catharina
een santeon ziekenhuis

SMPE/e SCHOOL OF MEDICAL PHYSICS AND ENGINEERING EINDHOVEN

Design of an Automated Process for Creating Breast Cancer Treatment Plans, with Artificial Intelligence

Development and Clinical Introduction of AI Models to Automate Segmentation and Planning for Breast Cancer Radiation Treatment Plans

Ir. N.L.M. (Nienke) Bakx

September, 2022

Design of an Automated Process for Creating Breast Cancer Treatment Plans, with Artificial Intelligence

Development and Clinical Introduction of AI Models to Automate Segmentation and Planning for Breast Cancer Radiation Treatment Plans

executed at

Catharina Hospital Eindhoven

By
Ir. N.L.M. (Nienke) Bakx

Guided by
Dr. Ing. C.W. (Coen) Hurkmans
Dr. M.A.J.M. (Maureen) van Eijnatten
Dr. Ir. I.M.M. (Ivonne) Lammerts

PDEng report number: 2022/029

Confidential

□ Yes

☑ No

The work described in this report is executed in accordance with the TU/e Code of Scientific Conduct.

One year project presented to Eindhoven University of Technology towards the degree of Engineering Doctorate in Qualified Medical Engineer.

This Project is partly funded by TKI HTSM

SMPE/e SCHOOL OF MEDICAL PHYSICS AND ENGINEERING EINDHOVEN

TU/e

The PDEng Thesis Evaluation Committee consists of:

Scientific supervisors:	Dr. Ir. I.M.M. (Ivonne) Lammerts
	Dr. M.A.J.M. (Maureen) van Eijnatten
Supervisor healthcare facility:	Dr. Ing. C.W. (Coen) Hurkmans
First external evaluator:	Dr. Ir. R.G.J. (Roel) Kierkels
Second external evaluator:	Dr. Ir. A.J.E. (Alexander) Raaijmakers
Other members:	Dr. M.J.C. (Maurice) van der Sangen
	Dr. Ir. E.J.E. (Ward) Cottaar
Chairman of the Evaluation Committee:	Dr. Ir. I.M.M. (Ivonne) Lammerts

Public summary

Breast cancer is one of the most common cancer types in The Netherlands. Treatment of breast cancer often consists of (breast conserving) therapy, followed by post-operative radiotherapy. In order to perform radiotherapy, a treatment plan needs to be created, for which clinical target volumes (CTVs) and organs at risk (OARs) need to be identified and a dose distribution is calculated. Both steps of the treatment planning process involve iterative and manual actions. Besides the cumbersome nature of these steps, they are prone to the experience of the Radiotherapy Technologist (RTT) and Radiation Oncologist (RO), resulting in inter- and intra-observer variability. The goal of this design project, performed at the department of radiotherapy in the Catharina Hospital Eindhoven (CZE), is to develop and clinically introduce Artificial Intelligence (AI) models to automate the delineation of contours (auto-segmentation) and creation of the dose distribution (auto-planning).

For auto-segmentation, two AI models were developed, trained and evaluated, for both left- and right-sided breast cancer including the lymph nodes. The model training framework was provided by RaySearch Laboratories, including a 3D U-Net architecture. In total, 80 patients were included for training of both models, of which the contours were all visually inspected on abnormalities and corrected by two experienced RTTs and ROs, when needed. In a retrospective study, both models were tested for 15 patients: they showed to fulfill the predefined quantitative requirements for most of the cases. Therefore, a clinical pilot was performed in which the automatically generated contours were qualitatively scored by several RTTs and ROs. Besides, the time needed to automatically create the contours and perform corrections when needed was measured, too. A mean reduction in time of 42% was found for the OARs, while an even larger reduction of 59% was found for the CTVs. Furthermore, 92% of the contours were scored as clinically acceptable or useful for correction, indicating a high usability for clinical practice.

For auto-planning, multiple models were developed, trained and validated for left-sided whole breast radiotherapy. During a previously performed project, two AI models had been trained and retrospectively validated for conventional breast irradiation (40.05 Gy in 15 fractions), using treatment plans of 90 patients. The first model was in-house developed, based on a 2D U-net architecture, whereas the second model was developed by RaySearch Laboratories, and based on a contextual Atlas Regression Forest (cARF). In this design project, both models were validated in a clinical pilot. Manually and automatically created plans were blindly scored by four experienced ROs, and the time to generate these plans was measured, too. Although there was a difference in preferences of the observers, 95% of the 2D U-Net plans were found to be clinically acceptable for all, which was the case in 90% of the manually generated and cARF plans. When only considering user-interaction time, both auto-planning methods showed time efficiency. Following the results of this study, a 3D U-Net was trained by RaySearch, based on the same dataset, and was successfully commissioned for use at the department of radiotherapye in CZE. Hence, since May 2022, the model is used in clinic to generate treatment plans. Besides, a 2D U-Net model was trained and retrospectively validated for fast-forward irradiation (2.6 Gy in 5 fractions). For this model, transfer learning was used, using the 2D U-Net for conventional irradiation as a starting point. This method proved to be promising, with good results while only using a dataset of 52 patients for training, and should be further investigated in the future for clinical use.

In conclusion, in this design project several AI models were successfully developed, trained and validated for delineation of contours and creation of dose distribution for breast cancer. While an auto-planning model was finally actually implemented in clinical practice, the auto-segmentation model showed promising results and will be clinically implemented in the near future.

TU/e EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

Declaration concerning the TU/e Code of Scientific Conduct for the PDEng thesis

I have read the TU/e Code of Scientific Conduct[i].

I hereby declare that my PDEng thesis has been carried out in accordance with the rules of the TU/e Code of Scientific Conduct

Date

26/07/2022……………………………………………..…………..

Name

N.L.M. Bakx……………………………………………..…………..

Signature

Nbakx
……………………………………………………..…………..

[i] See:https://www.tue.nl/en/our-university/about-the-university/organization/integrity/scientific-integrity/
The Netherlands Code of Conduct for Scientific Integrity, endorsed by 6 umbrella organizations, including the VSNU, can be found here also. More information about scientific integrity is published on the websites of TU/e and VSNU

Contents

List of abbreviations

ABS	Atlas Based Segmentation
AI	Artificial Intelligence
ARF	Atlas Regression Forest
cARF	Contextual Atlas Regression Forest
CI	Confidence Interval
CKI	Catharina Cancer Institute
CNN	Convolutional Neural Network
CRF	Conditional Random Field
CTV	Clinical Target Volume
CZE	Catharina Hospital Eindhoven
DL	Deep Learning
DSC	Dice Similarity Coefficient
DVH	Dose-Volume Histogram
FF	Fast-Forward
HD	Hausdorff Distance
HFMEA	Health Care Failure Mode and Effect Analysis
KBP	Knowledge Based Planning
MHD	Mean Heart Dose
ML	Machine Learning
MLC	Multi Leaf Collimator
MLD	Mean Lung Dose
MUs	Monitor Units
OARs	Organs at Risk
pRF	predict Regression Forest
PDF	Probability Distribution Function
PRA	Prospective Risk Analysis
PTV	Planning Target Volume
RF	Regression Forest
RTT	Radiotherapy Technologist
RO	Radiation Oncologist
ROI	Region of Interest
sDSC	Surface DSC
TPS	Treatment Planning System
TU/e	Eindhoven University of Technology

1 Introduction

1.1 Clinical problem

The Catharina Hospital Eindhoven (CZE) is a top-clinical hospital. The department of Radiotherapy, part of the Catharina Cancer Institute (CKI), treats about 4.000 patients a year. One of the most common cancer types in the Netherlands is breast cancer, with more than 18.000 new cases in 2021 [1]. Different stages of breast cancer can be defined, which are classified according to the TNM classification system [2]. This classification is determined by the size and growth of the tumor (T), the involved (regional) nodes (N) and the presence of metastasis (M).

Breast cancer treatment often consists of (breast conserving) surgery, followed by post-operative radiotherapy. During radiotherapy, the tumor and possibly lymph nodes are irradiated by ionising radiation, often photons. The treatment is based on the interaction of the radiation and the tissue, damaging the cells which eventually leads to cell death. Cells which are actively dividing, which is the case for cancer cells, are more sensitive to this effect than less active cells, such as healthy tissue cells [3]. Although the effect is thus less harmful for healthy cells, it is still of utmost importance to spare the surrounding healthy tissue as much as possible. One of the differences between healthy and malignant cells which can be utilized to spare healthy tissue, is the ability to repair DNA damage, which is greater in healthy cells. By splitting the total radiation dose into multiple fractions, healthy cells can repair damage while malignant cells are less able to recover from radiation damage. Besides, a fractionated dose is more effective to kill tumor cells, as the radio sensitivity of cells depend on their stage in the cell cycle, and not all tumor cells are in the same stage at a time. Therefore, the total dose needed is divided in multiple dose fractions [3]. The malignant and healthy regions are identified using imaging techniques, after which the treatment planning process is started. During this process, the target and healthy tissues are identified and a dose distribution is created. Section 1.2 further elaborates on this process. A lot of iterative and manual steps are involved in the planning process, of which the outcome is dependent on the experience of the Radiotherapy Technologist (RTT) and Radiation Oncologist (RO). By automating different steps, time saving can be achieved, while maintaining (or improving) quality, and decreasing intra- and inter-observer variability. Automation of the process can be achieved with the help of Artificial Intelligence (AI). The principle of AI will be further explained in section 1.3. Section 1.4 will elaborate on the possibilities of AI within the field of radiotherapy.

1.2 Radiotherapy treatment plans

Whenever a patient arrives at the radiotherapy department, several steps are followed before the actual irradiation treatment starts, which is visualized in Figure 1.1. In this project, the focus is on the segmentation and plan optimization steps, which are both executed in the Treatment Planning System (TPS) RayStation (RaySearch laboratories, AB, Sweden).

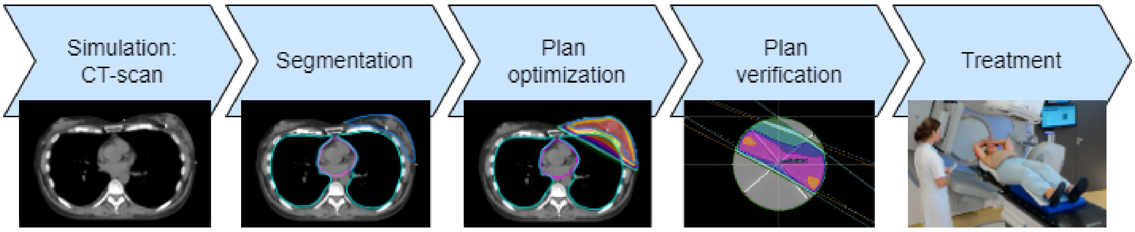


Figure 1.1: *The process of radiotherapy treatment planning.*

During the process of segmentation, the target volume(s) (Clinical Target Volume (CTV)) and surrounding organs (Organs at Risk (OARs)) are delineated, such that these can be used during plan optimization to calculate the dose delivered. CTVs are delineated by the ROs. Thereafter, the Planning Target Volume (PTV) is created by adding a margin to prevent under-dosage, caused by variations in patient position and movement during treatment. In the case of breast irradiation, the CTV of the full breast is expanded by 5 mm, and of the node levels with 7 mm. In addition, the PTVs are cropped 5 mm under the skin. The OARs are delineated by the RTTs. In case of breast irradiation including node levels, these OARs are the lungs, heart, (part of) the esophagus, the humerus (humeral head) and thyroid. Subsequently, the RTT creates a radiation treatment plan within the TPS in the plan optimization step. During this step, the dose distribution is calculated, and corresponding treatment machine parameters are determined. These machine parameters contain among others the gantry angle of the machine and the shape of the segments of the Multi Leaf Collimator (MLC). A MLC contains several leaves, which can be shifted to vary the shape and intensity of the beam, as can be seen in Figure 1.2. When a treatment plan is optimized, a trade-off is made between dose delivered to the CTVs and OARs. In order to do so, the RTT can tweak objectives about e.g. the dose homogeneity in the PTV or maximum allowed dose to an organ. The treatment plan is eventually evaluated with pre-defined clincial goals, which make demands on the coverage of the PTV and maximum allowed dose to the OARs.

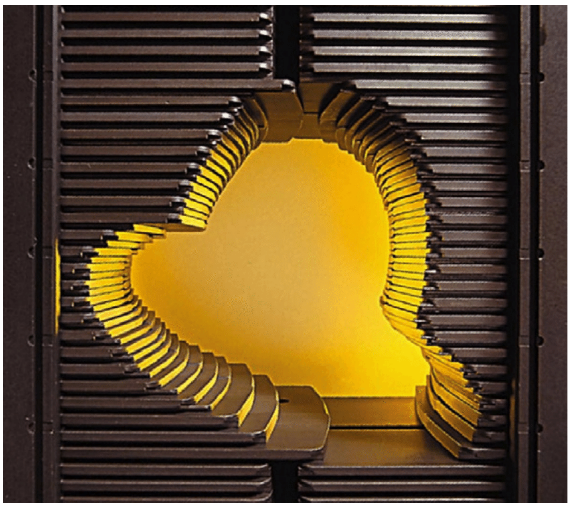


Figure 1.2: *Example of a multileaf collimator. Image courtesy of Varian Medical Systems, Inc. All rights reserved. Source: https://bit.ly/3FA0B56*

1.3 Artificial Intelligence (AI)

The past few years, the use of AI in health care has strongly increased. AI systems are trained to perform tasks on the level of human intelligence, and are furthermore able to keep developing by learning from former actions taken. Several examples of AI in daily clinical practice are support for diagnosis, selection of treatment, automated surgery or support in patient monitoring [4].
Two terms associated with AI are Machine Learning (ML) and Deep Learning (DL), of which the relation is visualized in Figure 1.3. ML is a subset of AI, in which patterns and relations in data are discovered with the help of computer algorithms, utilizing different statistical models. DL is a subset of ML, which discovers these relationships in large (raw) datasets, using (deep) neural networks. These networks mimic the way the human brain operates, which is by done by using connections between neurons [4, 5]. The most common used neural network is a Convolutional Neural Network (CNN). A CNN consists of several layers, which each perform a mathematical operation on the input of that

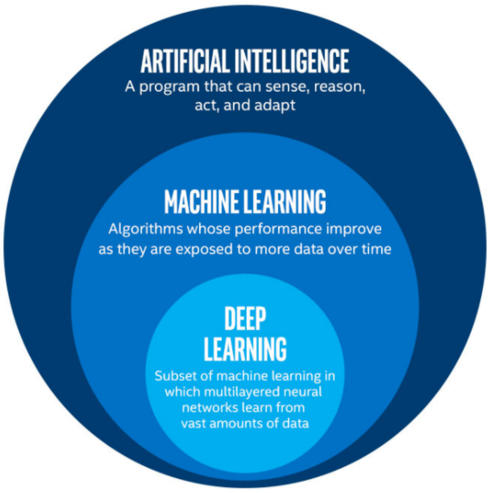

ARTIFICIAL INTELLIGENCE
A program that can sense, reason, act, and adapt
MACHINE LEARNING
Algorithms whose performance improve as they are exposed to more data over time
DEEP LEARNING
Subset of machine learning in which multilayered neural networks learn from vast amounts of data

__Figure 1.3:__ The relation between Artificial Intelligence, Machine Learning and Deep Learning.

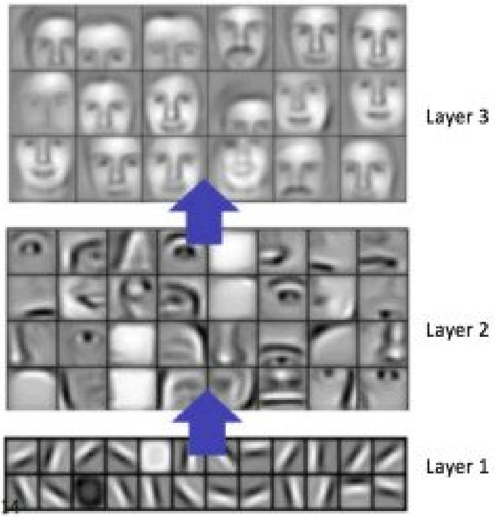


__Figure 1.4:__ Learned features from a Convolutional Neural Network. Image retrieved from [7]

layer, to find different features in the input. These features are important to determine patterns and relationships in the input data [6]. An example of different features, such as lines and transitions, that can be discovered in the different layers for face recognition is shown in Figure 1.4. However, in contrast to the features in Figure 1.4, a CNN will also discover features which cannot be interpreted by humans.

1.4 Artificial Intelligence in Radiotherapy

The increase of the use of AI is also reflected in the field of radiotherapy [8]. Figure 1.5 shows at which moments in the patient-workflow AI can be applied, i.e. nearly everywhere in the workflow. The focus in this QME design project lies on automatic delineation of targets and OARs (auto-segmentation) and automatic plan optimization (auto-planning), that is dose prediction, but also involves dose mimicking, which is not shown in the figure.

Auto-segmentation of contours is a well-studied subject in the field of medical imaging, and a lot of methods have been developed and evaluated in the past few years. A traditional ML method for automatic delineation in radiotherapy is Atlas Based Segmentation (ABS), which is available in several commercial software packages [9]. ABS utilizes prior knowledge by using an atlas of previously contoured images to automatically delineate contours for a new patient, with the help of a spatial transformation of the atlas images. However, the use of CNNs is increasing, as it has shown great potential [9, 10].

The plan optimization step can be automated by several methods. A widely-used term is Knowledge Based Planning (KBP), which includes all methods that use prior knowledge to generate a treatment plan. The outcome of a KBP method can be both a Dose-Volume Histogram (DVH) or a full dose distribution [10]. ML models based on patient geometry features are mostly used for predicting DVHs, wehereas dose distributions are typically predicted using either ML or DL models.

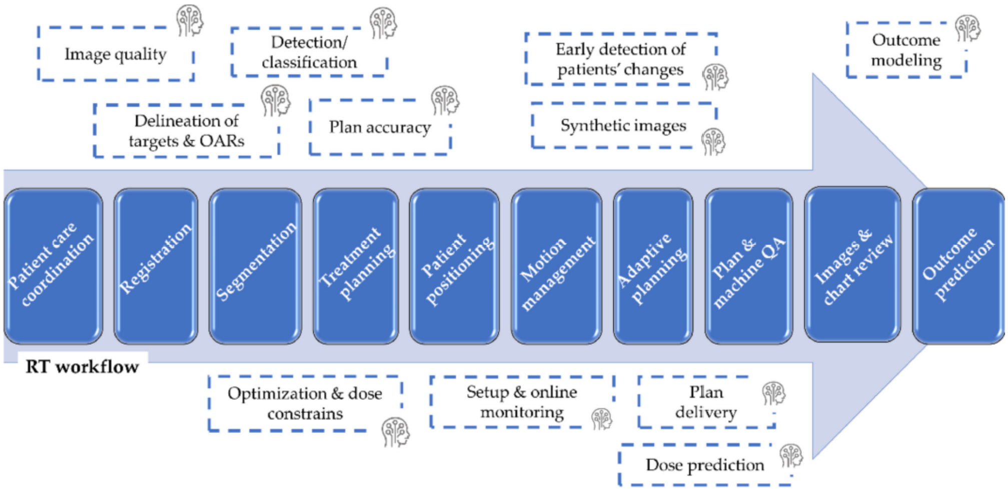


Figure 1.5: *Possible applications of AI within the process of radiation treatment planning. Image retrieved from [8]*

1.5 Outline

This report describes the design process that has been performed and completed in order to achieve a clinical process in which the breast cancer treatment plans are created automatically with AI. In Chapter 2, the project structure and its goal is clarified, after which Chapter 3 further elaborates on the requirements of the final design. Chapter 4 describes several choices made concerning data and AI models used for the purpose of the project, and the final choices and clinical implementation are described in Chapter 5. Thereafter, the product is validated and verified in Chapter 6 to test if it meets the set up requirements. Chapter 7 briefly summarizes the project and its outcomes and concludes about the implementation, followed by Chapter 8, that discusses the choices made in the project and gives recommendations for further research and projects for successful application of AI within the radiotherapy department. Finally, Chapter 9 contains the reflection of the executed project as well as a personal reflection on the process.

2 | Project Definition

2.1 Introduction and Project goal

Currently, several research projects to investigate the use of AI are running or recently finished at the radiotherapy department of CZE. The aim of these projects is to improve the process of creating a treatment plan. This improvement is mostly in terms of time efficiency, but is also aimed to decrease the intra- and interobserver variability. Several studies show that, both in delineation as in plan optimization, this variability is present due to e.g. the varying experience of the RTT and RO [11–14]. The studies at the department are executed in close collaboration with RaySearch. Because of this collaboration, it is possible to implement and evaluate in-house developed and trained AI models in the TPS. Moreover, AI models developed by RaySearch are evaluated.

One of the first studies at the CZE, in close collaboration with Eindhoven University of Technology (TU/e), comprised the evaluation of two ML models for dose prediction for conventional treatment of the whole breast [15]. One of these models was in-house developed, the other one was developed by RaySearch. This retrospective study showed that both models had potential for clinical use. More details can be found in Appendix A. However, more research was needed for clinical implementation. In addition, next to the conventional treatment, consisting of 15 fractions of 2.67 Gy, a new treatment fraction scheme was introduced in our clinic. This scheme, referred to as Fast-Forward (FF), consists of only 5 fractions of 5.2 Gy [16]. It was desired to study the possibility to extend the existing model to fit this new scheme. Finally, there was the desire to study the clinical potential of automatic segmentation for patients with breast cancer, including lymph nodes, as this is a frequent and time consuming step that is taken.

Hence, the goal of this design project is defined as *"Automation of the treatment planning process for breast irradiation using AI for segmentation of contours and treatment plan optimization using dose prediction"*. AI models will be developed and/or trained and validated with the help of clinical data. Furthermore, these models will be tested in a retrospective study, followed by a clinical pilot at the department of radiotherapy to evaluate the clinical potential. Finally, if good results are obtained in the clinical pilot, it is the intention to really implement them in clinical practice in the CZE.

2.2 Project organisation

This design project involved multiple stakeholders, which are all visualized in Figure 2.1. The roles of these stakeholders can be divided into the following categories:

- Project manager: responsible for planning, organizing and monitoring the project
  - Nienke Bakx
- Supplier: supplies knowledge, manpower and means needed for the project
  - Coen Hurkmans: supplier and also client on behalf of the CZE, as part of the management team of the department of radiotherapy
  - Fredrik Löfman: supplier of knowledge, means and manpower of RaySearch Laboratories in his role as Director of ML
- User: end-users of the design, as in the end the automation by AI models will be implemented in their workflow
  - RO: Maurice van der Sangen & Jacqueline Theuws: experienced ROs, specialized in breast cancer treatment
  - RTT: Thérèse van Nunen & Jorien van der Leer: experienced RTTs, specialized in breast cancer treatment

- Project team members: are involved during the project and perform tasks
  - Medical physicists: Coen Hurkmans & Hanneke Bluemink: support with knowledge of the workflow, take part in discussion of the different concepts and analyze and evaluate the results
  - Technical support: Els Hagelaar (project assistant) & Dave van Gruijthuijsen (Medical Engineer): support with knowledge and help with implementation of the AI models in the workflow

Of course, the project team members were involved from the start of the project and weekly meetings were held to discuss progress of the project. In these meetings, also other closely related projects were discussed. However, it is also important to involve the users early in the project, as you need to fulfill their needs. Therefore, several multi-disciplinary meetings were organized with the ROs, RTTs, medical physicists and the project manager. Also, in several other steps the end-users were involved. For example, work instructions which were provided for all ROs and RTTs to execute the clinical studies were first tested by the above-mentioned users. Furthermore, these users also checked the data used for the AI models, as will be further explained in Sections 4.2.1 and 4.3.1. In addition to the team members and users of the CZE, close contact was maintained with two ML engineers of RaySearch laboratories. They provided support with the training and validation of the AI models and enable fast implementation of the models in the TPS.

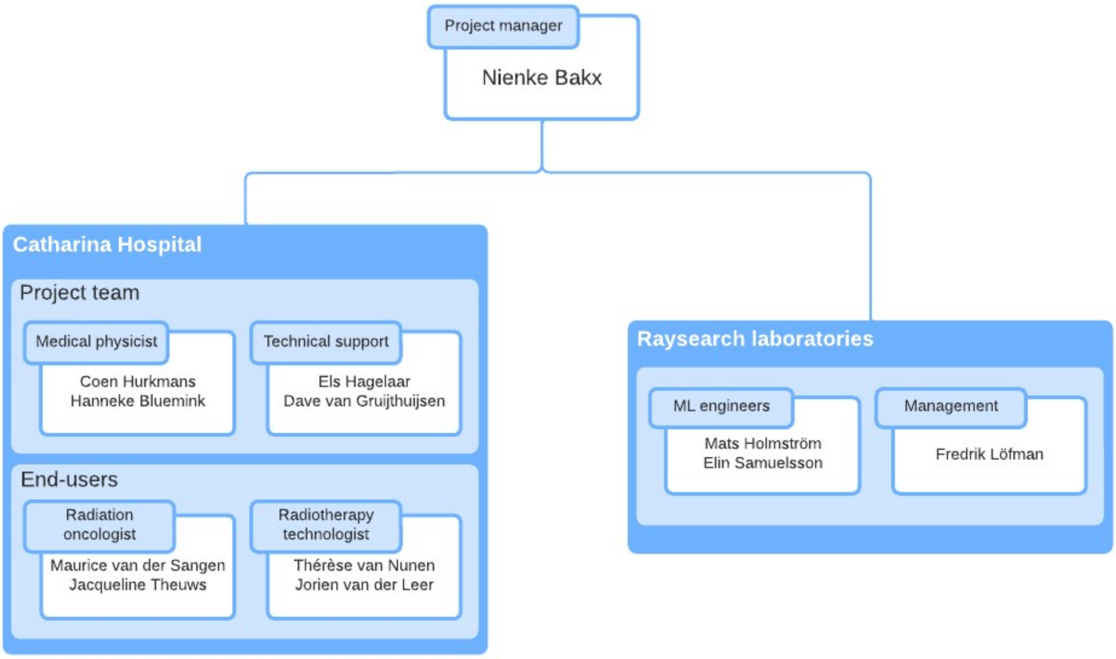


Figure 2.1: *Project organization*

2.3 Deliverables and non-deliverables

In this design project, the treatment planning process for breast irradiation will be automated, by using AI models for both auto-segmentation and auto-planning. To scope the project, the following deliverables and non-deliverables were defined at the start of the project:

Deliverables

- Auto-segmentation
  - Retrospective evaluation of AI model on clinical data
  - Clinical validation of AI model in a pilot study
- Auto-planning
  - Retrospective evaluation of AI dose prediction model for FF breast irradiation
  - Clinical validation of AI dose prediction model for conventional breast irradiation in a pilot study
  - Clinical implementation of AI dose prediction model at the department of radiotherapy in CZE

Non-deliverables

- Clinical implementation of AI dose prediction models for other patient groups than left-sided conventional breast irradiation
- Clinical implementation of AI auto-segmentation models for breast cancer
- A complete AI-based workflow for auto-segmentation and -planning for breast irradiation

2.4 Project planning

To keep an overview of the different sub-tasks that needed to be performed in order to successfully perform the project, a final project planning was made after some iteration steps, which is shown in Figure 2.2. In this planning, unforeseen delays are indicated by the lighter-colored boxes. One important cause for delay, for instance, was the late upgrade to RayStation v10B-SP1. This version was needed for clinical implementation of the AI dose prediction model. Also, it was desired to perform the clinical pilot for auto-segmentation in this version as this is the version in which it will be eventually implemented for clinical use. Besides, data collection and model training for auto-segmentation turned out to be more time consuming than expected. For data collection, all the data needed to be checked thoroughly. During model training, several settings and parameters were tested, and eventually the final model needed to be trained in RayStation v10B-SP1. Section 4.2.1 will elaborate further on this process of data collection and model training. On the other hand, as can be seen in Figure 2.2, data collection, model training and retrospective validation of the AI dose prediction model for FF irradiation was performed in a strict time frame without any difficulties. This was carried out by a student, under supervision of the project team, and the framework for the training and validation was already available from the previous project from the QME project manager for conventional irradiation, making it a straightforward sub-project of this overall design project.

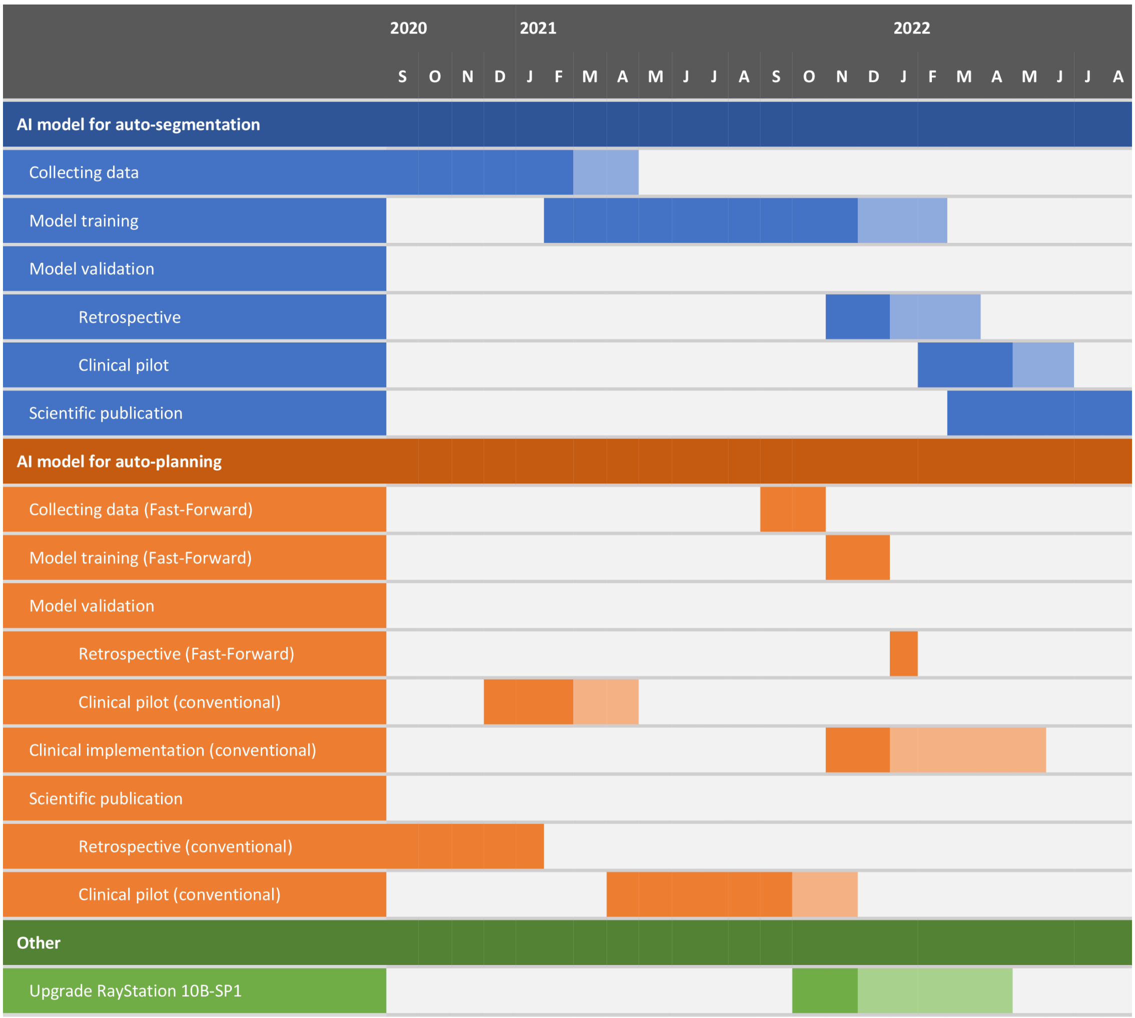


Figure 2.2: *Project planning. Unforeseen delays are indicated by lighter-colored boxes.*

3 | Detailing Project Goal

3.1 Introduction

The AI models that were developed in this design project to automate segmentation and plan optimization need to meet certain functional and technical requirements before clinical implementation can be realized. The primary endpoint of both models for auto-segmentation and auto-planning is time saving, compared to the manual process. Besides, quality of the delineations and treatment plans should be maintained or even increased. Lastly, implementing the models will lead to a decrease of intra- and inter-observer variability. Section 3.2 will elaborate further on these desired outcomes and how and when these will be measured. Next, in section 3.3 these requirements will be described in more detail and for each outcome the minimum requirement for successful clinical implementation is stated.

3.2 Functional requirements

3.2.1 Time saving

Time saving is the primary desired outcome of both AI models. In the current clinical workflow, the RTT starts the treatment planning process with loading a template, which contains empty structures for all contours (CTVs and OARs), needed for the treatment. Next, the RTT manually delineates the OARs. Thereafter, the RO manually delineates the CTV and checks the delineations made by the RTT. To complete the segmentation process, the other structures, such as the PTVs, are automatically generated, based on the existing delineations. Then, the RTT starts the plan optimization. Again, a template is used, which contains a standard beam set, optimization settings and a standard set of objectives to start the optimization. Then, the RTT can manually tweak these objectives to create a treatment plan that fulfills the clinical goals. Finally, the treatment plan is checked by another RTT, a RO and a medical physicist before the treatment starts.

When using the AI models, a similar workflow as usual will be maintained. First, the RTT will run the model to automatically delineate the targets and OARs, and then check the delineation of the OARs and make corrections whenever necessary. Then, the RO will do the same for the target volumes and check the OARs. Thereafter, the RTT will run the AI model to automatically generate a treatment plan, which can be manually adjusted afterwards as well whenever necessary. The final check of the plan is the same as in the current clinical workflow.

To measure the possible time saved with auto-segmentation, the following measurements will be taken:

- Baseline: time needed for manual delineation of CTVs and OARs
- AI model: time needed to automatically generate delineations by AI model
- AI model + correction: time needed to automatically generate delineations and manually correct for clinical use

These measurements only contain the actual time spent on the delineations, not the lead time of the full process. Timing of the tasks performed by the RTT and RO will thus be measured and logged independently.

For auto-planning, promising results were achieved concerning the number of fulfilled clinical goals during the previously performed retrospective study. Therefore, it was decided to not manually correct the plans during the clinical pilot, to explore the possibility of using auto-planning without any intervention, except for manually opening the leaves. Therefore, only the following measurements were performed:

- Baseline: time needed to manually create a treatment plan
- AI model: time needed to automatically generate a treatment plan

3.2.2 Quality

To generate clinically usable delineations, certain international guidelines are used [17, 18]. Hence, the automatically generated contours are checked, and can be corrected if necessary on the basis of these guidelines. Because the main goal of the automation is time saving, the quality of the automatically generated delineations has to be as good as possible, to minimize the amount of corrections needed. Therefore, it is needed to measure both the quality of the outcome of the AI model and the amount of correction needed. Besides, as mentioned before in Section 2.1, there exists intra- and inter-observer variability in manual delineations. As the model can be seen as a new observer, one could say that if the model performs as well as the interobserver variability, the model mimicks reality and performs adequately [10]. Therefore, the accuracy of the automatic delineations, will also be compared to the interobserver variability of a similar study. Several quantitative metrics will be measured, to be able to compare the model performance with results found in literature. Besides, these parameters can reveal potential consistent errors, such that they can be easily tracked down to improve the model [19]. To assess the clinical potential, it is important to also perform qualitative measurements [10]. The quantitative and qualitative measurements of the quality of auto-segmentation will be performed as follows:

- Quantitative measurement: the automatically generated contours (auto-contours) will be compared with a ground truth. This ground truth is the contour that was manually generated in clinic and was checked to see if it follows the guidelines.
- Qualitative measurement: the auto-contours will be checked and scored by RTTs and ROs during a clinical pilot.

To assess the quality of treatment plans, clinical goals are used. These goals contain requirements to e.g. the minimum coverage needed for the PTV and maximum dose allowed to the OARs. The main goal is always to spare the OARs as much as possible, while still maintaining adequate target coverage. Besides, DVHs can be used to extract several parameters, such as mean and maximum dose, to measure the quality. Similar as with auto-segmentation, qualitative measurement is important to assess clinical potential, as is mentioned in several studies [20–22]. The following quantitative and qualitative measurements will therefore be performed as follow:

- Quantitative measurement: the automatically generated plans (auto-plans) will be scored using the clinical goals. Besides, comparison with manually generated plans will be made by using several DVH parameters.
- Qualitative measurement: the auto-plans will be blindly scored and compared to manually generated plans by several ROs.

3.2.3 Intra- and interobserver variability

Intraobserver variability refers to differences in perception and action of one person in the same situation, while interobserver variability refers to different perceptions and actions of different persons in the same situation. For both types, it is known that they occur in both delineation as plan optimization, caused by differences in the experience of the ROs and RTTs, among other things [11, 12, 14, 23, 24].
An AI model always generates the same output for the same input, eliminating the intra- and interobserver variability. However, as the output of both models can be corrected manually, there will still be some variability. This variability will be lower than variability in completely manually generated delineations or treatment plans, since the starting point for correction is the same for all users. This decrease in variability in automatic delineation is -for example- found in the study of Byun *et al.*, where experts of multiple institutes evaluated an automatic delineation method [25]. Furthermore, Wang *et al.* found a decrease in variability between RTTs with different levels of experience, when using a KBP method [26]. The difference in variability between ROs or RTTs in our institute won't be measured in this project, due to limited time.

3.2.4 Study design

The above mentioned measurements with respect to time saving (Section 3.2.1) and quality (Section 3.2.2) will be performed in two different phases:

1. Retrospective validation:
   - validate AI model for auto-segmentation on 20 patients for each side, independent of training set
     - quantitative measurement of quality
   - validate AI model for FF auto-planning on 20 patients, independent of training set
     - quantitative measurement of quality
2. Clinical pilot
   - implement AI model for auto-segmentation in clinical workflow, validate on 10 patients for each side
     - measurement of time
     - quantitative measurement of quality
     - qualitative measurement of quality
   - implement AI model for conventional auto-planning in clinical workflow, validate on 20 patients
     - measurement of time
     - quantitative measurement of quality
     - qualitative measurement of quality

The clinical pilots have been executed when the AI model performed adequate according to the requirements specified in Section 3.3. In addition, this choice was based on the judgement of two experienced ROs, whether they think the outcomes could have clinical potential.

3.3 Technical requirements

3.3.1 Time saving

Several studies are performed in which auto-segmentation is used for (a part of) the contours as in this study. A study of Chung *et al.* reports a decrease in time of approximately 75% using auto-segmentation, compared to manual delineation [27]. Another similar study, using an atlas-based method, even shows a decrease in time of 93%. In addition, they report a decrease in time of 32% after correction of the auto-contours [28].

In contrast to auto-segmentation, only a few studies are performed where auto-planning is timed. Sheng *et al.* developed an auto-planning method using a random forest model. They report that the whole process is finished within 5 minutes, in contrast to 30 minutes to 4 hours for the manual process [29].

As mentioned above, several time measurements are performed:

1. time needed to manually generate contours/plan (T_{manual})
2. time needed to automatically generate contours/plan (T_{auto})
3. time needed to correct automatically generated contours/plan ($T_{correct}$)

For both AI models, in this project a decrease in time in 90% of the cases is demanded, including correction needed to get a clinically acceptable outcome ($T_{auto} + T_{correct} < T_{manual}$).

3.3.2 Quality

As mentioned before, the quality of the auto-contours and -plans will be measured with both quantitative as qualitative metrics. This section describes which methods will be used to measure those, and which outcomes are needed for a clinically feasible model.

Quantitative measurements

For auto-segmentation, several methods are known to measure the quality of the generated contours. Besides the comparison of the outcome with a ground truth, the outcomes will also be compared with interobserver variability reported in a study of Chung *et al.*, shown in Table 3.1 [27]. This variability was found by delineation of contours in one patient by three different ROs. When similar values are found with the model of this study, the clinical reality is reflected and the clinical pilot can be started to further assess the quality of the model. Of course, the variability reported by Chung *et al.* could differ from the variability within our institute, but as mentioned before, this variability will not be measured in our institute.

The auto-contours without correction, which are in the first study phase the only contours generated, don't have to comply with the guidelines. However, to get a better idea of clinical performance before starting the clinical pilot, two experienced ROs will visually inspect the auto-contours of some demo-patients during the retrospective phase. If they have enough trust that these contours will support the clinical workflow, and when the quantitative results are sufficient, the clinical pilot will be started to asses the actual clinical potential.

Eventually, the following parameters will be measured, which are also visualized in Figure 3.1:

- Dice Similarity Coefficient (DSC): $DSC = \frac{2|X \cap Y|}{|X|+|Y|}$, with X the volume of the ground truth and Y auto-contour, where a DSC score of 1 represents a perfect overlap between both volumes [30,31]. When the measured DSC scores are in the same order of magnitude as the values in Table 3.1, the model is found to be adequate for this criteria. However, the interobserver variability was not measured for the humerus, so a minimum DSC score of 0.7 is demanded, which is a widely used limit in several studies.
- 95^{th} percentile of Hausdorff Distance (HD) (95%HD): $HD = \max\{h(X,Y), h(Y,X)\}$, with $h(X,Y) = \max_{x \in X} \min_{y \in Y} \parallel x - y \parallel$ in mm, with a score of 0 representing perfect overlap. By using the 95^{th} percentile, a small subset of outliers will be eliminated. Again, values of the same order of magnitude as in Table 3.1 are found the be adequate performance. For the humerus, a maximum 95%HD of 7 mm is set.

ROI	**DSC (mean ± std)**	**95%HD [mm] (mean ± std)**
CTVp	0.85 ± 0.02	8.94 ± 2.86
CTVn1	0.69 ± 0.04	13.58 ± 3.00
CTVn2	0.47 ± 0.17	18.74 ± 8.15
CTVn3	0.56 ± 0.10	9.87 ± 3.61
CTVn4	0.45 ± 0.13	11.82 ± 4.88
Heart	0.91 ± 0.01	13.00 ± 5.10
Lung Left	0.99 ± 0.00	2.33 ± 0.95
Lung Right	0.98 ± 0.00	2.19 ± 0.66
Thyroid	0.72 ± 0.07	5.37 ± 1.70
Esophagus	0.78 ± 0.04	7.08 ± 3.52
Humerus	0.70 ± 0.05	7.00 ± 2.00

Table 3.1: *The technical requirements for the quantitative measurements for each ROI. The values are based on the interobserver variabilities found by Chung et al. [27]. NB: as no values were reported for the humerus, these requirements were set based on more widely used limits.*

Recently, a new method called Surface DSC (sDSC) gained more attention in evaluating the outcome of auto-segmentation model, which will also be measured in this study. The sDSC is the overlap between two surfaces, with a pre-defined tolerance for allowed deviation. By quantifying the deviation in contours rather than volumes, it better reflects the correction needed [32]. This method seems to correlate better with time needed for correction, and relative time saved, which is the primary outcome

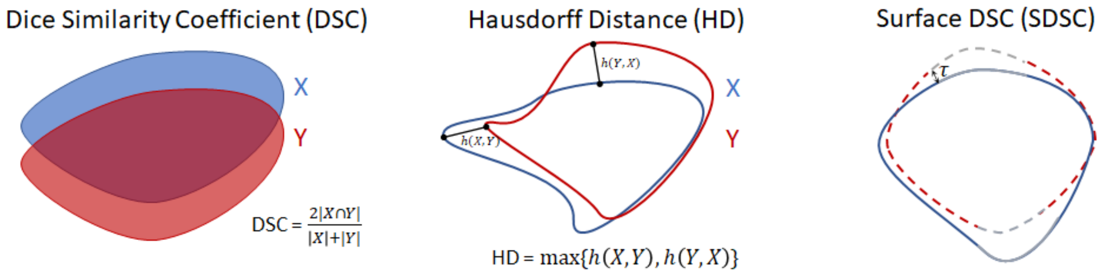

Dice Similarity Coefficient (DSC)
X
Y
DSC = $\frac{2|X\cap Y|}{|X|+|Y|}$
Hausdorff Distance (HD)
$h(Y,X)$
$h(X,Y)$
X
Y
HD = $\max\{h(X,Y), h(Y,X)\}$
Surface DSC (SDSC)
τ

Figure 3.1: *A visual representation of the different quantitative measurements used in this study.*

of this study [33]. Hereby a higher sDSC relates to less time needed for correction. However, this metric is not reported yet in similar studies, so no requirements were derived.

For auto-planning, the auto-plans are compared to manually created plans with several methods:

- Clinical goals: both automatically and manually generated plans are scored with the clinical goals. When a clinical goal is not met, the specific patient will be inspected to see if this is clinically relevant. Not all goals have the same priority, meaning they are target values and no hard constraints. For both conventional and FF plans, the clinical goals are displayed in Table 3.3.
- DVH parameters: the DVH parameters specified in Table 3.2 will be reported for both the manually and automatically generated plan. To compare these values, the Wilcoxon Signed Rank test will be used, (p-value $<$ 0.05 is significant). Quality is maintained when the Mean Heart Dose (MHD), Mean Lung Dose (MLD) and V5Gy to the lungs are not significant higher for the auto-plans, or when these differences are not considered as clinical relevant. For the PTV, a non-significant difference is strived for, but a difference is allowed when clinical goals are still met.
- Monitor Units (MUs): the number of MUs is compared between the automatically and manually generated plans, striving for a non-significant difference between both plans, calculated with the Wilcoxon Signed Rank test.

ROI	**DVH parameter**
PTVp	Average dose [cGy]
	Maximum dose ($D_{2\%}$) [cGy]
Lungs	Average dose (MLD) [cGy]
	Volume receiving 5 Gy [cc]
Heart	Average dose (MHD) [cGy]

Table 3.2: *DVH parameters used to compare quantitative performance of different treatment plans*

Qualitative measurements

The auto-contours are scored by 5 different RTTs and ROs, using a 3-point system:

1. Clinically acceptable contour, no correction needed
2. Not-clinically acceptable contour, but useful as starting point for correction
3. Not-clinically acceptable contour, not useful as starting point for correction

A contour is said to be useful as starting point, when the RTT or RO thinks he/she can save time using this contour as start point, with respect to starting over.

ROI	Conventional irradiation	Fast-Forward irradiation
PTV	At least 3805 cGy at 98%	At least 2470 cGy at 98%
PTV	At least 3965 cGy average dose	At least 2574 cGy average dose
PTV	At most 4045 cGy average dose	At most 2626 cGy average dose
PTV	At most 4285 cGy at 2% volume	At most 2782 cGy at 2% volume
Lungs	At most 600 cGy average dose *At most 400 cGy average dose*	At most 300 cGy average dose *At most 200 cGy average dose*
Heart	At most 300 cGy average dose *At most 200 cGy average dose*	At most 150 cGy average dose *At most 100 cGy average dose*
CL Breast	At most 100 cGy average dose	At most 100 cGy average dose
External-PTV	At most 10 cc at 4285 cGy	At most 10 cc at 2782 cGy

Table 3.3: *Clinical goals used to evaluate the plans for conventional and FF breast irradiation. Goals printed in italic are of less importance and no hard constraints.*

To qualitatively assess the auto-plans, 4 ROs will blindly score them according to the following two points:

1. Clinical acceptability
2. Ranking of plans (manual vs auto-plan); equal ranking is allowed

Up until now, no studies involving qualitative assessment of auto-plans for breast cancer are published. However, two studies concerning auto-planning for prostate cancer report clinical acceptability in respectively 80% and 89% of the cases [22, 34]. For this project, a minimum clinically acceptability rate of 90% is required. In addition, inter-observer variation will be present in this assessment of the ROs, which should not be higher for auto-plans compared to manually generated plans. Ranking of the different plans will be used to determine if the automatically generated plans are non-inferior to the manual plans, which is set as a requirement. This method is used by Cornell *et al.*, considering an auto-plan non-inferior if the lower limit of the 95% Confidence Interval (CI) of the success rate was greater than 45%. The success rate is defined as the number of times the auto-plan is considered equal to or better then the manually generated plans. Then, the Wilson score interval method will be used to calculate the CI. Figure 3.2 visualizes non-inferiority with Δ as a margin, with will be set to 5% in this study.

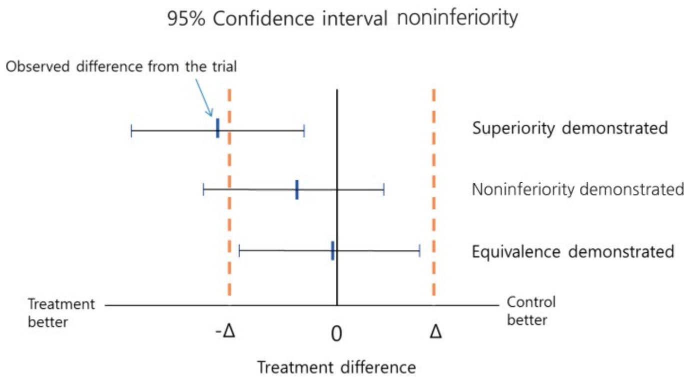
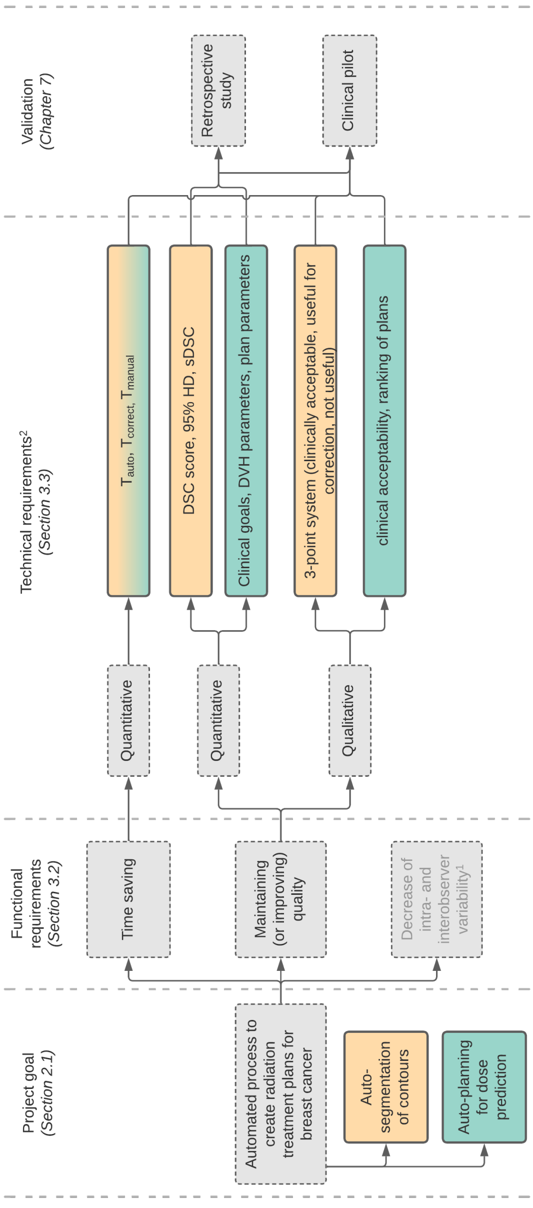


Figure 3.2: *A visual representation of non-inferiority and superiority tests. Δ is the margin, which is set to 5% in this project. Image retrieved from [35]*

3.4 Complete set of requirements

Figure 3.3 summarizes the complete set of requirements.


Project goal
(Section 2.1)
Functional requirements
(Section 3.2)
Technical requirements²
(Section 3.3)
Validation
(Chapter 7)
Automated process to create radiation treatment plans for breast cancer
Auto-segmentation of contours
Auto-planning for dose prediction
Time saving
Maintaining (or improving) quality
Decrease of intra- and interobserver variability¹
Quantitative
Quantitative
Qualitative
T_auto, T_correct, T_manual
DSC score, 95% HD, sDSC
Clinical goals, DVH parameters, plan parameters
3-point system (clinically acceptable, useful for correction, not useful)
clinical acceptability, ranking of plans
Retrospective study
Clinical pilot

[1] This functional requirement is out of scope for this project, see Section 3.2.3 for explanation

[2] The minimum requirements of these metrics can be found in Section 3.3

Figure 3.3: *Complete set of functional and technical requirements. Requirements related to auto-segmentation are indicated with yellow, whereas requirements related to auto-planning are indicated with green.*

4 | Design Process

4.1 Introduction

Several steps need to be taken before an AI model can be successfully implemented in clinical reality. The first important step is to create a suitable dataset for training, validating and testing. The data must be checked thoroughly on their quality, as the model will reproduce the quality of the input. Besides, the dataset should be large enough and reflect real-world variation in the patient group to prevent overfitting, which is a phenomenon where the model works perfect for a particular set of data, but doesn't work for unseen data. Next, a suitable AI model should be chosen for the problem. The focus in this project is on DL models, although one ML model is evaluated for auto-planning. For these DL models, an appropriate CNN architecture should be chosen and parameters should be tweaked during training. To successfully train and evaluate a model, the dataset is divided into three sets: training, validation and testing. Both the ML and DL models are trained on the training set and the model is real-time validated during training by the validation set, which contains unseen data for the model. During each step of training, also known as an epoch, the DL model uses an optimisation function to minimize the error between the generated output and the known ground truth. This error is computed using a loss function, which can thus be used to quantify the performance of the model and monitor its behaviour. The loss function is computed for the training set as well as for the validation set, such that the performance on an unseen set is also monitored and overfitting can be detected. When a final model is trained, it is evaluated on an independent test set, which was never seen before during training. In this project, multiple test sets are created for the retrospective studies and clinical pilots, as explained in 3.2.4. This chapter elaborates on data collection and the model training procedure for both the auto-segmentation as auto-planning model.

4.2 Auto-segmentation

4.2.1 Data collection

Patient data was retrieved from RayStation and the clinical archive, for patients treated between February 2017 and May 2022 for breast cancer. Two separate models will be trained for respectively left and right sided breast cancer, demanding two separate datasets. Several patient groups are included for both sides, treated to either node levels 1 and 2 or node levels 1 to 4. In total, 80 patients were used for training of both models. In the retrospective study and clinical pilot, respectively 15 and 10 patients were used for each side. For every patient, it has been noted which structures are present, and the structures were visually inspected as a first check and abnormalities were noted. Then the data statistics were checked to detect any outliers. This check was performed by plotting the bounding box per structure for each patient. Figure 4.1 shows an example plot. Outliers were again visually inspected and again abnormalities were noted when found. Ultimately, the marked structures were checked by two experienced ROs and corrected or removed from the set when needed.

The structures that were mainly corrected or removed where node level 3 and 4. Twenty patients were checked, and for only four patients the clinical structures were accepted without any correction. This observation confirms the presumption that these structures are hard to delineate, which is also reflected by the interobserver variability metrics in 3.1. Besides, the ROs indicated that in recent years special attention is paid to the delineations of these node levels, as more and more discussion is taking place about the correctness of the guidelines for these structures, again confirming the presumption. Another structure that needed to be corrected in multiple patients was the thyroid. The thyroid is susceptible to multiple anatomical variations, making it a difficult structure to delineate for the RTT.

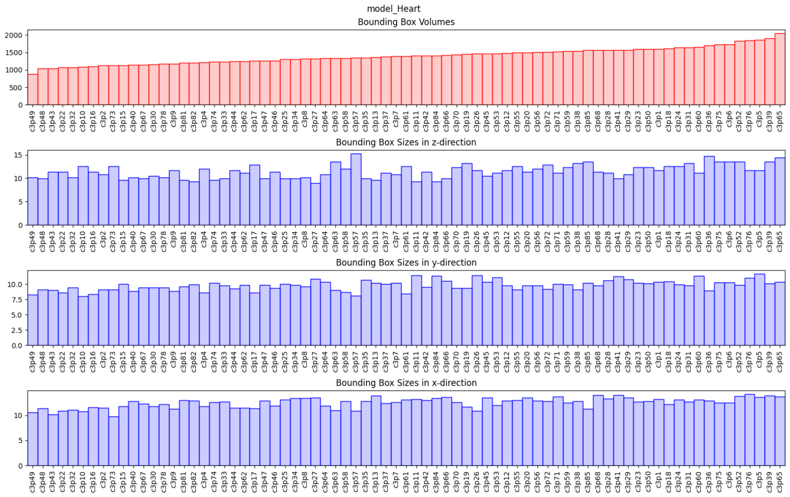


Figure 4.1: *An example of the visual representation of data statistics. The bounding box volumes and sizes of along each axis (x, y and z) for the heart of each patient are shown.*

In addition, the thyroid consists of two lobes that are connected. This connection is not always as easy to notice and thus not always delineated. However, for training of the model, it was desired that the lobes were connected and the thyroid consisted of only 1 contour instead of 2 separate contours. Therefore, two experienced RTTs checked and corrected thyroids consisting of 2 separate contours.

4.2.2 Model architecture

For auto-segmentation, a DL model architecture based on U-net was used [36]. This is a widely used CNN architecture in medical image analysis, where the information on different levels of resolution is combined, schematically shown in Figure 4.2. An adapted version of this architecture, 3D U-net, is implemented by RaySearch and the code was made available for training on the dataset of our clinic [37].

The auto-segmentation model within RaySearch consists of several submodels. Each submodel is a 3D U-net, which is trained on one or more Region of Interest (ROI)s. The input of these models is a volumetric image, defined by a bounding box centered around the ROI(s) of that submodel. By using these cropped volumes instead of the full volume, the required memory for training decreases and training can be performed on a higher resolution. Eventually, the process of auto-segmentation consists of the following 4 steps, of which the first 3 steps are visualized in Figure 4.3:

1. Atlas-based segmentation: a small atlas of 4 patients containing all ROIs is stored in the model. When starting segmentation of a new patient, these images are first rigidly registered to the input image. The best matching image is selected, and its delineations are propagated to the input image. These delineations are used as a starting point for the DL models.

2. Initial segmentation: a first segmentation is performed on a low-resolution input, where the bounding box is based on the output of the first step. The DL model used for this step is one of the submodels, trained on all ROIs.

3. Refined segmentation: for each ROI or subgroup of ROIs, segmentation is performed on high-resolution input by several sub-models. Input volumes are defined by bounding boxes resulting of the initial segmentation.

4. Post-processing: extra processing can be performed on the produced output, such as smoothing, adjusting the number of components per ROI and correct overlapping structures.

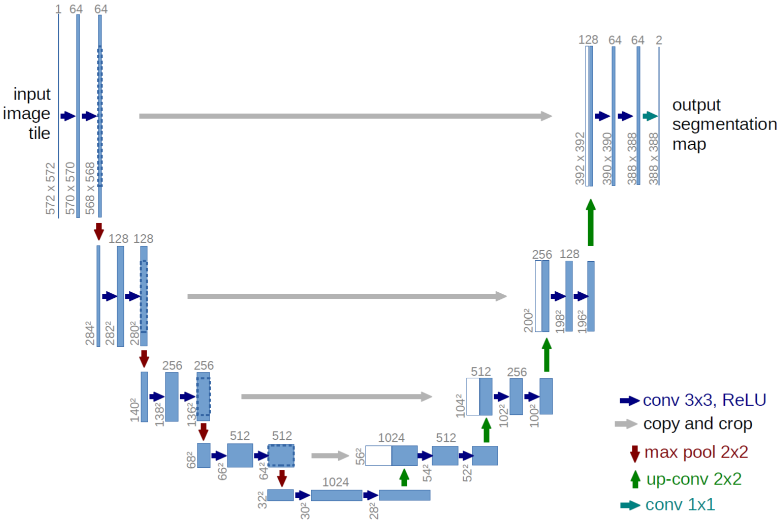


Figure 4.2: *The U-net architecture, image retrieved from [36]. Each blue box corresponds to a multi-channel feature map. The numbers on top of the boxes represent the number of channels, whereas the resolution of each box is provided at the lower left edge. White boxes represent copied feature maps and arrow denote the different operations*

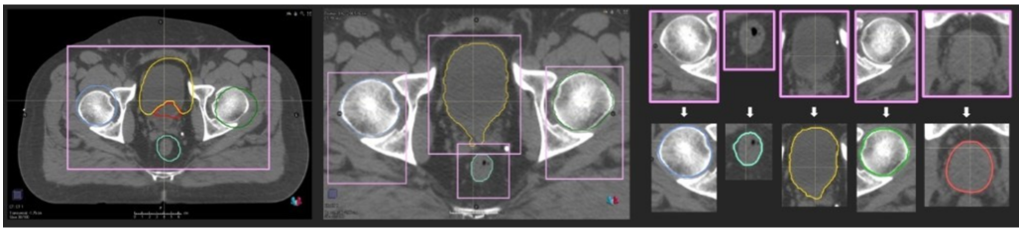


Figure 4.3: *The first three steps of the auto-segmentation method. Left image (step 1): atlas-based segmentation, resulting into approximate organ positions and the input window to initial deep learning model (pink rectangle). Middle image (step 2): initial segmentation, resulting in low resolution segmentations and the input windows for the refined deep learning models (pink rectangles). Right image (step 3): refined segmentation, resulting in high-resolution segmentations. Image retrieved from RaySearch reference manual*

4.2.3 Model training

As explained above, a training and validation set were used during model training. To further evaluate the performance of the model, cross-validation was used. This technique resamples the data in different training and validation splits during each iteration, referred to as a fold. For this study, 5-fold cross-validation was used, meaning the dataset is split into 5 groups, whereafter each group will be used as the validation set in one of the 5 folds. This method is especially useful for smaller datasets, to inspect the impact of the data split of the total training set. The results of this cross-validation and additional tests performed to assess the performance of some submodels are further explained in Appendix B.

During training, several parameters are involved that can be tuned, of which a complete overview can be found in Appendix B. The most relevant ones are listed below:

- Number of epochs: during one epoch, the full training set is passed one time through the model. The number of epochs should be sufficient for the model to converge, which can be monitored through visualization of the training and validation loss. The model can intermediately be saved during training after a pre-defined number of epochs, in order to compare the performance of the model at different stages of training.

- Batch size: the number of samples used in each update of the weights of the CNN. A higher number leads to a smoother training process, but also requires more memory. As the 3D input occupies a lot of memory, a batch size of 1 is used.

- Learning rate: during training, the Adam optimizer is used [38]. The learning rate controls the gradient step size used during weight update of the network. A high learning rate can lead to a diverging instead of converging learning process, and a low learning rate leads to a slow conversion. A learning rate of 1e-4 is used in this project, which was set as default by RaySearch.

- L2 regularization weight: L2 regularization is used to prevent overfitting, by preventing the weights from growing too large. The default value of 1e-5 is used in this study.

To further increase the model performance, data augmentation was performed during training. Data augmentation is a method to artificially increase the size of the dataset by adding adjusted copies of the existing dataset. This augmentation was performed on-the-fly during training, causing the model to be trained with a slightly different version of the dataset each epoch. It contains translation, where the image is moved along the x-, y- and/or z-axis, rotation, where the image is rotated in the x-, y- and/or z-axis and elastic deformation. To ensure that the ROI is still fully covered by the bounding box after augmentation, an extra margin is added to the original dataset.

4.3 Auto-planning

4.3.1 Data collection

Data for conventional breast irradiation was collected in the previously executed study, and more details can be found in Appendix A. In total, 90 patients were used for training of the ML models used for auto-planning for conventional breast irradiation. For the FF protocol, 52 patients were used for training. Of these patients, 19 patients were treated with the FF protocol in clinic. The other plans were generated for a part of the patients used in the conventional irradiation study. These plans were checked by experienced RTTs before training of the model. For all patients of both protocols, CT data with delineated ROIs was available.

4.3.2 Model architecture

RaySearch developed a framework for auto-planning which consists of several steps, visualized in Figure 4.4 and explained below:

1. Feature extraction: two methods can be used, in isolation or combination, to extract features from the CT images and delineations: (1) Filterbank, which uses a hand tuned filterbank to extract over 80 image features, such as isolation of certain shapes, tissue types or edges, and (2) Signed distance maps, which calculates signed distance maps based on specific ROIs selected during the training procedures, helping the model with depth perception in relation to the selected ROIs.

2. Spatial model: predicts a dose distribution per voxel. Several models are available, which will be further explained below.

3. Prior model: predicts a prior (DVH) for a ROI. As for the spatial model, several models are available.

4. Conditional Random Field (CRF) and strategy: a CRF optimization is performed on the ROIs to get a final dose distribution. In this step, certain ROI goals, called strategies, can be applied to the priors, e.g. to restrict the dose to a certain ROI. The predicted dose will then meet these goals in the most likely way. In this project, we only create one strategy with certain goals.

5. External model: in the default model, the dose in the external region is the dose predicted by the spatial model. In addition, a DL model can also be used to predict this region. However, in this study, the default method will be used.

The final outcome of this process is a predicted dose per voxel, which is not directly clinically applicable. To obtain a clinically deliverable plan, dose mimicking is needed. During dose mimicking, direct machine parameter optimization is used to approximate the predicted dose distribution, while taking dose constraints into account. These dose constraints can be defined in the model settings to tweak your model further to meet clinical goals.

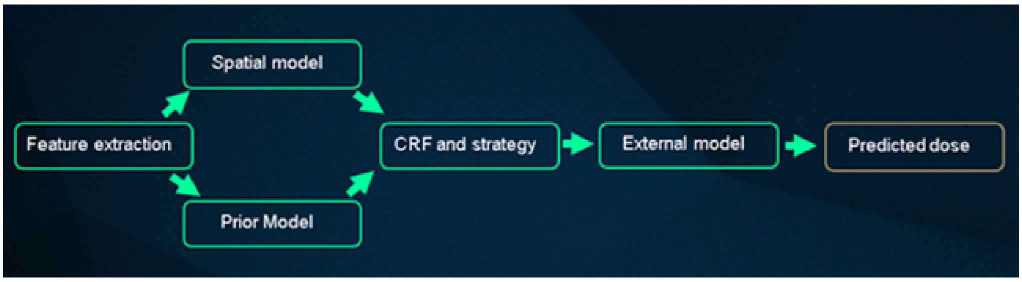


Figure 4.4: *Framework for auto-planning. Image retrieved from RaySearch reference manual*

Three different models were developed and trained for dose prediction:

- 2D U-net

  In the previously executed study for conventional breast irradiation, a 2D U-net model was developed and trained. This is a spatial model, based on the aforementioned U-net architecture from Ronneberger *et al.* [36]. The input of the model is a multi-channel matrix, containing masks of the PTV, heart, lungs and external. The masks for heart, lungs and external are binary, with zeroes outside and ones inside the structure, whereas the PTV mask has a value equal to the prescribed dose for voxels inside the structure. The output of the model is the spatial dose distribution. More details can be found in Appendix A. Moreover, this model was also used for training of the model for FF irradiation with the help of transfer learning, which will be further explained in Section 4.3.3.

- Contextual Atlas Regression Forest (cARF)

  The cARF model is an atlas-based ML method, where each patient in the atlas contains two ML models: an Atlas Regression Forest (ARF) and a predict Regression Forest (pRF). The ARF is used to predict a dose distribution using image features. The pRF is then used to predict for each patient in the atlas how well its ARF works for a new patient. This method is further explained by McIntosh *et al.* in [39, 40]. The cARF model is used for auto-planning of conventional breast irradiation, where the model is applied in the following ways for the spatial and prior model:

  - Spatial model: with the use of the image features, the pRFs of all atlas patients predict an estimate of the error for the corresponding ARF. Then, the 5 best ARFs are selected and merged into one Regression Forest (RF), which will predict a dose value for each voxel.
  - Prior model: with the use of the image features, the pRFs of all atlas patients predict an estimate of the error for the corresponding atlas DVH. The closest atlases are then selected, and for each ROI the Probability Distribution Function (PDF) of all atlases is merged to one weighted PDF per region, which is used to calculate the DVH.

- 3D U-net

  Similar as for auto-segmentation, RaySearch developed a 3D U-net for auto-planning, which is described in more detail in Appendix B. A multi-channel 3D volume is used as input, containing three channels containing a binary mask of the PTV, a mask of the union of the heart and lungs and a mask of the external. The same model can be used for both the spatial as the prior model, without any additional training, as the first step for both models is predicting the dose value for each voxel. This predicted volume is the final output for the spatial model. However, for the prior model an extra post-processing step is added by calculating the DVHs of the selected ROIs from the inferred dose and returning that as output.

4.3.3 Model training

Both for conventional and FF breast irradiation, several models were trained during this project as well as in a previous project, as mentioned in Section 2.1.

Conventional breast irradiation

For conventional breast irradiation, both the 2D U-net and cARF models were trained during the previous project [15]. While the 2D U-net was trained in-house, the cARF model was trained by RaySearch, using the same dataset. More details about these models and the results of this study can be found in Appendix A. In addition, a 3D U-net was trained during this project by RaySearch on the same dataset, of which details can be found in Appendix B.

FF breast irradiation

For FF breast irradiation, a 2D U-net model was trained using transfer learning during this design project. This method uses a pre-trained model to train a new model for a different, but related, problem. As the model leverages knowledge from the previously trained model, such as features and weights, the newly trained model already understands features and makes it faster and able to train on a smaller dataset [41]. As the input used for the AI models for conventional and FF breast irradiation is the same, the goal of this study is to investigate if transfer learning is applicable here. If so, a smaller dataset could be used. In this study, only 52 patients were used for training to assess the use of transfer learning. As a starting point, the 2D U-net was used. More details on the training parameters can be found in Appendix B and [42].

5 | Final Design and Implementation

In the previous chapter it has been described how several models were developed and trained for both auto-planning and auto-segmentation, with varying model architectures, parameters and datasets. This chapter summarizes the final datasets and models used for both purposes. Furthermore, it elaborates on the clinical implementation of both models in the current clinical workflow, including a risk analysis and change management procedure.

5.1 Auto-segmentation

In this design project, two final models were trained for auto-segmentation of respectively left- and right-sided breast cancer, based on the model architecture described in Section 4.2.2. Both models consist of the same submodels, listed in Table 5.1, with corresponding number of patients included for each submodel. For the final model, 90% of the patients were used in each submodel as training set, while 10% was used as validation set. Submodel 'All' is used for the initial segmentation, including all ROIs, whereas the other submodels are used during the refined segmentation. During post-processing, smoothing is applied and overlap between CTVp and CTVn1 is removed. For all ROIs, only one component is created, except for the thyroid, allowing the two lobes to be delineated without the connection being detected.

5.2 Auto-planning

As mentioned in Section 4.3.2, several models were trained during the project for auto-planning. The final models used in the different phases of this project are listed in Table 5.2. Although both the 2D U-net and cARF model were used in the first phases, the 3D U-net, trained by RaySearch, is used for clinical implementation due to regulatory reasons.

Submodel	Left-sided	Right-sided
All	30	36
CTVp	75	69
CTVn1-4	41	49
Heart	75	45
Lung_L	75	72
Lung_R	75	72
Esophagus	45	50
Thyroid	40	42
Humerus	70	69

Table 5.1: *Overview of submodels and the corresponding number of patients used to train these models.*

Protocol	Phase	Model
Conventional breast irradiation	Retrospective study	2D U-net, cARF
	Clinical pilot	2D U-net, cARF
	Clinical implementation	3D U-net
FF breast irradiation	Retrospective study	2D U-net

Table 5.2: *Overview of the different auto-planning models for the different irradiation protocols and study phases.*

5.3 Clinical introduction

5.3.1 Implementation

The end goal of both auto-planning and auto-segmentation is clinical implementation in the current daily workflow. For auto-segmentation, in this design project this implementation is tested during a clinical pilot, whereas the auto-planning model is actually implemented for clinical use (after being tested in a clinical pilot). In the current manual workflow, scripting is used in both processes to semi-automate some steps in the process. For both purposes, these scripts were slightly adapted to incorporate the AI models in the clinical workflow.

For segmentation, currently a script is used to create empty structures, based on the clinical protocol for treatment, which can then be manually delineated. In the new workflow, the same script is used, and again based on the clinical protocol for treatment, either automatic segmentation will be performed or empty structures will be created. After automatically creating the structures, the RO and RTT can use the same tools available in the TPS as for manual delineation, to correct the structures if desired. To ease the correction of large structures, in this design project an extra scripted tool to delete intermediate slices of a contour was created. As a result, the structure only contains a few slices, which can then be corrected, and interpolation can be used to create a whole structure again. Hence, this tool saves time, compared to adjusting every single slice.

For plan optimization, in the current workflow a similar script is used to create a new plan, to set up a standard set of beams, to load the clinical goals and to load the objectives and settings used for optimization, all linked to the clinical protocol. Again, in the new workflow the same script will be used, but now a plan will be set up which is suitable for auto-planning. Then, the RTT can start the auto-plan optimization, and extra objectives can be added afterwards to correct the plan, if desired.

5.3.2 Introduction in workflow

To successfully introduce the AI models in the clinical workflow, not only the technical implementation is of importance, but also the impact for the users should be taken into account. Therefore, another step needed for clinical implementation was to educate the ROs and RTTs about AI and how it can be used. To facilitate this, at several moments throughout the project, information about AI in general, and this project in particular, was provided during presentations to the whole department. More frequently, the involved ROs and RTTs of this project were updated about the status and results. In addition, work protocols were written and introduced for clinical implementation.

5.3.3 Commissioning

Before a model can be used in clinic, it needs to be commissioned. A general commissioning procedure is created by RaySearch, in which the performance of the model is evaluated on an independent dataset and the results are discussed with RaySearch. Besides, a commissioning report is created and shared with RaySearch, including information about the model training and validation process. For this project, the 3D U-net needed to be commissioned before actual implementation in clinical practice could be performed. For this commissioning process, a part of the test set of the clinical pilot study was used, including 10 patients. The results of the 3D U-net were not only compared with the manually created plans, but also with the auto-plans created during the clinical pilot, as those were qualitatively scored by the ROs. The 3D U-net plans were better regarding dose to OARs, while still maintaining adequate dose to PTV, when compared to the 2D U-net plans, which were already seen as clinically acceptable during the clinical pilot (Appendix E.2.2). Therefore, the 3D U-net was successfully commissioned and the model is now used in clinic since May 2022[1].

[1] https://www.tue.nl/en/news-and-events/news-overview/13-05-2022-breast-cancer-treatment-plans-at-the-touch-of-a-button/

5.4 Risk analysis

Before introducing a new method in clinical practice, possible risks should be understood and listed and mitigation should be performed. In this design project, an adapted version of the Health Care Failure Mode and Effect Analysis (HFMEA), used within the hospital, was used [43]. The HFMEA is a Prospective Risk Analysis (PRA) method, where possible failures and its causes for each step in a process are identified, after which a risk score is calculated, based on probability of occurrence and its impact. Critical moments are listed and the desired actions to be taken are identified. More information on how to identify and classify these actions can be found in Appendix C. For this project, a PRA was executed on the full workflow, from model training to implementation in the clinical workflow, and all the changes involved. A multidisciplinary team, consisting of the project leader, clinical physicists, a medical engineer and a RO and RTT, was composed to ensure all steps within the process were included. The final result of the PRA can be found in Table 5.3. Below, the moderate (risk score ≥ 5) and high (risk score ≥ 10) risks, and its actions to mitigate these risks, are further described below.

Moderate risks

- Designing model architecture: due to a lack of knowledge or skills, an architecture could be chosen and designed which is inappropriate for the desired purpose, resulting in insufficient performance of the model for clinical implementation. This can be overcome by a thorough literature study and testing in an early stage, which has been performed in this design project.
- Implementing model architecture: due to a lack of knowledge or skills, the architecture could be wrongly implemented, causing insufficient performance. To overcome this, several tools can be used to visualize the programmed architecture. In addition, another expert could check the code to prevent errors, demanding the code to be well written and well described with the help of comments. In this design project, the final architecture was visualized to check correctness and written codes were provided with comments and additional documents.
- Auto-planning; select strategy: as described before in Section 4.3.2, RaySearch included a functionality which enables to choose different strategies, to focus on e.g. extra sparing of OARs or target coverage. By choosing a wrong strategy, a sub-optimal plan could be created for the patient. However, for the treatment protocols included in this study, for all patients it is strived to spare the OARs as much as possible. Therefore only one strategy is created. This step is still included in this PRA, as for future models, trained for other target areas, multiple strategies might be applicable.

High risks

- Auto-segmentation; check contours: after the delineations are automatically generated, a check needs to be performed by the RTT and RO, to make sure they comply with the guidelines. An extra trigger will be build into the program Work4All to remind the RTT and RO to not only create the contours, but also check them. This program gives an overview of the steps taken in the process of a patient. Besides, during optimization and checking of the plan, the delineations are visible for the RTT and RO, so remarkable delineations can be noticed.
- Auto-planning; check plan and optimize further: after automatically generating the plan, it needs to be checked and can be optimized further if needed to comply with the clinical goals. In contrast to the previous step, a trigger is already built-in to not forget this step in the current workflow, as the plan will be checked by multiple persons in the plan approval step, so a non-optimal plan will be noticed.

Process step	Possible failure	Possible consequence	Impact	Possible causes	Probability	Risk score	Critical moment?	Controlled?	Detectable?	Action: eliminate, control or accept	Description of action	Responsible person(s)
1 Preparation												
data collection	not enough variation (bias in data)	non-optimal model	4	Human; misjudgement	1	4	NO					
data check and correction (pre-processing)	data not properly processed for purpose	non-optimal model	4	Human; misjudgement	1	4	NO					
designing model architecture	model architecture not suitable for purpose	model unsuitable for clinical purpose	3	Human; misjudgement	2	6	NO					
implementing model architecture	incorrectly programmed	programmed architecture doesn't match designed architecture	4	Human & technical	2	8	NO					
train model	no check on training progress	non-optimal model	4	Human & technical	1	4	NO					
validate model	validation protocol not sufficient f or clinical purpose	model not suitable for clinical use, but will be implemented	4	Human & technical	1	4	YES	YES				
2 Procedure												
auto-segmentation: select model	wrong model selected	wrong contours are delineated	2	Human; inattention	2	4	NO					
auto-segmentation: check contours	delineations are not checked	contours don't comply with guidelines	4	Human; organizational	3	12	YES	NO	NO	ELIMINATE	Extra check in W4A	Dave
auto-planning: beam set-up	Wrong beam set-up	non-optimal plan	3	Human	1	3	NO					
auto-planning: select model	wrong model selected	incorrect outcome	2	Human	2	4	NO					
auto-planning: select strategy	wrong strategy selected	plan suboptimal for patient	3	Human	2	6	NO					
auto-planning: open leaves	leaves of MLC not opened	plan not robust for swelling and patient movement	3	Human	1	3	YES	YES				
auto-planning: check plan and optimize further	plan is not checked	non-optimal plan	4	Human; organizational	3	12	YES	YES				
3 Closure												
auto-planning: plan approval	no connection with safe2treat	plan won't be approved	2	Technical	1	2	YES	YES				

Table 5.3: *Overview of the Prospective Risk Analysis, identifying critical moments and the desired actions to be taken. The risk score is calculated by multiplying the impact by the probability. A green risk score indicates low risk (risk score $\leq$ 4), yellow indicates moderate risks (risk score $\geq$ 5) and red indicates high risk (risk score $\geq$ 10)*

5.5 Change management

The introduction of AI models in clinical practice at the department of radiotherapy in CZE implies changes for the end-users. These changes are not only the actual changes in the workflow, but also involve the need for changes on other levels. First of all, for acceptance of a new workflow, it is important for the end-users to gain knowledge about the method, which is provided by for example presentations, as mentioned before in Section 5.3.1. Furthermore, the end-users need to gain trust in the AI models, which can also be achieved by involving them in an early stage and when performing the clinical pilot. Besides, by performing research in the fairly new field of AI, new research protocols are needed. It is of high performance to register the dataset used for the development and evaluation of AI models, as well as the tests performed to evaluate the performances. Therefore, a model registration sheet is developed during this design project as well, which can be found in Appendix D.

6 Results, Verification and Validation

Verification and validation are performed to check whether the implemented design meets its requirements. Verification is used to check whether the design meets the specifications, while validation is used to check if design generates an output conform the user needs and intended use (see Figure 3.3). For this project, the performance will be assessed by comparing the results of the AI models with the requirements set in Chapter 3.

6.1 Auto-Segmentation

6.1.1 Retrospective study

In the retrospective study, only quantitative measurements were performed. The resulting DSC scores and 95% HD values are listed in Table 6.1 and visualized in Figure 6.1, whereas the other quantitative results can be found in Appendix E. The DSC score and 95% HD are compared to the requirements, which were introduced in Section 3.3.2. Green boxes indicates that the requirements are fulfilled as a better result is achieved, while red boxes indicates that the model did not meet the requirements for that ROI. The yellow boxes indicate scores which did not completely meet the requirements as the mean scores were lower, but when taking the deviation of the requirements in account, one could say that it is still within the accepted range.

For the esophagus, the DSC score is too low for both models. After visual inspection, it was discovered that the mismatch is partly due to a difference in the length of the delineated contour. Therefore, the DSC score was also calculated for the overlapping parts, which resulted in a score (mean ± std) of 0.77 ± 0.08 for both models, approaching the requirement better. Therefore, although these scores still do not fully fulfil the requirements, it was still chosen to continue to the next phase without any additional changes for this ROI. The thyroid is the second ROI which does not meet the requirements. However, as explained in Section 4.2.1, a lot of variation existed within the dataset, which makes it a hard contour to train and evaluate. Therefore, for this ROI it was also chosen to

		DSC score			95% HD [mm]	
		(mean ± sd)			(mean ± sd)	
ROI	Left	Right	Requirements	Left	Right	Requirements
CTVp	0.94 ± 0.02	0.94 ± 0.02	0.85 ± 0.02	10.45 ± 12.26	9.26 ± 3.64	8.94 ± 2.86
CTVn1	0.78 ± 0.06	0.80 ± 0.04	0.69 ± 0.04	11.98 ± 5.33	10.54 ± 3.73	13.58 ± 3.00
CTVn2	0.74 ± 0.07	0.69 ± 0.07	0.47 ± 0.17	9.52 ± 4.95	9.32 ± 3.60	18.74 ± 8.15
CTVn3	0.74 ± 0.07	0.74 ± 0.05	0.56 ± 0.10	6.45 ± 1.94	7.56 ± 3.23	9.87 ± 3.61
CTVn4	0.58 ± 0.12	0.56 ± 0.13	0.45 ± 0.13	6.12 ± 2.18	6.78 ± 3.74	11.82 ± 4.88
Esophagus	0.70 ± 0.11	0.70 ± 0.09	0.78 ± 0.04	8.83 ± 5.58	10.34 ± 8.08	7.08 ± 3.52
Heart	0.94 ± 0.02	0.94 ± 0.01	0.91 ± 0.01	6.81 ± 2.94	8.10 ± 4.38	13.00 ± 5.10
Lung Left	0.98 ± 0.01	0.98 ± 0.01	0.99 ± 0.00	2.32 ± 3.02	2.81 ± 1.42	2.33 ± 0.95
Lung Right	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.00	1.42 ± 0.63	2.87 ± 1.37	2.19 ± 0.66
Thyroid	0.67 ± 0.12	0.58 ± 0.20	0.72 ± 0.07	7.11 ± 4.38	9.35 ± 9.03	5.37 ± 1.70
Humerus	0.86 ± 0.06	0.84 ± 0.05	0.70 ± 0.05	7.40 ± 4.44	8.97 ± 2.71	7.00 ± 2.00

Table 6.1: *Quantitative results of the retrospective study for auto-segmentation. Fulfilled requirements are indicated by green boxes, close-to-fulfilled requirements by yellow boxes, and not-fulfilled requirements by red boxes.*

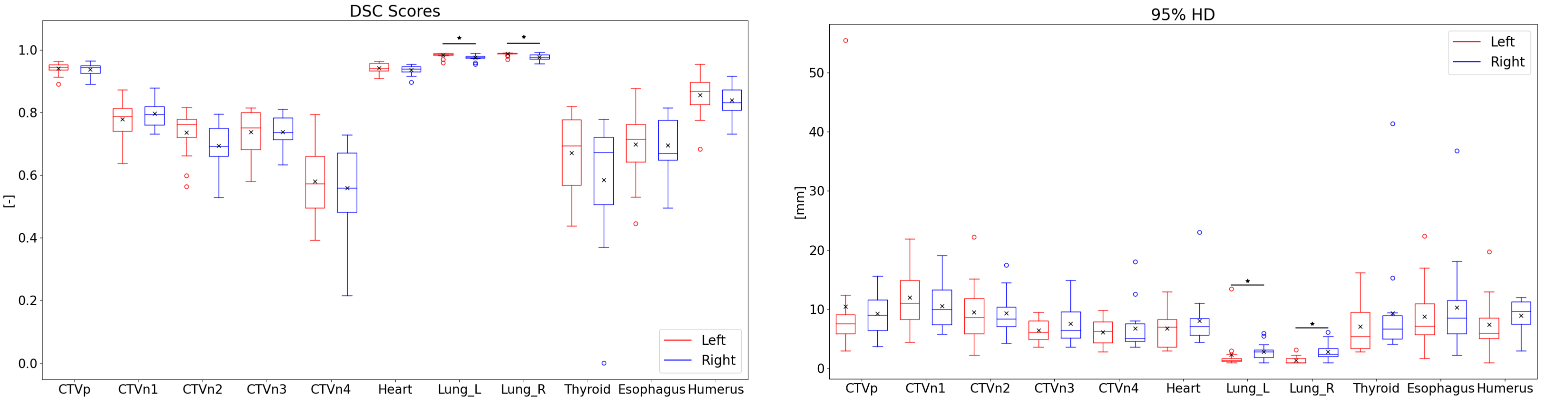


Figure 6.1: *Visualization of the quantitative results of the retrospective study for auto-segmentation. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences between the two models are indicated with with an asterisk (p < 0.05).*

continue without any additional changes. Moreover, Table 6.1 shows a slightly too high value for the 95% HD for the CTVp, with a high standard deviation for the left-sided model. However, as can be observed in Figure 6.1, this can be explained by the fact that there is one outlier for this ROI.

Figure 6.2 shows some axial slices for two example cases on which auto-segmentation is performed. For case 1, a high similarity between the auto-segmentation contours, visualized by the filled contours, and the manual contours, visualized by the lines, can be observed. However, in case 2 some faulty delineations are present. In the first two slices, the CTVp is wrongly delineated in the dorsal direction by delineating a larger area. In contrast, the automatically generated CTVn1 contour misses a region. On the last slice, two small orange spots can be observed, which are wrongly delineated spots belonging to the thyroid. However, in clinical practice these errors will be checked and corrected before use, which can still lead to time saving. For example, removing the small spots of the thyroid is quite easy, and the remaining part was well delineated.

Finally, the two models are statistically compared with the Wilcoxon Rank-Sum test to assess if there is a difference in performance. Only for both lungs, a p-value < 0.05 was found, indicating a better performance of the left-sided model for these ROIs. However, these differences are clinically irrelevant since the quantitative scores are still indicating almost perfect overlap for both sides. In addition, two experienced ROs evaluated the performance of the models on a set of 6 demo-patients, and concluded that the performance was sufficient to start the clinical pilot.

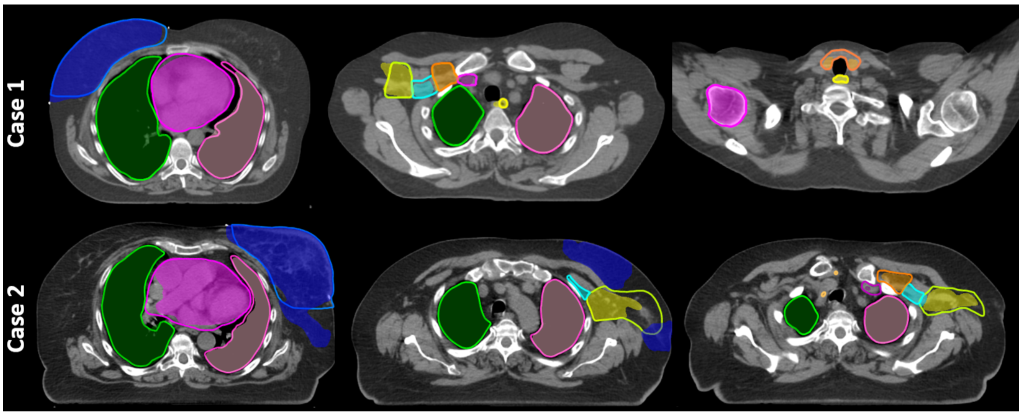


Figure 6.2: *Two example cases for auto-segmentation. Auto-segmentation delineations are represented by the filled contours, whereas manual delineations are represented by the lines.*

6.1.2 Clinical pilot

For the clinical pilot, both quantitative and qualitative measurements were performed. In total, 10 patients were used for both the left- and right-sided model, which were collected from clinical practice during the period of September 2021 - June 2022. For the left- and right sided set, respectively 4 and 7 patients included delineations of node levels 1 to 4, whereas the other patients only included levels 1 and 2, reflecting the variation within the patient group in clinical practice. The models showed comparable quantitative results before correction to the results found in the retrospective study, that can be found in Appendix E.

The most important outcome of this study is the time saved while using auto-segmentation. Only for one patient case out of 20, the time needed to correct the OARs took more time than the manual delineation, leading to a decrease in time in 95% of the patients, fulfilling the requirement set in Section 3.3.1. The mean time (hh:mm:ss) for manual delineation was 0:17:05 and 0:41:31 for the OARs and CTVs, respectively. While using auto-segmentation, the total time spent including correction was 00:08:47 and 0:15:43 for the OARs and CTVs, resulting in a reduction (mean ± std) of 42.4% ± 26.5% and 58.5% ± 19.1%, respectively.

The generated contours were also qualitatively scored by the RTTs and ROs, using a 3-point scale (clinically acceptable, corrections needed or not usable). For both models, the scores are visualized in Figure 6.3. While for the OARs, only the heart and thyroid both got scored as not usable in only one case, this was more often the case for one of the CTVs. The primary CTV (CTVp) was found not to be usable in 7 cases, the first node level (CTVn1) in 6 cases and the fourth node level (CTVn4) in 4 cases. A few observations can be made. First of all, the correction needed for the left and right lungs in respectively 20 and 50% of the cases for the right-sided model is remarkable, given the high quantitative score for these ROIs. Except for one case, these scores were assigned by the same observer, emphasizing the subjectivity of these scores. Moreover, it emphasizes the fact that inter-observer variability is present in the segmentation process. Something similar can be observed for CTVp and CTVn1, which were always assigned a score of 3 by one of the observers, which scored 4 cases in total. No correlation was found between the metrics and the assigned score, except for assigning score 3 to cases which were outliers in terms of DSC score and HD95%. However, score 3 was also assigned to ROIs which had high quantitative results.

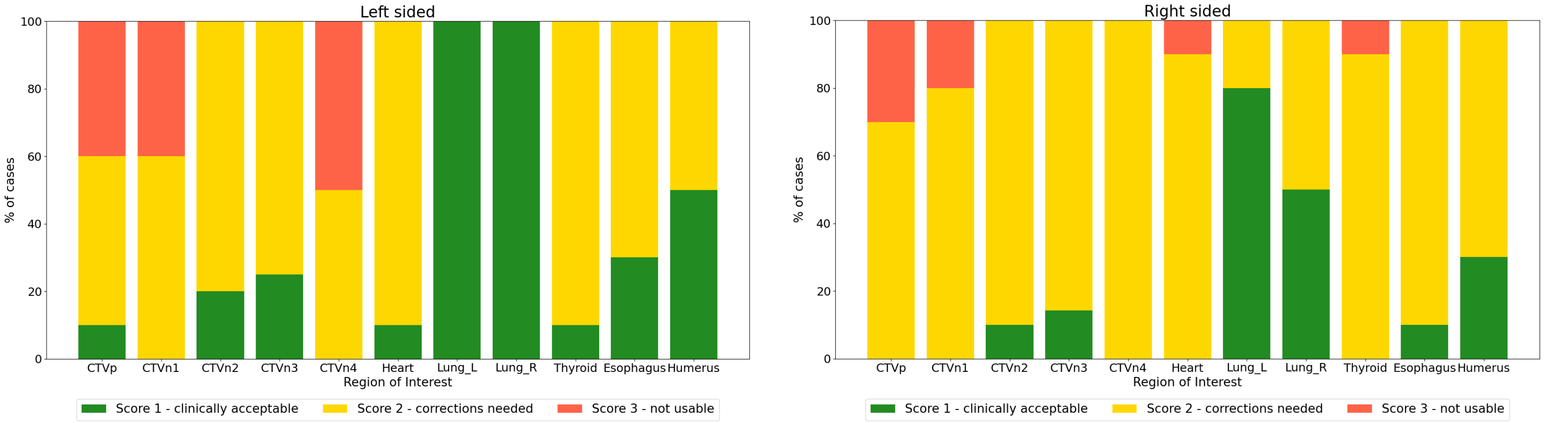


Figure 6.3: *Qualitative results of the clinical pilot for auto-segmentation. For each ROI, the percentage of cases receiving one of the scores is indicated. For both sides, 10 patients were included, which contain CTVn3 and CTVn4 in 4 and 7 cases, for respectively the left- and right-sided model.*

The impact of the corrections made was also quantitatively measured, by calculating the quantitative metrics for both the automatically generated and corrected structures, using the manually generated structure as ground-truth. A complete overview of these results can be found in Appendix E. The differences in DSC score and 95% HD for each structure are visualized in Figure 6.4. A positive difference in DSC score indicates a higher DSC score for the corrected structure and thus a better

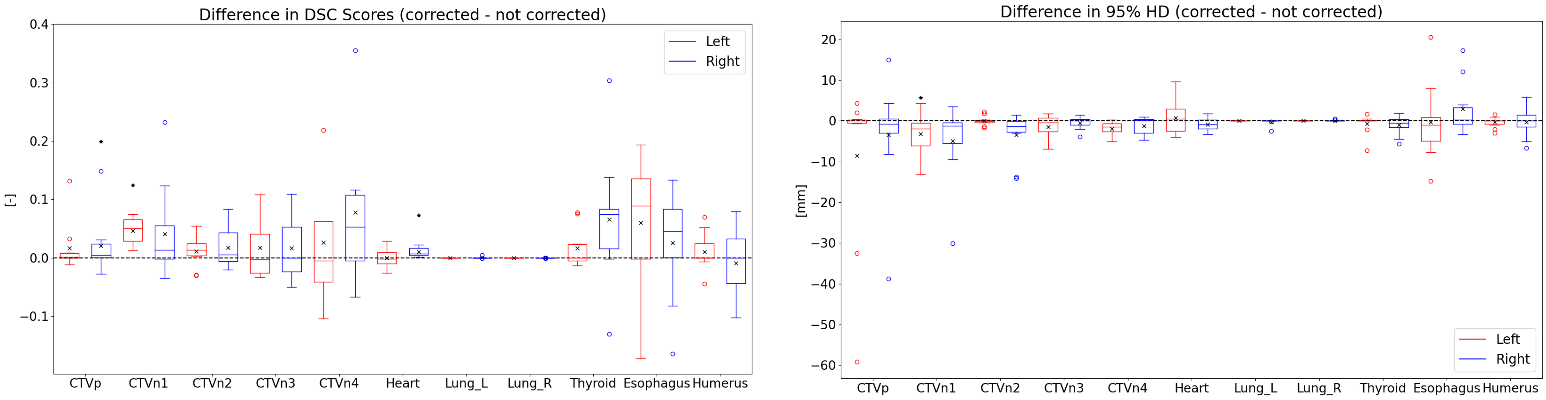


Figure 6.4: *The difference in quantitative measurements of the automatically generated contours and corrected contours in the clinical pilot. A better score for the corrected contours is indicated by a positive DSC value or a negative 95% HD value.*

agreement with the manual structure. The opposite is the case for the 95% HD, where a negative difference indicates a better agreement of the corrected with the manual structure. The Wilcoxon signed-rank test is used to assess statistically significant difference between the corrected and not-corrected structures. For the DSC score, a significant difference is found for the right-sided CTVp and heart and left-sided CTVn1, whereas for the 95% HD this was only the case for the left-sided CTVn1. These results indicate that, although corrections were made in several ROIs, only for a small number of cases these corrections are significantly different. Investigating the actual clinical relevance of these corrections was out of scope for this project.

6.2 Auto-Planning

6.2.1 Retrospective study

The retrospective results for conventional breast irradiation were analyzed in the previous project [15]. In Appendix A the method used, results and conclusion are briefly described. For FF irradiation, transfer learning was used to train a model and the predicted and mimicked dose distributions were evaluated. To compare the manually and automatically generated plans, all plans were scaled such that 98% of the PTV volume receives 95% of the prescribed dose, which is one of the clinical goals. The percentage of fulfilled clinical goals for each method can be found in Table 6.2. As can be seen, the predicted plans only failed one clinical goal for one patient, whereas more clinical goals were failed for the mimicked plans.

The DVH parameters are listed in Table 6.3. A Wilcoxon signed-rank test is performed to compare the doses of the predicted and mimicked plans with the manually generated doses, and show a significant difference for all DVH parameters of the mimicked plans. However, further tweaking of the mimick settings was not performed, which could improve these results. For the predicted plans, only a significant difference was found for the maximum dose the heart, indicating that transfer learning can be used successfully, requiring only a small dataset. Lastly, the Wilcoxon signed-rank test was used to compare the number of MUs used for the manual and auto-plan, resulting in a p-value of 0.2, meaning no significant difference was found, and thus fulfilling this requirement with regard to the number of MUs.

	Clinical goals met [%]		
	Manual	Auto-plan	
Clinical goal		**Predicted**	**Mimicked**
PTV: at most 2574 cGy average dose	100	100	100
PTV: at most 2626 cGy average dose	100	100	40
PTV: at most 2782 cGy at 2% volume	100	100	90
Lungs: at most 300 cGy average dose	100	100	100
Lungs: at most 200 cGy average dose	100	100	90
Heart: at most 150 cGy average dose	100	100	100
Heart: at most 100 cGy average dose	100	90	90
CL Breast: at most 100 cGy average dose	100	100	100
External-PTV: at most 10 cc at 2782 cGy	100	100	100

***Table 6.2:** The percentage of clinical goals met in the retrospective study for FF breast irradiation for the test set, containing 10 patients.*

		PTV		Heart	Lungs
		Dose [cGy]	**Difference wrt prescribed [%]**	**Dose [cGy]**	**Dose [cGy]**
Manual	**Average**	2609 ± 12	+0.34	64 ± 15	123 ± 33
	Maximum	2723 ± 23	+4.73	280 ± 141	1629 ± 418
Predicted	**Average**	2604 ± 10	+0.15	68 ± 18	126 ± 32
	Maximum	2708 ± 13	+4.13	*331 ± 197*	1664 ± 343
Mimicked	**Average**	*2633 ± 13*	+1.27	*72 ± 22*	*132 ± 32*
	Maximum	*2147 ± 17*	+5.65	*382 ± 264*	*1766 ± 364*

***Table 6.3:** Average and maximum doses in cGy to ROIs for the clinical plans, predicted and mimicked plans of the U-net model for FF irradiation (mean ± standard deviation). For PTV, the difference between mean average and maximum dose with respect to the prescribed dose (2600 cGy) is shown. Doses differing significantly from clinical doses are printed in italic.*

6.2.2 Clinical pilot

A clinical pilot was performed for conventional breast irradiation, in which the performance of the 2D U-net and cARF models was assessed by blindly scoring the resulting plans and a manually generated plan of 20 patients. An example case is shown in Figure 6.5, showing the dose distribution of a manual, cARF and U-net plan.

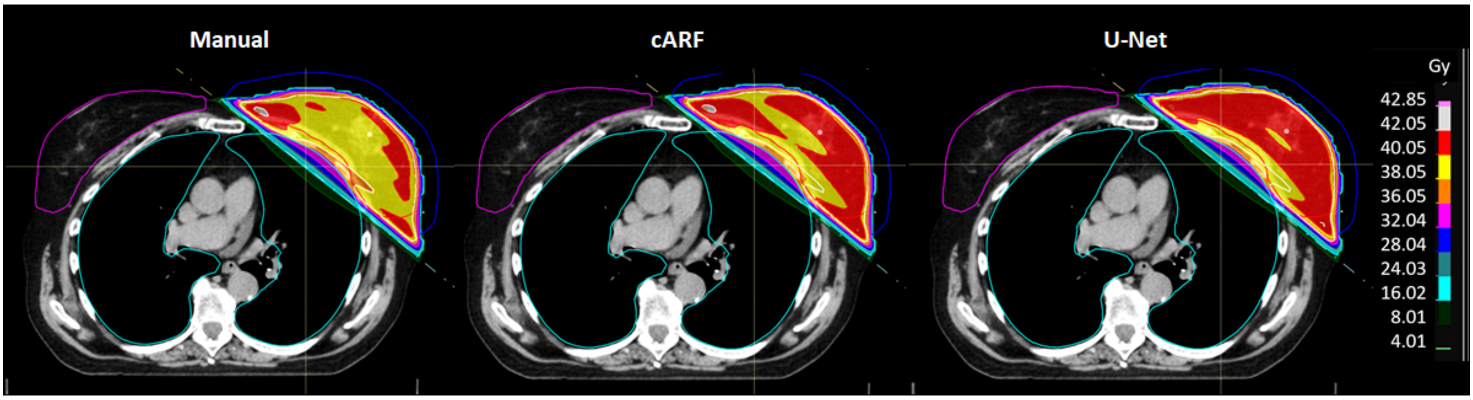


***Figure 6.5:** Example case for auto-planning, showing the manual, cARF and U-net plan.*

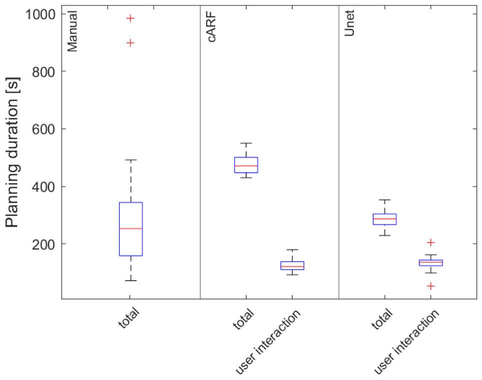


Figure 6.6: *Time needed for the plan generation. For the AI plans, the time spent on user interaction sis separately specified. The red crosses represent outliers.*

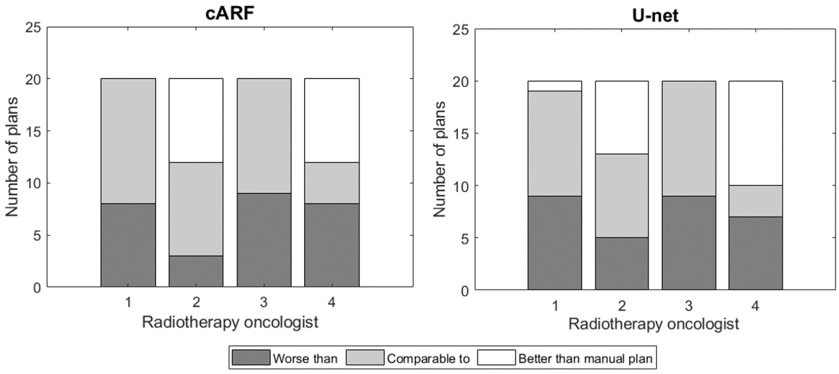


Figure 6.7: *Qualitative results of the clinical pilot for auto-planning, containing the ranking of the AI plans in comparison with the manual plan by the radiation oncologists on an individual basis*

The most important outcome measure is the time needed to manually and automatically generate the plans. The median time needed was 253 s (range 72-984 s) for the manual plans, 471 s (430-550 s, p=0.014) for the cARF plans, and 287 s (229-352 s, p=0.411) for the U-net plans. However, these times include computation times for the auto-plans, and the user interaction was only 121 s (92-180 s) for the cARF plans and 136 s (53-205 s) for the U-net plans. For the manual plans, the computation time was not recorded separately as it is often interleaved with manual adjustments. According to the technical requirements set in Section 3.3.1, a decrease in time in 90% of the cases was demanded. For the cARF and U-net plans, when considering user-interaction, a decrease in time was only achieved in 85% and 75% of the cases, respectively. However, the maximum observed increase of time was 65 s, whereas a decrease can be as much as 861 s. This observation is also reflected in Figure 6.6, where it can be observed that the time range for manual planning is much larger. So although the requirement is not fulfilled, it can be stated that auto-planning is a more time efficient way.

During the blind scoring of the plans, 90% of both the manual and cARF plans and 95% of the U-net plans were considered clinically acceptable by all 4 observers. As can be seen in Figure 6.7, the preference of the observers quite differed, resulting in only 35% of the cases where all observers independently agreed the auto-plan was equally suitable or better than the manual plan, and in 45-50% of the cases there was even no consensus (Table 6.4).

The quantitative measurements and its comparison to the requirements stated in Chapter 3.3.2 can be found in Appendix E. When comparing the qualitative results to the requirements, the minimum clinically acceptability rate of 90% is fulfilled for both models, and U-net even outperforms this requirement. Besides, the auto-plans should be non-inferior to the manual plans, using the Wilson score interval method, where the lower limit should be greater than 45%. For each observer, the CI is calculated and displayed in Table 6.5. Only for observer 2, the requirement of non-inferiority is fulfilled. However, when considering all 80 observations independently, both methods show non-inferiority.

6.2.3 Clinical implementation

As mentioned in Section 5.3.3, the 3D U-net needed to be commissioned before it could be used in clinic. In the commissioning process, 10 of the patients of the clinical pilot were used, enabling comparison of the 3D U-net plans with manual plans. The resulting DVH parameters and clinical goals can be found in Appendix E.2.2.

	Acceptable for all	Consensus autoplan worse	Consensus autoplan equal/better	No consensus
	[%]	[%]	[%]	[%]
Manual	90			
cARF	90	15	35	50
U-net	95	20	35	45

Table 6.4: *Qualitative results of the clinical pilot for auto-planning, evaluated by the radiation oncologists*

Observer	cARF	U-net
1	60% (39-79%)	55% (34-74%)
2	85% (64-95%)	75% (53-89%)
3	55% (34-74%)	55% (34-74%)
4	60% (39-79%)	65% (43-82%)
Total	65% (54-75%)	62.5% (52-72%)
(n = 80)		

Table 6.5: *Confidence intervals for each observer, calculated using the Wilson score interval method. Non-inferiority is achieved if the lower limit is higher than 45%.*

6.3 Summary

Several studies were performed to assess the performance of the design for both auto-segmentation and auto-planning. Figure 6.8 gives an overview of the most important outcomes, related to the set of requirements set in Chapter 3 and visualized in Figure 3.3.

For auto-segmentation, not all quantitative requirements were met in the retrospective study. However, after visual inspection of the results and evaluation by two experienced ROs, it was decided to start the clinical pilot. Similar values for the quantitative measures were found in the clinical pilot, indicating robustness of the model. The requirement of a decrease in time in 90% of the cases was fulfilled, with a mean time reduction of 42.4% and 58.5% for the OARs and CTVs, respectively. Besides, during qualitative measurement, the automatic structures got a score of 1 or 2 in 92% of the cases, indicating usefulness in clinic.

For auto-planning, not all requirements were met for the FF irradiation model after mimicking. However, the predicted dose did fulfill these requirements, and further tweaking of the mimick settings could improve the final results. For conventional irradiation, the cARF and U-net model resulted in a clinically acceptable plan in 90% to 95% of the cases, in a time efficient way. Finally, the 3D U-net model was successfully commissioned for clinical use. The results and follow-up steps will be further discussed in Chapter 8.

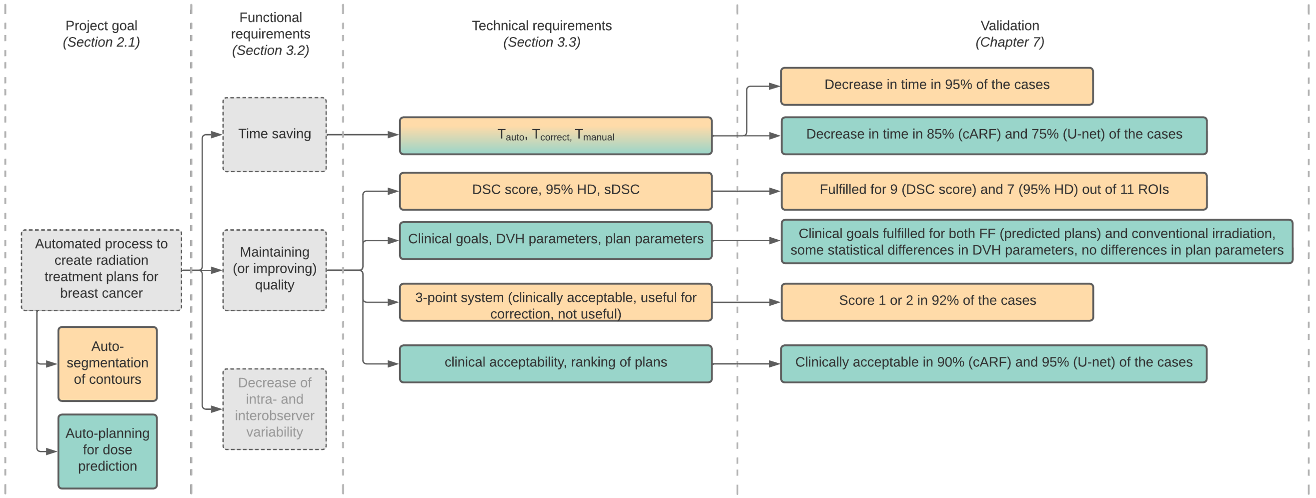

Project goal
(Section 2.1)
Functional requirements
(Section 3.2)
Technical requirements
(Section 3.3)
Validation
(Chapter 7)
Automated process to create radiation treatment plans for breast cancer
Auto-segmentation of contours
Auto-planning for dose prediction
Time saving
Maintaining (or improving) quality
Decrease of intra- and interobserver variability
T_{auto}, $T_{correct}$, T_{manual}
DSC score, 95% HD, sDSC
Clinical goals, DVH parameters, plan parameters
3-point system (clinically acceptable, useful for correction, not useful)
clinical acceptability, ranking of plans
Decrease in time in 95% of the cases
Decrease in time in 85% (cARF) and 75% (U-net) of the cases
Fulfilled for 9 (DSC score) and 7 (95% HD) out of 11 ROIs
Clinical goals fulfilled for both FF (predicted plans) and conventional irradiation, some statistical differences in DVH parameters, no differences in plan parameters
Score 1 or 2 in 92% of the cases
Clinically acceptable in 90% (cARF) and 95% (U-net) of the cases

Figure 6.8: *Overview of the set of functional and technical requirements and most important outcomes. Requirements and outcomes related to auto-segmentation are indicated with yellow, whereas requirements and outcomes related to auto-planning are indicated with green.*

7 | Conclusions

The goal of this design project was to automate the treatment planning process for breast irradiation using AI. Several models were developed and/or trained and validated for segmentation of contours (auto-segmentation) and dose prediction for treatment plan optimization (auto-planning). For both purposes, a retrospective study was performed, followed by a clinical pilot.

For auto-segmentation, a mean time reduction of 42% for OARs and 59% for CTVs was achieved, fulfilling the primary end goal of time saving. Besides, the contours were scored as directly usable or useful for correction in 92% of the cases, indicating clinical usefulness.

For auto-planning, transfer learning showed promising results in the retrospective study to create a FF irradiation model, based on the previously trained model for conventional irradiation. For conventional irradiation, two models were tested in a clinical pilot, where the 2D U-net model was scored as clinically acceptable for all observers in 95% of the cases, in a time efficient way.

Both models showed promising results and added value in the radiotherapy treatment planning process, with clinically acceptable outputs while saving time. These results already led to successful clinical implementation of the AI model for auto-planning at the radiotherapy department at CZE at May 2022 (during this design project), and preparations are started to implement the AI model for auto-segmentation in Q3 of 2022 (after the design project).

8 Discussion and Recommendations

In this design project, AI models were developed, trained and evaluated to automate segmentation and plan optimization for breast cancer. Throughout the process, several critical moments occurred and crucial choices were made, which influenced the outcome of the project. In this section, these moments and choices are discussed, as well as alternatives for these situations. Moreover, future recommendations for successful implementation of AI models in clinic are discussed.

8.1 Auto-segmentation

For auto-segmentation, a model provided by RaySearch was trained on the data from our own clinic, retrospectively obtained from the clinical archive. Even though these delineations were checked before treatment of the patient, not all delineations fulfilled the guidelines used in our clinic, causing the need for extensive review of the dataset. Several causes can be distinguished for this problem. First of all, some delineations are adapted to be patient specific. For example, for some patients a bigger region around the lymph nodes was delineated, as the PET scan indicated that those regions were also involved in the malignant region. Second, the importance of accuracy is different for the CTVs and OARs volumes, where CTVs need higher accuracy and anatomical individualisation. Besides, throughout the years, the interpretation of the guidelines can change, causing a difference in delineations. For example, for high risk patients, recently the lymph node areas were enlarged. At last, there is a known interobserver variability, which was already discussed in Section 3.2.3. The review of the dataset needs to be performed in close collaboration with experienced RTTs and ROs, to ensure a high quality dataset. However, available time of these users can be scarce, resulting in the need of a clear procedure on how to review this data.

During the training process of the model, several choices were made, which are further explained in Appendix B. These choices resulted in an increase of model performance. However, for some structures, in particular the esophagus and thyroid, the desired result was still not achieved, as is discussed in Section 6.1.1. These outcomes could probably be improved by using a larger and more uniform dataset for training. In this project, it was chosen to not invest in collecting more data, as this is a time consuming and user intensive process. It was decided to continue to the clinical pilot, to assess the clinical importance of the deviations from the requirements. Although these structures needed correction in most of the cases, time saving was still achieved when compared to manual delineation, therefore meeting the primary endpoint. In the future, more data could be included to improve the model performance, leading to a decrease in corrections needed.

During analysis of the results of the clinical pilot, it was observed that different scores were assigned by different RTTs and ROs for comparable corrections made. For example, one RO would assign a score 2 if it had to correct 8 slices of the CTVp, while another RO would assign a score 3. Therefore, it is important to further investigate the actual corrections made during the pilot. Moreover, the results should be discussed with the RTTs and ROs involved. Besides, the pilot only involved auto-contours, which could induce subjectivity. Users could judge automatically generated contours differently, then they would judge manually generated contours by a change in perception towards the use of AI. This difference could be overcome by performing a head-to-head comparison, such as the Turing test, in which the user has to identify the origin of the contour [44]. However, as the execution of the clinical pilot is already time-consuming, and the primary outcome of the model is to evaluate time-saving and not quality, this test was not performed.

Recently, a study was published by Almberg *et al.* in collaboration with RaySearch, training a model for the same target area. Similar results were achieved, except for CTVn4, thyroid and humerus, which scored better in the other study. These regions were known to perform less in the model trained for our project as explained above. In a future study, an extensive comparison could be made to assess the difference in performance in more depth and explore the origin of these differences. For example, their model could be tested on our data and vice versa, which tests the generalisability of the model. Moreover, this comparison could expose differences present in the dataset, which would be valuable information.

An evaluation method of the model performance which was not performed in this study, but is used in other studies involving auto-segmentation, is dosimetric evaluation [10, 45]. This method gives an indication of the clinical relevance of variations in contours and could therefore quantify the clinical relevance of corrections made. It can be performed by comparing the dose delivered to automatically and manually generated contours. For a fair comparison, variability existing in creating treatment plans should be eliminated, which can be fulfilled by using auto-planning. Besides, it is also important to only evaluate the dosimetric parameters that correlate with clinical outcomes, which is not always the case for parameters used in clinical trials and routine practice [45]. In a future study, the dosimetric impact of this auto-segmentation model could be evaluated. Ideally, this information could be incorporated in the auto-segmentation tool, indicating in which regions the uncertainty of the auto-contour is large, and in which regions corrections would be clinical relevant.

8.2 Auto-planning

Throughout the course of this design project, several models were developed, trained and evaluated for auto-planning, including two different fraction schemes. A retrospective study was performed for the FF irradiation scheme. Although the predicted dose showed comparable results in terms of average and maximum dose to the PTV and OARs, the mimicked dose lead to significant higher doses. Thereby, the average dose to PTV was too high in 60% of the mimicked plans, which could lead to clinically unacceptable plans. However, the mimick settings could be changed to improve these results, which was not investigated in this study, since the main goal was to investigate if transfer learning would be feasible to develop a model with less data available. The results of the predicted plans indicate that this is the case. The concept of transfer learning was also discussed with RaySearch, which confirmed the possibility of this method using their own model architecture and framework. Therefore, the 3D U-net, developed by RaySearch, could be trained using the relatively small dataset, using the outcome of the trained 3D U-net for conventional irradiation as starting point.

For conventional irradiation, model development and retrospective evaluation was already performed. This design project involved the clinical pilot, which showed promising results. However, some issues had to be overcome. The most important issue was a change in the evaluation of the dose to PTV, demanding an average dose between 99% and 101% of the prescribed dose. However, this requirement was not represented in the plans involved in training of the model, which often contained a higher average dose. To assess the performance of the model, it was chosen to use the evaluation criteria which were used when collecting the data, and thus exclude the new PTV criteria. Therefore, the RTTs were asked to create manual plans as if this criteria was not introduced. However, since the RTTs were already used to the new criteria in clinical practice, and also the ROs were used to judge the plans based on these criteria, a bias was introduced in creation the manual plans and blind scoring of all plans. In addition to this bias, a difference in preference was observed in the scoring of the ROs, resulting in a low consensus rate. This difference in preference is insightful when developing and validating a model, and stresses that a model should not only be validated by quantitative measures. In addition, it should be noted that these difference in perception could be considered as a more general issue, and not only related to the performance of the model. Since the auto-plans and manual plans perform comparably in terms of clinical goals, it should be discussed whether the differences in preference are considered relevant, which calls for peer review, education and possible new guidelines.

While in the clinical pilot the new evaluation criteria were ignored, they should of course be covered to enable clinical implementation. First, RaySearch trained the 3D U-net with our dataset, generating comparable predictions. To further tweak the model outcome, multiple sessions were held with the ML engineers of RaySearch to finalize the mimick settings. The changes in the settings lead to a decrease in the average dose to PTV, while still remaining a low dose to the OARs. It is important to gain more knowledge of the effect of these mimick settings on the model outcomes, to be able to test more settings, independent of RaySearch.

As mentioned in Section 5.3.3, the 3D U-net was successfully commissioned and is used in clinic since May 2022. Although a thoroughly test procedure was performed, including the retrospective study, clinical pilot and commissioning process, new questions and issues arose. RTTs were unsure if and what adaptions they were allowed to make, and had issues to judge if the plan was optimal for the patient, or if, for example, the dose to the OARs could be further decreased. Therefore, it is of utmost importance to be present at clinic after clinical implementation to answer such questions and support with the first patient cases. Moreover, it stresses the importance of involving some users in an early-stage of the project. In this project, the involved RTTs could support in clinic, which was very valuable. Regarding the adaptations made in the plans, it was observed that as the model was used more often, less adaptations were made. This indicates an increase in trust in the model. Some RTTs created a manual plan, next to the automatic plan, and concluded that the automatic plan was comparable and thus usable. So although this finding was already proved in several studies, it is important that the end-user gets familiar with the AI model and its outcomes.

8.3 Future recommendations

Next to the future recommendations regarding auto-segmentation and -planning, mentioned above, more aspects need to be explored to successfully integrate AI in daily clinical practice. An advisory report is written, in which the recently drafted vision of the CZE on the use of AI is reviewed, and needed changes and concrete next steps are set up for the department of radiotherapy to meet up with this vision. The report can be found in Appendix F. One of the key-points of this report is the need for an "AI responsible person", who will monitor the different AI projects of the department and the documentation and registration of these projects, as mentioned earlier in Section 5.5. Besides, this person will be in contact with other departments involved in the AI process, such as the legal and ICT department and the, yet to be set up, AI competence center. Besides, the need for multidisciplinary teams within the department is stressed, to involve the end-users at an early stage. Finally, the added value of an AI model should be defined, including the target profit, and eventually measuring the achieved profit.

9 | Reflection

9.1 Project reflection

During the QME program, several tools and resources are provided to support the design process and project management. This project involved a larger research aspect than most design projects, which made it sometimes hard at first sight to directly apply these tools and resources. However, at several moments these tools definitely supported in this project.

First of all, the initial project planning differed from the final project planning due to unforeseen delays, as explained in Section 2.4. However, by working in an iterative way and evaluating the project progress at several moments throughout the project, the project team could be flexible and all deliverables could be completed in time. Improvements regarding to time management could have been made regarding efficiency of data collection. A lot of work was involved in checking and adjusting the data, executed in multiple iterations. However, when more knowledge of the data at the start of the project would have been gained, by for example a start-up meeting with the end-users to highlight the key features to pay attention to, less iterations would have been needed.

The importance of these end-users was also highlighted by creating a stakeholder map and identifying users and suppliers. When performing the clinical pilot for auto-planning, there was a mismatch between the evaluation criteria of the design and clinical reality. After involving the end-users more in the project, the execution of the clinical pilot for auto-segmentation went well and in an efficient way, while receiving positive feedback and also more involvement and enthusiasm of other RTTs and ROs that were not directly involved.

Regarding the design cycle provided by the QME program, the emphasis on the functional and technical requirements, and how to validate and verify these, benefited the final design. It was not straightforward to find and set hard requirements, which led to fruitful discussions with the project team and during the project reviews, improving the quality of the executed studies and final design.

In conclusion, although at first sight this project was not a regular design project, applying the design cycle and other tools had added value in succesfully completing the project.

9.2 Personal reflection

This design project gave me the opportunity to clinically validate my previously conducted research and actually implement own developed and trained models in clinical practice. This opportunity was really valuable, as it is quite rare to be involved in such a project from start to end. These next steps meant that the clinical side became more important, and more discussion was needed with the end-users to make sure the end product has added value. Unfortunately, Covid-19 made us work from home for a part of the project, mainly during the first year. This situation made it difficult to get in contact with the RTTs and ROs. Besides, the clinical workload increased, making the available time of the RTTs and ROs even more scarce. Luckily, during the clinical pilot, contact with the RTTs and ROs was good and valuable feedback was given during the clinical implementation of the auto-planning model.

Personally, I have experienced growth on several aspects during this design project. First of all, I have gained knowledge about processes at the department, but also about more overarching processes of the hospital, for example with regard to the legal aspects of conducting research. During the project, I gained more trust from the end-users by intermediately discussing results and evaluating the performance, which was important for the acceptance of the design in clinical practice. Although I could have been more active in promoting my project at the department, which was also harder due to the working form home situation, I already grew in communicating my results. This growth was also reflected in multiple presentations and talks that I gave during the design project. Although it is always a bit uncomfortable for the first few times, I am now able to explain my project, results and other relevant topics in a clear and calm way, with more confidence. Another important point of growth is taking more initiative to propagate my ideas and solutions. My professional network has been expanded over the past few years, especially regarding AI researchers and other people involved in those projects in our hospital. I took the initiative to discuss my experiences and ideas regarding my AI project. This lead to the ability to participate as one of the first use-cases for the launch of the AI platform of the hospital. Furthermore, I became involved in a hospital-wide initiative to further professionalize AI projects. In the future, I look forward to contribute to implementing AI in a safe and value adding manner, to improve healthcare outcomes for patients.

Bibliography

[1] Borstkanker in Nederland, kerncijfers uit de Nederlandse Kankerregistratie. `https://iknl.nl/borstkankercijfers`. Geraadpleegd op: 03-05-2022.

[2] Gábor Cserni, Ewa Chmielik, Bálint Cserni, and Tibor Tot. The new tnm-based staging of breast cancer. *Virchows Archiv*, 472(5):697–703, 2018.

[3] Elaine M Zeman, Eric C Schreiber, and Joel E Tepper. Basics of radiation therapy. In *Abeloff's Clinical Oncology*, pages 431–460. Elsevier, 2020.

[4] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. Artificial intelligence in healthcare. *Nature biomedical engineering*, 2(10):719–731, 2018.

[5] Trishan Panch, Peter Szolovits, and Rifat Atun. Artificial intelligence, machine learning and health systems. *Journal of global health*, 8(2), 2018.

[6] Saad Albawi, Tareq Abed Mohammed, and Saad Al-Zawi. Understanding of a convolutional neural network. In *2017 international conference on engineering and technology (ICET)*, pages 1–6. Ieee, 2017.

[7] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616, 2009.

[8] Miriam Santoro, Silvia Strolin, Giulia Paolani, Giuseppe Della Gala, Alessandro Bartoloni, Cinzia Giacometti, Ilario Ammendolia, Alessio Giuseppe Morganti, and Lidia Strigari. Recent applications of artificial intelligence in radiotherapy: Where we are and beyond. *Applied Sciences*, 12(7):3223, 2022.

[9] Philip MP Poortmans, Silvia Takanen, Gustavo Nader Marta, Icro Meattini, and Orit Kaidar-Person. Winter is over: the use of artificial intelligence to individualise radiation therapy for breast cancer. *The Breast*, 49:194–200, 2020.

[10] Liesbeth Vandewinckele, Michaël Claessens, Anna Dinkla, Charlotte Brouwer, Wouter Crijns, Dirk Verellen, and Wouter van Elmpt. Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance. *Radiotherapy and Oncology*, 153:55–66, 2020.

[11] Vikneswary Batumalai, Michael G Jameson, Dion F Forstner, Philip Vial, and Lois C Holloway. How important is dosimetrist experience for intensity modulated radiation therapy? a comparative analysis of a head and neck case. *Practical radiation oncology*, 3(3):e99–e106, 2013.

[12] Sean L Berry, Amanda Boczkowski, Rongtao Ma, James Mechalakos, and Margie Hunt. Interobserver variability in radiation therapy plan output: results of a single-institution study. *Practical radiation oncology*, 6(6):442–449, 2016.

[13] Delia Ciardo, Angela Argenone, Genoveva Ionela Boboc, Francesca Cucciarelli, Fiorenza De Rose, Maria Carmen De Santis, Alessandra Huscher, Edy Ippolito, Maria Rosa La Porta, Lorenza Marino, et al. Variability in axillary lymph node delineation for breast cancer radiotherapy in presence of guidelines on a multi-institutional platform. *Acta oncologica*, 56(8):1081–1088, 2017.

[14] Shalini K Vinod, Myo Min, Michael G Jameson, and Lois C Holloway. A review of interventions to reduce inter-observer variability in volume delineation in radiation oncology. *Journal of medical imaging and radiation oncology*, 60(3):393–406, 2016.

[15] Nienke Bakx. Optimization and automation of radiotherapy treatment plans for breast cancer (master thesis report). 2020.

[16] Adrian Murray Brunt, Joanne S Haviland, Duncan A Wheatley, Mark A Sydenham, Abdulla Alhasso, David J Bloomfield, Charlie Chan, Mark Churn, Susan Cleator, Charlotte E Coles, et al. Hypofractionated breast radiotherapy for 1 week versus 3 weeks (fast-forward): 5-year efficacy and late normal tissue effects results from a multicentre, non-inferiority, randomised, phase 3 trial. *The Lancet*, 395(10237):1613–1626, 2020.

[17] Birgitte V Offersen, Liesbeth J Boersma, Carine Kirkove, Sandra Hol, Marianne C Aznar, Albert Biete Sola, Youlia M Kirova, Jean-Philippe Pignol, Vincent Remouchamps, Karolien Verhoeven, et al. Estro consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer. *Radiotherapy and oncology*, 114(1):3–10, 2015.

[18] Birgitte V Offersen, Liesbeth J Boersma, Carine Kirkove, Sandra Hol, Marianne C Aznar, Albert Biete Sola, Youlia M Kirova, Jean-Philippe Pignol, Vincent Remouchamps, Karolien Verhoeven, et al. Estro consensus guideline on target volume delineation for elective radiation therapy of early stage breast cancer, version 1.1. *Radiotherapy and oncology*, 118(1):205–208, 2016.

[19] Jordan Wong, Vicky Huang, Derek Wells, Joshua Giambattista, Jonathan Giambattista, Carter Kolbeck, Karl Otto, Elantholi P Saibishkumar, and Abraham Alexander. Implementation of deep learning-based auto-segmentation for radiotherapy planning structures: a workflow study at two cancer centers. *Radiation Oncology*, 16(1):1–10, 2021.

[20] Mingqing Wang, Qilin Zhang, Saikit Lam, Jing Cai, and Ruijie Yang. A review on application of deep learning algorithms in external beam radiotherapy automated treatment planning. *Frontiers in oncology*, 10:2177, 2020.

[21] Mariel Cornell, Robert Kaderka, Sebastian J Hild, Xenia J Ray, James D Murphy, Todd F Atwood, and Kevin L Moore. Noninferiority study of automated knowledge-based planning versus human-driven optimization across multiple disease sites. *International Journal of Radiation Oncology* Biology* Physics*, 106(2):430–439, 2020.

[22] Chris McIntosh, Leigh Conroy, Michael C Tjong, Tim Craig, Andrew Bayley, Charles Catton, Mary Gospodarowicz, Joelle Helou, Naghmeh Isfahanian, Vickie Kong, et al. Clinical integration of machine learning for curative-intent radiation treatment of patients with prostate cancer. *Nature Medicine*, 27(6):999–1005, 2021.

[23] X Allen Li, An Tai, Douglas W Arthur, Thomas A Buchholz, Shannon Macdonald, Lawrence B Marks, Jean M Moran, Lori J Pierce, Rachel Rabinovitch, Alphonse Taghian, et al. Variability of target and normal structure delineation for breast cancer radiotherapy: an rtog multi-institutional and multiobserver study. *International Journal of Radiation Oncology* Biology* Physics*, 73(3):944–951, 2009.

[24] 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.

[25] Hwa Kyung Byun, Jee Suk Chang, Min Seo Choi, Jaehee Chun, Jinhong Jung, Chiyoung Jeong, Jin Sung Kim, Yongjin Chang, Seung Yeun Chung, Seungryul Lee, et al. Evaluation of deep learning-based autosegmentation in breast cancer radiotherapy. *Radiation Oncology*, 16(1):1–8, 2021.

[26] Juanqi Wang, Weigang Hu, Zhaozhi Yang, Xiaohui Chen, Zhiqiang Wu, Xiaoli Yu, Xiaomao Guo, Saiquan Lu, Kaixuan Li, and Gongyi Yu. Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? *Radiation Oncology*, 12(1):1–8, 2017.

[27] Seung Yeun Chung, Jee Suk Chang, Min Seo Choi, Yongjin Chang, Byong Su Choi, Jaehee Chun, Ki Chang Keum, Jin Sung Kim, and Yong Bae Kim. Clinical feasibility of deep learning-based auto-segmentation of target volumes and organs-at-risk in breast cancer patients after breast-conserving surgery. *Radiation Oncology*, 16(1):1–10, 2021.

[28] Ahmed R Eldesoky, Esben S Yates, Tine B Nyeng, Mette S Thomsen, Hanne M Nielsen, Philip Poortmans, Carine Kirkove, Mechthild Krause, Claus Kamby, Ingvil Mjaaland, et al. Internal and external validation of an estro delineation guideline–dependent automated segmentation tool for loco-regional radiation therapy of early breast cancer. *Radiotherapy and Oncology*, 121(3):424–430, 2016.

[29] Yang Sheng, Taoran Li, Sua Yoo, Fang-Fang Yin, Rachel Blitzblau, Janet K Horton, Yaorong Ge, and Q Jackie Wu. Automatic planning of whole breast radiation therapy using machine learning models. *Frontiers in Oncology*, page 750, 2019.

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

[31] Abdel Aziz Taha and Allan Hanbury. Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool. *BMC medical imaging*, 15(1):1–28, 2015.

[32] Stanislav Nikolov, Sam Blackwell, Alexei Zverovitch, Ruheena Mendes, Michelle Livne, Jeffrey De Fauw, Yojan Patel, Clemens Meyer, Harry Askham, Bernardino Romera-Paredes, et al. Deep learning to achieve clinically applicable segmentation of head and neck anatomy for radiotherapy. *arXiv preprint arXiv:1809.04430*, 2018.

[33] 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.

[34] P Meyer, M-C Biston, C Khamphan, T Marghani, J Mazurier, V Bodez, L Fezzani, PA Rigaud, G Sidorski, L Simon, et al. Automation in radiotherapy treatment planning: Examples of use in clinical practice and future trends for a complete automated workflow. *Cancer/Radiothérapie*, 25(6-7):617–622, 2021.

[35] Seokyung Hahn. Understanding noninferiority trials. *Korean journal of pediatrics*, 55(11):403, 2012.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *arXiv preprint arXiv:1505.04597*, pages 234–241, 2015.

[37] Özgün Çiçek, Ahmed Abdulkadir, Soeren S Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *arXiv preprint arXiv:1606.06650*, pages 424–432, 2016.

[38] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[39] Chris McIntosh and Thomas G Purdie. Contextual atlas regression forests: multiple-atlas-based automated dose prediction in radiation therapy. *IEEE transactions on medical imaging*, 35(4):1000–1012, 2015.

[40] Chris McIntosh and Thomas G Purdie. Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. *Physics in Medicine & Biology*, 62(2):415, 2016.

[41] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International conference on artificial neural networks*, pages 270–279. Springer, 2018.

[42] Simone Dietzenbacher. Training and evaluation of a model predicting high-dose radiotherapy treatment plans for breast cancer patients (8ZM00 internship report). 2021.

[43] Joseph DeRosier, Erik Stalhandske, James P Bagian, and Tina Nudell. Using health care failure mode and effect analysis™: the va national center for patient safety's prospective risk analysis system. *The Joint Commission journal on quality improvement*, 28(5):248–267, 2002.

[44] 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.

[45] Michael V Sherer, Diana Lin, Sharif Elguindi, Simon Duke, Li-Tee Tan, Jon Cacicedo, Max Dahele, and Erin F Gillespie. Metrics to evaluate the performance of auto-segmentation for radiation treatment planning: A critical review. *Radiotherapy and Oncology*, 160:185–191, 2021.

Appendices

A | Retrospective study auto-planning

A.1 Introduction

Prior to the QME design project, in the master thesis project of the same author a retrospective study was performed in which two auto-planning models were validated for conventional breast irradiation [15]. This chapter briefly describes the method used, the results and the conclusion of the study.

A.2 Materials and methods

For this study, 105 patients were included, all diagnosed with left-sided node-negative breast cancer and treated in 15 fractions with a prescribed total dose of 40.05 Gy. Two AI models were used for dose prediction. The first model was in-house developed, based on an adapted version of the U-net architecture[1]. The input of the U-net model contains 4 channels, consisting of binary masks of the PTV, the OARs (heart and lungs) and the external. For the PTV mask, the value of the prescribed dose is assigned to voxels within the structure, whereas ones are assigned to voxels within the structure for the other masks. Training was performed using a batch size of 24 slices (8 slices of 3 patients), using a Gaussian sampling scheme with a standard deviation equal to one-third of the distance from the central to the end slice. This sampling scheme was used, as central slices are mode important for dose prediction, as they contain the PTV. The second model was based on a cARF algorithm, developed and integrated by RaySearch, which is further described by McIntosh and Purdie[2,3]. Both models were trained with the treatment plans of 90 patients (training/validation ratio of 80/20) and independently tested with the treatment plans of 15 additional patients. Dose mimicking was used for both models to create clinically deliverable plans after voxel-wise dose prediction by the model. The resulting treatment plans are evaluated with the use of clinical goals, after scaling all plans such that 98% of the PTV volume receives 95% of the prescribed dose. In addition, the average and maximum doses of the ROIs were calculated, as well as the percent error relative to the prescribed dose ($\frac{predicted\ dose - clinical\ dose}{prescribed\ dose} * 100\%$). Significance between the average and maximum doses of the clinical plans and auto-plans was investigated with the Wilcoxon signed rank test.

A.3 Results

Table A.1 show the number of clinical goals met for both models. After mimicking, three patient plans exceeded the allowed volume of 2% receiving more than 4285 cGy for the U-net model. For the cARF model, three different patient cases also failed to meet all clinical goals, with one patient having a lower average PTV dose than allowed, one patient receiving more than 4285 cGy in the external and one patient having a heart dose higher than 200 cGy.

[1] Nguyen, D., Long, T., Jia, X., Lu, W., Gu, X., Iqbal, Z., & Jiang, S. (2019). A feasibility study for predicting optimal radiation therapy dose distributions of prostate cancer patients from patient anatomy using deep learning. Scientific reports, 9(1), 1-10.

[2] McIntosh, C., & Purdie, T. G. (2015). Contextual atlas regression forests: multiple-atlas-based automated dose prediction in radiation therapy. IEEE transactions on medical imaging, 35(4), 1000-1012.

[3] McIntosh, C., & Purdie, T. G. (2016). Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning. Physics in Medicine & Biology, 62(2), 415.

	Number of patients achieving goal			
	Manual	**U-net**		**cARF**
Clinical goal		**Predicted**	**Mimicked**	
PTV: average dose ≥ 4005 cGy	15	15	15	14
PTV: at most 4285 cGy at 2% volume	15	15	12	15
Heart: at most 300 cGy average dose	15	15	15	15
Heart: at most 200 cGy average dose	15	15	15	14
Lungs: at most 600 cGy average dose	15	15	15	15
Lungs: at most 400 cGy average dose	15	15	15	15
External-PTV: at most 10 cc at 4285 cGy	15	15	15	14

Table A.1: *Number of patients of the test set (n = 15) achieving the clinical goals, for the clinical plans, predicted and mimicked plans of the U-net model and conditional Atlas Regression Forest (cARF) plans. Goals printed in italic are of lower priority than the others, meaning they are target values, not hard constraints.*

The average and maximum doses to the ROIs are listed in Table A.2. After mimicking, both average and maximum doses increased for the U-net model, thereby significantly exceeding the clinical doses. The percent errors on these doses are shown in Figure A.1. The U-net model did not differ significantly from the clinical dose distribution before mimicking, but did afterwards for all ROIs. For the cARF model, this was the case for both heart and lungs. Overall, the percent error was lower for the average doses compared to the maximum doses, and the spread of the error was decreased after mimicking. The average doses were within a range of 1.0% to 1.5% compared to the clinical plans, whereas the maximum doses to heart and lungs deviated more, within a range of 6.6% and 3.3%. However, these differences were not found clinically relevant, since the clinical accepted average doses were not exceeded.

		PTV		**Heart**	**Lungs**
		Dose [cGy]	**Difference wrt prescribed [%]**	**Dose [cGy]**	**Dose [cGy]**
Clinical	**Average**	4054 ± 24	+1.2	96 ± 37	189 ± 53
	Maximum	4206 ± 32	+5.0	441 ± 377	2616 ± 465
U-net-pred	**Average**	4050 ± 11	+1.1	101 ± 25	197 ± 37
	Maximum	4222 ± 19	+5.4	493 ± 311	3684 ± 390
U-net-mim	**Average**	*4081 ± 25*	+1.9	*102 ± 36*	*199 ± 49*
	Maximum	*4243 ± 39*	+5.9	*489 ± 375*	*2749 ± 433*
cARF	**Average**	4044 ± 24	+1.0	*109 ± 47*	*208 ± 53*
	Maximum	4207 ± 37	+5.0	*594 ± 615*	*2881 ± 368*

Table A.2: *Average and maximum doses in cGy to ROIs for the clinical plans, predicted and mimicked plans of the U-net model and mimicked plans generated by the cARF model (mean ± standard deviation). For PTV, the difference between mean average and maximum dose with respect to the prescribed dose (4005 cGy) is shown. Doses differing significantly from clinical doses are printed in italic*

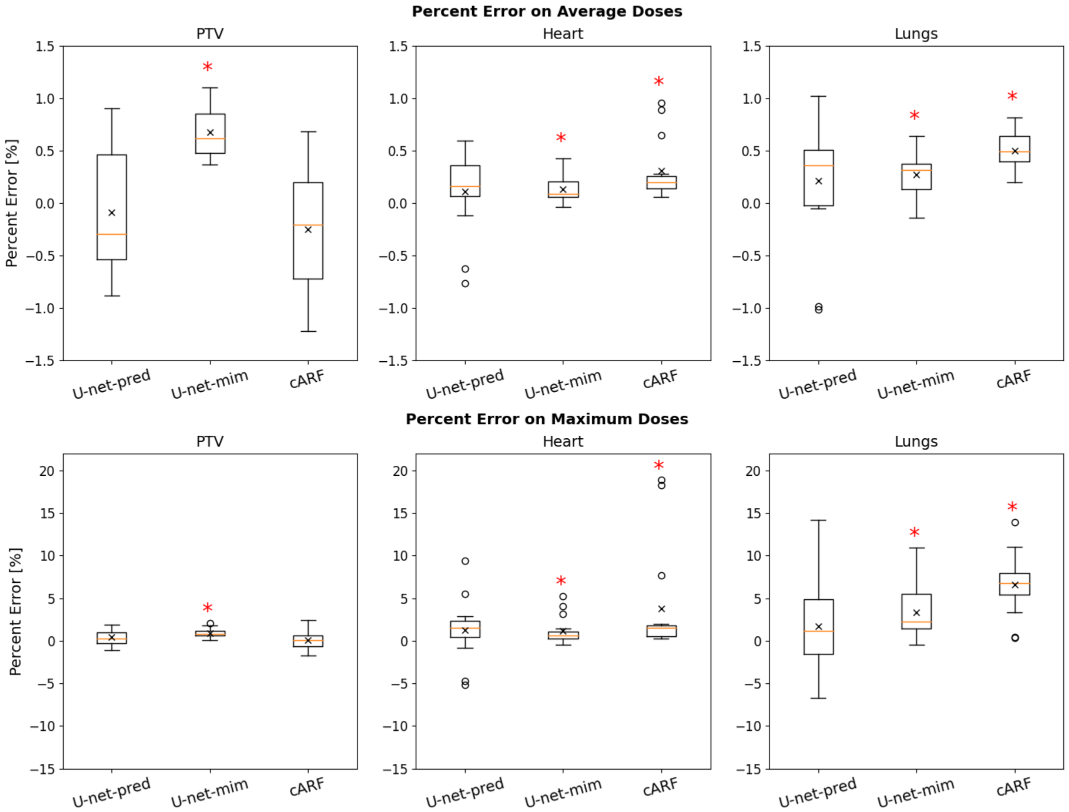


Figure A.1: *Percent error of predicted and mimicked dose distributions of the U-net model and mimicked does distribution of the cARF model, when compared to clinical plans, with respect to the prescribed doses. Average (upper row) and maximum (lower row) doses of different ROIs are evaluated. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences with respect to clinical dose are marked with a red asterisk ($p < 0.05$).*

A.4 Conclusion

In this study, a U-net model was developed and tested to predict dose distributions, as well as an already existing commercially available cARF model. When comparing the outcomes of both models with clinical plans, small differences in predicted doses to OARs were found, but were not considered as clinically relevant. Results of both model are promising for automatic plan generation.

B | Model training details

This section contains more details about model development, training and tests performed during this design project for both auto-segmentation and auto-planning.

B.1 Auto-segmentation

B.1.1 Training procedure

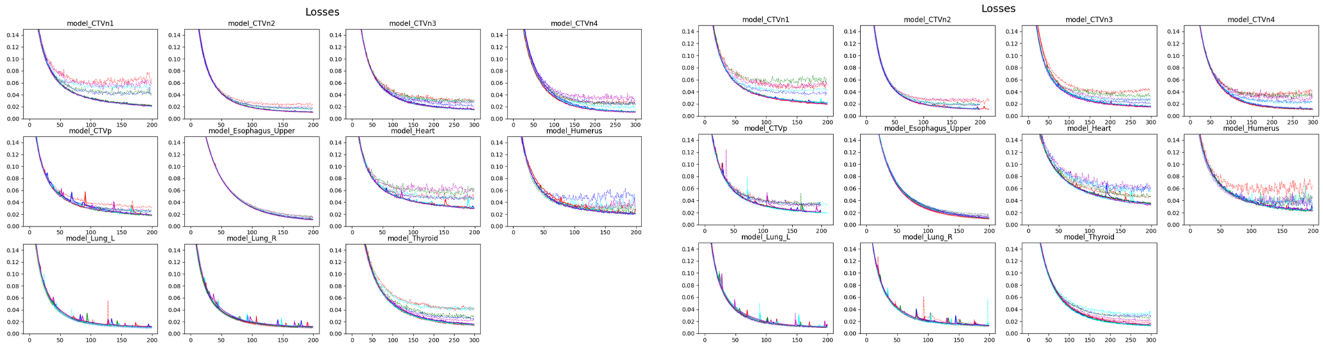
Several steps were taken to assess the model performance during training:

- Cross validation

  As mentioned in Section 4.2.3, cross validation is used to inspect the impact of the data split. This is especially useful for smaller datasets, which in this case include the node levels 3-4 and thyroid. Both models were trained on 5 folds (training/validation split 80/20) and evaluated on six independent test patients. The models were trained for 200 epochs for each ROI, except for node levels 3 and 4 and the thyroid, which were trained for 300 epochs after visual inspection of the loss. Figure B.1 shows the losses of both models for each fold. Although some variation can be observed in the validation loss, almost no variation is seen in the training loss, which indicates that the dataset is adequate and the model performance is not too dependent on the split.

- Node level combination

  After visual inspection of the delineations of the node levels, it was observed that they were not connected when each node level was trained in a separate submodel. Therefore, it was decided to test the effect of combining node levels in a submodel to the full independent test set of 15 patients for each side. To create PTV volumes of the node levels, nodes level 1 and 2 and node levels 3 and 4 are combined. That is why it was chosen to create a submodel for each of these combinations. Moreover, more data is available for node levels 1 and 2, which would go to waste when only patients would be included which contain all 4 node levels. Table B.1 shows the DSC scores for the node levels using the separate and combined models. As can be seen, combining the node levels gives a slight decrease of the DSC score for level 1, but improves the scores for all other levels. On an even more important note, two experienced ROs were asked to blindly score the sets, and both preferred the delineations resulting from the combined submodels, since



***(a)** A subfigure*

***(b)** A subfigure*

***Figure B.1:** A figure with two subfigures*

Region of Interest	Left-sided		Right-sided	
	Not combined	Combined	Not combined	Combined
CTVn1	0.81 ± 0.05	0.78 ± 0.06	0.79 ± 0.04	0.77 ± 0.06
CTVn2	0.70 ± 0.09	0.74 ± 0.07	0.70 ± 0.06	0.70 ± 0.07
CTVn3	0.73 ± 0.08	0.74 ± 0.07	0.74 ± 0.06	0.75 ± 0.10
CTVn4	0.52 ± 0.12	0.59 ± 0.12	0.56 ± 0.13	0.60 ± 0.19

Table B.1: *table*

the connection within the two sets of node levels were restored. However, they also pointed out the importance of the connection between all node levels. Therefore, for the final model it was decided to include all node levels within one model, even though this decreases the datasize for node level 1 and 2.

- Overfitting

  While inspecting the losses during the training phase, it was observed that node level 4 might be overfitting, which could be caused by the relatively small dataset. Therefore, two models were created, trained with respectively 50 and 300 epochs. When evaluating these models on the small testset of 6 patients, no large differences were found (DSC score (mean ± std) of 0.66 ± 0.02 vs 0.64 ± 0.03 over 5 folds), indicating no sign of overfitting.

B.1.2 Training parameters

During training, several parameters can be tuned to improve performance of the model. The parameters are related to the input data for the models, the data augmentation process and the actual training loop. Below, each of these parameters and its used (default) value are explained:

Input data

- Dimension: physical size in centimeters used for the input region of the model; by default set to the largest bounding box found in the training data for each ROI.
- Max number of voxels: the default resolution will be decreased automatically so that the number of voxels in the input does not exceed this parameter, to decrease the GPU-memory used during training and inference. The default value of $4 * 10^6$ was used.
- Resolution: the voxel size used in the input, to which the image data is resampled to during training and inference. In this project, the resolution was 0.3x0.16x0.16 cm for the lungs, and 0.3x0.12x0.12 cm for other structures.
- Padding: allows for a margin to be added so that the input-region increases, making the model robust against translational variance and larger ROIs then expected. In this project, a padding of 2 cm in all directions was used for all ROIs.
- Padding for augmentation: additional padding is needed to account for loss of image data along the edges, as a result of data augmentation. After augmentation has been applied, the center is cut out and fed to the model. The used padding ranges from 1.5 to 3.0 cm in this project.

Data augmentation

- Additive pixelwise noise: adds Gaussian noise of zero mean to the intensities in the input volume. By default, no noise is applied.
- Translation: the amount of translation in each direction is sampled from a normal distribution with zero mean, with a default value of 2 cm for each axis.
- Rotation: the amount of rotation for each axis is sampled from a normal distribution with zero mean, with a default value of 5 degrees for each axis.
- Elastic deformation: a coarse 3D grid is placed over the volume, where for each grid point a 3D vector is sampled from a normal distribution, resulting in a 3D deformation field. The amount of elastic deformation is specified by the spacing of the coarse grid (8 cm by default) and the standard deviation (0.5 cm by default).

Training loop

- Number of epochs: during one epoch, the full training set is passed one time through the model. The number of epochs should be sufficient for the model to converge, which can be monitored through visualization of the training and validation loss. The model can intermediately be saved during training after a pre-defined number of epochs, in order to compare the performance of the model at different stages of training. For this project, 750 epochs were used for the initial segmentation model, containing all ROIs. For the other submodels, 300 epochs were used for for heart and lungs, and 400 epochs for the other ROIs.
- Batch size: the number of samples used in each update of the weights of the CNN. A higher number leads to a smoother training process, but also requires more memory. As the 3D input occupies a lot of memory, a batch size of 1 is used.
- Learning rate: during training, the Adam optimizer is used [38]. The learning rate controls the gradient step size used during weight update of the network. A high learning rate can lead to a diverging instead of converging learning process, and a low learning rate leads to a slow conversion. A learning rate of 1e-4 is used in this project, which was set as default by RaySearch.
- L2 regularization weight: L2 regularization is used to prevent overfitting, by preventing the weights from growing too large. The default value of $1 * 10^{-}5$ is used in this study.

B.2 Auto-planning

B.2.1 Model architecture 3D U-net

RaySearch created a 3D U-net, based on the work of Çiçek *et al.*[1], which is an adapted version of the 2D U-net by Ronneberger *et al.*[2], and visualized in Figure B.2. During training layer normalization was used, which normalizes the distributions of intermediate layers. This technique enables smoother gradients, faster training and better generalization accuracy[3].

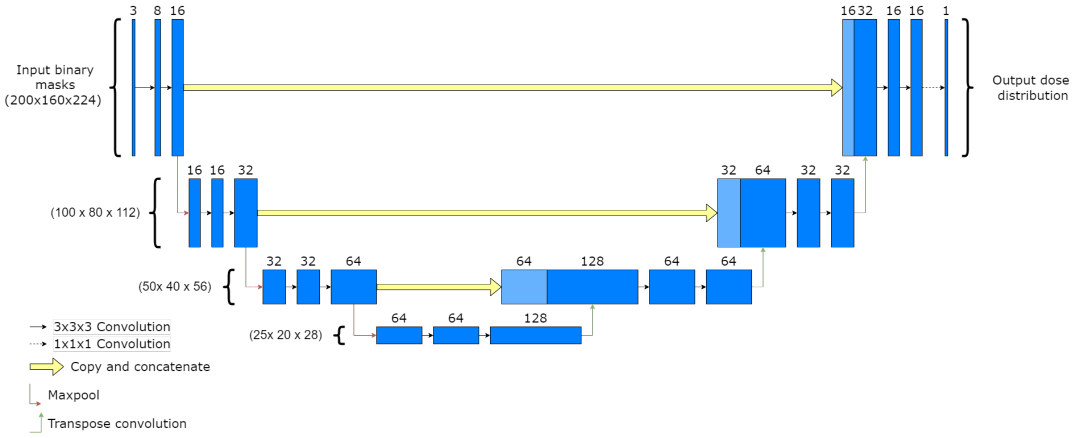


Figure B.2: *Architecture of the 3D U-net, implemented by RaySearch for auto-planning.*

[1] Çiçek, Ö., Abdulkadir, A., Lienkamp, S. S., Brox, T., & Ronneberger, O. (2016, October). 3D U-net: learning dense volumetric segmentation from sparse annotation. In International conference on medical image computing and computer-assisted intervention (pp. 424-432). Springer, Cham.

[2] Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

[3] Xu, J., Sun, X., Zhang, Z., Zhao, G., & Lin, J. (2019). Understanding and improving layer normalization. Advances in Neural Information Processing Systems, 32.

C | Prospective Risk Analysis

In CZE, the decision tree and matrix in Figure C.1 are used to decide which action should be taken for a failure.

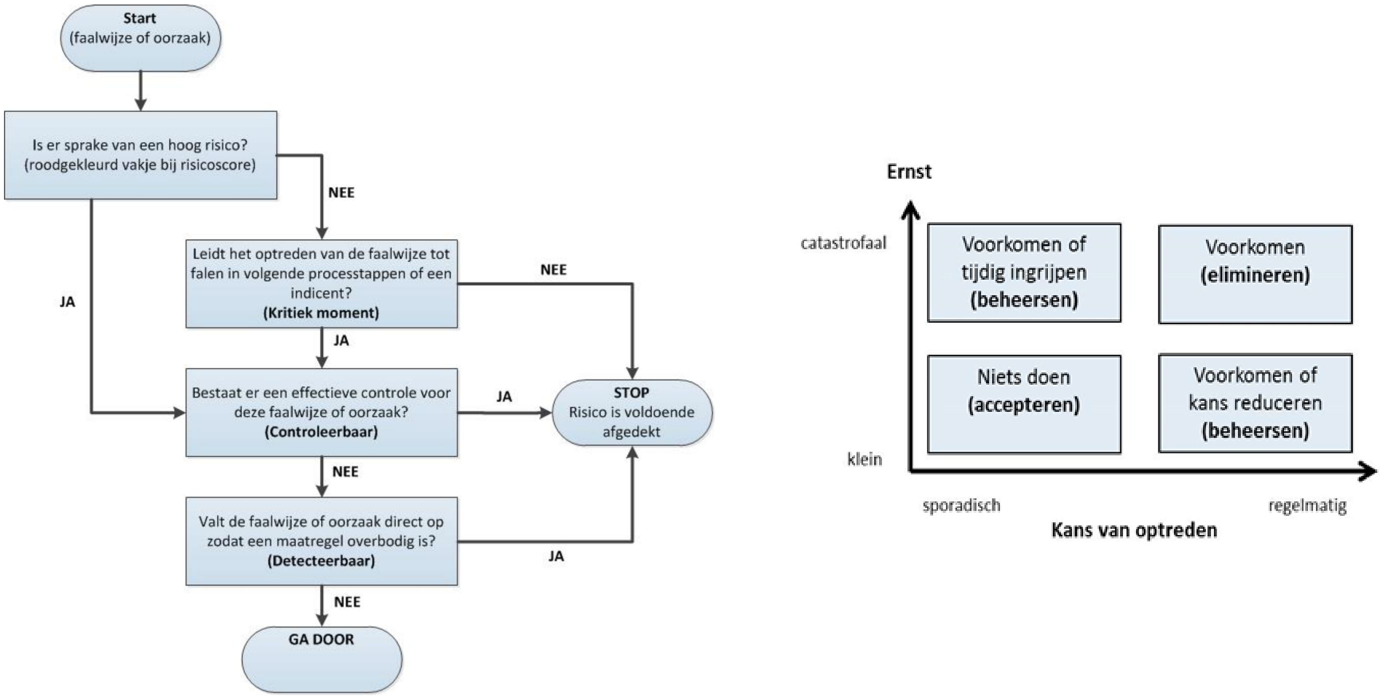

Start
(faalwijze of oorzaak)
Is er sprake van een hoog risico?
(roodgekleurd vakje bij risicoscore)
NEE
JA
Leidt het optreden van de faalwijze tot falen in volgende processtappen of een indicent?
(Kritiek moment)
NEE
JA
Bestaat er een effectieve controle voor deze faalwijze of oorzaak?
(Controleerbaar)
JA
STOP
Risico is voldoende afgedekt
NEE
Valt de faalwijze of oorzaak direct op zodat een maatregel overbodig is?
(Detecteerbaar)
JA
NEE
GA DOOR
Ernst
catastrofaal
klein
Voorkomen of tijdig ingrijpen
(beheersen)
Voorkomen
(elimineren)
Niets doen
(accepteren)
Voorkomen of kans reduceren
(beheersen)
sporadisch
regelmatig
Kans van optreden

Figure C.1: *The decision tree (left) and matrix (right) used in the Prospective Risk Analysis procedure of the CZE.*

D | Model registration sheet

Model overzicht

Vermeld MICT projectnummer, type model (planning, segmentatie, beeldverbetering, QA, ….), ontwikkelaar (in-huis of 3^e^ partij) en eindverantwoordelijke voor het model.

Projectnummer	
Type	
Algoritme	
Ontwikkelaar	
Eindverantwoordelijke	

Zelf-ontwikkeling

Alleen van toepassing wanneer het model zelf-ontwikkeld is; onder documentatie valt eventuele verslaglegging /publicatie. Scripts dienen te zijn voorzien van comments en documentatie.

Ontwikkelaar	
Documentatie	

Scripting

Voeg locatie van ontwikkelde scripts in, deze dienen te zijn voorzien van comments en documentatie. Noteer daarnaast relevante packages/libraries en bijbehorende versie (zoals keras/tensorflow etc.). Wanneer transferlearning is gebruikt, dient aangegeven te worden welk model hiervoor gebruikt is.

Scripting locatie	
Programmeertaal + versie	
Gebruikte packages/libraries	
Transfer learning model	

Model training

Noteer doelgebied en bijbehorende protocolnummer(s) welke zijn geïncludeerd in de studie. Gebruikte dataset dient geregistreerd te worden, waarvan hieronder bestand(locatie) bijgevoegd dient te worden. Limitaties van de patiëntenset, zoals bijvoorbeeld exclusie van bepaalde gevallen, dienen te worden vermeld. De oorsprong van de data heeft betrekking op zijnde enkel CZE data, externe data, of een combinatie van beide. Daarnaast dient ook software + versie waarin training is uitgevoerd vermeld te worden wanneer van toepassing.

Doelgebied	

Versie 0.1
Auteur: nbx

Protocolnummer(s)	
Patiëntregistratie	
Aantal patiënten	
Tijdsperiode	
Limitaties	
Oorsprong data	
Datum afronden	
Uitgevoerd door	
Software versie	

Model validatie

Een overzicht van de uitgevoerde testen (retrospectief/klinische validatie), bijbehorende documentatie (verslaglegging/publicatie) en gebruikte data voor het model. Bij update naar een nieuwe software versie dient het model opnieuw gevalideerd te worden, en dit dan ook geregistreerd te worden.

Soort test	Documentatie	Aantal patiënten	Datum afronden	Uitgevoerd door	Software versie
Kies een item.					
Kies een item.					
Kies een item.					

RayStation

Wanneer het model klinisch geïntroduceerd wordt in RayStation, graag een link toevoegen naar het commissioning bestand(locatie), de laatste versie waarin commissioning uitgevoerd is, en een overzicht van de dataset voor commissioning

Commissioning bestand	
Commissioning versie	
Commissioning dataset	

Versie 0.1
Auteur: nbx

E Results

This section contains additional results of the different studies for auto-segmentation and auto-planning.

E.1 Auto-segmentation

E.1.1 Retrospective study

In the retrospective study, several quantitative measures were computed for the test set. Table E.1 and Figure E.1 visualize the results.

	sDSC score (mean ± sd)	
ROI	**Left**	**Right**
CTVp	0.86 ± 0.05	0.85 ± 0.04
CTVn1	0.68 ± 0.12	0.71 ± 0.08
CTVn2	0.84 ± 0.07	0.81 ± 0.09
CTVn3	0.83 ± 0.09	0.84 ± 0.08
CTVn4	0.79 ± 0.14	0.77 ± 0.15
Esophagus	0.88 ± 0.10	0.87 ± 0.09
Heart	0.86 ± 0.08	0.84 ± 0.05
Lung Left	0.98 ± 0.02	0.97 ± 0.02
Lung Right	0.99 ± 0.01	0.97 ± 0.02
Thyroid	0.85 ± 0.12	0.76 ± 0.22
Humerus	0.82 ± 0.09	0.79 ± 0.08

Table E.1: *sDSC scores of the auto-segmentation models, found in the retrospective study.*

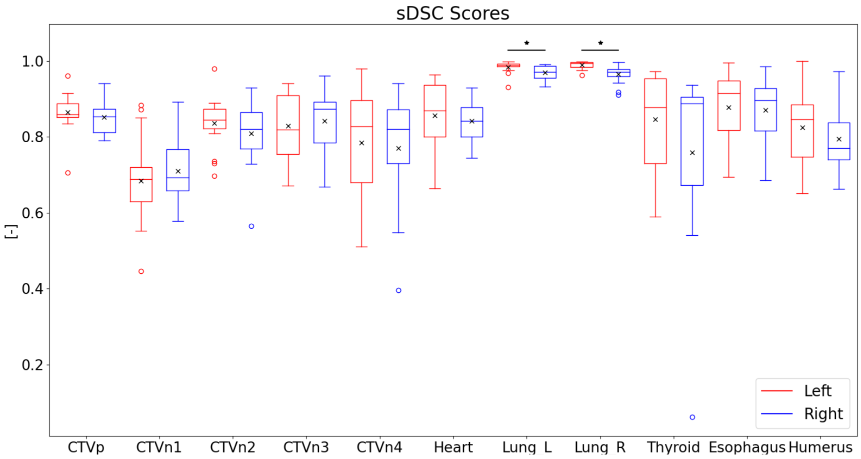


Figure E.1: *Visualization of the sDSC scores of the retrospective study for auto-segmentation. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences between the two models are indicated with with an asterisk ($p < 0.05$).*

E.1.2 Clinical pilot

The automatically generated structures are compared to the manually generated structures before correction, to assess robustness of the model. The results of the different quantitative parameters are shown in Table E.2. For most ROIs, comparable results are found as in the retrospective study, summarized in Table 6.1. Only for CTVp, worse results are found for the 95% HD, caused by some outliers, which can be observed in the boxplots in Figure E.2.

The quantitative parameters of the corrected structures, with respect to the manually delineated structures, were also calculated and can be found in Table E.3. Large improvements can be observed for the 95% HD for the CTVp when compared to the results in Table E.2. However, these differences were not found to be significant, as was stated in Section 6.1.2. Moreover, the requirements were still not (completely) fulfilled for the esophagus and thyroid, indicating that these structures are hard to delineate and prone to interobserver variability. Besides, the fact that almost no structure significantly improved, when compared to the manually delineated structure, also indicates that the AI model results in structures within the range of the interobserver variability of our clinic.

	DSC score (mean ± sd)			95% HD [mm] (mean ± sd)			sDSC score (mean ± sd)	
ROI	**Left**	**Right**	**Requirements**	**Left**	**Right**	**Requirements**	**Left**	**Right**
CTVp	0.93 ± 0.04	0.92 ± 0.06	0.85 ± 0.02	15.62 ± 20.55	13.18 ± 12.58	8.94 ± 2.86	0.83 ± 0.11	0.82 ± 0.11
CTVn1	0.77 ± 0.05	0.79 ± 0.09	0.69 ± 0.04	12.66 ± 3.65	14.43 ± 10.61	13.58 ± 3.00	0.68 ± 0.05	0.70 ± 0.15
CTVn2	0.71 ± 0.07	0.71 ± 0.06	0.47 ± 0.17	8.47 ± 3.04	12.32 ± 6.03	18.74 ± 8.15	0.84 ± 0.06	0.79 ± 0.06
CTVn3	0.71 ± 0.11	0.75 ± 0.05	0.56 ± 0.10	7.90 ± 3.99	5.66 ± 2.20	9.87 ± 3.61	0.79 ± 0.13	0.85 ± 0.09
CTVn4	0.60 ± 0.12	0.53 ± 0.13	0.45 ± 0.13	7.28 ± 2.13	7.04 ± 2.11	11.82 ± 4.88	0.77 ± 0.12	0.72 ± 0.13
Esophagus	0.66 ± 0.08	0.74 ± 0.06	0.78 ± 0.04	14.22 ± 7.57	6.63 ± 3.83	7.08 ± 3.52	0.84 ± 0.08	0.92 ± 0.06
Heart	0.95 ± 0.02	0.93 ± 0.02	0.91 ± 0.01	6.64 ± 4.33	7.48 ± 2.25	13.00 ± 5.10	0.89 ± 0.07	0.75 ± 0.11
Lung Left	0.99 ± 0.00	0.98 ± 0.01	0.99 ± 0.00	1.37 ± 0.37	3.06 ± 2.02	2.33 ± 0.95	0.99 ± 0.01	0.97 ± 0.02
Lung Right	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.00	1.24 ± 0.60	3.12 ± 1.91	2.19 ± 0.66	0.99 ± 0.01	0.96 ± 0.03
Thyroid	0.68 ± 0.08	0.58 ± 0.14	0.72 ± 0.07	7.86 ± 2.99	8.52 ± 3.96	5.37 ± 1.70	0.85 ± 0.07	0.76 ± 0.14
Humerus	0.88 ± 0.05	0.87 ± 0.05	0.70 ± 0.05	5.89 ± 3.67	9.28 ± 6.91	7.00 ± 2.00	0.88 ± 0.09	0.84 ± 0.09

Table E.2: *Quantitative results of the clinical pilot for auto-segmentation, of not-corrected contours. Fulfilled requirements are indicated by green boxes, close-to-fulfilled requirements by yellow boxes, and not-fulfilled requirements by red boxes.*

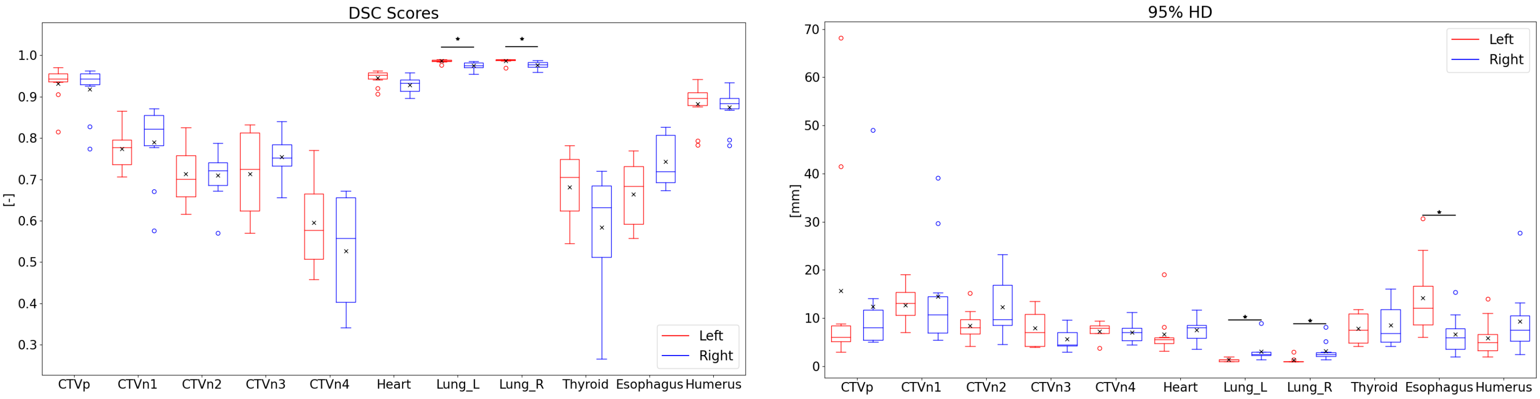


Figure E.2: *Visualization of the quantitative results of the clinical pilot for auto-segmentation. Horizontal lines in boxes are medians, crosses are means, dots are outliers. Statistically significant differences between the two models are indicated with with an asterisk ($p < 0.05$).*

	DSC score (mean ± sd)			95% HD [mm] (mean ± sd)			sDSC score (mean ± sd)	
ROI	**Left**	**Right**	**Requirements**	**Left**	**Right**	**Requirements**	**Left**	**Right**
CTVp	0.95 ± 0.01	0.94 ± 0.03	0.85 ± 0.02	7.08 ± 2.46	9.41 ± 4.74	8.94 ± 2.86	0.87 ± 0.05	0.87 ± 0.05
CTVn1	0.78 ± 0.04	0.79 ± 0.09	0.69 ± 0.04	9.49 ± 4.69	9.49 ± 4.15	13.58 ± 3.00	0.77 ± 0.05	0.76 ± 0.08
CTVn2	0.72 ± 0.05	0.73 ± 0.04	0.47 ± 0.17	8.51 ± 3.09	8.94 ± 5.17	18.74 ± 8.15	0.85 ± 0.04	0.83 ± 0.06
CTVn3	0.73 ± 0.08	0.77 ± 0.04	0.56 ± 0.10	6.40 ± 1.52	5.15 ± 0.87	9.87 ± 3.61	0.84 ± 0.09	0.88 ± 0.05
CTVn4	0.62 ± 0.16	0.61 ± 0.11	0.45 ± 0.13	5.37 ± 2.10	5.80 ± 1.92	11.82 ± 4.88	0.82 ± 0.17	0.80 ± 0.13
Esophagus	0.72 ± 0.08	0.77 ± 0.09	0.78 ± 0.04	13.87 ± 8.08	9.55 ± 7.09	7.08 ± 3.52	0.88 ± 0.07	0.91 ± 0.07
Heart	0.95 ± 0.01	0.94 ± 0.02	0.91 ± 0.01	7.40 ± 4.29	6.62 ± 1.89	13.00 ± 5.10	0.88 ± 0.05	0.81 ± 0.10
Lung Left	0.99 ± 0.00	0.98 ± 0.01	0.99 ± 0.00	1.37 ± 0.37	2.78 ± 1.29	2.33 ± 0.95	0.99 ± 0.01	0.97 ± 0.02
Lung Right	0.99 ± 0.01	0.98 ± 0.01	0.98 ± 0.00	1.24 ± 0.60	3.20 ± 1.89	2.19 ± 0.66	0.99 ± 0.01	0.96 ± 0.03
Thyroid	0.70 ± 0.08	0.65 ± 0.13	0.72 ± 0.07	7.16 ± 3.42	7.48 ± 4.09	5.37 ± 1.70	0.86 ± 0.07	0.82 ± 0.13
Humerus	0.89 ± 0.07	0.86 ± 0.07	0.70 ± 0.05	5.53 ± 6.34	5.00 ± 4.24	7.00 ± 2.00	0.90 ± 0.11	0.84 ± 0.12

Table E.3: *Quantitative results of the clinical pilot for auto-segmentation, of corrected contours. Fulfilled requirements are indicated by green boxes, close-to-fulfilled requirements by yellow boxes, and not-fulfilled requirements by red boxes.*

E.2 Auto-planning

E.2.1 Clinical pilot

During the clinical pilot, the auto-plans, generated by both the cARF and U-Net model, were compared with the manual plans by assessing the clinical goals fulfilled after mimicking and comparing the DVH parameters. Table E.4 shows the percentage of clinical goals met for the three methods, after scaling such that 98% of the volume receives 95% of the prescribed dose, which is one of the clinical goals. The average dose of the PTV is not taken into account in this evaluation, as the clinical goals changed during the project. Currently, an average dose between 99% and 101% is demanded for the PTV, which was no requirement before, resulting in higher average doses for most of the plans used for training of the models. The results reflect this as well, as can be seen in E.5, where both models result in significant higher doses to the PTV according to the Wilcoxon signed rank test. Before clinical implementation can be realized, this should be solved by tweaking the mimick settings. Furthermore, both models result in a significant higher average and maximum dose to the heart, which is also the case for the lungs for the cARF model. However, the ROs still agreed that 90% of the cARF and 95% of the U-Net models is clinically acceptable.

	Clinical goals met [%]		
Clinical goal	**Manual**	**cARF**	**U-Net**
PTV: at most 4285 cGy at 2% volume	100	90	95
Lungs: at most 600 cGy average dose	100	100	100
Lungs: at most 400 cGy average dose	100	100	100
Heart: at most 300 cGy average dose	95	95	95
Heart: at most 200 cGy average dose	95	90	90
CL Breast: at most 100 cGy average dose	100	100	100
External-PTV: at most 10 cc at 4285 cGy	95	95	95

Table E.4: *The percentage of clinical goals met in the clinical pilot for conventional breast irradiation for the test set, containing 20 patients.*

		PTV		**Heart**	**Lungs**
		Dose [cGy]	**Difference wrt prescribed [%]**	**Dose [cGy]**	**Dose [cGy]**
Manual	**Average**	4007 ± 40	+ 0.1	117 ± 77	192 ± 62
	Maximum	4169 ± 57	+ 4.1	720 ± 872	2713 ± 628
cARF	**Average**	*4037 ± 35*	+ 0.8	*123 ± 76*	*198 ± 64*
	Maximum	*4201 ± 56*	+ 4.9	*853 ± 969*	2767 ± 616
U-Net	**Average**	*4053 ± 36*	+ 1.2	*124 ± 87*	196 ± 64
	Maximum	*4202 ± 56*	+ 4.9	798 ± 905	2744 ± 637

Table E.5: *Average and maximum doses in cGy to ROIs for the clinical plans, predicted and mimicked plans of the U-net and cARF model for conventional breast irradiation (mean ± standard deviation). For PTV, the difference between mean average and maximum dose with respect to the prescribed dose (4005 cGy) is shown. Doses differing significantly from clinical doses are printed in italic.*

E.2.2 Clinical implementation

During commissioning, 10 test patients of the clinical pilot are used to compare the result of auto-planning with manual plans. The only difference is the auto-plans are not scaled, to reflect clinical reality. The manual plans are scaled, as the un-scaled data was not available. Table E.6 shows the percentage of patients in which the clinical goals are met, whereas Table E.7 shows the DVH parameters. The Wilcoxon signed-rank test was performed to assess significant differences between both methods, which was only found for both average and maximum dose to PTV ($p < 0.05$). For the average dose, the p-value was very close to 0.05, questioning the significance. In addition, although the PTV dose is slightly higher than the manual dose, the clinical goal on the maximum average dose only failed for one patient. After manual adjustments to this auto-plan, this could easily be improved.

	Clinical goals met [%]	
Clinical goal	**Manual**	**U-Net**
PTV: at least 3965 cGy average dose	100	100
PTV: at most 4045 cGy average dose	100	90
PTV: at most 4285 cGy at 2% volume	100	100
Lungs: at most 600 cGy average dose	100	100
Lungs: at most 400 cGy average dose	100	100
Heart: at most 300 cGy average dose	90	90
Heart: at most 200 cGy average dose	90	80
CL Breast: at most 100 cGy average dose	100	100
External-PTV: at most 10 cc at 4285 cGy	100	100

Table E.6: *The percentage of clinical goals met during commissioning for clinical implementation of conventional breast irradiation, containing 10 patients.*

		PTV		**Heart**	**Lungs**
		Dose [cGy]	**Difference wrt prescribed [%]**	**Dose [cGy]**	**Dose [cGy]**
Manual	**Average**	4007 ± 39	+ 0.1	150 ± 98	219 ± 76
	Maximum	4164 ± 51	+ 4.1	1123 ± 1102	2948 ± 689
U-Net	**Average**	*4038 ± 90*	+ 0.8	163 ± 123	220 ± 71
	Maximum	*4216 ± 39*	+ 5.3	1208 ± 1173	2999 ± 747

Table E.7: *Average and maximum doses in cGy to ROIs for the clinical and automatically generated plans, during commissioning for clinical implementation of conventional breast irradiation (mean $/pm$ standard deviation). For PTV, the difference between mean average and maximum dose with respect to the prescribed dose (4005 cGy) is shown. Doses differing significantly from clinical doses are printed in italic.*

F Advisory report for AI implementation

F.1 Inleiding

Binnen het Catharina Ziekenhuis Eindhoven (CZE) is een werkgroep opgezet welke een visie ontwikkelt omtrent Artificial Intelligence (AI). Op dit moment is deze werkgroep aan het verkennen welke mogelijkheden er zijn binnen het CZE. Ook op de afdeling radiotherapie vindt in sterk toenemende mate onderzoek plaats naar het gebruik van AI om processen te automatiseren. Dit betreft o.a. auto-planning, auto-segmentatie en het genereren van synthetische CTs vanuit cone-beam CTs. Bij auto-planning wordt er een dosisverdeling voorspeld van de af te stralen dosis op het doelgebied en omliggend weefsel. Deze dosisverdeling kan vervolgens gebruikt worden om tot een definitief bestralingsplan te komen. Auto-segmentatie omvat het automatisch intekenen van contouren, zowel doelgebied als omliggende organen, welke gebruikt worden bij het creëren van de bestralingsplannen. De synthetische CTs worden gegenereerd om zo beelden van voldoende kwaliteit te hebben tijdens de behandeling, vergeleken met laag-kwaliteit cone-beam CTs. De daadwerkelijke invoering van AI in de klinische praktijk brengt echter veel veranderingen met zich mee, zowel op technisch als organisatorisch vlak. In dit rapport zal de opgestelde visie op AI worden toegelicht, en gebruikt worden om de huidige situatie en eventueel benodigde veranderingen op de afdeling radiotherapie voor succesvolle implementatie te analyseren.

F.2 Toekomstvisie

F.2.1 Visie op AI in het CZE

De algemene missie van het CZE luidt 'met onze zorg elke dag de kwaliteit van leven merkbaar verbeteren', waarbij het een meer-jaren strategie heeft gedefinieerd om deze missie te vervullen. Omdat AI in de toekomst sterk zal bijdragen aan het nemen van betere beslissingen, en zo kan bijdragen aan het merkbaar verbeteren van patiënt-waarde, zal AI een belangrijk onderdeel van de strategie moeten uitmaken. AI is sterk in ontwikkeling, en dus zal het CZE moeten afstemmen met zowel regionale als landelijke initiatieven om de AI-visie naar te kunnen realiseren. Bij het definiëren van een AI-visie is het ook van belang om de verschillende kernkwaliteiten van het CZE in overweging te nemen:

- Technologische innovaties worden ontwikkeld en geïmplementeerd samen met partners, zoals bijvoorbeeld binnen e/MTIC verband maar ook andere (internationale) bedrijven in de med-tech industrie, wat leidt tot vroege fase van marktintroductie.
- Patiënten aantallen; het grote volume biedt mogelijkheden op het terrein van onderzoek en innovatie.
- Efficiëntie van processen; zorgprocessen worden efficiënt georganiseerd en innovaties worden snel toegepast, wat een goede basis vormt om AI op een efficiënte en verantwoorde manier verder te introduceren.

F.2.2 Afdeling radiotherapie

Als onderdeel van het Catharina Kanker Instituut maakt de afdeling radiotherapie een belangrijk deel uit van het topklinische aspect van het ziekenhuis, waar hooggespecialiseerde zorg wordt geleverd. Op de afdeling radiotherapie vindt veel onderzoek plaats, waarbij dit in toenemende mate in samenwerking is met de TU/e, aansluitend op de eerste kernkwaliteit. Daarnaast zijn er ook verschillende samenwerkingen met andere partners, zoals de leveranciers van radiotherapie medische technologie RaySearch en Elekta.

F.3 Juridisch kader

F.3.1 Visie op AI in het CZE

Verschillende nieuwe vraagstukken omtrent privacy en wet- en regelgeving komen aan bod bij de invoer van AI, welke aandacht nodig hebben:

- AVG: aspecten als grondslag, doelbinding en dataminimalisatie, rechten van betrokkenen en plichten voor verantwoordelijken voor verwerking moeten in acht worden genomen. Er ontstaan vraagstukken zoals;
  - Toestemming van patiënten voor gebruik en hergebruik van data
  - Anonimiseren en pseudonimiseren; welke techniek en uiteindelijke verantwoordelijkheid
  - Consequenties voor huidige manier van werken via Data Sharing Agreement, Data Transfer Agreements en (D)PIA's (Data Protection Impact Assessment)
- Intellectual property (IP): tot heden heeft het CZE gekozen om geen aanspraak te maken op IP van concepten die in consortia worden ontwikkeld. De IP landt bij andere partijen, die zo bereid zijn om te investeren in de onderzoeks- & ontwikkelingsprojecten met het CZE. Het voorstel vanuit de werkgroep is om deze koers ook te volgen ten aanzien van AI.
- Transparantie: op het gebied van aansprakelijkheden kunnen complexe situaties ontstaan, door de complexiteit van AI-toepassingen en het gebrek aan transparantie. Er zijn verschillende zaken van toepassing;
  - Patiënten rechten blijven onveranderd; gegevensbescherming, recht op informatie etc.
  - Transparantieplicht: patiënt moet geïnformeerd worden over onderliggende logica van het algoritmische advies/besluit.
  - Discriminatierisico: uitkomsten kunnen vooroordelen representeren wanneer de data om het algoritme te trainen niet voldoende representatief is.

Het is verstandig om een onderscheid te maken tussen implementatie in de context van onderzoek en implementatie in de routine zorgprocessen. Binnen deze twee dimensies worden andere eisen gesteld vanuit externe kaders, waardoor er aanbevolen wordt om binnen het CZE voor beide dimensies een duidelijk kader te definiëren.

F.3.2 Afdeling radiotherapie

AVG

Vanzelfsprekend wordt er in de AI-projecten veelvuldig gebruik gemaakt van patiëntdata. Aan het begin van elk project dient de PIA voor AI in wetenschappelijk onderzoek (WO) ingevuld te worden, voor beoordeling door functionaris gegevensbescherming. Data kan eenvoudig worden geanonimiseerd binnen de RaySearch software waarbij verschillende DICOM-headers worden verwijderd of vervangen. Echter wordt binnen de verschillende projecten geregistreerd welke patiënten geïncludeerd zijn in de dataset, en wordt hier een sleutel aan gekoppeld, waardoor de data pseudoniem is. Deze informatie is nodig zodat een patiënt bijvoorbeeld niet meerdere malen aan een dataset wordt toegevoegd of om later extra informatie (nieuwe CT-scan e.d.) toe te voegen aan de juiste patiënt. Daarnaast wordt met deze informatie bij de start van de studie ook gecontroleerd of een van de geïncludeerde patiënten bezwaar heeft gemaakt tegen gebruik van data t.b.v. wetenschap en scholing.

Benodigde veranderingen:

- Bij de start van elk AI-project dient in theorie altijd een PIA voor AI in WO ingevuld te worden. Echter zou het idealiter anders kunnen worden aangepakt, waarbij voor nieuwe projecten, waarin data uit een vorig project wordt gebruikt, geen nieuwe PIA wordt ingevuld. Deze doorontwikkeling is namelijk een belangrijk aspect van AI en zou niet beperkt moeten worden door dit papierwerk, daar het gebruik van de data eerder al is goedgekeurd met dezelfde intenties. Daarnaast zal de data ook bewaard moeten worden voor periodieke evaluatie, waarvoor ook niet telkens een PIA ingevuld zou moeten worden.

- Wanneer data worden verstuurd naar een server, zowel intern als extern, dient deze volledig geanonimiseerd te zijn. De functionaris gegevensbescherming kan oordelen of de anonimisatie voldoet aan de eisen.
- Patiëntsleutels dienen altijd beveiligd te zijn, en alleen te benaderen door de verantwoordelijke onderzoeker. Deze sleutels dienen niet verspreid te worden naar externe partijen.

PIA

In samenwerking met het AI competence center zal er overlegd moeten worden met de jurist en functionaris gegevensbescherming over hoe om te gaan met hergebruik van data binnen AI. Hiervoor zal er binnen de afdeling een verantwoordelijke moeten worden aangewezen, welke o.a. een aanspreekpunt is vanuit het AI competence center. Deze persoon krijgt hierbij binnen de afdeling een rol die op meerdere vlakken verantwoordelijkheden omtrent AI zal dragen, welke ook verder in dit document genoemd zullen worden. Op dit moment lopen er meerdere AI projecten in verschillende gebieden (bv. auto-planning, auto-segmentatie, beeldverbetering, projecten omtrent MR Linac) maar ontbreekt er samenhang tussen deze projecten. Om uniformiteit te creëren zal de verantwoordelijke binnen alle projectgroepen mbt AI betrokken zijn (in meer of mindere mate). Door een vaste medewerker van de afdeling deze rol te geven, behoudt de afdeling de benodigde kennis en vaardigheden, in tegenstelling tot wanneer een onderzoeker/student in tijdelijk dienstverband deze verantwoordelijkheden draagt.

Anonimisatie

De functionaris gegevensbescherming is gevraagd om een oordeel over de anonimisatie van data uit RayStation. Hij heeft een geanonimiseerd DICOM-file geanalyseerd en kwam met de volgende conclusie: *"Van alle velden die direct herleidbaar zijn tot een patiënt, zijn de gegevens verwijderd dan wel blanco gemaakt. Van de indirect herleidbare velden zijn bij een aantal wel en bij een aantal geen acties voorzien. Om te bepalen of deze leiden tot één van de criteria worden deze hieronder beschreven. 1.Singling out criterium: door de aard van de gegevens, en het feit dat alle herleidbare velden gemuteerd zijn, is het herleiden van een individu niet langer mogelijk. 2.Linkability-criterium: alle indirect herleidbare velden, waaronder datum/tijd velden, zijn aangepast, waardoor niet langer de mogelijkheid bestaat om de patiënt in andere systemen (zoals het RIS) te achterhalen. 3.Inference-criterium: uit de aard van de gegevens zijn geen kenmerkende karakteristieken van een patiënt te herleiden. De conclusie van bovenstaande is, dat deze dataset voldoende geanonimiseerd tegen huidige en toekomstig te verwachten mogelijkheden om anonimisering teniet te doen."*

Daarnaast is het van belang dat de onderzoeker verantwoord omgaat met geanonimiseerde data en patiënt sleutels. Natuurlijk wordt er bij aanvang van een dienstverband op gewezen dat de onderzoeker verplicht is tot absolute geheimhouding ten aanzien van alle informatie, en moet verslaglegging altijd goedgekeurd worden door de instelling. Ten slotte is er een werkinstructie opgesteld, waarin o.a. wordt toegelicht hoe anonimisatie wordt uitgevoerd en hoe en waar patiëntenregistratie plaats vindt.

Intellectual property

Bij de verschillende projecten zijn er overeenkomsten met externe partners (RaySearch/Elekta), waarbij zij investeren doormiddel van funding en licenties en ligt de IP bij de externe partij, wat voldoet aan het advies. Transparantie De AI-modellen die op korte termijn klinisch geïmplementeerd worden (auto-planning en auto-segmentatie), leiden tot uitkomsten die altijd nog gecontroleerd worden door de artsen. Zij kunnen ze goedkeuren of eventueel aanpassingen doorvoeren om te zorgen dat de uitkomst klinisch acceptabel is. De AI-modellen dienen dus enkel als ondersteuning en patiënten zullen hierover niet ingelicht worden. Om discriminatierisico te voorkomen is het belangrijk dat de dataset voldoende variatie aan patiënten bevat en het model zorgvuldig getest wordt.
Benodigde veranderingen:

- Centrale registratie van gebruikte dataset per model: hoeveel patiënten, welke patiëntengroep, inhoud (CT/contouren/planningen etc.).
- Centrale registratie van uitgevoerde tests per model: welke patiënten zijn gebruikt, welke criteria zijn gebruikt.

Concrete vervolgstappen:

Registratie data:

Binnen de AI projecten mbt mamma patiënten, is in het verleden al een patiëntenregistratie document opgesteld, waarbij voor ieder medisch protocol per geïncludeerde patiënt de volgende zaken worden geregistreerd:
- Pseudo ID
- Data verantwoordelijke (aanmaker van data, eventueel extra bewerkt tov klinische data)
- Project waarbij gebruikt + fase in project (training/validatie/test/pilot)
- Inhoud patiënt (contouren/klinisch plan/. . .)
- Opslaglocatie (klinische database (wanneer niet bewerkt) of research database)

Een soortgelijk document kan op de langere termijn ook voor andere doelgebieden worden opgesteld, zodat er een centraal en uniforme manier van registratie is.

Registratie testen:

De uitgevoerde testen zullen, samen met andere model informatie, geregistreerd gaan worden in een model sheet. Hiervoor is, naar aanleiding van deze scriptie, een eerste versie opgesteld in overleg met de medical engineer, klinisch fysici en andere onderzoekers. De verantwoordelijke voor AI, zal deze documenten controleren op volledigheid en uniformiteit.

MDR

Een juridisch kader dat niet in de visie opgenomen is maar wel degelijk van belang is, is de MDR. Software, waaronder dus ook AI modellen, vallen namelijk onder de MDR wanneer het een medisch hulpmiddel is. Er zijn 3 varianten met bijbehorende eisen te onderscheiden:

(A) Extern ontwikkeld, in-huis gebruikt: software moet over CE-markering beschikken, afgegeven onder MDD of MDR. Wanneer de software anders dan beoogd gebruik wordt of een extra module wordt toegevoegd, moet eerst gekeken worden of er geen commerciële software verkrijgbaar is die voldoet aan de door de gebruik opgestelde eisen, of dat de fabrikant zelf deze aanpassingen wil doorvoeren. Wanneer dit niet het geval is, mogen de veranderingen in-huis ontwikkeld worden, en valt de software onder de variant B.

(B) In-huis ontwikkeld, in-huis gebruikt: bij zelf-ontwikkeling moet voldaan worden aan verschillende eisen (MDR-Artikel 5.5). Een aantal belangrijke punten hieruit;

  (a) De software wordt niet overgedragen aan een andere rechtspersoon
  (b) De software wordt ontwikkeld in een passend kwaliteitsmanagementsysteem
  (c) De software moet voldoen aan de algemene veiligheids- en prestatie-eisen
  (d) De software moet worden geëvalueerd en vereiste corrigerende acties moeten ondernomen worden
  (e) De software wordt niet op industriële schaal vervaardigd
  (f) De software moet beschikken over documentatie die voldoet aan de daarvoor opgestelde eisen in de MDR.

(C) In-huis ontwikkeld, extern gebruikt: ziekenhuis-breed is afgesproken dat het ziekenhuis geen fabrikant is/wordt van software, daarmee is deze variant niet van toepassing.

F.4 AI Platformen

F.4.1 Visie op AI in het CZE

Op dit moment zijn er een aantal platformen relevant voor het CZE:

1. Health Intelligence Platform Santeon (HIPS): gezamenlijk dataplatform van de 7 Santeon ziekenhuizen, waar informatie uitgewisseld en vergeleken wordt, om zo te onderzoeken, verbeteren en innoveren. Op dit moment niet geschikt voor AI initiatieven.

2. e/MTIC Health Data Platform (HDP): aangesloten organisaties leveren hun data aan op een datalaag. In het portaal, geleverd door softwarepakket AnDREa (RadboudMC) kunnen onderzoekers hun AI-omgeving bouwen en deze trainen op de data uit de data-laag (pay-per-use model).

3. AI platform CZE (Ludwig): in eerste instantie bedoeld voor interne onderzoeksprojecten met eigen data, mogelijkheden om op te schalen naar organisatie overstijgend onderzoek. Bestaat uit een dataplatform en portaal met een gestandaardiseerde AI-omgeving (MLOPS).

F.4.2 Afdeling radiotherapie

Binnen de AI-onderzoeken zijn er op dit moment verschillende mogelijkheden waarvan gebruik wordt gemaakt om AI modellen te trainen:

- Binnen RaySearch software: voor de auto-segmentatie modellen biedt RaySearch een trainingswijze die enkel binnen RayStation uitgevoerd kan worden. Hiervoor dienen verschillende scripts gerund te worden voor dataextractie, pre-processing, training en post-processing. Voor training van deze modellen is een GPU nodig, wat betekent dat deze voor bepaalde tijd niet beschikbaar is voor andere werkzaamheden. Daarnaast duurt de training meerdere dagen, waardoor er meerdere dagen een RayStation licentie wordt gebruikt en de desktop niet afgesloten mag worden.
  - Een Research desktop is ingericht waarbinnen RayStation te gebruiken is, en welke niet dagelijks afgesloten wordt, zoals bij reguliere desktops, zodat de training kan plaatsvinden
  - Bij grootschalige of meerdere projecten dient gedocumenteerd te worden wie en wanneer er gebruik gemaakt wordt van deze desktop
- Externe server: via samenwerking met de TU/e wordt er bij enkele projecten gebruik gemaakt van rekenclusters op de TU/e. Echter is dit ongewenst, daar er data het CZE moet verlaten. In de toekomst zal er dan ook geen gebruik van gemaakt moeten worden, maar van het CZE platform Ludwig.
- Ludwig: momenteel wordt een van de radiotherapie AI projecten gebruikt als use-case op het platform van het CZE. Hierop worden veel mogelijkheden geboden, zoals de mogelijkheid om externe installaties uit te voeren (zoals pakketten geleverd door RaySearch). Dit platform zal bij grootschalige uitrol dan ook geschikt zijn om te gebruiken voor model training.
- Lokaal: voor enkele projecten is er binnen de afdeling hardware aangeschaft om lokaal modellen te trainen

Benodigde veranderingen:
Geen gebruik meer maken van externe servers.
Concrete vervolgstappen:
Idealiter zou je alle AI modellen trainen binnen het eigen ziekenhuis. Dit kan met behulp van een special-pc met extra rekencapaciteit, maar zal grotendeels op het platform Ludwig moeten. Echter is de capaciteit van Ludwig niet onbeperkt en is deze ook nog in pilot fase. Het is op dit moment nog onbekend hoe de beschikbaarheid van Ludwig uiteindelijk verdeelt gaat worden over de verschillende projecten in het ziekenhuis. Het is in ieder geval van belang dat de AI verantwoordelijke het aanspreekpunt wordt bij het AI competence center, zodat deze aanvragen kan doen over eventuele benodigde server ruimte. Op de korte termijn zal er daarom nog gebruik gemaakt worden van externe servers. Zoals eerder vermeld, wordt er uitsluitend gebruik gemaakt van gepseudonimiseerde

data. Wanneer deze data naar een externe server wordt verstuurd, zullen de pseudo-sleutels niet meegestuurd worden, wat de data anoniem maakt voor de externe server. Door deze werkwijze is het op korte termijn werkbaar, totdat de capaciteit op eigen servers toereikend zal zijn.

F.5 AI Capabilities

F.5.1 Visie op AI in het CZE

Er wordt voorgesteld om de 'AI capabilities', oftewel kennis vanuit eigen onderzoeksprojecten en bij samenwerkingspartijen, langs 4 assen te ontwikkelen:

1. AI-data engineering (data kwaliteit, data beschikbaarheid)
2. AI-infrastructuur (rekenkracht, opslag, delen, security)
3. AI-analyse (modellen bouwen)
4. AI-applicatie/toepassing

Doormiddel van een pragmatische en projectgebasseerde aanpak, met een aantal use-cases, wordt er gewerkt aan ontwikkeling en opschaling van deze capabilities.

F.5.2 Afdeling radiotherapie

1. AI-data engineering: retrospectieve data wordt opgehaald uit het archief JiveX, en kan vanuit daar ingeladen worden in RayStation. Vervolgens kan deze data geanonimiseerd worden door een anonieme back-up te maken, en deze vervolgens weer in te laden in RayStation. Hierdoor worden voor veel projecten losse back-up files opgeslagen, mogelijk van dezelfde patiënten. Er wordt gewerkt aan een werkwijze waarbij back-up files die niet meer nodig zijn en tijdig worden opgeschoond, en back-up files die bewaard moeten worden (bijvoorbeeld omdat de data bewerkt is nadat het uit het archief is opgehaald) op een centrale plaats worden opgeslagen. Registratie is hier wederom belangrijk. Voor enkele projecten rondom auto-planning wordt een ander data format gebruikt (pickle files). Deze kunnen op elk moment opnieuw gegenereerd worden vanuit RayStation wanneer de juiste patiënten beschikbaar zijn, en zouden dus niet naast de back-up files bewaard moeten worden.

2. AI-infrastructuur: zoals eerder benoemd worden er verschillende manieren gebruikt om modellen te trainen. Wanneer dit binnen RayStation verloopt, is de AI-infrastructuur eenvoudig omdat de data in het systeem blijft. Bij de andere opties is het echter ingewikkelder. Zoals bij punt 1 genoemd, zullen bestanden moeten worden opgeslagen, om deze naar de betreffende servers te sturen. Registratie van opslag en het tijdig opschonen is hier wederom van belang. Op dit moment kan de data uit RayStation worden opgeslagen als pickle-files doormiddel van scripting binnen RayStation. In de toekomst is de wens om een directe dataontsluiting te realiseren vanuit RayStation naar het CZE platform Ludwig. De mogelijkheden hiervan zullen moeten worden onderzocht, daar het om een externe partij gaat.

3. AI-analyse: bij een aantal projecten wordt het bouwen van modellen uitgevoerd door studenten vanuit TU/e. Zij kunnen opgebouwde kennis in de praktijk toepassen. Bij een aantal andere projecten worden modellen aangeleverd door de leverancier (RaySearch/Elekta) en kan hieraan in meer of mindere mate aanpassingen worden gedaan om modellen te trainen. Het is niet van belang dat alle betrokkenen de exacte werking van het model weten, maar enkele basiskennis is wel van belang (wat is de input en output data, hoe wordt het model geëvalueerd e.d.). Daarnaast is het belangrijk om bij zelf-ontwikkeling goed te documenteren hoe het model is getraind (architectuur/parameters e.d.), en om voor geleverde modellen deze documentatie op te vragen.

4. AI-applicatie/toepassing: binnen de afdeling groeit de kennis en kunde omtrent het gebruik van AI. Regelmatig worden er bij bijeenkomsten van de fysici (research club /refereer-lunches) presentaties gegeven over de lopende AI projecten en andere relevante onderzoeken uit het werkveld. Ook binnen het gehele werkveld krijgt AI steeds meer aandacht en worden er verschillende cursussen en lezingen aangeboden, en ook in vooropleidingen krijgt het een rol. In mindere mate is dit het geval bij afdelings-brede overleggen (even bijpraten/scholingsdagen). Het is belangrijk om een aantal laboranten en radiotherapeuten actief te betrekken tijdens de ontwikkeling, om tijdig de klinische toepasbaarheid te toetsen en via hun de rest van de afdeling te enthousiasmeren. Daarnaast zal er bij klinische introductie van een AI toepassing scholing moeten plaatsvinden en documentatie in de vorm van werkinstructies worden geschreven.

Benodigde veranderingen:
- Beleid opstellen omtrent data opslag (RayStation back-up files/pickle files)
- Onderzoek naar mogelijkheden directe data ontsluiting RayStation naar Ludwig
- Centrale documentatie modellen, zowel zelf-ontwikkeld als door leverancier
- Scholing van de afdeling met bijbehorende documentatie (werkinstructies e.d.)

Concrete vervolgstappen:

Data opslag

Het probleem van losse back-up files en pickle files dient opgelost te worden. Ten eerste zijn pickle files overbodig om te bewaren, daar deze op elk moment gereproduceerd kunnen worden vanuit RayStation. De patiënt informatie die hiervoor nodig is, komt uit het archief, de database, of opgeslagen back-up files (RSBAK files). Er is daarom een beleid opgesteld dat deze files na afronding van een project verwijderd worden, welke via de werkinstructie voor onderzoekers wordt opgenomen in het stappenplan bij afronding. Wanneer patiënten worden geanonimiseerd uit de klinische database/archief, worden ze opgeslagen als RSBAK files. Deze kunnen vervolgens in de research omgeving worden ingelezen, waarvan uit verder wordt gewerkt voor de ontwikkeling van een AI model. De opgeslagen RSBAK files kunnen vervolgens verwijderd worden, want er zijn vervolgens 2 mogelijkheden:
1. De patiëntdata wordt niet bewerkt voor gebruik; dit betekent dat voor reproductie de originele data uit het archief kan worden gebruikt; geen extra data opslag nodig
2. De patiëntdata wordt bewerkt voor gebruik; dit betekent dat de bewerkte data opgeslagen dient te worden. Dit is mogelijk als RSBAK file, maar neemt veel schijfruimte in. Naar aanleiding hiervan is besproken met de technisch applicatiebeheerder dat ook van de research database periodiek een back-up gemaakt wordt.

Data ontsluiting

Op dit moment is directe data ontsluiting vanuit RayStation niet mogelijk, en dit lijkt ook op korte termijn niet mogelijk binnen de huidige infrastructuur. Ook directe ontsluiting van data uit het archief en omzetting in pickle-files is niet mogelijk, daar het creëren van pickle-files met behulp van RayStation scripting environments gebeurd. Deze scripting environments zijn niet te 'decompilen', wat wil zeggen dat de inhoud van deze environments niet bekend/aan te passen is. Deze verandering zal dus op korte termijn niet gerealiseerd worden en heeft daarnaast ook geen hoge prioriteit.

Registratie

De eerder genoemde registratie omtrent modellen en uitgevoerde tests is ook hier van toepassing. Bij ontwikkeling door derden dient deze sheet ook ingevuld te worden, al dan niet met behulp van de leverancier.

Scholing

Ongeveer 4 à 5 maal per jaar is er een 'even bijpraten' moment voor de gehele afdeling. Hier is onlangs een presentatie gegeven over AI, welke op een begrijpelijk niveau voor niet-technici (laboranten/artsen) de hoofdlijnen uitlegt. Deze sessie is opgenomen en kan voor eventuele bijscholing worden gebruikt. Bij de klinische live-gang van het eerste AI model voor auto-planning, zijn erg veel vragen ontstaan vanuit zowel de klinisch fysici als laboranten. Naar aanleiding hiervan heeft overleg plaatsgevonden met fysici en laboranten, waaruit de volgende punten zijn geconcludeerd:
- Invoering van AI model dient te worden gecommuniceerd met fysici en artsen in staf-overleg, en met laboranten in dagstart. Uitgebreide voorlichting is hierbij niet nodig, maar ondersteunend materiaal dient aanwezig te zijn (zoals bijvoorbeeld opgenomen scholing van even bijpraten over AI).

- Invoering van AI model voor auto-planning wordt gezien als verandering in het protocol; metingen dienen te worden uitgevoerd voor invoering én bij de eerste 20 patiënten na invoer
- Een werkinstructie dient, analoog met andere werkinstructies, te worden opgesteld en gepubliceerd op intranet. Daarnaast zal tijdens de eerste fase in de werkinstructie een link worden geplaatst naar een extra document waarin vragen/opmerkingen verzameld kunnen worden.

F.6 Financiële inbedding

F.6.1 Visie op AI in het CZE

Uitgangspunten bij toekomstige AI-toepassing in de dagelijkse zorg kunnen de principes van waarde gedreven zorg zijn; bijdragen aan verbetering van de zorg (bv. betere uitkomsten) en efficiëntie van de zorg (bv. kortere ligduur). Investeren in een AI-omgeving kan overwogen worden om een koplopersrol te vervullen en daardoor rendement te behalen in de vorm van onderscheidende kwaliteit en efficiëntie ten opzichte van andere ziekenhuizen. Op korte termijn kan een onderscheid worden gemaakt tussen investeringen die nodig zijn voor AI in een onderzoeks-context en AI binnen de dagelijkse zorg. Voor onderzoeksprojecten die extern gefinancierd worden kunnen dezelfde regels gehanteerd worden die nu ook gelden voor het gebruik van infrastructuur van het CZE (pay per use). Voor AI binnen de dagelijkse zorg kunnen businesscases beoordeeld worden op basis van kosten en opbrengsten (waarde).

F.6.2 Afdeling radiotherapie

AI projecten binnen de afdeling hebben verschillende doelen, zoals verbetering van de kwaliteit (genereren van synthetische CTs) en versnellen van het werkproces (auto-planning en auto-segmentatie). Bij de start van een onderzoeksproject is het belangrijk om de beoogde winst te definiëren en na afloop van het project de behaalde winst te kwantificeren. Bij introductie van projecten in de dagelijkse zorg, kunnen deze uitkomsten gebruikt worden om een businesscase op te stellen.

Benodigde veranderingen:

Registratie beoogde en behaalde winst

Concrete vervolgstappen:

Klinische pilots

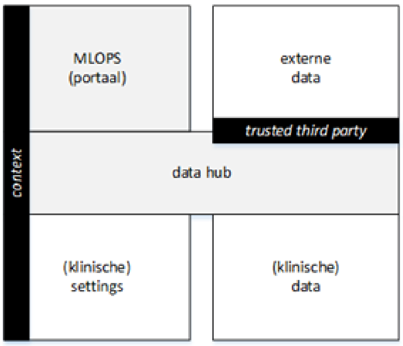
Uiteraard dient bij start van het project het doel geformuleerd te worden, welke in AI projecten vaak afname van tijd, afname van variabiliteit of verbetering van kwaliteit is. Er dient te worden opgesteld hoe deze wordt gemeten. Verschillende studies hebben aangetoond dat kwantitatieve metingen niet voldoende zijn, en daardoor dient er altijd een klinische pilot te worden uitgevoerd met kwalitatieve metingen. Dit houdt in dat ervaringen van laboranten en artsen worden gescoord met betrekking tot klinische acceptatie en bruikbaarheid in de workflow. Bij registratie van het model dient eveneens de informatie omtrent deze uitgevoerde pilot te worden geregistreerd. Ten slotte dient na invoering van een AI model monitoring plaats te vinden van de uitkomsten van de eerste 20 patiënten. Onder deze monitoring vallen o.a. metingen, evaluatie van handmatige aanpassingen na automatische generatie van plannen/segmentaties, ervaringen van laboranten/artsen.

F.7 AI infrastructuur

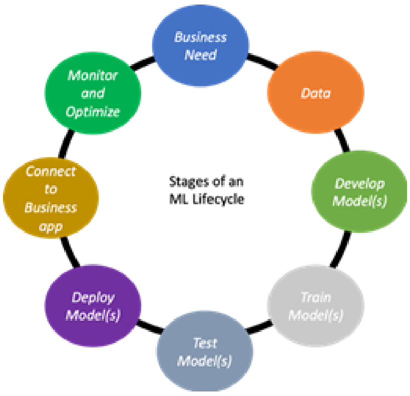
F.7.1 Visie op AI in het CZE

De bovengenoemde AI capabilities zijn verbonden via een AI infrastructuur, weergeven in Figuur F.1a, bestaande uit de volgende elementen:

- MLOps: standaardisatie en stroomlijning van het ontwikkeling van AI-modellen, weergeven in Figuur F.1b, maakt het mogelijk om het CZE iteratief te laten groeien in AI-volwassenheid. Het model en de bijbehorende data moeten blijven overeenkomen met de verwachtingen van de oorspronkelijke doelstelling. Met de best-practices uit MLOps kunnen AI-teams versiebeheer



(a) *Beoogde AI infrastructuur CZE*



(b) *MLOps Cyclus*

uitvoeren, begrijpen of opnieuw getrainde modellen beter zijn dan vorige versies en overzicht houden van de status van elk model door standaardisatie van het ontwikkelproces.

- Data hub: gestreefd wordt om binnen de data hub zoveel mogelijk te standaardiseren op basis van (internationale) standaarden om aan te kunnen sluiten op andere dataplatformen en concepten. Gezien de grote hoeveelheden aan diverse patiëntgevens is het borgen van privacy van patiënten (trusted third party) een belangrijke randvoorwaarde die geleverd wordt vanuit de data hub.
- (Klinische) data: naast de alfanumerieke gegevens uit het EPD worden ook in toenemende mate (ruwe) complexere datasets vereist. Bij aanschaf van nieuwe modaliteiten en (klinische) systemen zal de koppelbaarheid van een AI-data pipeline in de toekomst een vereiste moeten zijn, zodat er geen tijd verloren gaat aan het telkens opnieuw handmatig verzamelen van data.
- Externe data: data acquisitie uit consumer en homemonitoring devices; kunnen ook opgenomen worden in de datahub en betrokken worden in het klinische proces.
- (Klinische) settings: gebruikers dienen over kennis en vaardigheden te beschikken om op een verantwoorde manier AI-toepassingen te gebruiken; men dient na implementatie continue op zowel meerwaarde als (potentiële) risico's te monitoren.

F.7.2 Afdeling radiotherapie

- MLOps: op dit moment worden AI projecten opgezet vanuit onderzoeksdoeleinden, maar ook vanuit een klinische vraagstelling (business need). Echter kan tijdens het uitvoeren van het onderzoek nog beter de verbinding met klinische praktijk gezocht worden, om zo probleem en oplossing te blijven monitoren en valideren. Hiervoor is het belangrijk dat er een multidisciplinair team wordt samengesteld (onderzoeker, fysicus, radiotherapeut, laborant). Daarnaast dient versiebeheer bijgehouden te worden, door betere registratie zoals eerder al gesuggereerd.
- Data hub en klinische data: zoals eerder genoemd wordt benodigde data momenteel lokaal opgeslagen en vervolgens geüpload naar de server, maar is het de wens om dit in de toekomst direct vanuit de software RayStation naar een datahub te kunnen uploaden.
- Klinische settings: zoals eerder genoemd is scholing van gebruikers (radiotherapeuten en laboranten) van belang bij klinische introductie van een AI model. Daarnaast zal monitoring moeten plaatsvinden, om verschillende evaluaties uit te voeren. Denk hierbij aan vergelijking met retrospectieve handmatig gegenereerde data.

Benodigde veranderingen:
- Multidisciplinaire teams bij AI projecten
- Monitoring na klinische introductie AI model

Concrete vervolgstappen:

Multidisciplinaire teams

Binnen elke tumorgroep zijn er verschillende fysici, artsen en laboranten welke gespecialiseerd zijn, welke een werkgroep vormen. Binnen deze werkgroepen dienen lopende AI projecten bekend te zijn,

en bij overleg van deze werkgroepen dient de onderzoeker een update te geven. Daarnaast kunnen laboranten en artsen uiteraard ook bij andere project meetings worden uitgenodigd, maar in praktijk blijkt dit lastiger te plannen dan de werkgroep overleggen. *Monitoring*
Zoals eerder genoemd, dienen de eerste 20 patiënten waarbij een nieuwe AI model is gebruikt te worden gemonitord. Naast de uitkomst (plan/segmentatie etc.), dient zeker ook de ervaring van de gebruikers gemonitord te worden.

F.8 Samenvatting

Binnen het CZE is er door de werkgroep een visie op AI opgesteld, waarin op verschillende vlakken een aantal lijnen zijn geformuleerd om het gebruik van AI binnen het CZE te professionaliseren. Ook op de afdeling radiotherapie vinden verschillende onderzoeken naar het gebruik van AI plaats. In dit adviesrapport is de huidige standv van zaken geanalyseerd. Het beschrijft tevens eventueel benodigde veranderingen op de afdeling radiotherapie aan de hand van de CZE visie, wat heeft geleid tot een aantal concrete vervolgstappen:
Verantwoordelijkheid AI
Binnen de afdeling zal een verantwoordelijke AI moeten worden aangesteld, welke het aanspreekpunt over AI wordt, zowel vanuit projecten binnen de afdeling als naar buiten toe.
Juridisch vlak
Op juridisch vlak is afstemming nodig binnen het hele ziekenhuis over het (her)gebruik van data met betrekking tot de AVG en opstellen van een PIA, wat door de verantwoordelijke AI zal worden opgepakt, in samenwerking met het AI competence center. De anonimisatie van data binnen de afdeling is gecontroleerd door de functionaris gegevensbescherming, waaruit blijkt dat de huidige anonimisatie voldoende is. Ten slotte is de werkwijze omtrent data, anonimisatie en versleuteling voor onderzoekers vastgelegd in een werkinstructie. Om transparantie te garanderen is een registratie document opgesteld waarin verschillende informatie over AI modellen, zoals training- en testprocedure, kan worden opgenomen.
AI platform
Met betrekking tot het trainen van AI modellen is binnen het CZE het Ludwig platform gelanceerd, waar idealiter ook door de afdeling radiotherapie gebruik van gemaakt gaat worden. Wanneer de capaciteit niet toereikend is, kan er ook gebruik gemaakt worden van training servers op de TU/e, maar enkel wanneer de data volledig geanonimiseerd is.
AI capabilities
AI capabilities dienen vanuit 4 assen te worden ontwikkeld. Allereerst AI-data engineering, welke betrekking heeft op data kwaliteit en beschikbaarheid. Binnen de afdeling is nu een beleid opgesteld met betrekking tot data opslag en data back-up binnen een research database, welke is opgenomen in de werkinstructie voor onderzoekers. AI-infrastructuur heeft betrekking op de rekenkracht, zoals het Ludwig platform, maar ook data ontsluiting. Binnen de afdeling is geconcoludeerd dat op dit moment directe dataontsluiting naar het platform niet mogelijk is, daar er veel gewerkt wordt met derde partijen. Vervolgens is er AI-analyse, over het ontwikkelen van AI modellen, welke geregistreerd dient te worden in het opgestelde registratie document. Ten slotte is er AI-applicatie, over de uiteindelijke implementatie in de klinische workflow. Hiervoor zijn afspraken gemaakt met fysici, laboranten en artsen (dus de eind-gebruikers) over het opstellen van werkinstructies en communicatie binnen werkgroepen.
Financiële inbedding
Om de waarde van een AI project te definiëren, is het van belang om de beoogde winst en uiteindelijk behaalde dienst te meten en af te stemmen. Zo dienen er goede studies uitgevoerd te worden, en ook na klinische implementatie is het van belang om de resultaten te monitoren.
AI infrastructuur
Ten slotte wordt in het CZE advies een AI infrastructuur voorgesteld, bestaande uit een portaal, externe en/of klinische data, een datahub en de klinische setting. In toevoeging tot bovengenoemde zaken die betrekking hebben op het portaal, data en de datahub, is het binnen de afdeling radiotherapie van belang om de klinische setting te betrekken bij AI projecten door middel van multidisciplinaire teams, met fysici, artsen, laboranten en onderzoekers.