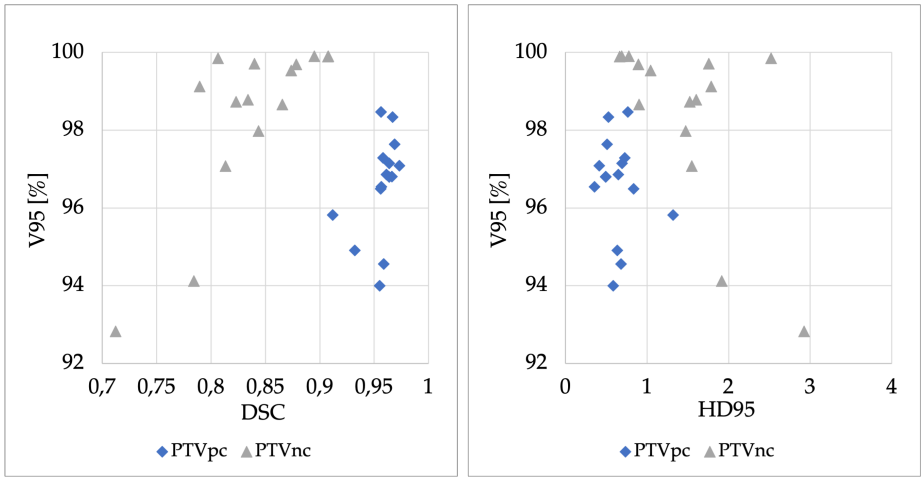
A dosimetric evaluation is more clinically relevant than using just geometric metrics for evaluating segmentations for radiotherapy. Simões et al. [60] writes that "any clinically meaningful evaluation of auto-contouring performance should include a dosimetric assessment of geometrical differences". Dosimetric evaluation has become more common the recent years, but only a geometric evaluation is still mostly used. This is probably due to the geometric metrics being easier to extract. It would be interesting to find a relationship between the geometric and dosimetric metrics, to know which is more clinically relevant. The relationship between the dosimetric metric V95 and the geometric metrics DSC and HD95 can be seen plotted based on data from this thesis in figures 43 and 44 for the CTVs and PTVs, respectively. For the local model, V95 to the manual delineations is used as the plans were based on the automatic segmentations. While for the Siemens model, V95 to the automatic segmentations were used as the plans were based on the manual delineations.

Spearman's rank correlation coefficient was calculated to find relationships between V95 and these geometric metrics. The coefficient should be found for CTV and PTV separately, as these should have different coverage. Knowing DSC's strong correlation with volume, the correlation for primary and nodal target volumes should be found separately. The correlation coefficients for the CTVps are calculated together for both models. The correlation coefficient for DSC with V95 was 0.82 and



(a) V95 plotted against DSC **(b)** V95 plotted against HD95

Figure 43: V95 for the manual delineations plotted against DSC and HD95 for CTVp and CTVn segmented by the local model, and V95 for the automatic segmentations plotted against DSC and HD95 for the CTVp segmented by the Siemens model.



(a) V95 plotted against DSC **(b)** V95 plotted against HD95

Figure 44: V95 for the manual delineations plotted against DSC and HD95 for PTVpc and PTVnc for the local model.

0.81 for the CTVp and CTVn, respectively, indicating a strong correlation. While the correlation coefficient for HD95 with V95 was -0.84 and -0.55 for CTVp and CTVn, respectively, indicating strong and moderate correlation. For the PTVs (only local model), on the other hand, it is more difficult to see a relationship. This is probably because the clearest relationship for the CTVs comes from the Siemens model, as seen in figure 43. The correlation between DSC and V95 was 0.59 and 0.70 for PTVPc and PTVnc, respectively. There was no correlation between HD95 and V95 for the PTVpc, but for the PTVnc this coefficient was -0.56.

An HD95 value lower than 1 cm, seems to be a good indication for a coverage higher than 98 % for the primary CTVs. HD95 might have a more robust correlation with V95, but both geometric metrics correlate with coverage and can therefore be seen as useful for evaluating the coverage of the target volumes.

5.4 Methods used for qualitative evaluation

An automatic method cannot be implemented clinically without a proper qualitative evaluation, since quantitative methods are not necessarily connected directly to clinical acceptability.

Huyskens et al. [69] conducted a clinical evaluation similar to the clinical scoring in this project. This same scoring system was also used by Schreier et al. [70], but in a blinded manner where the observers did not know if it was a manual delineation or automatic segmentation. As seen in the modified Turing test done for the Siemens model, the manual delineations do not necessarily receive a top score. It might therefore be a good idea to blindly score the segmentations, to get a reference score for the manual delineations. This is, however, a more time-consuming process and was not done for the clinical scoring of the Siemens model.

For the clinical scoring, the radiation therapist should, ideally, not have access to the manual delineations, as done by Huyskens et al. [69] and Schreier et al. [70]. This could perhaps lessen the need for a manual delineation reference score, as the observer is less biased by not having the "ground truth" delineation in view. Especially given that the manual delineations were the ones of high quality in the COBRA dataset. However, having received an average score of the top score, this did not seem to be a problem when evaluating the OARs to use for breast irradiation only.

Due to clinical evaluations being time-consuming and requiring experienced resources, the time-saving aspect was not evaluated directly. Instead, the time-saving aspect was integrated into the grading scheme for the scores and might therefore be more subjective than actually measuring the time used to edit the automatic segmentations. Nevertheless, this was a very time-effective method to evaluate the time-saving aspect of automatic segmentation.

Ideally, more than one radiation therapist or physician should do the evaluation, to avoid personal bias, and this could be considered for the next evaluation. Perhaps it would be better to divide the test patients between different observers or reduce the number of patients, rather than reduce the number of observers. Huyskens et al. [69] used six different physicians, and Schreier et al. [70] used 10 observers from three different clinics.

The modified Turing test is a less comprehensive clinical evaluation and was performed as method-testing for St. Olavs Hospital. The time used to answer this test was 10 minutes and 17 minutes by the two oncologists for three organs, i.e., heart, left breast, and prostate, while for the clinical scoring the time was 2-3 hours for 5 organs. Seeing how little time the modified Turing test took, perhaps one could include more observers and more images for each organ, thereby including a larger variety of patients and organ slices. With more observers, it would also be interesting to use Cohen's kappa to evaluate the inter-rater reliability [71].

An observation made when making the modified Turing test was that the automatic segmentations in general were a bit more pixeled than the manual delineations. However, this was not commented by the oncologists but could have been a problem for the Turing question and general bias for the other two questions.

Comments from the oncologists on the modified Turing test were that "there should be an alternative for uncertain" and that "it is difficult to evaluate a single slice for whether the segmentation is correct or not - especially for the prostate". If the observer is uncertain about which segmentation they prefer or which is made by model or human, this will show in the results whether there is an alternative for uncertain or not. Not having this "uncertain" alternative makes the observer choose anyway, and they might have an opinion on more of the segmentations than if there was an alternative option. When it comes to the difficulty of evaluating an organ in single slices, this was not as big a problem for the organs included in this thesis, but it could be considered to include all three directional views to make this easier.

The modified Turing test is more efficient and also includes different types of information than clinical scoring. The two first questions are interesting for comparison of the manual delineations and the automatic segmentations. While the quality assurance question is approximately the same as the clinical scoring and is a necessary question to know if the segmentations are good enough. The modified Turing test also gives a reference score for the manual delineations, like the blinded

scoring done by Schreier et al. [70]. A classical clinical scoring is perhaps the safest choice for a qualitative evaluation, but there are also benefits of using the modified Turing test instead. It may for example make the physicians more aware of the uncertainties that exist in manual delineation as well.

5.5 Further work

The work with automatic segmentation will continue at St. Olavs. The main focus will be on the local model, which is the most promising model. The final version of the local model is now being trained after adjustments, and when it is done it will be evaluated for all structures. The inter-observer variability will also be quantified for all structures and based on more physicians than included in the project thesis [7]. This will be used to evaluate the effect in standardization by also quantifying the inter-observer variability after implementing the model for clinical use.

The evaluation of resource usage would also be interesting. This could be done by measuring the time it takes to manually delineate versus to segment with the model and any manual editing necessary, This would give a specific result for the time-saving aspect of using segmentation models. The models might also need retraining with updated guidelines, and this should also be taken into consideration.

Further evaluations of the local model that could be interesting include whether the model can be used for other diagnoses than breast cancer and if the model can be validated at another clinic with their guidelines. For example, the model could be tested on OARs for lung cancer or esophageal cancer. Validating a segmentation model at another clinic can be seen as the ultimate seal of approval, and therefore cooperation with another clinic could be interesting.

The script for automatic plan optimization has, as mentioned, already been implemented in the clinic. However, it will be used parallel to manual planning for a period of time. This is to improve both the manual planning and the script. Eventually, one believes that mainly the script will be used. Cooperation with RaySearch to make AI-based automatic treatment planning based on the script is also under consideration. It would then be interesting to compare the script-based treatment planning with the AI model and see which performs best and which is fastest. Through cooperation with RaySearch, other centers will also have access to automatic VMAT-planning for breast cancer, without the need for implementing and maintaining advanced scripting.

6 Conclusion

One locally trained and one pre-trained DL segmentation model for OARs and target volume(s) have been evaluated. The quantitative evaluations for breast cancer radiotherapy are promising, especially for the local model. The local model segmentations of the heart, spinal canal, esophagus, and both breasts have significantly better correspondence with the manual delineations than the pre-trained Siemens model. These organs are also on the level of local or external inter-observer variability for the local model. For the Siemens model, larger differences were seen for the breasts. Furthermore, the lymph node areas are not available in the Siemens model. The local model is, therefore, the preferred model for the target volumes and will likely be good enough for clinical use when the final version is trained. The lungs and sternum had similar results for both models. The dosimetric differences and clinical scoring of most OARs indicate that the quality of both models is adequate for breast radiotherapy. However, in some cases, manual adjustments might be required, especially when using the Siemens model. This emphasizes the importance of automatic segmentations being quality assured by competent personnel before use.

A script for automatic VMAT planning has been dosimetrically evaluated and compared to clinical hybrid and clinical VMAT plans. The evaluation indicates that the target volume coverage and treatment quality are preserved when using automatic planning. OAR doses were generally reduced with the automatic VMAT plans. Compared to the hybrid plans, large dose reductions were found for the heart and left lung, which are assumed to be clinically significant. Automatic planning is now being implemented in the clinic and will probably improve the standardization, efficiency, and quality of treatment planning.

Bibliography

[1] Serena Gianfaldoni, Roberto Gianfaldoni, Uwe Wollina, Jacopo Lotti, Georgi Tchernev, and Torello Lotti. An overview on radiotherapy: from its history to its current applications in dermatology. *Open access Macedonian journal of medical sciences*, 5(4):521, 2017.

[2] Barbara Segedin and Primoz Petric. Uncertainties in target volume delineation in radiotherapy–are they relevant and what can we do about them? *Radiology and oncology*, 50 (3):254–262, 2016.

[3] Vandewinckele Liesbeth, Claessens Michaël, M Dinkla Anna, L Brouwer Charlotte, Crijns Wouter, Verellen Dirk, et al. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, 2020.

[4] J Van der Veen, S Willems, S Deschuymer, D Robben, W Crijns, F Maes, and S Nuyts. Benefits of deep learning for delineation of organs at risk in head and neck cancer. *Radiotherapy and Oncology*, 138:68–74, 2019.

[5] Kazuki Kubo, Hajime Monzen, Kentaro Ishii, Mikoto Tamura, Ryu Kawamorita, Iori Sumida, Hirokazu Mizuno, and Yasumasa Nishimura. Dosimetric comparison of rapidplan and manually optimized plans in volumetric modulated arc therapy for prostate cancer. *Physica Medica*, 44:199–204, 2017.

[6] Loyce MH Chua, Eric PP Pang, Zubin Master, Rehena Sultana, Jeffrey KL Tuan, and Christopher M Bragg. Dosimetric comparison of rapidplan and manually optimised volumetric modulated arc therapy plans in prostate cancer. *Journal of Radiotherapy in Practice*, pages 1–8, 2020.

[7] Jeanette Bonden Isachsen. Deep learning image segmentation of organs at risk in breast cancer radiotherapy. Unpublished project thesis, 2020.

[8] Mary Feng, Gilmer Valdes, Nayha Dixit, and Timothy D Solberg. Machine learning in radiation oncology: opportunities, requirements, and needs. *Frontiers in oncology*, 8:110, 2018.

[9] Eric J Hall, Amato J Giaccia, et al. *Radiobiology for the Radiologist*, volume 6. Philadelphia, 2006.

[10] Willi A Kalender. *Computed tomography: fundamentals, system technology, image quality, applications*. John Wiley & Sons, 2011.

[11] Euclid Seeram. *Computed tomography: physical principles, clinical applications, and quality control*. Elsevier Health Sciences, 2015.

[12] Philip Mayles, Alan Nahum, and Jean-Claude Rosenwald. *Handbook of radiotherapy physics: theory and practice*. CRC Press, 2007.

[13] M Baker. Medical linear accelerator celebrates 50 years of treating cancer. *Stanford Report*, 5: 689–694, 2007. URL `https://news.stanford.edu/news/2007/april18/med-accelerator-041807.html`. Accessed 14.05.2021.

[14] Faiz M Khan, John P Gibbons, and Paul W Sperduto. *Khan's treatment planning in radiation oncology*. Lippincott Williams & Wilkins, 2016.

[15] Ward van Rooij, Max Dahele, Hugo Ribeiro Brandao, Alexander R Delaney, Berend J Slotman, and Wilko F Verbakel. Deep learning-based delineation of head and neck organs at risk: geometric and dosimetric evaluation. *International Journal of Radiation Oncology* Biology* Physics*, 104(3):677–684, 2019.

[16] Gillian A Whitfield, P Price, Gareth J Price, and Christopher J Moore. Automated delineation of radiotherapy volumes: are we going in the right direction? *The British journal of radiology*, 86(1021):20110718–20110718, 2013.

[17] International Bureau of Weights, Measures, Barry N Taylor, and Ambler Thompson. *The international system of units (SI)*. US Department of Commerce, Technology Administration, National Institute of . . . , 2001.

[18] CF Njeh. Tumor delineation: The weakest link in the search for accuracy in radiotherapy. *Journal of medical physics/Association of Medical Physicists of India*, 33(4):136, 2008.

[19] Lisanne V van Dijk, Lisa Van den Bosch, Paul Aljabar, Devis Peressutti, Stefan Both, Roel JHM Steenbakkers, Johannes A Langendijk, Mark J Gooding, and Charlotte L Brouwer. Improving automatic delineation for head and neck organs at risk by deep learning contouring. *Radiotherapy and Oncology*, 142:115–123, 2020.

[20] Benjamin E Nelms, Wolfgang A Tomé, Greg Robinson, and James Wheeler. Variations in the contouring of organs at risk: test case from a patient with oropharyngeal cancer. *International Journal of Radiation Oncology* Biology* Physics*, 82(1):368–378, 2012.

[21] Gregory Sharp, Karl D Fritscher, Vladimir Pekar, Marta Peroni, Nadya Shusharina, Harini Veeraraghavan, and Jinzhong Yang. Vision 20/20: perspectives on automated image segmentation for radiotherapy. *Medical physics*, 41(5), 2014.

[22] Rajit Rattan, Tejinder Kataria, Susovan Banerjee, Shikha Goyal, Deepak Gupta, Akshi Pandita, Shyam Bisht, Kushal Narang, and Saumya Ranjan Mishra. Artificial intelligence in oncology, its scope and future prospects with specific reference to radiation oncology. *BJR—Open*, 1(xxxx):20180031, 2019.

[23] Jinzhong Yang, Samuel G Armato Iii, Justin S Kirby, and Bruno Oliveira. Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. .

[24] Filippo Pesapane, Marina Codari, and Francesco Sardanelli. Artificial intelligence in medical imaging: threat or opportunity? radiologists again at the forefront of innovation in medicine. *European radiology experimental*, 2(1):35, 2018.

[25] Jinhan Zhu, Jun Zhang, Bo Qiu, Yimei Liu, Xiaowei Liu, and Lixin Chen. Comparison of the automatic segmentation of multiple organs at risk in ct images of lung cancer between deep convolutional neural network-based and atlas-based techniques. *Acta Oncologica*, 58(2): 257–264, 2019.

[26] Sang Hee Ahn, Adam Unjin Yeo, Kwang Hyeon Kim, Chankyu Kim, Youngmoon Goh, Shinhaeng Cho, Se Byeong Lee, Young Kyung Lim, Haksoo Kim, Dongho Shin, et al. Comparative clinical evaluation of atlas and deep-learning-based auto-segmentation of organ structures in liver cancer. *Radiation Oncology*, 14(1):1–13, 2019.

[27] Klaus Mainzer. *Artificial intelligence-When do machines take over?* Springer Nature, 2019.

[28] Stuart Russell and Peter Norvig. Artificial intelligence: a modern approach. 2002.

[29] Sergios Theodoridis. *Machine learning: a Bayesian and optimization perspective*. Academic press, 2015.

[30] Arthur L Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.

[31] Alvin Rajkomar, Jeffrey Dean, and Isaac Kohane. Machine learning in medicine. *New England Journal of Medicine*, 380(14):1347–1358, 2019.

[32] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[33] Ton J Cleophas, Aeilko H Zwinderman, and Henny I Cleophas-Allers. *Machine learning in medicine*, volume 9. Springer, 2013.

[34] Hyunseok Seo, Masoud Badiei Khuzani, Varun Vasudevan, Charles Huang, Hongyi Ren, Ruoxiu Xiao, Xiao Jia, and Lei Xing. Machine learning techniques for biomedical image segmentation: An overview of technical aspects and introduction to state-of-art applications. *Medical physics*, 47(5):e148–e167, 2020.

[35] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[36] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[37] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958, 2014.

[38] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: learning dense volumetric segmentation from sparse annotation. In *International conference on medical image computing and computer-assisted intervention*, pages 424–432. Springer, 2016.

[41] Kang-Ping Lin, Ratko Magjarevic, and Paulo De Carvalho. *Future Trends in Biomedical and Health Informatics and Cybersecurity in Medical Devices: Proceedings of the International Conference on Biomedical and Health Informatics, ICBHI 2019, 17-20 April 2019, Taipei, Taiwan*, volume 74. Springer Nature, 2019.

[42] Tim Lustberg, Johan van Soest, Mark Gooding, Devis Peressutti, Paul Aljabar, Judith van der Stoep, Wouter van Elmpt, and Andre Dekker. Clinical evaluation of atlas and deep learning based automatic contouring for lung cancer. *Radiotherapy and Oncology*, 126(2):312–317, 2018.

[43] Mark J Gooding, Annamarie J Smith, Maira Tariq, Paul Aljabar, Devis Peressutti, Judith van der Stoep, Bart Reymen, Daisy Emans, Djoya Hattu, Judith van Loon, et al. Comparative evaluation of autocontouring in clinical practice: a practical method using the turing test. *Medical physics*, 45(11):5105–5115, 2018.

[44] Amy TY Chang, Albert WM Hung, Fion WK Cheung, Michael CH Lee, Oscar SH Chan, Helen Philips, Yung-Tang Cheng, and Wai-Tong Ng. Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy. *International Journal of Radiation Oncology* Biology* Physics*, 95(3):981–990, 2016.

[45] Norwegian Breast Cancer Group. Nasjonalt handlingsprogram med retningslinjer for diagnostikk, behandling og oppfølging av pasienter med brystkreft, 2020.

[46] Sonali Pandya and Richard G Moore. Breast development and anatomy. *Clinical obstetrics and gynecology*, 54(1):91–95, 2011.

[47] National Breast Cancer Foundation. Breast cancer anatomy and how cancer starts. URL `https://nbcf.org.au/about-breast-cancer/diagnosis/breast-cancer-anatomy/`. Accessed 14.05.2021.

[48] Breastcancer.org. Non-invasive or invasive breast cancer. URL `https://www.breastcancer.org/symptoms/diagnosis/invasive`. Accessed 14.05.2021.

[49] Helsedirektoratet. Brystkreftdiagnostiserte kvinner som fikk brystbevarende operasjon, 2018. URL `https://www.helsedirektoratet.no/statistikk/kvalitetsindikatorer/kreft-behandling-og-overlevelse/brystbevarende-operasjon-for-kvinner-diagnostisert-med-brystkreft`. Last updated 03.12.2020, accessed 14.04.2021.

[50] Michael Jones-Lee and Terje Aven. Alarp—what does it really mean? *Reliability Engineering & System Safety*, 96(8):877–882, 2011.

[51] Acute and long-term cardiovascular toxicity after modern radiotherapy for breast cancer. URL `https://www.clinicaltrials.gov/ct2/show/NCT02541435?cond=breast&cntry=NO&draw=4&rank=25`. Accessed 19.05.2021.

[52] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945.

[53] Daniel P Huttenlocher, Gregory A. Klanderman, and William J Rucklidge. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence*, 15(9):850–863, 1993.

[54] Jun Wang and Ying Tan. Hausdorff distance with k-nearest neighbors. In *International Conference in Swarm Intelligence*, pages 272–281. Springer, 2012.

[55] MJ Gooding. Assessment of thoracic auto-contouring using a modified turing test.

[56] Frank Wilcoxon. Individual comparisons by ranking methods. In *Breakthroughs in statistics*, pages 196–202. Springer, 1992.

[57] Xue Dong, Yang Lei, Tonghe Wang, Matthew Thomas, Leonardo Tang, Walter J Curran, Tian Liu, and Xiaofeng Yang. Automatic multiorgan segmentation in thorax ct images using u-net-gan. *Medical physics*, 46(5):2157–2168, 2019.

[58] Jinzhong Yang, Samuel G Armato Iii, Justin S Kirby, and Bruno Oliveira. Autosegmentation for thoracic radiation treatment planning: A grand challenge at aapm 2017. .

[59] Ji Zhu, Xinyuan Chen, Bining Yang, Nan Bi, Tao Zhang, Kuo Men, and Jianrong Dai. Evaluation of automatic segmentation model with dosimetric metrics for radiotherapy of esophageal cancer. *Frontiers in Oncology*, 10:1843, 2020.

[60] Rita Simões, Geert Wortel, Terry G Wiersma, Tomas M Janssen, Uulke A van der Heide, and Peter Remeijer. Geometrical and dosimetric evaluation of breast target volume auto-contouring. *Physics and Imaging in Radiation Oncology*, 12:38–43, 2019.

[61] Mary Feng, Jean M Moran, Todd Koelling, Aamer Chughtai, June L Chan, Laura Freedman, James A Hayman, Reshma Jagsi, Shruti Jolly, Janice Larouere, et al. Development and validation of a heart atlas to study cardiac exposure to radiation following treatment for breast cancer. *International Journal of Radiation Oncology* Biology* Physics*, 79(1):10–18, 2011.

[62] Yabo Fu, Shi Liu, H Harold Li, and Deshan Yang. Automatic and hierarchical segmentation of the human skeleton in ct images. *Physics in Medicine & Biology*, 62(7):2812, 2017.

[63] Mette H Nielsen, Martin Berg, Anders N Pedersen, Karen Andersen, Vladimir Glavicic, Erik H Jakobsen, Ingelise Jensen, Mirjana Josipovic, Ebbe L Lorenzen, Hanne M Nielsen, et al. Delineation of target volumes and organs at risk in adjuvant radiotherapy of early breast cancer: national guidelines and contouring atlas by the danish breast cancer cooperative group. *Acta oncologica*, 52(4):703–710, 2013.

[64] Vishruta A Dumane, James Tam, Yeh-Chi Lo, and Kenneth E Rosenzweig. Rapidplan for knowledge-based planning of malignant pleural mesothelioma. *Practical radiation oncology*, 11(2):e219–e228, 2021.

[65] Chifang Ling, Xu Han, Peng Zhai, Hao Xu, Jiayan Chen, Jiazhou Wang, and Weigang Hu. A hybrid automated treatment planning solution for esophageal cancer. *Radiation Oncology*, 14 (1):1–7, 2019.

[66] Annika Reinke, Matthias Eisenmann, Minu D Tizabi, Carole H Sudre, Tim Rädsch, Michela Antonelli, Tal Arbel, Spyridon Bakas, M Jorge Cardoso, Veronika Cheplygina, et al. Common limitations of image processing metrics: A picture story. *arXiv preprint arXiv:2104.05642*, 2021.

[67] Femke Vaassen, Colien Hazelaar, Ana Vaniqui, Mark Gooding, Brent van der Heyden, Richard Canters, and Wouter van Elmpt. Evaluation of measures for assessing time-saving of automatic organ-at-risk segmentation in radiotherapy. *Physics and Imaging in Radiation Oncology*, 13: 1–6, 2020.

[68] Stanislav Nikolov, Sam Blackwell, Ruheena Mendes, Jeffrey De Fauw, Clemens Meyer, Cían Hughes, Harry Askham, Bernardino Romera-Paredes, Alan Karthikesalingam, Carlton Chu, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[69] Dominique P Huyskens, Philippe Maingon, Luc Vanuytsel, Vincent Remouchamps, Tom Roques, Bernard Dubray, Benjamin Haas, Patrik Kunz, Thomas Coradi, René Bühlman, et al. A qualitative and a quantitative analysis of an auto-segmentation module for prostate cancer. *Radiotherapy and Oncology*, 90(3):337–345, 2009.

[70] Jan Schreier, Angelo Genghi, Hannu Laaksonen, Tomasz Morgas, and Benjamin Haas. Clinical evaluation of a full-image deep segmentation algorithm for the male pelvis on cone-beam ct and ct. *Radiotherapy and Oncology*, 145:1–6, 2020.

[71] Mary L McHugh. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282, 2012.

[72] Kaja S Øvrelid. Artificial intelligence-based automatic segmentation for breast cancer radiotherapy. Master's thesis, NTNU, Trondheim, Norway, 2020.

Appendices

A Script for extracting geometric metrics for model evaluation

The script written for the project thesis [7] to calculate and extract the chosen geometric metrics. It has not been edited for the master thesis. It is written for Ironpython and to use in RayStation 9B. The script is used for both segmentation models, but the details specific for the local model are inserted in the script attached, i.e, patient info and region of interests specified.

The code will make a spreadsheet in Microsoft Excel that contains the results for each patient. The code uses roughly 30 minutes for each patient. The function for calculating Hausdorff distances goes through point by point in each organ, therefore the lungs especially increase the run-time because they have relatively large volumes.

The script from Øvrelid [72] was used as basis for the code written.

```
from connect import *
import clr, sys, math


clr.AddReference('Office')
clr.AddReference('Microsoft.Office.Interop.Excel')
import Microsoft.Office.Interop.Excel as interop_excel
import System.Array


def create_array(m, n):
    """Create two-dimensional array."""
    dims = System.Array.CreateInstance(System.Int32, 2)
    dims[0] = m
    dims[1] = n
    return System.Array.CreateInstance(System.Object, dims)


def set_up_header_rows(first_row, second_row):
    """Set up header rows"""
    rows = create_array(2, len(second_row) * len(first_row) + 1)
    rows[0, 0] = 'Patient'
    for i in range(len(first_row)):
        rows[0, i * len(second_row) + 1] = first_row[i]

        for j in range(1, len(second_row) + 1):
            rows[1, j + i * len(second_row)] = second_row[j - 1]
    return rows


def directed_hausdorff_distances(contour_1, contour_2):
    """Return list of directed hausdorff distances for two different contours
    (not sorted)
    Iterates through all the contours in point_list_1, then all the points in
    the contours and finally x,y,z in each of the points.
    Reach a point by using print(contour_1[i][j].z).
    """
    min_distances = []
    for level_nr_in_contour_1 in contour_1:
        for p1 in level_nr_in_contour_1:
            distance_p1_to_points_in_2 = []
            for level_nr_in_contour_2 in contour_2:
                for p2 in level_nr_in_contour_2:
                    distance = math.sqrt(math.pow(p1.x - p2.x, 2)
```

```
                                        + math.pow(p1.y - p2.y, 2)
                                        + math.pow(p1.z - p2.z, 2))
                    distance_p1_to_points_in_2.append(distance)
            min_distances.append(min(distance_p1_to_points_in_2))
    return min_distances


def hausdorff_distances(contour_1, contour_2):
    """Return list of hausdorff distances for two different contours
    (sorted from min to max)
    """
    hd = (directed_hausdorff_distances(contour_1, contour_2)
          + directed_hausdorff_distances(contour_2, contour_1))
    hd.sort()
    return hd


def get_percentile(dist, percentile):
    """Return specific percentile from a list of hausdorff distances"""
    index = int(float(percentile) / 100 * len(dist)) - 1
    return dist[index]


patient_db = get_current('PatientDB')  # Load patient database

try:
    info = patient_db.QueryPatientInfo(Filter={'FirstName': 'Breast model final',
                                               'LastName': 'BM patient_*',
                                               'PatientID': '181096_*'})
except:
    print("Could not find patient info")

data_array = create_array(500, 500)

roiA = ['Heart_manual', 'Lung_L_manual', 'Lung_R_manual', 'SpinalCanal_manual',
        'Esophagus_manual', 'Sternum_manual', 'Breast_R_manual', 'Breast_L_manual',
        'LN_Ax_L1_L_manual', 'LN_Ax_L2_L_manual', 'LN_Ax_L3_L_manual',
        'LN_Ax_L4_L_manual', 'LN_Ax_Pectoral_L_manual']
roiB = ['Heart_Modell', 'Lung_L_Modell', 'Lung_R_Modell', 'SpinalCanal_Modell',
        'Esophagus_Modell', 'Sternum_Modell', 'Breast_R_Modell', 'Breast_L_Modell',
        'LN_Ax_L1_L_Modell', 'LN_Ax_L2_L_Modell', 'LN_Ax_L3_L_Modell',
        'LN_Ax_L4_L_Modell', 'LN_Ax_Pectoral_L_Modell']

metrics = ['Volume', 'Volume_AI', 'HD95', 'HD99', 'HD100', 'DSC', 'AVD']
number_of_metrics = len(metrics)

for p in range(len(info)):
    # iterates through the patients

    patient = patient_db.LoadPatient(PatientInfo=info[p])
    structure_set = (patient.Cases['For Organs RT and Breast model'].PatientModel
        .StructureSets[0])

    data_array[p, 0] = patient.Name

    # Extract evaluation metrics for the organs
    for r in range(len(roiA)):
        # iterates through the organs and extracts the metrics for each organ
```

```
103
104         structure_set.RoiGeometries[roiA[r]].SetRepresentation(Representation = "Contours")
105         structure_set.RoiGeometries[roiB[r]].SetRepresentation(Representation = "Contours")
106         contour_a = structure_set.RoiGeometries[roiA[r]].PrimaryShape.Contours
107         contour_b = structure_set.RoiGeometries[roiB[r]].PrimaryShape.Contours
108
109         hausdorff_dist = hausdorff_distances(contour_a, contour_b)
110
111         # volume of contoured organ is in cm^3
112         data_array[p, 1 + r * number_of_metrics] = (structure_set.RoiGeometries[roiA[r]]
113                                                     .GetRoiVolume())
114         data_array[p, 2 + r * number_of_metrics] = (structure_set.RoiGeometries[roiB[r]]
115                                                     .GetRoiVolume())
116         # H95
117         data_array[p, 3 + r * number_of_metrics] = get_percentile(hausdorff_dist, 95)
118         # HD99
119         data_array[p, 4 + r * number_of_metrics] = get_percentile(hausdorff_dist, 99)
120         # HD100
121         data_array[p, 5 + r * number_of_metrics] = get_percentile(hausdorff_dist, 100)
122         # DSC
123         data_array[p, 6 + r * number_of_metrics] = structure_set.ComparisonOfRoiGeometries(
124             RoiA=roiA[r],
125             RoiB=roiB[r],
126             ComputeDistanceToAgreementMeasures=False)['DiceSimilarityCoefficient']
127         # AVD
128         data_array[p, 7 + r * number_of_metrics] = sum(hausdorff_dist) / (len(hausdorff_dist))
129     patient.Save()
130
131 file_path = None
132 close_excel = True
133
134 try:
135     # Open Excel with new worksheet
136     excel = interop_excel.ApplicationClass(Visible=True)
137     workbook = excel.Workbooks.Add(interop_excel.XlWBATemplate.xlWBATWorksheet)
138     worksheet = workbook.Worksheets[1]
139
140     header_row = set_up_header_rows(roiA, metrics)
141
142     # Add header row to work sheet
143     startcell = worksheet.Cells(1, 1)
144     header_range = worksheet.Range(startcell, startcell.Cells(header_row.GetLength(0),
145                                                               header_row.GetLength(1)))
146     header_range.Value = header_row
147
148     # Add ROI data array to work sheet
149     startcell = worksheet.Cells(3, 1)
150     data_range = worksheet.Range(startcell, startcell.Cells(data_array.GetLength(0),
151                                                             data_array.GetLength(1)))
152     data_range.Value = data_array
153
154     # Auto-fit the width of all columns
155     worksheet.Columns.AutoFit()
156
157 finally:
158     # The following is needed for the excel process to die when user closes worksheet
159     if file_path != None and close_excel:
160         excel.Quit()
```

```
161     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(worksheet)
162     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(workbook)
163     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(excel)
164     seriesCollection = None
165     chart = None
166     worksheet = None
167     workbook = None
168     excel = None
169     System.GC.WaitForPendingFinalizers()
170     System.GC.Collect()
```

B Script for extracting dosimetric metrics for model evaluation

The script used to extract clinical goals from RayStation 9B for the dosimetric evaluation of the models. It is written for Ironpython. The script is used for both segmentation models, but the details specific for the local model are inserted in the script attached, i.e. patient info.

All clinical goals to be extracted need to be added for each patient in RayStation. The clinical goals need to be the same for all patients. The script will make a spreadsheet in Microsoft Excel that contains the results for each patient.

It is based on the script "get_plan_current_clinical_goals" by Marit Funderud. It has been modified so it can be run for several patients at the same time.

```
from connect import *
import clr, sys, math

clr.AddReference('Office')
clr.AddReference('Microsoft.Office.Interop.Excel')
import Microsoft.Office.Interop.Excel as interop_excel
import System.Array


def create_array(m, n):
    """Create two-dimensional array."""
    dims = System.Array.CreateInstance(System.Int32, 2)
    dims[0] = m
    dims[1] = n
    return System.Array.CreateInstance(System.Object, dims)


def clinical_goals(da, ef, row, start_col, patient_number):
    for i, v in enumerate(ef):
        try:
            if ef[i].PlanningGoal.Type == 'DoseAtVolume':
                goal = ef[i].ForRegionOfInterest.Name + " : D" + str(round(
                    ef[i].PlanningGoal.ParameterValue * 100))
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'DoseAtAbsoluteVolume':
                goal = ef[i].ForRegionOfInterest.Name + " : D" + str(
                    ef[i].PlanningGoal.ParameterValue) + "cm3"
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'VolumeAtDose':
                goal = ef[i].ForRegionOfInterest.Name + " : V" + str(
                    round(ef[i].PlanningGoal.ParameterValue / 100)) + "Gy"
```

```
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) * 100
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) * 100
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'AverageDose':
                goal = ef[i].ForRegionOfInterest.Name + " : Dmean"
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'HomogeneityIndex':
                goal = ef[i].ForRegionOfInterest.Name + " : HI"
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'ConformityIndex':
                goal = ef[i].ForRegionOfInterest.Name + " : CI"
                if patient_number == 1:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                elif goal == da[0, i + 1]:
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
        except:
            print("Hei")
    return da


patient_db = get_current('PatientDB')  # Load patient database

try:
    info = patient_db.QueryPatientInfo(Filter={'LastName': 'BM patient_',
                                               'FirstName': 'Breast model final',
                                               'PatientID': '181096*'})
except:
    print("Could not find patient info")

case0 = 'For Organs RT and Breast model'
treatmentplan0 = 'Based on automatic segmentations'

data_array = create_array(500, 500)

data_array[0, 0] = 'PatientID'

for p in range(len(info)):
    # iterates through the patients
```

```
    patient = patient_db.LoadPatient(PatientInfo=info[p])

    eval_funcs = patient.Cases[case0].TreatmentPlans[
        treatmentplan0].TreatmentCourse.EvaluationSetup.EvaluationFunctions

    data_array[p + 1, 0] = patient.PatientID
    clinical_goals(data_array, eval_funcs, p + 1, 1, p + 1)

    print('Patient number: ' + str(p + 1))

# Select path where the Excel file should be saved
# Set file_path = None if the file should not be automatically saved
file_path = None

# Should the Excel file be closed after it is created?
# If no file path is selected, the Excel application will not be closed
close_excel = True
# Create an Excel file

try:
    # Open Excel with new worksheet
    excel = interop_excel.ApplicationClass(Visible=True)
    workbook = excel.Workbooks.Add(interop_excel.XlWBATemplate.xlWBATWorksheet)
    worksheet = workbook.Worksheets[1]
    l = 0
    # Set up header row
    # Edit this if other dose statistics are desired

    # Add ROI data array to work sheet
    startcell = worksheet.Cells(1, 1)
    data_range = worksheet.Range(startcell, startcell.Cells(data_array.GetLength(0),
                                                            data_array.GetLength(1)))
    data_range.Value = data_array

    # Auto-fit the width of all columns
    worksheet.Columns.AutoFit()

    if file_path != None:
        # File name is PatientNamePlanNameDoseStatistics
        # Edit this if another file name is desired
        filename = r"{0}\{1}DoseStatistics.xlsx".format(file_path, patient.PatientName)
        excel.DisplayAlerts = False
        workbook.SaveAs(filename)
finally:
    # The following is needed for the excel process to die when user closes worksheet
    if file_path != None and close_excel:
        excel.Quit()
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(worksheet)
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(workbook)
    System.Runtime.InteropServices.Marshal.FinalReleaseComObject(excel)
    seriesCollection = None
    chart = None
    worksheet = None
    workbook = None
    excel = None
    System.GC.WaitForPendingFinalizers()
    System.GC.Collect()
```

C Script for extracting dosimetric metrics and DVH-curves for treatment plan comparison

The script used to extract clinical goals and average DVH-curves from RayStation 9B for the evaluation of the script for automatic plan optimization, i.e. for comparison of two treatment plans. It is written for Ironpython. For this script, the automatic VMAT plan is the main plan and the plan it is compared with needs to be added as an "Evaluation dose" in RayStation.

All clinical goals to be extracted need to be added for each patient in RayStation. The clinical goals need to be the same for all patients. The function for extracting clinical goals is, like the script in appendix B, based on a script by Marit Funderud.

The DVH-curves are extracted as dose and volume values and can be plotted based on this. Four different volumes are extracted for each structure. These are the average curves for the manual hybrid plan and the automatic VMAT plan based on hybrid plan patients and the average curves for the the manual VMAT plan and the automatic VMAT plan based on the VMAT plan patients. The DVH-difference-curves can also be extracted, i.e. clinical plan DVH minus the automatic plan DVH.

The script will make a spreadsheet in Microsoft Excel that contains the results for each patient with either the clinical goals, DVH-curves or DVH-difference-curves. What is extracted depends on which is set to "true" in the script.

```
from connect import *
import clr, sys, math

clr.AddReference('Office')
clr.AddReference('Microsoft.Office.Interop.Excel')
import Microsoft.Office.Interop.Excel as interop_excel
import System.Array


def create_array(m, n):
    """Create two-dimensional array."""
    dims = System.Array.CreateInstance(System.Int32, 2)
    dims[0] = m
    dims[1] = n
    return System.Array.CreateInstance(System.Object, dims)


def clinical_goals(da, ef, row, start_col, compare_to_found, compare_to, patient_number):

    for i, v in enumerate(ef):
        try:
            if ef[i].PlanningGoal.Type == 'DoseAtVolume':
                goal = ef[i].ForRegionOfInterest.Name + " : D" + str(round(
                    ef[i].PlanningGoal.ParameterValue * 100))
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = ((ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                                                  / 100)
                    da[row + 2, i + start_col] = da[row + 1, i +
```

```
                                                start_col] - da[row, i + start_col]
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'DoseAtAbsoluteVolume':
                goal = ef[i].ForRegionOfInterest.Name + " : D" + str(
                    ef[i].PlanningGoal.ParameterValue) + "cm3"
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = ((ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                                                  / 100)
                    da[row + 2, i + start_col] = da[row + 1, i +
                                                    start_col] - da[row, i + start_col]
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'VolumeAtDose':
                goal = ef[i].ForRegionOfInterest.Name + " : V" + str(round(
                    ef[i].PlanningGoal.ParameterValue / 100)) + "Gy"
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) * 100
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) * 100
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = ((ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                                                  * 100)
                    da[row + 2, i + start_col] = da[row + 1, i +
                                                    start_col] - da[row, i + start_col]
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'AverageDose':
                goal = ef[i].ForRegionOfInterest.Name + " : Dmean"
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = (ef[i].GetClinicalGoalValue()) / 100
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = ((ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                                                  / 100)
                    da[row + 2, i + start_col] = da[row + 1, i +
                                                    start_col] - da[row, i + start_col]
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'HomogeneityIndex':
```

```
                goal = ef[i].ForRegionOfInterest.Name + " : HI"
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = (ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                    da[row + 2, i + start_col] = da[row + 1, i +
                                                    start_col] - da[row, i + start_col]
        except:
            print("Hei")
        try:
            if ef[i].PlanningGoal.Type == 'ConformityIndex':
                goal = ef[i].ForRegionOfInterest.Name + " : CI"
                if patient_number == 0:
                    da[row - 1, i + start_col] = goal
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                elif goal == da[0, i+start_col]:
                    da[row, i + start_col] = ef[i].GetClinicalGoalValue()
                if compare_to_found and goal == da[0, i+start_col]:
                    da[row + 1, i + start_col] = (ef[i]
                        .GetClinicalGoalValueForEvaluationDose(
                        DoseDistribution=compare_to, ScaleFractionDoseToBeamSet=False))
                    da[row + 2, i + start_col] = da[row + 1, i +
                                                    start_col] - da[row, i + start_col]
        except:
            print("Hei")
    return da


def add_list_to_list(l1, l2):
    for element in range(len(l1)):
        l1[element] += l2[element]
    return l1

def subtract_list_from_list(l1, l2):
    for element in range(len(l1)):
        l1[element] -= l2[element]
    return l1


def divide_list_with_number(list1, number):
    for li in range(len(list1)):
        list1[li] /= number
    return list1


def make_dict_for_rois(rois, length_of_lists):
    dictionary_of_rois = {}
    for r in range(len(rois)):
        dictionary_of_rois[rois[r]] = [0] * length_of_lists
    return dictionary_of_rois


# Choose what the script should extract.
```

```
get_clinical_goals = False
get_DVH = True
get_DVH_difference = False

list_of_rois = ['CTVp', 'CTVn', 'PTVpc', 'PTVnc', 'PTV', 'Body', 'Heart', 'Breast_R',
                'Lung_L', 'Lung_R', 'HumeralHead_L', 'Esophagus', 'Thyroid',
                'Trachea', 'SpinalCanal']


list_of_doses = []
for i in range(0, 4400, 22):
    list_of_doses.append(float(i))


patient_db = get_current('PatientDB')  # Load patient database

try:
    info = patient_db.QueryPatientInfo(Filter={'LastName': 'Lokomamma',
                                               'PatientID': '01012021 10*'})
except:
    print("Could not find patient info")

case0 = 'For auto-script validation'

data_array = create_array(500, 500)
if get_clinical_goals:
    data_array[0, 0] = 'PatientID'
    data_array[0, 1] = 'Plan'
elif get_DVH:
    vmat_patient_nr = 0
    hybrid_patient_nr = 0
    dict_of_rois_auto_hybrid = make_dict_for_rois(list_of_rois, len(list_of_doses))
    dict_of_rois_auto_vmat = make_dict_for_rois(list_of_rois, len(list_of_doses))
    dict_of_rois_hybrid = make_dict_for_rois(list_of_rois, len(list_of_doses))
    dict_of_rois_vmat = make_dict_for_rois(list_of_rois, len(list_of_doses))
elif get_DVH_difference:
    vmat_patient_nr = 0
    hybrid_patient_nr = 0
    dict_of_diff_vmat = make_dict_for_rois(list_of_rois, len(list_of_doses))
    dict_of_diff_hybrid = make_dict_for_rois(list_of_rois, len(list_of_doses))

total_patients = 0
for p in range(len(info)):
    # iterates through the patients
    patient = patient_db.LoadPatient(PatientInfo=info[p])
    if patient.Name == "Lokomamma_1":
        continue

    eval_funcs = patient.Cases[case0].TreatmentPlans[
        'Automatisk VMAT'].TreatmentCourse.EvaluationSetup.EvaluationFunctions


    tp = patient.Cases[case0].TreatmentPlans
    clinical_plan_name = "none"
    for i, v in enumerate(tp):
        if tp[i].Name == "Klinisk VMAT":
            clinical_plan_name = "Summed Dose Klinisk VMAT"
        elif tp[i].Name == "Klinisk hybrid":
```

```
            clinical_plan_name = 'Summed Dose Klinisk hybrid'

    clinical_plan = (patient.Cases[case0].TreatmentDelivery.FractionEvaluations[0]
        .DoseOnExaminations[0].DoseEvaluations[0])
    if clinical_plan.Name == clinical_plan_name:
        clinical_plan_found = True
    else:
        clinical_plan_found = False

    if get_clinical_goals:
        data_array[3 * total_patients + 1, 0] = patient.PatientID
        data_array[3 * total_patients + 1, 1] = 'Automatisk VMAT'
        if clinical_plan_name == "Summed Dose Klinisk VMAT":
            data_array[3 * total_patients + 2, 1] = 'Klinisk VMAT'
        elif clinical_plan_name == 'Summed Dose Klinisk hybrid':
            data_array[3 * total_patients + 2, 1] = 'Klinisk hybrid'
        data_array[3 * total_patients + 3, 1] = "Klinisk - Automatisk"
        clinical_goals(data_array, eval_funcs, 3*total_patients+1, 2, clinical_plan_found,
                       clinical_plan, total_patients)

    elif get_DVH:
        tc = patient.Cases[case0].TreatmentPlans[
            'Automatisk VMAT'].TreatmentCourse
        if clinical_plan_name == "Summed Dose Klinisk VMAT":
            vmat_patient_nr += 1
            for r in range(len(list_of_rois)):
                # Automatic
                volume_at_dose = tc.TotalDose.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r],  DoseValues=list_of_doses)
                dict_of_rois_auto_vmat[list_of_rois[r]] = add_list_to_list(
                    dict_of_rois_auto_vmat[list_of_rois[r]], volume_at_dose)
                # Clinical VMAT
                volume_at_dose_clinical = clinical_plan.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r],  DoseValues=list_of_doses)
                dict_of_rois_vmat[list_of_rois[r]] = add_list_to_list(
                    dict_of_rois_vmat[list_of_rois[r]], volume_at_dose_clinical)

        elif clinical_plan_name == "Summed Dose Klinisk hybrid":
            hybrid_patient_nr += 1
            for r in range(len(list_of_rois)):
                # Automatic
                volume_at_dose = tc.TotalDose.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r],  DoseValues=list_of_doses)
                dict_of_rois_auto_hybrid[list_of_rois[r]] = add_list_to_list(
                    dict_of_rois_auto_hybrid[list_of_rois[r]], volume_at_dose)
                # Clinical hybrid
                volume_at_dose_clinical = clinical_plan.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r],  DoseValues=list_of_doses)
                dict_of_rois_hybrid[list_of_rois[r]] = add_list_to_list(
                    dict_of_rois_hybrid[list_of_rois[r]], volume_at_dose_clinical)

    elif get_DVH_difference:
        tc = patient.Cases[case0].TreatmentPlans[
            'Automatisk VMAT'].TreatmentCourse
        if clinical_plan_name == "Summed Dose Klinisk VMAT":
            vmat_patient_nr += 1
            for r in range(len(list_of_rois)):
                volume_at_dose_auto = tc.TotalDose.GetRelativeVolumeAtDoseValues(
```

```
                    RoiName=list_of_rois[r],  DoseValues=list_of_doses)
                volume_at_dose_clinical = clinical_plan.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r], DoseValues=list_of_doses)

                difference = subtract_list_from_list(volume_at_dose_clinical,
                                                     volume_at_dose_auto)

                dict_of_diff_vmat[list_of_rois[r]] = add_list_to_list(
                    dict_of_diff_vmat[list_of_rois[r]], difference)

        elif clinical_plan_name == "Summed Dose Klinisk hybrid":
            hybrid_patient_nr += 1
            for r in range(len(list_of_rois)):
                volume_at_dose_auto = tc.TotalDose.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r], DoseValues=list_of_doses)
                volume_at_dose_clinical = clinical_plan.GetRelativeVolumeAtDoseValues(
                    RoiName=list_of_rois[r], DoseValues=list_of_doses)

                difference = subtract_list_from_list(volume_at_dose_clinical,
                                                     volume_at_dose_auto)

                dict_of_diff_hybrid[list_of_rois[r]] = add_list_to_list(
                    dict_of_diff_hybrid[list_of_rois[r]], difference)

    total_patients += 1
    print('Patient number: ' + str(total_patients))


if get_DVH:

    print('VMAT pasienter: ' + str(vmat_patient_nr))
    print('Hybrid pasienter: ' + str(hybrid_patient_nr))
    # dele på antall lister summert/pasienter
    for r in range(len(list_of_rois)):
        if vmat_patient_nr != 0:
            dict_of_rois_auto_vmat[list_of_rois[r]] = divide_list_with_number(
                dict_of_rois_auto_vmat[list_of_rois[r]], vmat_patient_nr)
            dict_of_rois_vmat[list_of_rois[r]] = divide_list_with_number(
                dict_of_rois_vmat[list_of_rois[r]], vmat_patient_nr)
        if hybrid_patient_nr != 0:
            dict_of_rois_auto_hybrid[list_of_rois[r]] = divide_list_with_number(
                dict_of_rois_auto_hybrid[list_of_rois[r]], hybrid_patient_nr)
            dict_of_rois_hybrid[list_of_rois[r]] = divide_list_with_number(
                dict_of_rois_hybrid[list_of_rois[r]], hybrid_patient_nr)

        data_array[4*r+1, 0] = list_of_rois[r] + ' (Automatic VMAT (VMAT))'
        data_array[4*r+2, 0] = list_of_rois[r] + ' (Automatic VMAT (Hybrid))'
        data_array[4*r+3, 0] = list_of_rois[r] + ' (VMAT)'
        data_array[4*r+4, 0] = list_of_rois[r] + ' (Hybrid)'
        for d in range(len(list_of_doses)):
            if r == 0:
                data_array[0, 0] = 'Dose (Gy)'
                data_array[0, 1+d] = list_of_doses[d]/100
            data_array[4*r+1, 1 + d] = dict_of_rois_auto_vmat[list_of_rois[r]][d]*100
            data_array[4*r+2, 1 + d] = dict_of_rois_auto_hybrid[list_of_rois[r]][d] * 100
            data_array[4*r+3, 1+d] = dict_of_rois_vmat[list_of_rois[r]][d]*100
            data_array[4*r+4, 1+d] = dict_of_rois_hybrid[list_of_rois[r]][d]*100

```

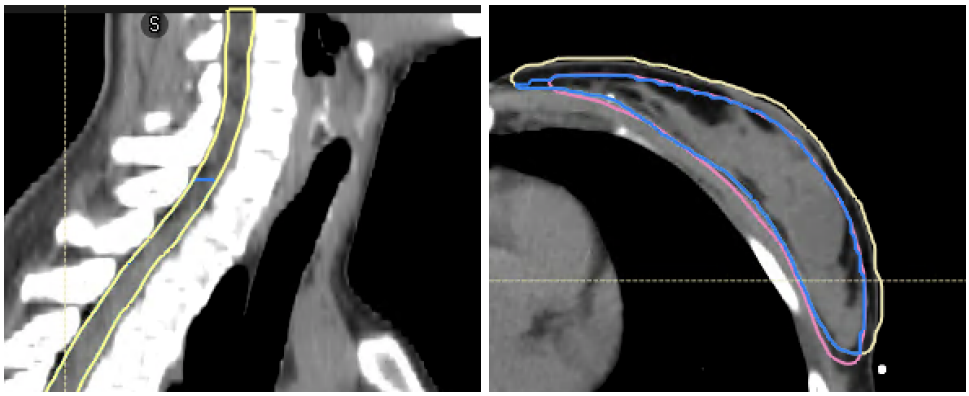
```
if get_DVH_difference:

    print('VMAT pasienter: ' + str(vmat_patient_nr))
    print('Hybrid pasienter: ' + str(hybrid_patient_nr))
    # dele på antall lister summert/pasienter
    for r in range(len(list_of_rois)):
        if vmat_patient_nr != 0:
            dict_of_diff_vmat[list_of_rois[r]] = divide_list_with_number(
                dict_of_diff_vmat[list_of_rois[r]], vmat_patient_nr)
        if hybrid_patient_nr != 0:
            dict_of_diff_hybrid[list_of_rois[r]] = divide_list_with_number(
                dict_of_diff_hybrid[list_of_rois[r]], hybrid_patient_nr)

        data_array[2*r+1, 0] = list_of_rois[r] + ' VMAT'
        data_array[2*r+2, 0] = list_of_rois[r] + ' Hybrid'
        for d in range(len(list_of_doses)):
            if r == 0:
                data_array[0, 0] = 'Dose (Gy)'
                data_array[0, 1+d] = list_of_doses[d]/100
            data_array[2*r+1, 1 + d] = dict_of_diff_vmat[list_of_rois[r]][d]*100
            data_array[2*r+2, 1 + d] = dict_of_diff_hybrid[list_of_rois[r]][d] * 100


# Select path where the Excel file should be saved
# Set file_path = None if the file should not be automatically saved
file_path = None

# Should the Excel file be closed after it is created?
# If no file path is selected, the Excel application will not be closed
close_excel = True
# Create an Excel file

try:
    # Open Excel with new worksheet
    excel = interop_excel.ApplicationClass(Visible=True)
    workbook = excel.Workbooks.Add(interop_excel.XlWBATemplate.xlWBATWorksheet)
    worksheet = workbook.Worksheets[1]
    l = 0
    # Set up header row
    # Edit this if other dose statistics are desired

    # Add ROI data array to work sheet
    startcell = worksheet.Cells(1, 1)
    data_range = worksheet.Range(startcell, startcell.Cells(data_array.GetLength(0),
                                                            data_array.GetLength(1)))
    data_range.Value = data_array

    # Auto-fit the width of all columns
    worksheet.Columns.AutoFit()

    if file_path != None:
        # File name is PatientNamePlanNameDoseStatistics
        # Edit this if another file name is desired
        filename = r"{0}\{1}DoseStatistics.xlsx".format(file_path, patient.PatientName)
        excel.DisplayAlerts = False
        workbook.SaveAs(filename)
finally:
    # The following is needed for the excel process to die when user closes worksheet
```

```
384     if file_path != None and close_excel:
385         excel.Quit()
386     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(worksheet)
387     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(workbook)
388     System.Runtime.InteropServices.Marshal.FinalReleaseComObject(excel)
389     seriesCollection = None
390     chart = None
391     worksheet = None
392     workbook = None
393     excel = None
394     System.GC.WaitForPendingFinalizers()
395     System.GC.Collect()
```

D Additional results from evaluation of automatic segmentation models

Additional results from the evaluation of the segmentation models can be found in this appendix, both for geometric and dosimetric evaluations.

The manual editing to the segmentations before the evaluations can be seen in figure D.1. An example of a difference in length can be seen in figure D.1a for the spinal canal segmented by the Siemens model. While figure D.1b shows how the breast was cropped for the Siemens model.



(a) Spinal canal in saggital plane **(b)** Left breast (CTVp) in transversal plane

Figure D.1: (a) Shows the difference in length of the spinal canal segmented by the Siemens model before and after removing slices so the length is the same as the manual delineation. Yellow is the original segmentation and blue is the edited one. (b) Shows the left breast before and after cropping 5 mm below the body surface. Yellow is the original segmentation, blue is the edited one, and pink is the manual delineation.

D.1 Geometric evaluation

Separate lymph node areas were segmented by the local model and were evaluated individually, in addition to as the union CTVn. Mean DSC and HD95 for the individual lymph node areas are presented in table D.1. DSC and HD95 are also presented for the lymph node areas as boxplots in figure D.2. L1, L2, L3, and L4 indicate axillary lymph nodes levels 1-4. The volumes of the lymph nodes are plotted in figure D.3. Statistically significant differences were found for L1, L2, and pectoral lymph nodes.

Additional metrics, i.e., HD99, HD100 and average distance (AVD), can be found in table D.2 and figures D.4 and D.5 for all structures.

Table D.1: Mean DSC and HD95 for the lymph node areas segmented by the local model. SD is denoted as the ± value.

	DSC	HD95 [cm]
L1	0.75 ± 0.09	1.3 ± 0.6
L2	0.68 ± 0.08	1.7 ± 0.9
L3	0.80 ± 0.07	0.6 ± 0.1
L4	0.79 ± 0.08	0.5 ± 0.2
Pectoral	0.5 ± 0.1	2 ± 1

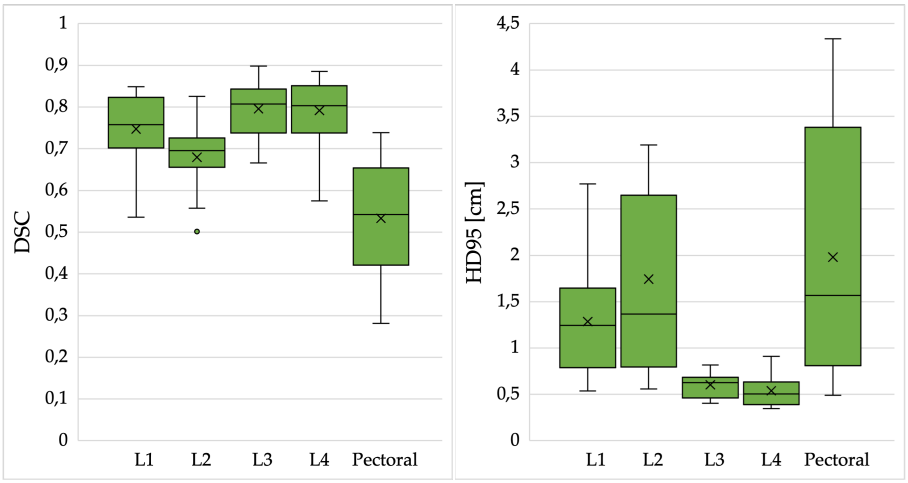


Figure D.2: DSC and HD95 for the lymph node areas segmented by the local model.

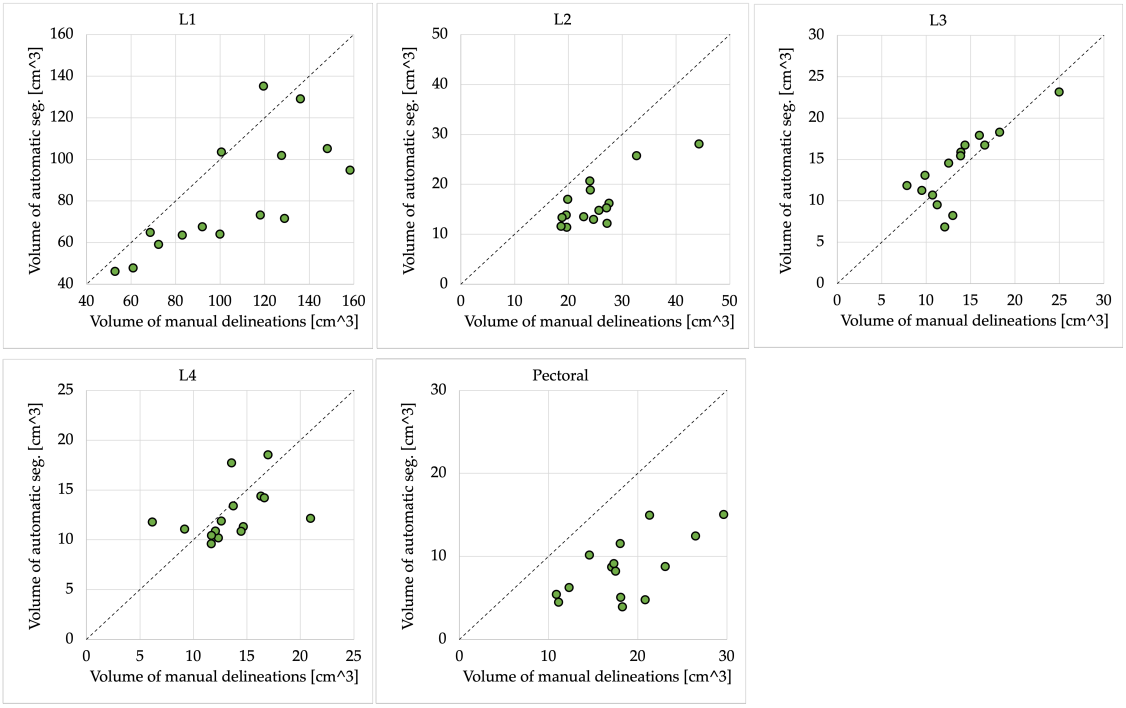


Figure D.3: Volume of automatic segmentation by the local model plotted against the volume of the manual delineation for all patients. Dotted line represents equality.

Table D.2: Mean HD99, HD100, and AVD obtained by both segmentation models for all structures. SD is denoted as the ± value.

	HD99 [cm]		HD100 [cm]		AVD [cm]	
	Local model	Siemens model	Local model	Siemens model	Local model	Siemens model
Heart	1.0 ± 0.3	1.3 ± 0.4	2.5 ± 0.5	2.6 ± 0.6	0.20 ± 0.04	0.3 ± 0.1
Left lung	0.7 ± 0.2	1.0 ± 0.3	2.0 ± 0.5	3.0 ± 0.5	0.16 ± 0.03	0.16 ± 0.01
Right lung	0.8 ± 0.4	1.2 ± 0.4	2.6 ± 0.8	3.6 ± 0.7	2.6 ± 0.8	3.6 ± 0.7
Spinal canal	0.25 ± 0.07	0.27± 0.03	0.5 ± 0.2	0.35 ± 0.05	0.079 ± 0.009	0.13 ± 0.02
Esophagus	0.39 ± 0.06	1 ± 1	0.7 ± 0.1	1 ± 2	0.103 ± 0.008	0.13 ± 0.04
Sternum	2 ± 1	1.0 ± 0.8	2 ± 1	1 ± 1	0.17 ± 0.06	0.15 ± 0.04
Right Breast	1.3 ± 0.3	2.4 ± 0.8	2.1 ± 0.4	3.6 ± 0.8	0.21 ± 0.03	0.4 ± 0.1
CTVp	1.0 ± 0.3	2.5 ± 0.9	1.6 ± 0.5	4 ± 1	0.20 ± 0.04	0.4 ± 0.1
CTVn	2 ± 1		3 ± 1		0.4 ± 0.1	
L1	1.7 ± 0.7		2.1 ± 0.7		0.4 ± 0.2	
L2	2.5 ± 1.0		3.0 ± 0.9		0.8 ± 0.2	
L3	0.8 ± 0.2		1.0 ± 0.2		0.24 ± 0.04	
L4	0.7 ± 0.2		0.9 ± 0.2		0.22 ± 0.04	
Pectoral	3 ± 1		3 ± 1		0.6 ± 0.4	

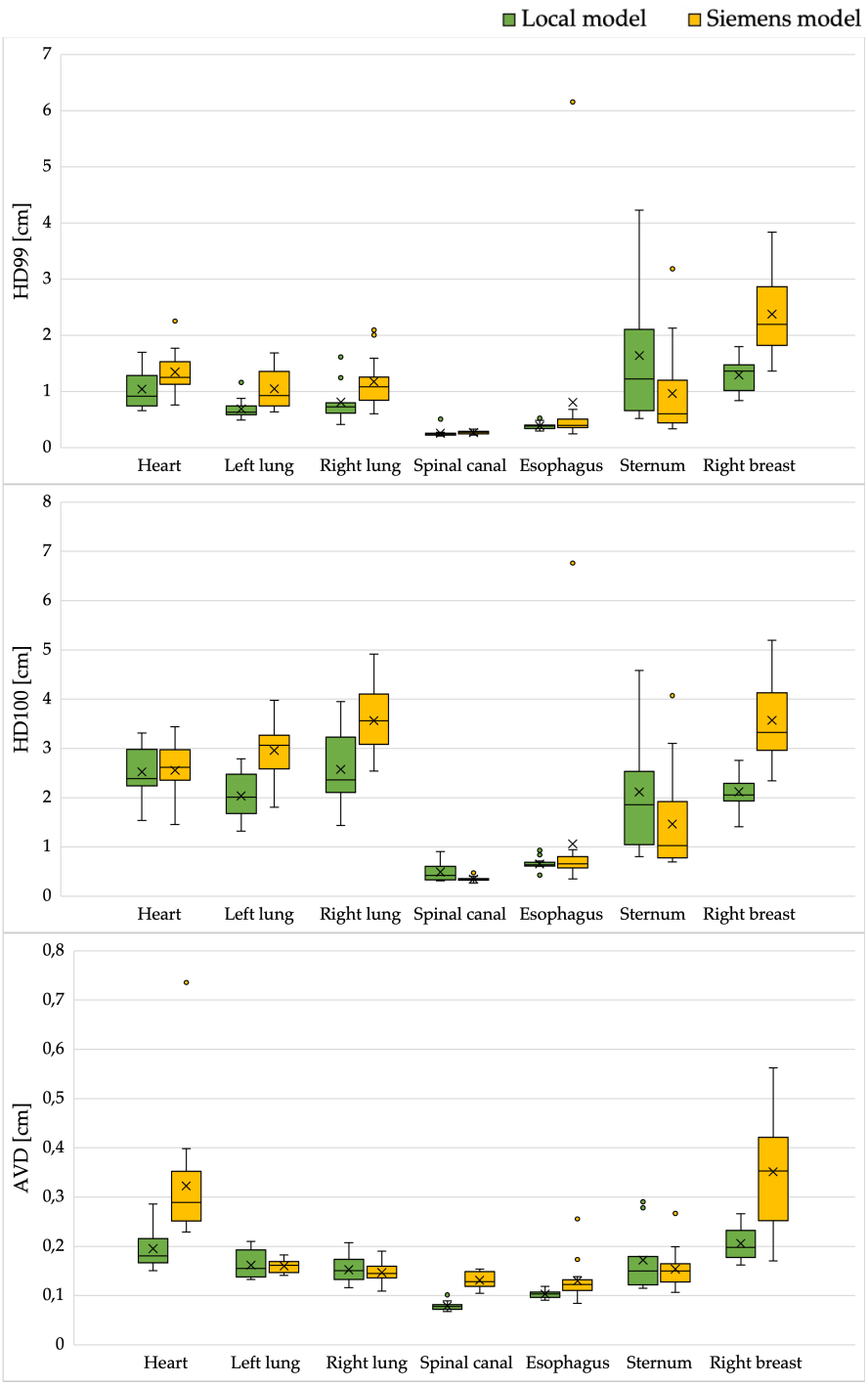


Figure D.4: HD99, HD100, and AVD obtained by both segmentation models for all organs at risk.

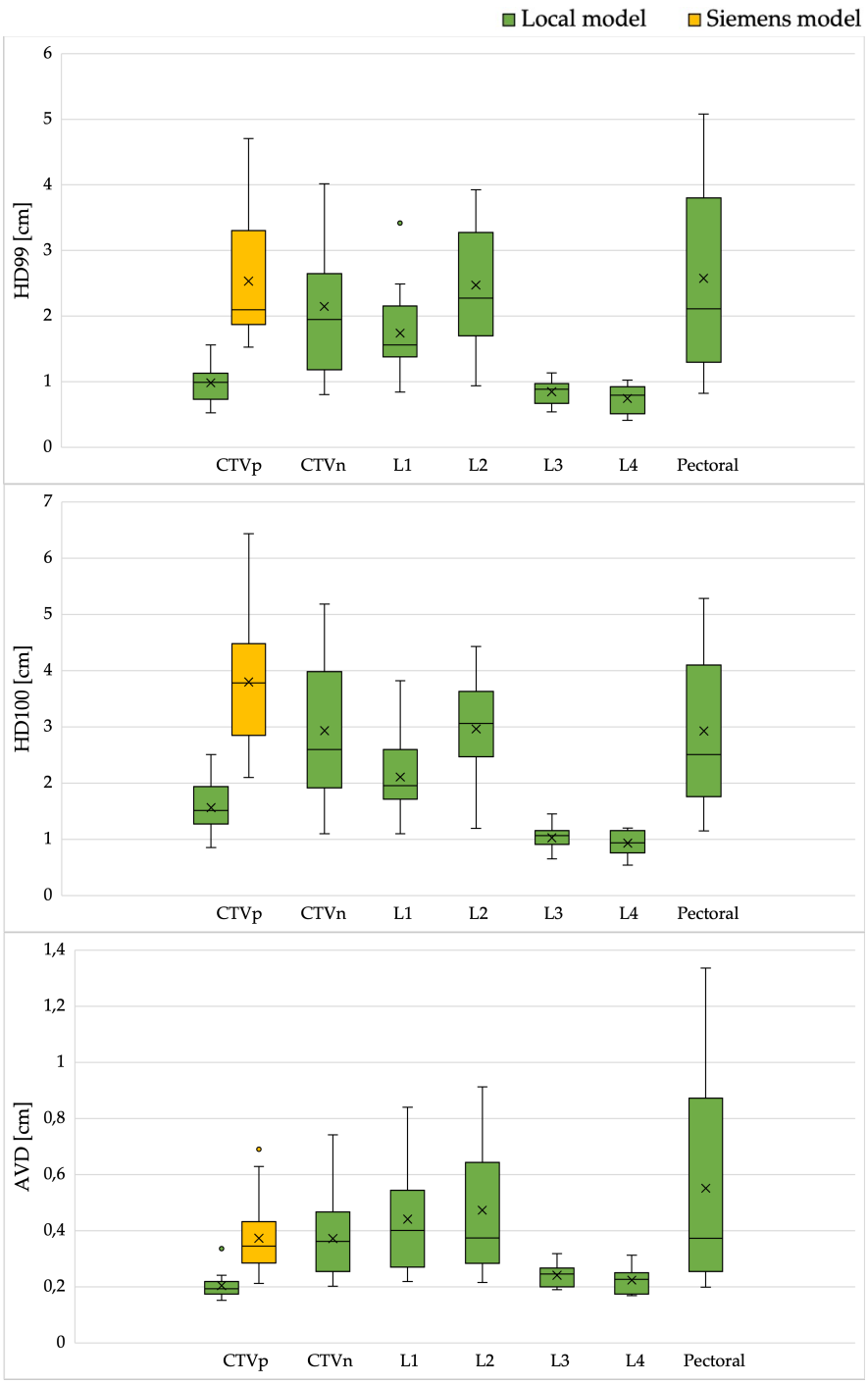


Figure D.5: HD99, HD100, and AVD obtained by both segmentation models for all target volumes.

D.2 Dosimetric evaluation

The mean values with SD for the additional dosimetric metrics can be found in table D.3 and D.4 for the local model and Siemens model, respectively. The dosimetric metrics for the individual lymph node areas can also be found in figure D.5.

Table D.3: Additional mean values of the dosimetric metrics for the local model. The treatment plans are made based on automatic segmentations by the local model. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Manual del.	Local model	Pairwise difference (Manual - model)	P-value
Heart	D2 [Gy]	4 ± 1	4 ± 1	-0.2 ± 0.2	**0.005**
	V20Gy [%]	0.0 ± 0.1	0.0 ± 0.1	0.00 ± 0.01	-
Left lung	D1cm3 [Gy]	40.9 ± 0.3	40.6 ± 0.4	0.3 ± 0.4	**0.017**
Right lung	D1cm3 [Gy]	5 ± 1	4.3 ± 0.9	0.3 ± 0.2	**0.001**
Right breast	V3Gy [%]	2 ± 2	2 ± 2	0 ± 1	0.078
	D1cm3 [Gy]	8 ± 4	12 ± 6	-3 ± 6	0.088
Esophagus	Dmean [Gy]	2.2 ± 0.2	2.2 ± 0.2	0.0 ± 0.1	0.427
CTVp	Dmean [Gy]	40.14 ± 0.04	40.12 ± 0.03	0.02 ± 0.02	**0.023**
	D2 [Gy]	41.4 ± 0.1	41.4 ± 0.2	0.03 ± 0.05	**0.011**
	V90 [%]	99.98 ± 0.04	99.9999 ± 0.0001	-0.02 ± 0.04	**0.041**
	V105 [%]	0.05 ± 0.06	0.04 ± 0.05	0.01 ± 0.02	**0.009**
	HI [1]	1.07 ± 0.01	1.068 ± 0.008	0.000 ± 0.004	1.000
CTVn	Dmean [Gy]	39.9 ± 0.1	40.03 ± 0.06	-0.1 ± 0.1	**0.023**
	D2 [Gy]	41.4 ± 0.1	41.1 ± 0.2	0.3 ± 0.2	**0.001**
	V90 [%]	99.5 ± 0.9	100.0 ± 0.0	-0.5 ± 0.9	**0.001**
	V105 [%]	0.05 ± 0.06	0.04 ± 0.05	0.02 ± 0.02	**0.009**
	HI [1]	1.09 ± 0.03	1.052 ± 0.005	0.04 ± 0.03	**0.001**
PTVpc	D2 [Gy]	41.59 ± 0.06	41.59 ± 0.06	-0.008 ± 0.007	**0.003**
	V90 [%]	98.6 ± 0.8	99.76 ± 0.06	-1.1 ± 0.9	**0.001**
	V105 [%]	0.1 ± 0.1	0.1 ± 0.1	-0.004 ± 0.008	0.069
	CI [1]	0.61 ± 0.07	0.62 ± 0.08	-0.01 ± 0.02	0.069
PTVnc	D2 [Gy]	41.61 ± 0.07	41.65 ± 0.06	-0.04 ± 0.06	**0.031**
	V90 [%]	96 ± 2	99.83 ± 0.04	-3 ± 2	**0.001**
	V105 [%]	0.24 ± 0.08	0.4 ± 0.1	-0.1 ± 0.1	**0.001**
	CI [1]	0.31 ± 0.06	0.28 ± 0.07	0.03 ± 0.02	**0.001**

Table D.4: Additional mean values of the dosimetric metrics for the Siemens model. The treatment plans are made based on manual delineations. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Manual del.	Siemens model	Pairwise difference (Manual - model)	P-value
Heart	D2 [Gy]	4 ± 2	5 ± 3	-1 ± 2	**0.001**
	V20Gy [%]	0.1 ± 0.2	0.2 ± 0.3	-0.1 ± 0.2	0.169
Left lung	D1cm3 [Gy]	41.0 ± 0.3	40.9 ± 0.2	0.2 ± 0.3	**0.012**
Right lung	D1cm3 [Gy]	4.1 ± 0.9	3.9 ± 0.9	0.2 ± 0.1	**0.002**
Right breast	V3Gy [%]	2 ± 2	4 ± 2	-2 ± 2	**0.006**
	D1cm3 [Gy]	9 ± 5	14 ± 7	-4 ± 7	**0.015**
Esophagus	Dmean [Gy]	2.3 ± 0.4	2.4 ± 0.4	-0.1 ± 0.1	**0.020**
CTVp	Dmean [Gy]	40.11 ± 0.03	39.7 ± 0.3	0.4 ± 0.3	**0.001**
	D2 [Gy]	41.4 ± 0.1	41.4 ± 0.1	0.00 ± 0.04	0.570
	V90 [%]	100 ± 0	98 ± 2	2 ± 2	**0.001**
	V105 [%]	0.02 ± 0.02	0.03 ± 0.04	-0.02 ± 0.03	**0.019**
	HI [1]	1.065 ± 0.005	1.2 ± 0.3	-0.2 ± 0.3	**0.001**

Table D.5: Mean values of the dosimetric metrics for the lymph node areas segmented by the local model. The treatment plans are made based on automatic segmentations by the local model. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Manual del.	Local model	Pairwise difference (Manual - model)	P-value
L1	Dmean [Gy]	39.9 ± 0.2	40.06 ± 0.05	-0.2 ± 0.2	**0.002**
	D2 [Gy]	41.2 ± 0.2	41.1 ± 0.2	0.1 ± 0.2	0.078
	V90 [%]	99 ± 1	100 ± 0	-1 ± 1	**0.003**
	V105 [%]	0.03 ± 0.07	0.01 ± 0.05	0.01 ± 0.04	0.263
	HI [1]	1.09 ± 0.05	1.046 ± 0.006	0.04 ± 0.05	**0.001**
L2	Dmean [Gy]	40.1 ± 0.1	40.01 ± 0.09	0.10 ± 0.09	**0.002**
	D2 [Gy]	41.5 ± 0.2	41.0 ± 0.2	0.4 ± 0.3	**0.001**
	V90 [%]	99.9 ± 0.4	100 ± 0	-0.1 ± 0.4	-
	V105 [%]	0.04 ± 0.09	0.001 ± 0.003	0.04 ± 0.08	**0.003**
	HI [1]	1.07 ± 0.02	1.048 ± 0.009	0.02 ± 0.02	**0.001**
L3	Dmean [Gy]	39.9 ± 0.1	39.9 ± 0.1	-0.01 ± 0.05	0.570
	D2 [Gy]	40.8 ± 0.3	40.7 ± 0.3	0.02 ± 0.09	0.650
	V90 [%]	100 ± 0	100 ± 0	0 ± 0	-
	V105 [%]	0 ± 0	0 ± 0	0 ± 0	-
	HI [1]	1.05 ± 0.01	1.05 ± 0.01	0.004 ± 0.009	0.173
L4	Dmean [Gy]	40.2 ± 0.2	40.1 ± 0.1	0.06 ± 0.09	**0.011**
	D2 [Gy]	41.7 ± 0.2	41.5 ± 0.2	0.2 ± 0.2	-
	V90 [%]	99.9 ± 0.3	100 ± 0	-0.1 ± 0.3	-
	V105 [%]	0.5 ± 0.4	0.1 ± 0.2	0.4 ± 0.4	**0.004**
	HI [1]	1.08 ± 0.01	1.068 ± 0.008	0.01 ± 0.01	0.053
Pectoral	Dmean [Gy]	40.1 ± 0.2	39.9 ± 0.1	0.2 ± 0.1	**0.001**
	D2 [Gy]	41.6 ± 0.3	41.0 ± 0.3	0.7 ± 0.4	**0.001**
	V90 [%]	100 ± 0	100 ± 0	0 ± 0	-
	V105 [%]	0.1 ± 0.2	0.0 ± 0.1	0.1 ± 0.2	**0.043**
	HI [1]	1.08 ± 0.01	1.06 ± 0.01	0.02 ± 0.01	**0.001**

E Additional results from validation of automatic plan optimization

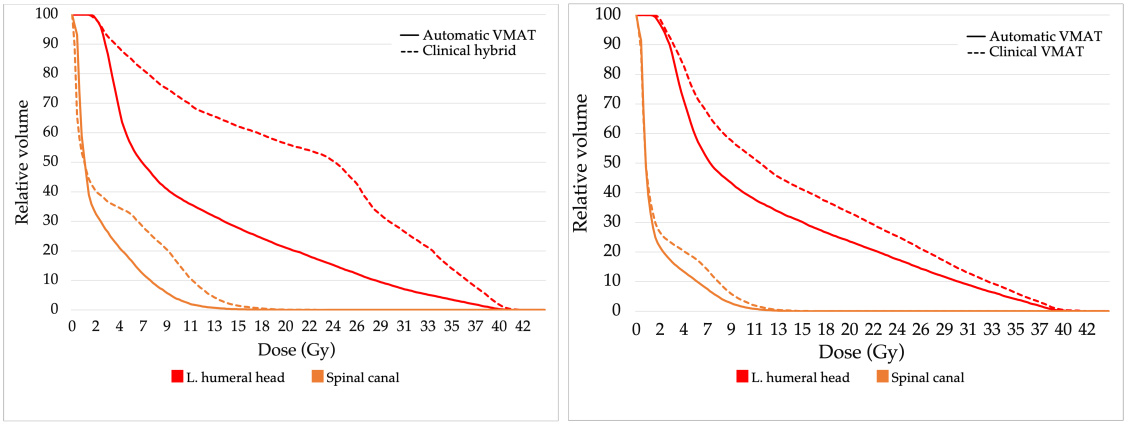
Additional dosimetric metrics for the validation of the automatic plan optimization can be found in this appendix. The mean values with SD can be found in tables E.1 and E.2 for the hybrid and VMAT plan patients, respectively. The DVH curves for the left humeral head and spinal canal can be seen in figure E.1.

Table E.1: Additional mean dosimetric metrics for the automatic VMAT plans and the clinical hybrid plans. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Automatic VMAT	Clinical hybrid	Pairwise difference (Clinical - Automatic)	P-value
CTVp	Dmean [Gy]	40.13 ± 0.06	40.16 ± 0.09	0.0 ± 0.1	0.237
	D2 [Gy]	41.3 ± 0.2	41.5 ± 0.2	0.2 ± 0.3	0.176
	V90 [%]	99.999 ± 0.002	99.997 ± 0.004	-0.002 ± 0.004	0.500
	HI [1]	1.065 ± 0.009	1.072 ± 0.006	0.007 ± 0.009	0.063
CTVn	Dmean [Gy]	40.06 ± 0.02	40.2 ± 0.1	0.1 ± 0.1	0.128
	D2 [Gy]	41.2 ± 0.2	41.4 ± 0.3	0.2 ± 0.4	0.310
	V90 [%]	100 ± 0	99.99 ± 0.01	-0.01 ± 0.01	-
	HI [1]	1.053 ± 0.007	1.06 ± 0.01	0.01 ± 0.02	0.398
PTVpc	D2 [Gy]	41.6 ± 0.2	41.52 ± 0.05	-0.1 ± 0.2	0.612
	V90 [%]	99.72 ± 0.03	99.92 ± 0.08	0.20 ± 0.09	**0.018**
	CI [1]	0.6 ± 0.1	0.4 ± 0.1	-0.11 ± 0.03	**0.018**
PTVnc	D2 [Gy]	41.7 ± 0.1	41.5 ± 0.2	-0.2 ± 0.2	**0.043**
	V90 [%]	99.86 ± 0.04	99.99 ± 0.02	0.13 ± 0.03	**0.018**
	CI [1]	0.30 ± 0.05	0.24 ± 0.03	-0.06 ± 0.02	**0.018**
Body	Dmax [Gy]	42.5 ± 0.3	42.4 ± 0.4	-0.1 ± 0.5	0.237
Heart	D2 [Gy]	4 ± 2	12 ± 6	7 ± 5	**0.018**
	V20Gy [%]	0.1 ± 0.2	1.1 ± 0.7	1.0 ± 0.6	-
Right Breast	V3Gy [%]	1 ± 2	3 ± 3	1 ± 2	0.612
	D1cm3 [Gy]	7 ± 8	9 ± 12	2 ± 6	0.398
Left Lung	D1cm3 [Gy]	40.5 ± 0.9	40.9 ± 0.3	0.4 ± 0.8	0.176
Right Lung	D1cm3 [Gy]	4 ± 2	8 ± 2	5 ± 1	**0.018**
Esophagus	Dmean [Gy]	3 ± 1	8 ± 6	6 ± 5	**0.018**
Thyroid	Dmean [Gy]	12 ± 6	21 ± 8	9 ± 5	**0.018**
L. humeral head	Dmean [Gy]	11 ± 4	21 ± 4	10 ± 4	**0.018**
	D1cm3 [Gy]	33 ± 6	39 ± 2	6 ± 5	**0.018**
Spinal canal	Dmax [Gy]	14 ± 4	17 ± 5	3 ± 5	0.176

Table E.2: Additional mean dosimetric metrics for the automatic VMAT plans and the clinical VMAT plans. SD is denoted as the ± value. P-values in bold font are considered statistically significant and "-" indicates not enough data points to calculate p-value.

Region	Metric	Automatic VMAT	Clinical VMAT	Pairwise difference (Clinical - Automatic)	P-value
CTVp	Dmean [Gy]	40.14 ± 0.04	40.16 ± 0.04	0.02 ± 0.07	0.515
	D2 [Gy]	41.5 ± 0.2	41.4 ± 0.2	-0.1 ± 0.2	0.594
	V90 [%]	99.999 ± 0.001	99.998 ± 0.005	-0.002 ± 0.004	0.575
	HI [1]	1.0727 ± 0.0093	1.07 ± 0.01	-0.01 ± 0.01	0.214
CTVn	Dmean [Gy]	40.06 ± 0.03	40.0 ± 0.1	0.0 ± 0.1	0.214
	D2 [Gy]	41.2 ± 0.2	40.9 ± 0.2	-0.3 ± 0.3	**0.021**
	V90 [%]	99.998 ± 0.006	99.998 ± 0.007	0.000 ± 0.001	-
	HI [1]	1.053 ± 0.009	1.04 ± 0.01	-0.01 ± 0.01	0.066
PTVpc	D2 [Gy]	41.6 ± 0.1	41.6 ± 0.2	-0.1 ± 0.2	0.314
	V90 [%]	99.69 ± 0.08	99.7 ± 0.3	0.0 ± 0.3	0.515
	CI [1]	0.68 ± 0.06	0.64 ± 0.05	-0.03 ± 0.01	**0.008**
PTVnc	D2 [Gy]	41.64 ± 0.09	41.4 ± 0.2	-0.2 ± 0.2	**0.021**
	V90 [%]	99.83 ± 0.03	99.9 ± 0.2	0.1 ± 0.2	0.110
	CI [1]	0.23 ± 0.06	0.22 ± 0.06	-0.009 ± 0.008	**0.021**
Body	Dmax [Gy]	42.6 ± 0.1	42.5 ± 0.3	0.0 ± 0.3	0.767
Heart	D2 [Gy]	8 ± 5	8 ± 5	0 ± 2	0.859
	V20Gy [%]	0.4 ± 0.4	0.3 ± 0.5	0.0 ± 0.2	0.735
Right Breast	V3Gy [%]	5 ± 4	8 ± 10	3 ± 6	0.139
	D1cm3 [Gy]	15 ± 7	15 ± 9	1 ± 4	0.515
Left Lung	D1cm3 [Gy]	39.9 ± 0.8	40.3 ± 0.5	0.4 ± 0.8	0.139
Right Lung	D1cm3 [Gy]	2.9 ± 0.5	6 ± 2	3 ± 2	**0.008**
Esophagus	Dmean [Gy]	3 ± 1	8 ± 5	4 ± 4	**0.008**
Thyroid	Dmean [Gy]	10 ± 4	16 ± 4	6 ± 3	**0.008**
L. humeral head	Dmean [Gy]	12 ± 5	15 ± 5	3 ± 3	**0.021**
	D1cm3 [Gy]	32 ± 10	34 ± 8	2 ± 2	**0.011**
Spinal canal	Dmax [Gy]	13 ± 3	12 ± 3	0 ± 3	0.441



(a) Compared to hybrid plans

(b) Compared to VMAT plans

Figure E.1: Cumulative DVH for the left humeral head and spinal canal.

F Main results from the project thesis

Results from the evaluation of the first version of the local segmentation model from the project thesis in 2020, together with the inter-observer variability found. The model was trained on 45 patients, tested on 17 patients, and the model segmentations were geometrically compared to manual delineations. Inter-observer variability was calculated pairwise between four physicians for five patient cases. One physician did not delineate the last three cases, giving a total of 21 combinations of delineations to calculate the inter-observer variability from. Inter-observer variability was not evaluated for the lungs. The mean values are presented in table F.1, and boxplots of DSC and HD95 are presented in figures F.1 and F.2, respectively.

Table F.1: DSC and HD95 for the comparison of segmentations made by the previous local model and manual delineations and the inter-observer variability (IOV). Mean values ± standard deviation.

	DSC		HD95 [cm]	
	IOV	Local model	IOV	Local model
Heart	0.961 ± 0.008	0.96 ± 0.02	0.39 ± 0.06	0.5 ± 0.2
Spinal canal	0.90 ± 0.02	0.94 ± 0.02	0.21 ± 0.03	0.21 ± 0.02
Esophagus	0.85 ± 0.03	0.88 ± 0.02	0.25 ± 0.05	0.24 ± 0.03
Left lung		0.985 ± 0.002		0.31 ± 0.01
Right lung		0.987 ± 0.001		0.305 ± 0.007

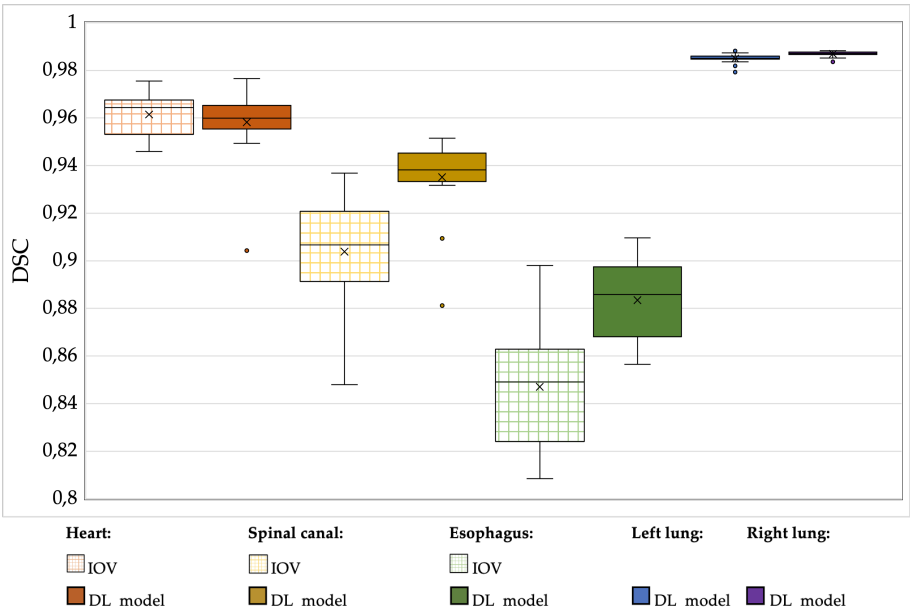


Figure F.1: DSC for inter-observer variability (IOV) and the first version of the local.

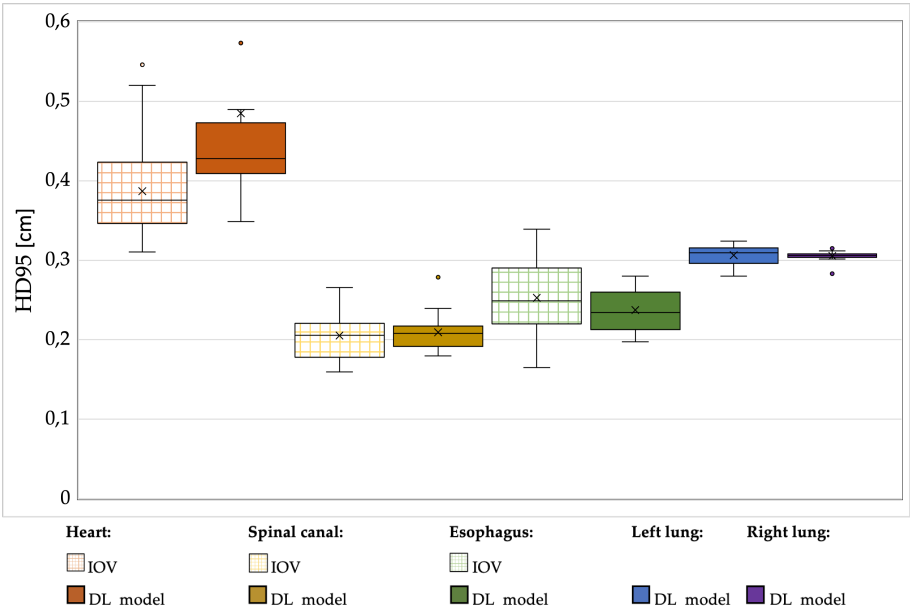


Figure F.2: HD95 for inter-observer variability (IOV) and the first version of the local.

Jeanette Bonden Isachsen

DL image segmentation and automatic treatment planning in breast cancer radiotherapy

NTNU
Norwegian University of
Science and Technology

ST. OLAVS HOSPITAL
TRONDHEIM UNIVERSITY HOSPITAL