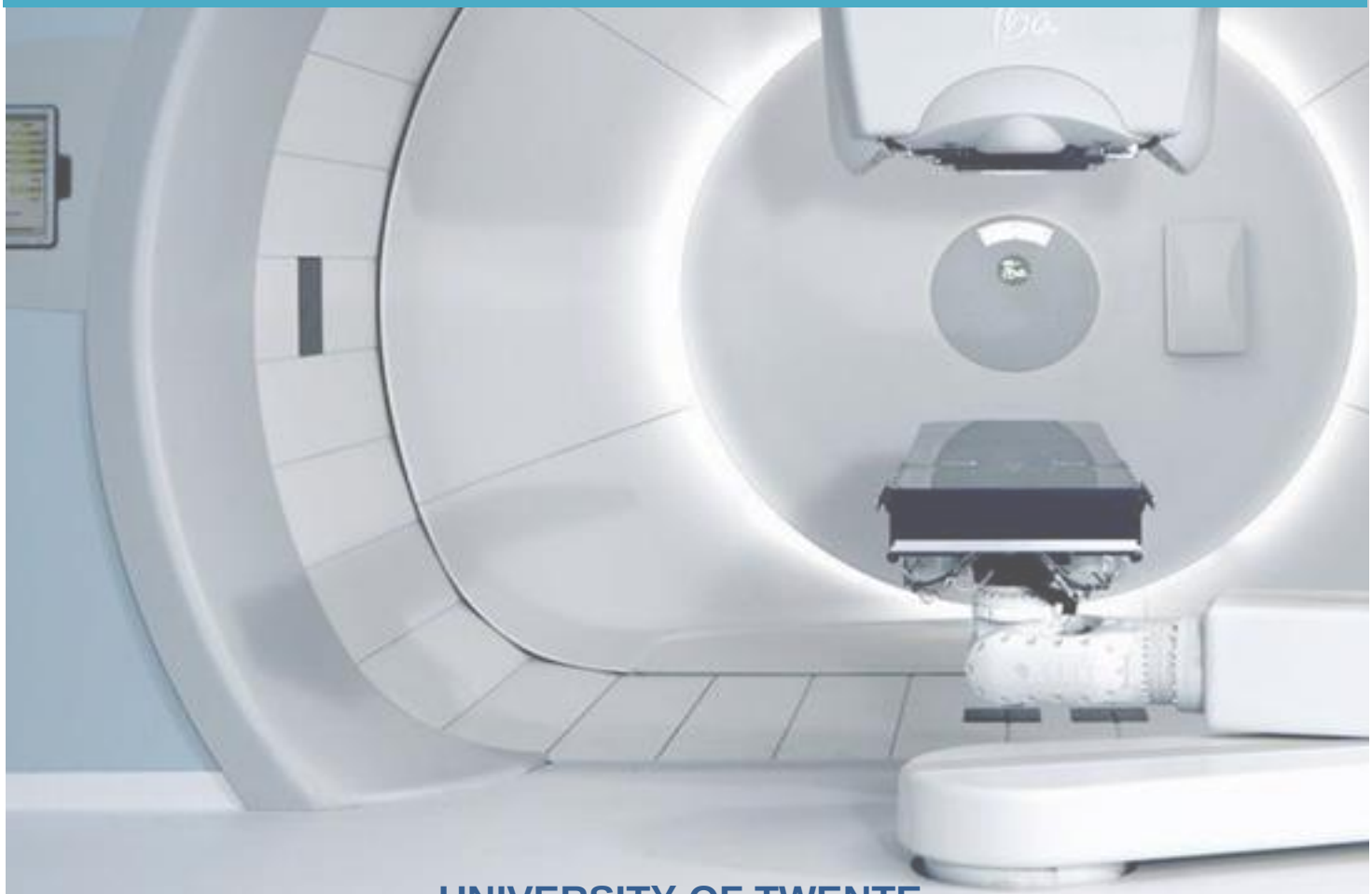
Automated robust planning for IMPT in oropharyngeal cancer patients using machine learning



UNIVERSITY OF TWENTE.

March 10, 2020

M. Huiskes

Automated robust planning for IMPT in oropharyngeal cancer patients using machine learning

Master Thesis

M. Huiskes

10-03-2020

University of Twente

Faculty of Science and Technology

Master Technical Medicine, Medical Imaging and Interventions

Graduation committee

Chairman:	Prof. dr. ir. C.H. Slump (UT)
Medical supervisor:	Dr. R.J.H.M. Steenbakkers (UMCG)
Technical supervisor:	Prof. dr. ir. C.H. Slump (UT)
Process supervisor:	Drs. P.A. van Katwijk (UT)
Extra member:	Dr. ir. R.G.J. Kierkels (UMCG)
Extra member:	Dr. ir. E.W. Korevaar (UMCG)
Extern member:	M.E. Kamphuis, MSc (UT)

Contents

List of abbreviations

ALARA	As low as reasonably achievable
ARF	Atlas regression forest
CI	Conformity index
CRF	Conditional random field
CRT	Chemoradiotherapy
CT	Computed tomography
CTV	Clinical target volume
DVH	Dose volume histogram
GTV	Gross tumor volume
HI	Homogeneity index
HNC	Head and neck cancer
HPV	Human papilloma virus
IMPT	Intensity modulated proton therapy
IMRT	Intensity modulated radiotherapy
LIPP	Dutch National Indication Protocol for Proton therapy
ML	Machine learning
MLO	Machine learning optimization
MRI	Magnetic resonance imaging
NTCP	Normal tissue complication probability
OAR	Organ at risk
PBS	Pencil beam scanning
PCM	Pharyngeal constrictor muscle
PET	Positron emission tomography
pRF	Prediction random forest
PTV	Planning target volume
RBE	Relative biological effectiveness
ROI	Region of interest
RT	Radiation therapy
SCC	Squamous cell carcinoma
SOBP	Spread-out Bragg peak
TCP	Tumor control probability
TNM	Tumor-node-metastasis
VMAT	Volumetric arc therapy
Vw-min	Voxel-wise minimum

Preface

In this thesis, 'Automated robust treatment planning for IMPT in oropharyngeal cancer patients using machine learning', I will present the results of using machine learning for robust proton therapy treatment planning in oropharyngeal cancer patients. I have been working on this project the past year at the radiotherapy department of the University Medical Center Groningen (UMCG). This thesis was written to obtain the master's degree of Technical Medicine at the University of Twente.

In Chapter 1, background information is given about head and neck cancer, (proton) radiotherapy and machine learning. The main set up and final results of this study are described in Chapter 2. An overview of the development process which have led to the final results of this thesis and further recommendations are given in Chapter 3. The references are listed in Chapter 4.

I hope you enjoy reading this thesis.

Merle Huiskes

Chapter 1 Background

1.1 Head and neck cancer

Head and neck cancer (HNC) consists of a heterogeneous group of cancers where the affected anatomical regions reach from the skull base to the clavicles. Cancer in the head and neck area can give rise to serious complaints due to involvement in important functions such as breathing, eating, drinking and speaking.[1][2] The annual incidence of HNC is approximately 3000 in the Netherlands, resulting in an mortality of 700 per year.[1] Males are more often diagnosed with HNC than females with a ratio of 2:1 and the most common age at time of diagnosis is between 60-80 years.[1][3]

1.1.1 Functional anatomy of the head and neck

Different anatomical structures are located in the head and neck region. The paranasal sinuses are a group of different air-filled sinuses located within the bones of the skull and face. The largest paranasal sinuses are the maxillary sinuses, which are located under the eyes. Within the frontal bone and above the eyes, the frontal sinuses are positioned. The ethmoid sinuses are located within the ethmoid bones between the nose and the eyes. In the center of the skull base, the sphenoid sinuses are positioned superior to the nasopharynx and inferior-medial to the cavernous sinuses. The cavernous sinuses contain several important structures such as cranial nerves. The function of the paranasal sinuses is debated; however they are implicated in several roles. They decrease the weight of the skull and can absorb shocks from outside in order to protect structures such as the brain and eyes. Also, they humidify and heat inspired air and increase the resonance of the voice.[4][5]

The oral cavity is separated from the nasal cavity by the hard palate.[4] The primary function of the oral cavity is to serve as an entrance of the alimentary tract where it initiates the digestive process.[6] Besides, it serves as a secondary respiration conduit and speech articulation. The lips form the anterior part of the oral cavity and the palate tonsils the posterior part. The floor of the oral cavity is formed by the lower gingiva and the anterior two thirds of the tongue. The mandible, the teeth and the retromolar trigone are also located in the oral cavity.[4]

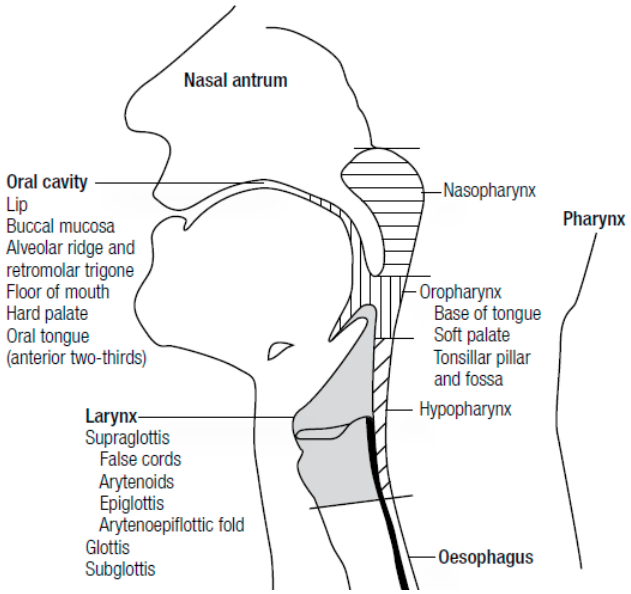


Figure 1 Anatomical structures in the head and neck [3]

The pharynx is part of the respiratory as well the digestive system. The pharynx can be divided into three parts: the nasopharynx, oropharynx and hypopharynx. The nasopharynx reaches from the skull base to the junction of the hard and soft palates. The nasopharynx allows conducting air through the nose. The pharyngeal tonsils are located in the posterior wall of the nasopharynx. The superior wall of the oropharynx is partly formed by the superior surface of the soft palate and the uvula. The anterior wall is formed by the base of the tongue and glossoepilottic folds and the lateral wall by the tonsils. The oropharynx continues inferiorly into the hypopharynx. The hypopharynx consists of three areas: the paired pyriform sinuses, the posterior pharyngeal wall and the post cricoid area. Both oropharynx and hypopharynx serve as a passageway for food and air.[4][7]

The larynx is a hollow muscular organ which is part of the respiratory system. The area of the larynx can be divided into the supraglottis, glottis and subglottis. In the supraglottis larynx, the epiglottis and false vocal cords are located. The epiglottis closes during swallowing to prevent aspiration and forcing the food to the esophagus. The true vocal cords are located in the glottis and are involved by the speech.[4][7] An overview of the structures mentioned above is given in figure 1.

The salivary glands are exocrine glands that secrete saliva. The sublingual glands are located underneath the mucosa of the anterior floor of the oral cavity. The submandibular glands are positioned in the upper neck and the parotic glands are located in the subcutaneous tissue of the face, anterior inferior to the external ear.[4]

The lymph nodes responsible for the lymphatic drainage of the mucosal surfaces of the head and neck are located in the fibro adipose tissues in the neck. These are the submandibular and submental nodes at level I, the upper, middle and lower jugular nodes at levels II, II and IV respectively, the posterior triangle nodes at level V, and the anterior neck nodes at level VI.[4] For an overview of these nodes, see figure 2.

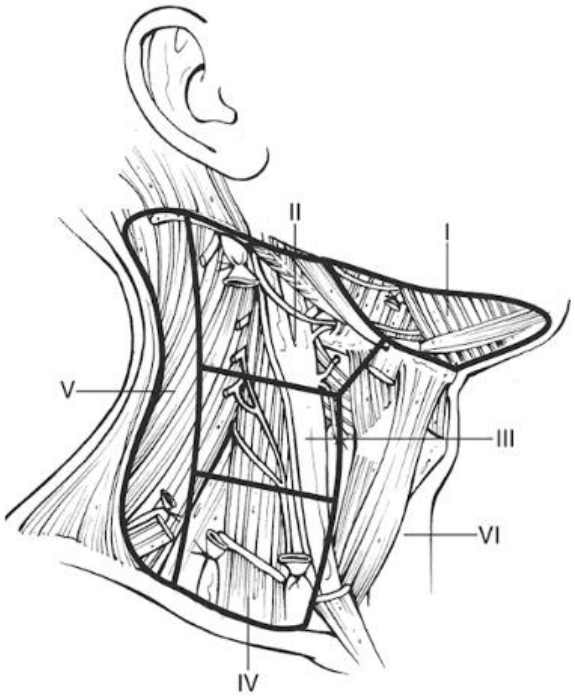


Figure 2 The different anatomical levels of the lymph nodes in the neck[4]

1.1.2 Pathophysiology

More than 90% of all HNC cases are squamous cell carcinomas (SCC) that arise from the mucosal surfaces of the head and neck area.[8] The tumor sites in decreasing incidence in The Netherlands includes; the oral cavity, larynx, oropharynx, salivary glands, hypopharynx, paranasal sinuses and nasopharynx.[1]

The main risk factors for development of SCC are smoking, alcohol use and the presence of human papilloma virus (HPV). The combination of smoking and alcohol use has synergistic effect. Overall, HPV positive patients have better prognosis.[9] Other contributing risk factors are exposure to ultraviolet light for the development of SCC of the lips, diets deficient in antioxidants, age and family history of SCC.[8]

Patients with HNC present with a variety of symptoms, mostly referable loco regional at the affected site. Examples of commonly symptoms are an ulceration, pain, bleeding, dysphagia, odynophagia, changes in articulation, otalgia, hoarseness, nasal of ear congestion, epistaxis, diplopia or a lump in the neck.[4] The curative rate for early stage HNC is high. However, two-third of the patients present with locally advanced disease.[10]

1.1.3 Diagnosis and staging

The first step to diagnosis of HNC is physical examination of clinical site, often followed by an endoscopy of the head and neck region. Histological examination is needed for confirming the diagnosis of HNC, after tissue is obtained by a biopsy. The tumor extent and eventual metastatic spread is determined by (contrast enhanced) computed tomography (CT) and magnetic resonance imaging (MRI). Positron emission tomography (PET) gives functional added information about the tumor and eventual affected lymph nodes.[4]

The tumor-node-metastasis (TNM) classification is used to assess the extent of disease. The tumor stage is specific to the different anatomical sites in the head in neck, while the nodule stage is common to all head and neck sites except the nasopharynx. For a detailed description of the TNM staging for HNC, see [11].

1.1.4 Treatment

The aim of treatment in HNC patients is to maximize loco regional control while minimizing functional loss and negative cosmetic effects. Depending on the type and TNM-stage of HNC, for patients without distant metastasis, curative treatment options are surgery and radiotherapy, possibly combined with systemic therapy.[12][13]

In surgical resection of the tumor, the overarching goal is to obtain pathologically negative margins. If this is not achieved, postoperative chemoradiotherapy (CRT) or radiation therapy (RT) is indicated. In early stage (I or II) HNC, either resection or RT alone provides effective treatment. In more advanced stages, a multimodality approach, e.g. concurrent or CRT is indicated.[12][13]

1.2 Radiation therapy

In RT, a specific amount of radiation dose is delivered to the tumor and causes DNA damage, resulting in cell apoptosis. The aim is to deliver enough radiation dose to the target to achieve a high probability of tumor control; the tumor control probability (TCP), while sparing the normal tissue to achieve a low risk of normal tissue complications; the normal tissue complication probability (NTCP).[14] RT is based on the concept that the DNA repair capacity in healthy cells is generally larger than in cancerous cells.[15]

In HNC, a curative treatment dose of 70 Gy is delivered in 35 daily fractions of 2 Gy to the primary tumor and the eventual affected lymph nodes. For the low risk lymph nodes, an amount of 54.25 Gy is given in daily fractions of 1.55 Gy. The daily fractions are given five or six days a week and the total radiation treatment has a duration of several weeks, where the radiation effect continues several weeks after the last treatment.[16]

1.2.1 Treatment planning

Prior to radiation treatment of HNC, a radiotherapy plan of the dose distribution is made based on a CT scan of the patient, where needed with additional information from the MRI and PET scan. During these scan acquisitions, a patient specific head neck mask is used to provide immobilization, to limit day to day position changes during treatment.

CT is the imaging modality which is the gold standard for the initial delineation of the tumor and eventual lymph nodes. The CT offers inherent information of the electron density, which is used for photon and proton dose calculation. The use of an iodinated contrast CT for the delineation increases the sensitivity, where the use of MRI can give more details of the soft tissues. PET gives additional metabolic information of the tumor and lymph nodes.[17][18]

The visible target volume (i.e. primary tumor or lymph nodes) is delineated as the gross tumor volume (GTV). This GTV is expanded with a margin of five millimeters accounting for microscopic extensions, which will form the clinical target volume (CTV). Next to this, an extra margin of three millimeters is added to correct for geometrical errors such as variability in in tissue positions, sizes, and shapes, as well as for variations in patient position and beam geometries, resulting in the planning target volume (PTV). The PTV concept will make sure that 95% of the prescribed dose is as least delivered to the CTV in 90% of the patients.[19]

1.2.2 OARs

Due to the delineation margins and the inherent lateral penumbra of photon and proton beams, the surrounding healthy tissue will receive also substantial radiation dose. Damage to certain organs can cause side-effects. These organs are called organs at risks (OARs) and specific dose limits are described for these organs. The OARs in the head neck region include the spinal cord, submandibular and parotic glands, oral cavity, swallowing muscles, brain(stem), eyes and the cochlea. The incidence of complications is related to the dose delivered to these organs.[20]

Dose delivery to the OARs is aimed to be as low as possible. Constraints are established as allowed doses delivered to the OARs. Serial organs, such as the spinal cord, can lose their functionality if only a small volume of the organ receives a dose above the tolerance limit. A maximal dose (Dmax) delivered in one voxel of the OAR is defined as constraint for serial organs. So-called ‘parallel’ organs, such as the parotic glands, are damaged only if a larger volume is included in the irradiation. For

parallel organs, a mean dose (Dmean) delivered in the total volume of the OAR is defined as constraint. An overview of the OARs in HNC and their dose constraints is given in table 1.

Table 1 Organs at risk and their corresponding dose constraint [21]

OAR	Dose constraint (Gy)
Spinal cord	Dmax = 54.25
Parotic glands	Dmean = ALARA*
Submandibular glands	Dmean = ALARA*
Brain	Dmax = 60.00
Brainstem	Dmax = 63.00
Optic nerve, optic chiasm, retina	Dmax = 60.00
Eyes	Dmax = 5.00
Cochlea	Dmax = 52.50

*** As low as reasonably achievable**

1.2.3 Complications

During and after RT treatment, patients can suffer from different complications. These complications can be subdivided into acute complications, which can be present during and a 3-6 months after the treatment, and late complications, which will be present 6 months and later after the treatment. Complication rates increase in case of concurrent chemotherapy is given. The RT related complications include mucositis of the throat or mouth, dermatitis, xerostomia, alternation in taste, dysphagia, hypothyroidism, aspiration, pain and hoarseness. Xerostomia is the most common radiation-related complication and still persists after completion of the treatment.[16]

1.2.4 Treatment delivery

For radiation therapy in HNC, the radiation type can be photons or protons. The most common type of radiotherapy used is photons, while a part of the HNC patients are eligible for proton therapy.[22]

In the past decades, several advances are made in the RT delivery techniques. In earlier times, three-dimensional conformal radiation therapy (3D-CRT) was used, where beams were shaped around the tumor using multi-leaf collimators.[23]

A more advanced form of 3D-CRT is intensity modulated radiotherapy (IMRT), which was introduced in 1998 in the Netherlands.[24] In IMRT, a computer-optimized inverse treatment planning and a computer-controlled multi-leaf collimator are used to determine the optimal fluency of each field. With these techniques, the intensity of the beams can be modulated so that a higher dose can be delivered to the target with a conformal target volume coverage, while the dose to the surrounding normal tissues is reduced.[16][23]

Later, volumetric arc therapy (VMAT) was developed in addition to IMRT, which uses a 360 degrees continuously rotating intensity modulated beam around the patient. Compared to IMRT, VMAT has a reduced delivery time and improved target volume coverage and healthy tissue sparing.[23] At the moment, in the UMCG, VMAT technique is used for photon therapy delivery in HNC. Intensity modulated proton therapy (IMPT) is the current state-of-the-art treatment modality in the UMCG available for proton therapy.

1.2.5 Proton therapy

Opposite to photons, protons are heavy particles that will stop after travelling a specific distance. Depending on the energy of the proton beam, the particles deposit their energy at a specific depth, which is characterized by the Bragg peak. After the Bragg peak, there is a steep dose gradient, resulting in practically no exit dose.[25]

Due to this Bragg peak, a high dose can be delivered very precisely in the target volume and very low doses are delivered in the surrounded tissues. However, the characteristic Bragg peak is too narrow for practical clinical applications. Therefore, the beam energy is modulated and the different proton beams will result in spread-out Bragg peaks (SOBP) to encompass the entire target. In figure 3, the Bragg peak, SOBP and photon dose-depth relation graphs are shown.

Proton therapy delivery is performed with pencil beam scanning (PBS) proton therapy. With PBS, a small pencil beam of protons is scanned across the predefined target volume, which allows for IMPT.[25]

Due to proton therapy is more expensive and not broad spread available, it is necessary to select which patients are likely to have significant clinical benefit from proton therapy relative to photon therapy. The clinically relevant benefit of proton therapy has still not been demonstrated consistently due to limited resource.[26] Therefore, in the Netherlands, before patients are eligible for proton therapy, a radiotherapy plan for both photons as well as protons is made and compared. The target coverage and NTCP reduction are calculated, which forms the basis for the decision for proton therapy. If a minimal amount of NTCP reduction is achieved, e.g. a delta NTCP of 10% for grade 2 complications and 5% for grade 3 complications, the patient will be eligible for proton therapy. The first version of this Dutch National Indication Protocol for Proton therapy (LIPP 1) for this procedure can be found in [27]. Since August 2019, an updated version of the indication protocol for protons (LIPP 2) is used, see [28]. The main changes between those versions are that more weights are given to the oral cavity in the calculation of the NTCPs. Also, for the calculation of the NTCP for xerostomia, the submandibular glands and ipsilateral parotid are included.

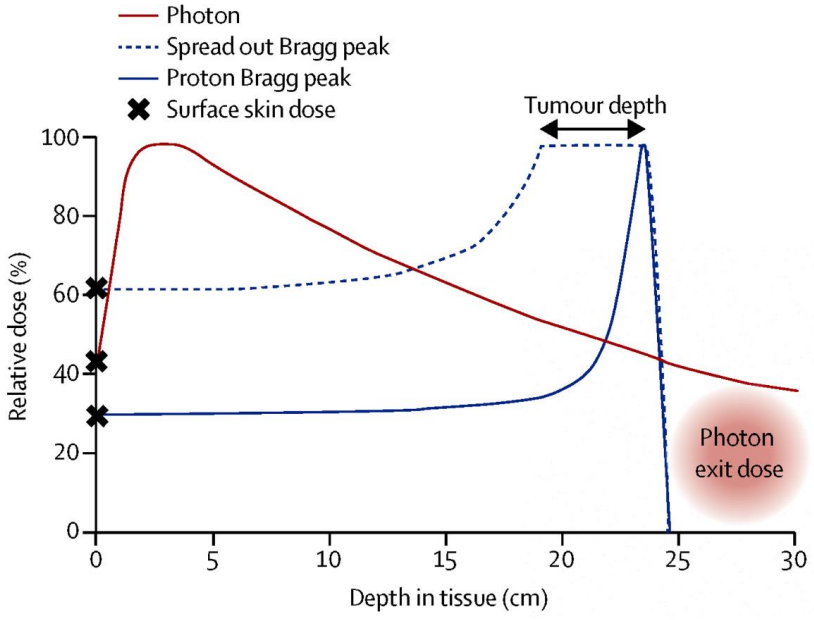


Figure 3 Dose-depth curves for photons and protons.[29]

1.2.6 CTV robustness

Different to the treatment planning of photons, in proton therapy the concept of a robust dose distribution of the CTV against geometric uncertainty is applied, because for protons the additional margin from CTV to PTV is not suitable. In photon therapy this concept works well because dose distributions of photons are relatively insensitive to density changes in the beam path. These density changes are the result of patient set-up or anatomy differences, and an extra marge will solve these mentioned geometrical errors.[30][31]

However, in proton therapy, the proton depth range is dependent of the stopping power, and thus of the beam path density changes. This makes the additional marge from CTV to PTV not suitable for proton therapy. Instead of this, a robust optimization principle is used.[31]

In robust optimization, multiple dose distributions are simulated for different scenarios of range and set-up uncertainties. Different methods for robust optimization are possible. Nowadays, in the UMCG planning software of RaySearch, a minimax optimization approach is applied. In this method, the worst-case dose distribution is one of 21 perturbed scenario dose distributions with the worst value of a certain dosimetric parameter, e.g. lowest near minimum dose in the CTV. Then, this worst-case scenario is optimized until it fulfills the predefined criteria. With this method, it is assured that even in the worst case scenario, adequate dose (D98>94%) is delivered to the target, and also OAR sparing is achieved.[30][31][32] Also, other mimicking strategies are possible, i.e. a more stochastic approach than minimax approach. A stochastic robust optimization approach optimizes the expected plan quality more than taking the plan quality for the worst error into account.[30]

Due to the competition between target coverage and OAR sparing, after a CTV robustness optimization it is necessary to perform a robustness evaluation. This evaluation quantifies whether coverage criteria have been satisfied under the specified uncertainty conditions for even more error scenarios than used in the optimization step. In the robustness evaluation, a voxel-wise-minimum and -maximum dose is constructed from all evaluation dose distributions. This represents the worst-case dose distribution by the minimum of each voxel inside the CTV and the maximum of each voxel outside the CTV taken from all evaluated dose distribution scenario.[31]

1.3 Machine learning

In machine learning (ML), algorithms implement a mathematical model based on training data to make predictions or decisions. In supervised learning, the mathematical model is built from a set of data that contains both the input and the desired output.[33] ML has promising applications in the workflow of RT, for example in the treatment planning.[34]

Manual clinical RT planning is a time-consuming process, which takes hours to days for one patient. In addition, manually created plans are susceptible for inter-observer plan quality variability, due to the skills, experience and time investment in one plan of the planners. ML is an option to overcome these shortcomings by automatically generating RT plans.[34] Previous studies on in the incorporation of ML for (semi-)automatic generation of treatment plans has already shown successful for different cancer types, also for head and neck cancer.[35] Currently, in RayStation (RaySearch Laboratories AB), a ML method is available which is based on the approach of McIntosh, see [35]. The ML method consists of different steps; a training, prediction and mimicking optimization. An overview of these different steps is given in figure 4, which will be explained in more detail in the next sections.

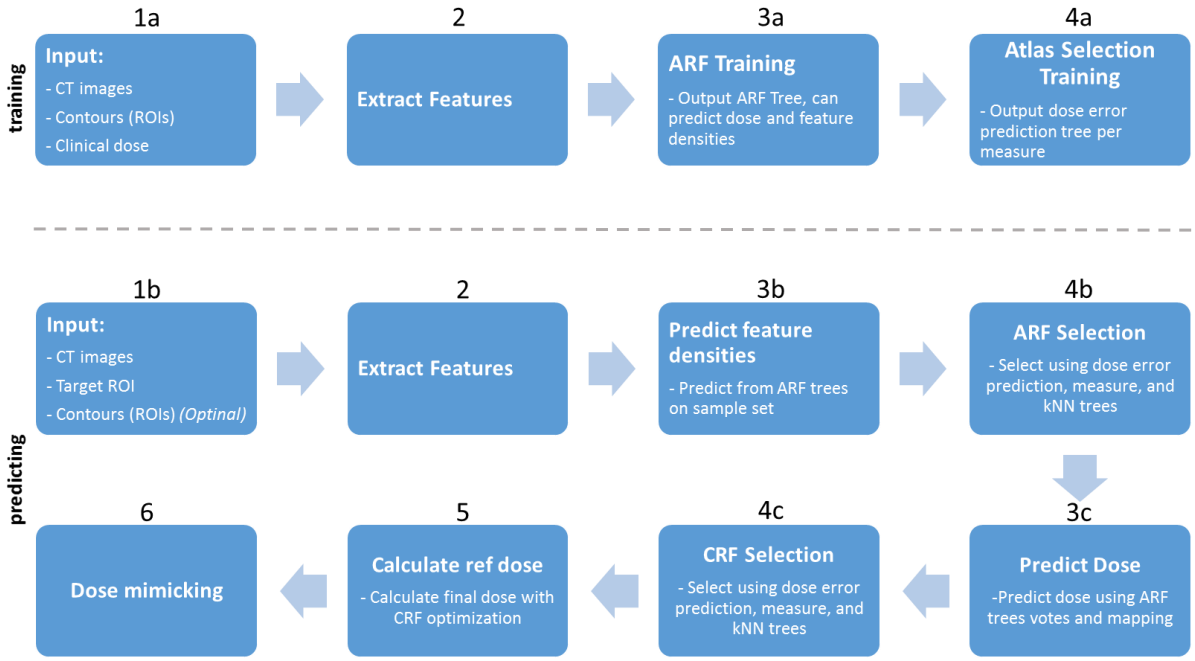


Figure 4 Overview of the different steps of the Machine Learning method, which consist of a training and a prediction part.[36]

1.3.1 Machine learning training

In the first phase of the ML model shown in figure 4, a training based on the input is performed, which can be subdivided into four steps. In step 1a, the planning CT including the delineated target and OAR structures and the clinical dose distribution are added. In the second step, the voxel-wise features are extracted by convolution with 3D-Gauission filters. Region of interest (ROI) specific features are extracted by calculating the distance of an OAR to the nearest target ROI. In step 3a, for every input CT of each patient, an atlas regression forest (ARF) training is performed. This is followed with step 4a, where a prediction random forest (pRF) is trained based on atlas selection accuracy measures from the predicted features densities of the other trained ARFs.[36]

1.3.2 Machine learning prediction

In the second phase of the ML model shown in figure 4, a dose prediction is made. Step 1b and 2 are similar to these steps in the ML training phase. This is followed by step 3b, where the feature density in new patients is predicted from the ARF trees in the training. Thereafter, the pRF trees are used to predict the accuracy of an ARF by selection of the closest ARF for dose prediction (4b). Then, in step 3c, the dose in a voxel is predicted by dose votes. Each voxel has a predicted dose distribution and the highest dose vote of five selected ARFs is selected (3c). In the conditional random field (CRF) selection in step 4c, the ARF is selected based on the mean dose error prediction with the dose volume histogram (DVH) of all ROIs. Then, CRF optimization is applied in step five to calculate the final dose where it fulfills the predefined clinical criteria.[36]

1.3.3 Machine learning mimicking

The final step of the ML model (step 6), is mimicking the predicted dose to create a deliverable treatment plan. For photon plans, this mimicking step was already available. However, for protons, a mimicking step has been become available in the RayStation development version 9A, to be able perform a robust optimization mimicking.[36]

Chapter 2 Automated robust treatment planning for IMPT in oropharyngeal cancer patients using machine learning

Merle Huiskes[1,3], Roel G.J. Kierkels[1], Erik W. Korevaar[1], Roel J.H.M. Steenbakkers[1], Mats Holmström[2], Hanna Gruselius[2], David Andersson[2], Albin Fredriksson[2], Karl Berggren[2], Cornelis H. Slump[3], Stefan Both[1], Johannes A. Langendijk[1], Fredrik Löfman[2].
[1]*University Medical Center Groningen, Department of Radiation Oncology, Groningen, The Netherlands.*
[2]*RaySearch Laboratories AB, Machine Learning Department, Stockholm, Sweden.*
[3]*University of Twente, Enschede, The Netherlands.*

2.1 Abstract

Background: Radiotherapy treatment planning is a complex and time-consuming process. In intensity modulated proton therapy (IMPT) treatment planning steep in-field dose gradients are used, which allows for dosimetric benefits, but making IMPT also sensitive for density- and set-up errors. Robust planning methods have been developed to account for these errors, but robust treatment planning is even more time consuming. In this study, we combined a robust dose mimicking optimization with a machine learning based dose prediction (machine learning optimization (MLO)) to automatically generate robust IMPT plans. We aimed to automatically generate robust IMPT plans for oropharyngeal cancer patients with similar quality as clinically available plans.

Methods and materials: A total of 79 robust IMPT plans of oropharyngeal cancer patients were included. Dose distributions, contours and CT image features of 66 patients were used to train a model to predict dose distributions for novel patients. The target coverage during training was based on primary and elective clinical target volume (CTV). Dose prediction was based on a random forest model. Hence, the predicted dose was converted into a deliverable plan using robust voxel-wise dose mimicking optimization, including 21 perturbed dose scenarios with 3 mm isocenter shifts and ±3% density uncertainty. IMPT plans from 8 patients were used for (cross)-validation, and subsequently testing was performed with 9 patients. Targets were assessed in terms of robust target coverage (voxel-wise minimum D98>94%), conformity and homogeneity indices. Organ at risk (OAR) dose was assessed by Dmax, Dmean and (sum) normal tissue complication probability (NTCP)s. All parameters were compared to the clinical plan.

Results: Robust target coverage was achieved in 8/9 MLO plans (89%) and voxel-wise minimum dose (D98) was statistically significant higher in the high dose CTV compared to the clinical plans. The sum NTCP was lower or did not increase >2% compared to the clinical plans in 8 of the 9 MLO plans (89%) for xerostomia and in 5 of the 9 MLO plans (56%) for dysphagia. The Dmax constraints to the brain, brainstem, spinal cord and eyes were not exceeded in the MLO plans and comparable to the clinical plan. The mean dose to the parotids, right submandibular and pharyngeal constrictor muscles (PCMs) were on average comparable to the clinical plans, i.e. differences were <1Gy. The average Dmean in the oral cavity was statistically significant higher in the MLO plans (3402±1504 cGy) compared to the clinical plans (3102±1515 cGy) with a p-value of 0.008.

Conclusion: The MLO algorithm was able to generate robust IMPT plans for oropharyngeal cancer patients with clinically acceptable robust target coverage, and with similar dose to most OARs as compared to the clinical plans. Further reduction of the dose to specific OARs and improvement of the plan consistency between patients have to be investigated.

Key words: Automated Treatment Planning, IMPT, Head and Neck Cancer, Machine Learning

2.2 Introduction

Radiation therapy (RT) is one of the main curative treatment options for head and neck cancer (HNC) patients without distance metastasis, possibly combined with surgery or concurrent chemotherapy.[12][13] The aim of RT is to deliver enough radiation dose to the target to achieve a high probability of tumor control; the tumor control probability (TCP), while sparing the normal tissues to achieve a low risk of normal tissue complications; the normal tissue complication probability (NTCP).[14] Intensity modulated proton therapy (IMPT) has shown lower dose to normal tissue compared to photon therapy while preserving target coverage.[37]

Especially in HNC, where multiple target dose levels are defined close to organs at risk (OARs), compromising between target coverage and OAR sparing makes RT treatment planning complex. In manual RT treatment planning, many steps are involved, and multiple iterations are required for plan optimization, making it a time-consuming process which takes up to two days for an IMPT plan. Also, manual RT treatment planning is susceptible for inter-observer plan quality variability, due to the skills, experience and time investment of the radiation therapy technician.[10][11]

In addition to this, IMPT uses multiple beams with steep in-field dose gradients, which allows for dosimetric benefits i.e. sparing surrounded healthy tissue whilst preserving target coverage. However, due to the steep in-field dose gradients, IMPT becomes sensitive for CT-density and setup errors. Mathematical robust planning methods, such as stochastic programming and minimax optimization, have been developed to account for these errors. Robust planning methods take multiple scenario dose distributions into account during the optimization process, which makes robust planning time consuming in comparison with traditional margin-based planning methods.[30]

In order to decrease the time required for RT treatment planning and to improve the overall plan quality and consistency, semi-automated planning algorithms have been developed.[10][11] With a machine learning (ML) approach, an algorithm can learn how features are related to and predictive for the outcomes from a training dataset. Previous studies on the incorporation of ML for (semi-) automatic generation of RT treatment plans have already shown successful for different cancer types, including HNC.[34][35][38][39][40]–[43]

McIntosh et al. have developed a method for multi-patient atlas-based dose prediction based on the assumption that patients with identical geometry and appearance should be treated in the same way.[35][44][45] A model is trained with features from a data set of computed tomography (CT) images with delineated regions of interest (ROIs) which are related to the spatial dose distribution. For a novel patient, most similar patients from this model are automatically selected with ML and used to predict the desired dose per voxel.[13][19][20] This atlas-based dose prediction was extended by applying a dose mimicking optimization, to create deliverable intensity modulated radiation therapy (IMRT) and volumetric modulated arc therapy (VMAT) plans. Subsequently testing in 12 right sided oropharyngeal patients resulted in fully automated generated IMRT or VMAT treatment plans with better target coverage and OAR sparing compared to manual optimized plans.[35]

In the treatment planning software system of RayStation (RaySearch Laboratories AB, Stockholm, Sweden) a machine learning optimization (MLO) method was implemented based on the method of McIntosh et al. Early institutional investigation of the application of this MLO method in photons showed promising results for automatically generating VMAT plans for oropharyngeal cancer.[46]

However, for IMPT this approach is not directly applicable, since a robust mimicking optimization step has to be performed to simulate a deliverable dose which is robust against certain set-up and range errors. Therefore, in the present study, a ML based dose prediction is combined with a robust mimicking optimization algorithm to automatically generate robust IMPT plans. The aim of this study is to automatically generate IMPT plans for oropharyngeal patients with at least comparable target coverage robustness as clinically optimized robust IMPT plans. The second goal is to reduce the dose to the OARs to achieve IMPT plans with at least similar quality as clinically available plans.

2.3 Methods and materials

2.3.1 Study population

A total of 79 oropharyngeal HNC patients planned for IMPT between January 2018 and December 2019 were included in this study. All patients were treated with curative intention and a prescribed dose of 7000 cGy (relative biological effectiveness (RBE)) to the primary tumor and 5425 cGy (RBE) to the bilateral elective lymph node regions. Clinical target volumes (CTVs) were delineated on the CT scan by the radiation oncologist and OARs were delineated by the treatment planners. After this, dose distributions were robustly optimized by the treatment planners to cover the CTVs robust for 0.5 cm isotropic position uncertainty and 3% density uncertainty for plans made in 2018. For plans made since the beginning of 2019, dose distributions were planned robust to cover the CTV for 0.3 cm isotropic position uncertainty and 3% density uncertainty.[47] Plans were optimized until they fulfilled the voxel-wise minimum (vw-min) criterion of D98>94% and minimizing dose on NTCP organs according to the Dutch National Indication Protocol for Proton therapy (LIPP) version 2 (since August 2019) or LIPP version 1 (before August 2019). An overview of the patients and plan characteristics is given in table 2. Treatment plans were generated with the RayStation version 8B clinical treatment planning system. In total, 4 beam directions with a range shifter of 4 cm were used. The dose grid was $0.3 \times 0.3 \times 0.3$ cm^3 for all plans and final dose was computed using Monte Carlo simulations with 1.0% uncertainty.[48] The final IMPT dose distributions were deliverable with the IBA ProteusPlus (Ion Beam Applications SA, Louvain-la-Neuve, Belgium) treatment machine.

2.3.2 Model training

For the training of a ML model, input data consisted of CT scans, delineated ROIs and the dose distribution of each treatment plan. ROIs included in the training were the planning target volumes (PTVs), CTVs, brain, brainstem, spinal cord, parotid and submandibular glands, oral cavity, larynx, cricopharyngeal muscle, pharyngeal constrictor muscles (PCMs), cervical esophagus, mandible and thyroid. From the CT images and these ROIs, voxel-wise features that characterize patient geometry, ROI shape, and image appearances, such as gradient and texture, were extracted with first- and second order 3D Gaussian filters. This resulted in a total of 88 image features, and per patient, one atlas regression forest (ARF) was trained to relate the features in a voxel to the corresponding dose.[5][13] The ARF consisted of 96 trees with a maximum depth of 10 node levels for each tree. The classifier was trained starting from each node with a random subset of features with replacement. For each set of features used in a split node, the information gain was calculated. Then, the features in the nodes with the highest information gain were used for the final ARF.[36]

With the trained ARFs, a prediction random forest (pRF) model of 96 trees with a maximal depth of 10 node levels was trained to predict the dose accuracy of an ARF. For all trained ARFs, leave-one-out cross-validation with given feature density differences between the ARFs were used to train the pRF,

resulting in a dose prediction accuracy. The Bhattacharyya distances of the ARFs were calculated from the feature values for each node in level 7 of the ARF. Using node level 7 instead of level 10 was a good trade-off between calculation time and information gain.[36] This distance was related to the dose prediction accuracy to train the pRF. Then, the pRF was able to predict the ARFs with the highest accuracy for a novel image, only based on image features and thus without knowing the dose distribution of the novel CT image.[35]

2.3.3 Dose prediction

For the prediction of the dose distribution for a novel CT image, first voxel-wise features were extracted from the CT and ROIs. For each ARF in the training, the Bhattacharyya distances from the novel feature values at node level 7 were calculated. Subsequently, the pRF could select the ARFs with the highest dose accuracy based on the novel image features, and thus the best matching ARFs from the training. It is recommended to not select more ARFs than 10% of the number of patients in the training set, to prevent overfitting.[49][50] So, depending on the size of the trained model in this study, three or five best matching ARFs were selected. The selected ARFs were used to predict a joint probability per voxel. So far, each voxel was treated independently and finally, a conditional random field (CRF) was added to the predicted dose. This CRF was able to find the most likely spatial dose distribution to each voxel that adheres to the dose prior, and generated the final predicted dose distribution.[36][35]

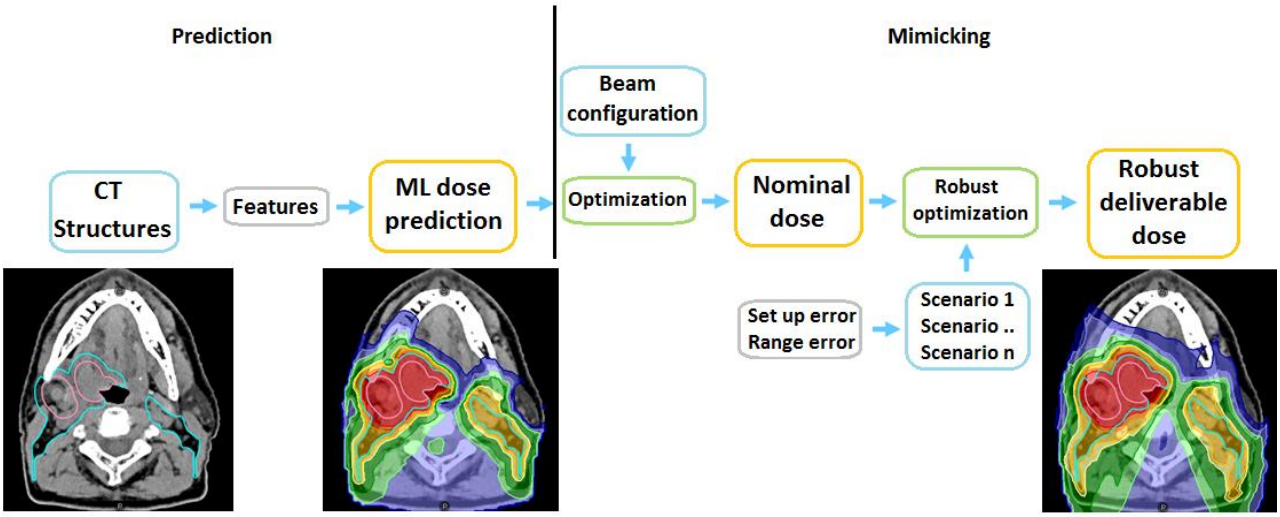


Figure 5 Overview of the different steps of the MLO to create a robust deliverable dose for a novel patient, divided into the prediction and mimicking part.

2.3.4 Robust mimicking optimization

To generate actual deliverable treatment plans, a robust dose mimicking optimization approach was applied to the predicted dose distributions. The dose mimicking could be voxel-based or dose volume histogram (DVH)-based. In voxel-based mimicking, the mimicking equals or improves the predicted dose at each voxel. In DVH-based mimicking, the dose as a distribution per ROI is taken into consideration and spatial dose information is lost.[35] In our method, a voxel-based mimicking method was applied.

The predicted dose distributions were mimicked by taking the beam geometry, scatter and attenuation into considering, and final dose calculations were performed using a Monte Carlo dose engine. The same beam arrangement as the clinical plans were used for the dose mimicking

optimization step. For the robustness optimization, 21 scenarios were taken into account where the same density- and position uncertainty parameters as in the clinical robustly optimized plans were used. With these settings, a total of three rounds with 60 iterations were used for the robust mimicking optimization. A schematic overview of the input and output of the prediction and robust mimicking steps is shown in figure 5.

2.3.5 MLO tuning

During the prediction and mimicking steps, various settings (as described in Chapter 3) could be added and tuned which would be taken into consideration while expressing the voxel dose value. In the prediction, based on the selected ARFs, for each voxel the dose value with the highest probability was selected. But in addition to this, each voxel had more possible values, which were less likely based on the pRF. The selection of a dose value for each voxel could be influenced by adjustments of the settings in the prediction phase, i.e. by adding ROI goals.

Also, in the mimicking step it was possible to influence the optimization process. Adjustments in the mimicking settings would consider to maintain the predicted dose in the voxels or to adjust the dose value, i.e. due to robustness parameters or settings which involve reduce dose to OARs. By the combination of tuning the settings in the prediction and mimicking optimization, it was possible to generate a MLO dose distribution with a specific strategy, for example a strategy with more sparing in specific OARs. In this study, the main focus was to tune on robust target coverage. An overview of the main development steps to achieve this goal is described in Chapter 3. When this goal was achieved, an OAR dose reduction step was added as a second goal next to robust target coverage.

2.3.6 Trained models

A total of 6 ML models with different amount of atlas patients or plan characteristics were trained for development purposes during this study. The final model used in this study consists of 66 treatment plans, further referred as Model-66. Four plans were used for initial tuning until robust target coverage was achieved. Cross-validation was performed by training a second model consisting of 66 treatment plans, including the 4 initial tuning plans of Model-66 and excluding 4 training plans, which were now used for tuning this new model. The remaining 9 plans formed the independent test group. More information about the size and plan characteristics in these models can be found in table 2.

2.3.7 Evaluation

To evaluate the performance of the ML models, the mimicked MLO plans were compared to the clinical optimized plans. To evaluate target coverage robustness, a multi-scenario plan evaluation method comprising 16 dose recalculations with 8 positional isocenter shifts of 0.3 cm and a ±3% density uncertainty was performed. The scenario dose distributions were combined into a vw-min dose distribution, and targets were considered robust if the D98 was at least 94% of the prescribed dose, i.e. 5100 cGy for the CTV5425 and 6580 cGy for the CTV7000.[51] Also, the conformity index (CI) and homogeneity index (HI) were calculated as:

$$CI = \frac{TV_{95}}{V_{95}}$$

where TV_{95} is the part of the target volume which receives 95% of the prescribed dose, and V_{95} is the total volume which receives 95% of the prescribed dose, and:

$$HI = \frac{D_2 - D_{98}}{D_{50}}$$

where D_2, D_{98} and D_{50} are the doses received by 2%, 98% and 50% of the target volume respectively.

The dose to the OARs in the MLO plans was assessed by the Dmean or Dmax and compared to the clinical plans. Also, the sum of the NTCPs for xerostomia grade 2 and xerostomia grade 3 (sum xerostomia) and dysphagia grade 2 and dysphagia grade 3 (sum dysphagia), were determined according the LIPP2 for both the clinical as well as the MLO plans. All dosimetric parameters were extracted from RayStation and visualization was done in Tableau (Tableau Software Inc., Seattle, Washington).

To asses if differences between the generated MLO plans and clinical plans were statistically significant ($p<0.05$), a Wilcoxon signed rank test was performed in SPSS (IBM SPSS Statistics 23, Armonk, New York, United States).

Next to this, also the time to train a model and to create a MLO dose distribution was recorded.

Table 2 Overview of the patient and plan characteristics per model.

		Total (n=79)	Model-66		Model-66 Cross		Independent test set (n=9)
			Train (n=66)	Tune (n=4)	Train (n=66)	Tune (n=4)	
Tumor Location	Oro-pharynx	51	42	3	43	2	6
	Base of tongue	15	13	1	12	1	2
	Tonsil	13	11	-	11	1	1
T-stage	T1	15	12	1	12	1	2
	T2	11	11	-	11	-	-
	T3	14	11	2	12	1	1
	T4*	39	32	1	31	2	6
N-stage	N0	11	9	-	8	1	2
	N1	17	15	1	16	-	1
	N2**	37	28	3	28	3	6
	N3***	14	14	-	14	-	-
LIPP version	1	66	60	2	60	2	4
	2	13	6	2	6	2	5
Robustness position uncertainty (cm)	0.3	36	23	4	23	4	9
	0.5	43	43	-	43	-	-

***T4a, T4b, T4NOS**

**** N2a, N2b, N2c, N2NOS**

*****N3a, N3b**

2.4 Results

The time it took to train a model depended on the size and was increased by including more plans. For the model with 66 plans, the training had a duration of 78 hours, and for the model with 32 plans, it took 28 hours to train the model. Prediction of a dose distribution for a novel image took approximately 15 minutes. The duration of a creating a deliverable treatment plan from the predicted dose distribution was dependent of the mimicking settings and had a duration of approximately 75 minutes. The MLO method was thus able to create a deliverable plan in roughly 1.5 hours, offline time. An example of a clinical, predicted and mimicked dose distribution for one patient is shown in figure 6. It can be seen that the ML predicted and mimicked dose distributions were very similar to each other and also show very similar dose to the targets compared to the clinical plan, but increased dose to the spinal cord in the prediction and oral cavity in the mimicked plan.

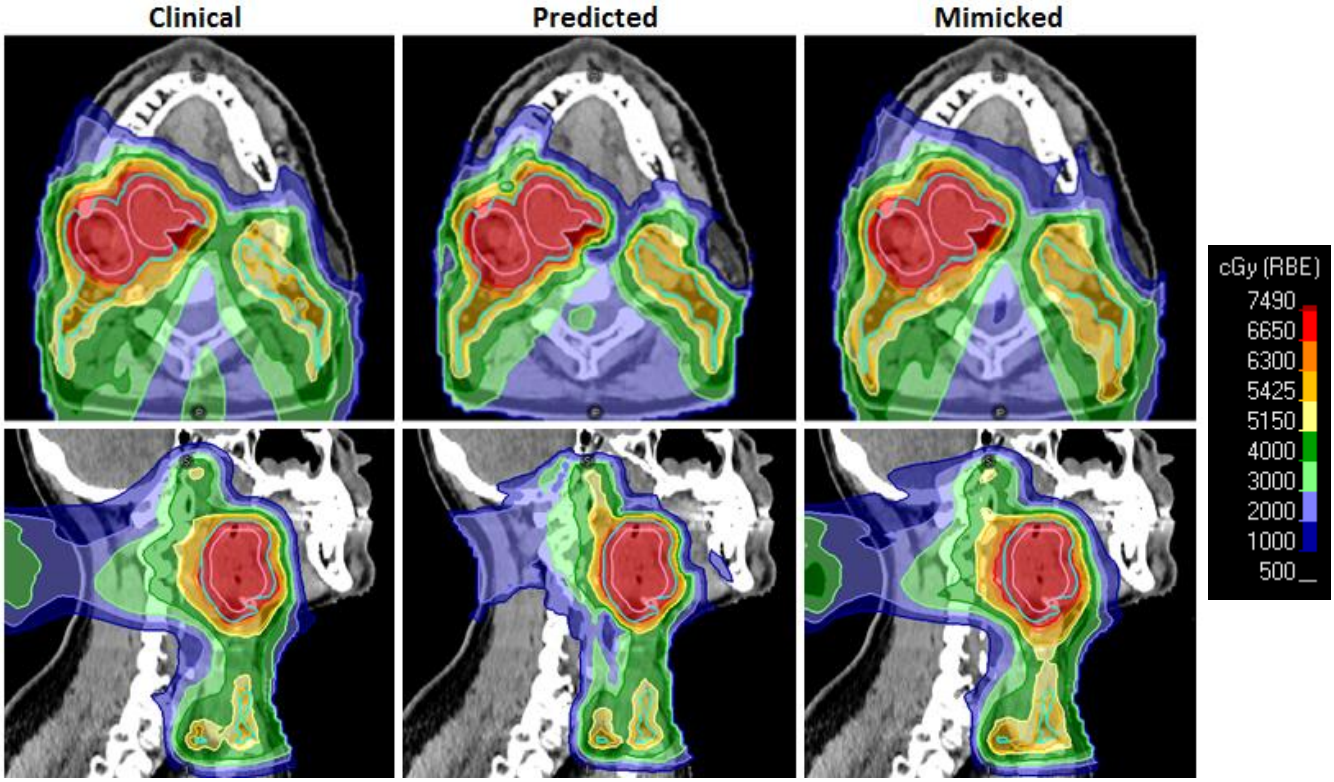


Figure 6 Transversal (row above) and sagittal (row below) images of the spatial dose distribution of a clinical deliverable plan, predicted ML plan and mimicked deliverable ML plan for one patient, respectively. Also, the CTV7000 (light pink) and CTV5425 (light blue) are shown.

Initial tuning

In figure 7, the resulting mean DVHs of the targets and several OARs from the initial tuning on 4 patients which has led to robust target coverage and of the according clinical plans are shown. Here, it can be seen that the DVHs of the CTVs of the MLO plans were nearly the same as the clinical plans. The average DVH of the CTV5425 of the MLO plans was higher compared to the clinical plans, with a D50 of 5752 cGy in the MLO and 5590 cGy in the clinical plans respectively. For the OARs it can be seen that the dose to the OARs in the MLO plans was lower than or very similar to the clinical plans.

A boxplot from the result of the robustness evaluation for these 4 patients is shown in figure 8. All MLO plans achieved robust target coverage since the D98 is above the threshold of 94% of the prescribed dose. In patient 1 and patient 3, for the CTV7000 it is striking that the scenario doses and vw-min were all above the scenario doses and vw-min of the clinical plan.

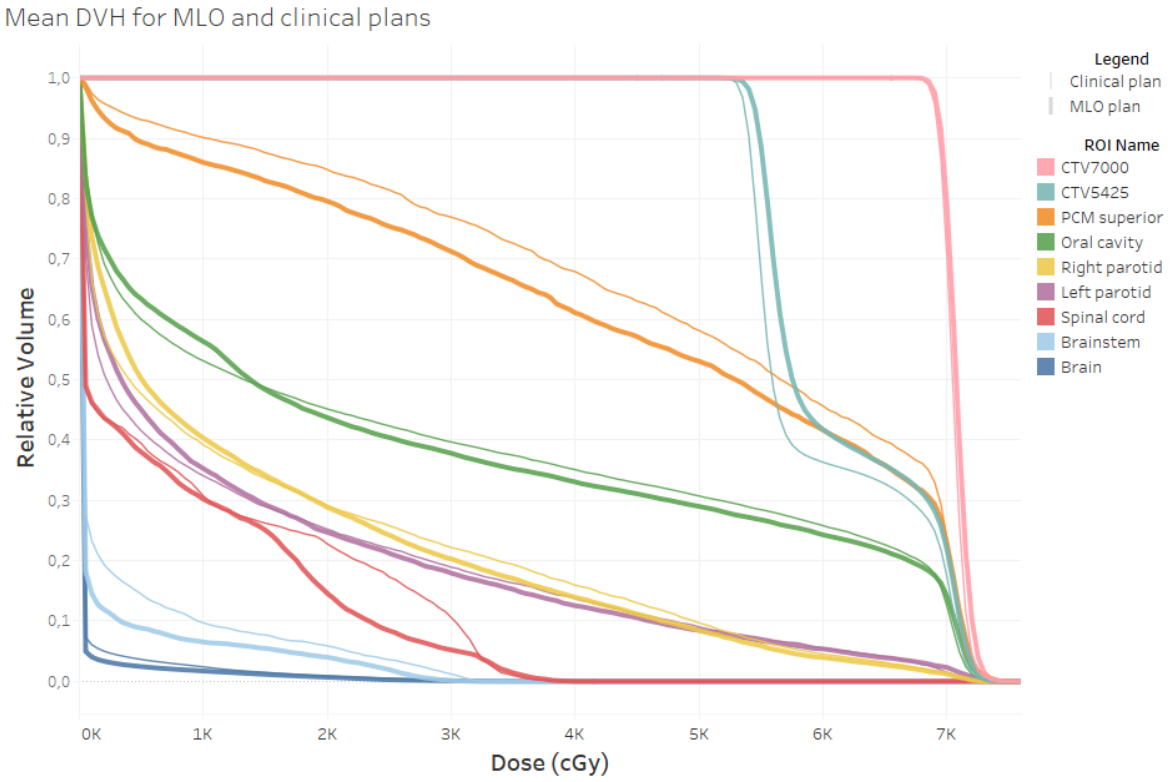


Figure 7 The mean DVH of 4 patients of the clinical plan and the MLO plan where initial tuning was performed on.

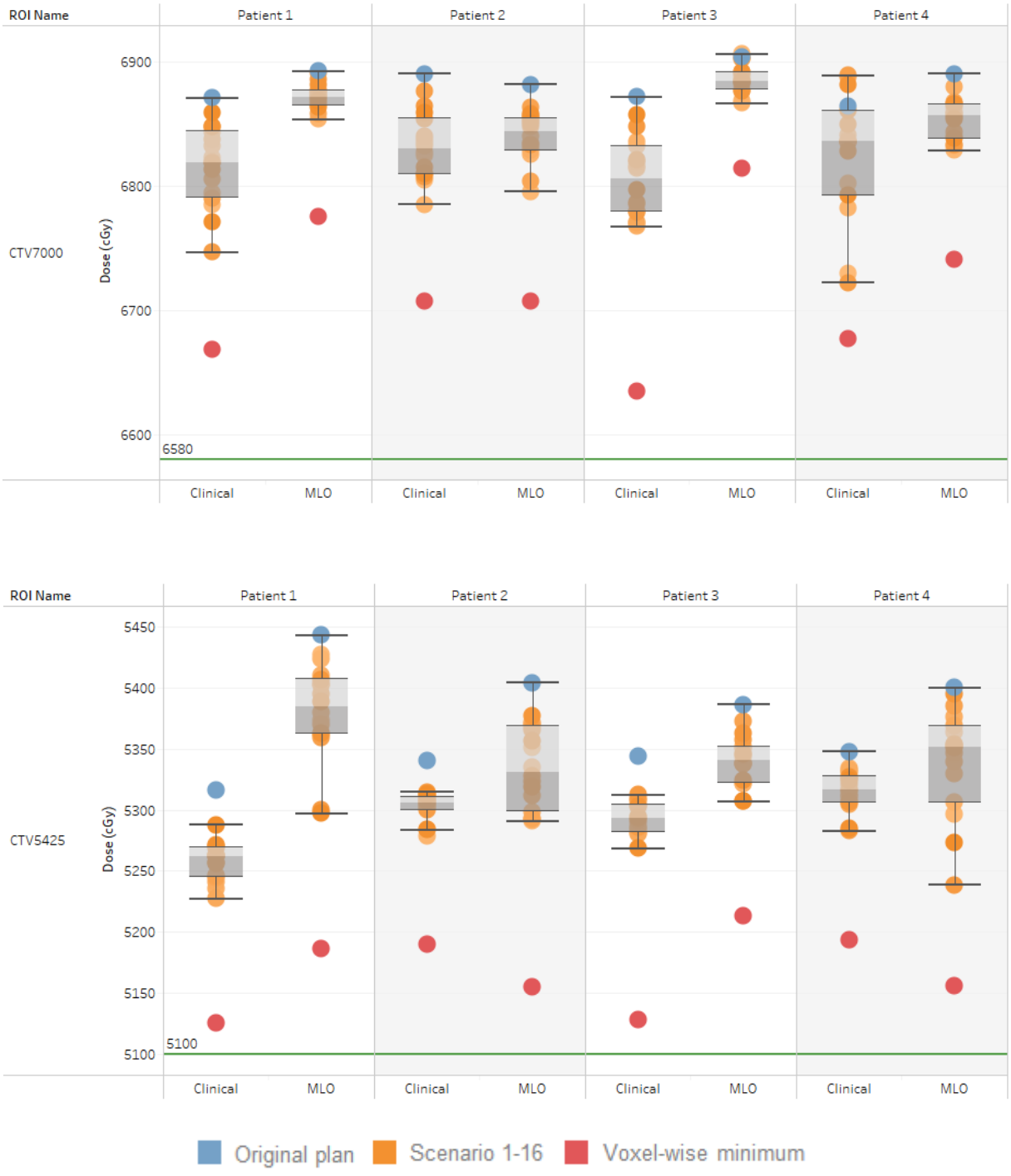


Figure 8 Boxplot of the 16 scenario doses from the robustness evaluation (orange), nominal dose (blue) and vw-min (red) of 4 initial tune patients of the CTV7000 (above) and CTV5425 (below). The green line indicates the threshold for 94% of the prescribed dose.

Cross-validation

To evaluate if the strategy used for the initial tuning would also perform well in another set of novel images, cross-validation was performed. In figure 9 the mean DVHs of the 4 cross-validation patients for the targets and several OARs from the MLO and according clinical plans are shown. It can be seen that on average the DVHs of the CTVs were nearly the same, but the DVH of the CTV5425 was somewhat higher in the MLO plans (D50 of 5709 cGy in the clinical plans and 5871 cGy in the MLO plans). The dose to the OARs was higher in the MLO plans compared to the clinical plans, except for the spinal cord which dose was on average lower in the MLO plans.

In figure 10, the result of the robustness evaluation is shown. It can be seen that for the CTV7000, the D98 of the vw-min in the MLO plans were higher compared to the clinical plans, and all were above the 94% threshold. Also, the distribution of the perturbed scenarios doses was smaller compared to the clinical plans. For the CTV5425, all plans were above the threshold of 94% of the prescribed dose.

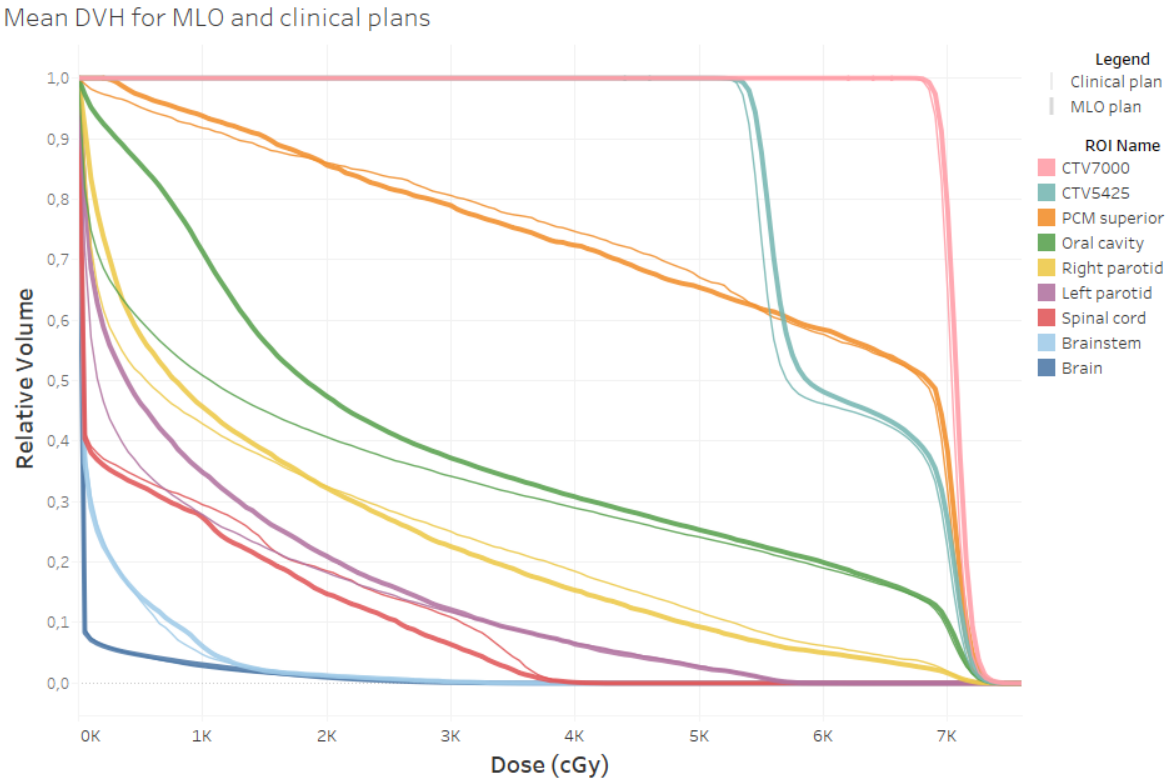


Figure 9 The mean DVH of 4 patients of the clinical plan and the MLO plan where cross-validation was performed on.

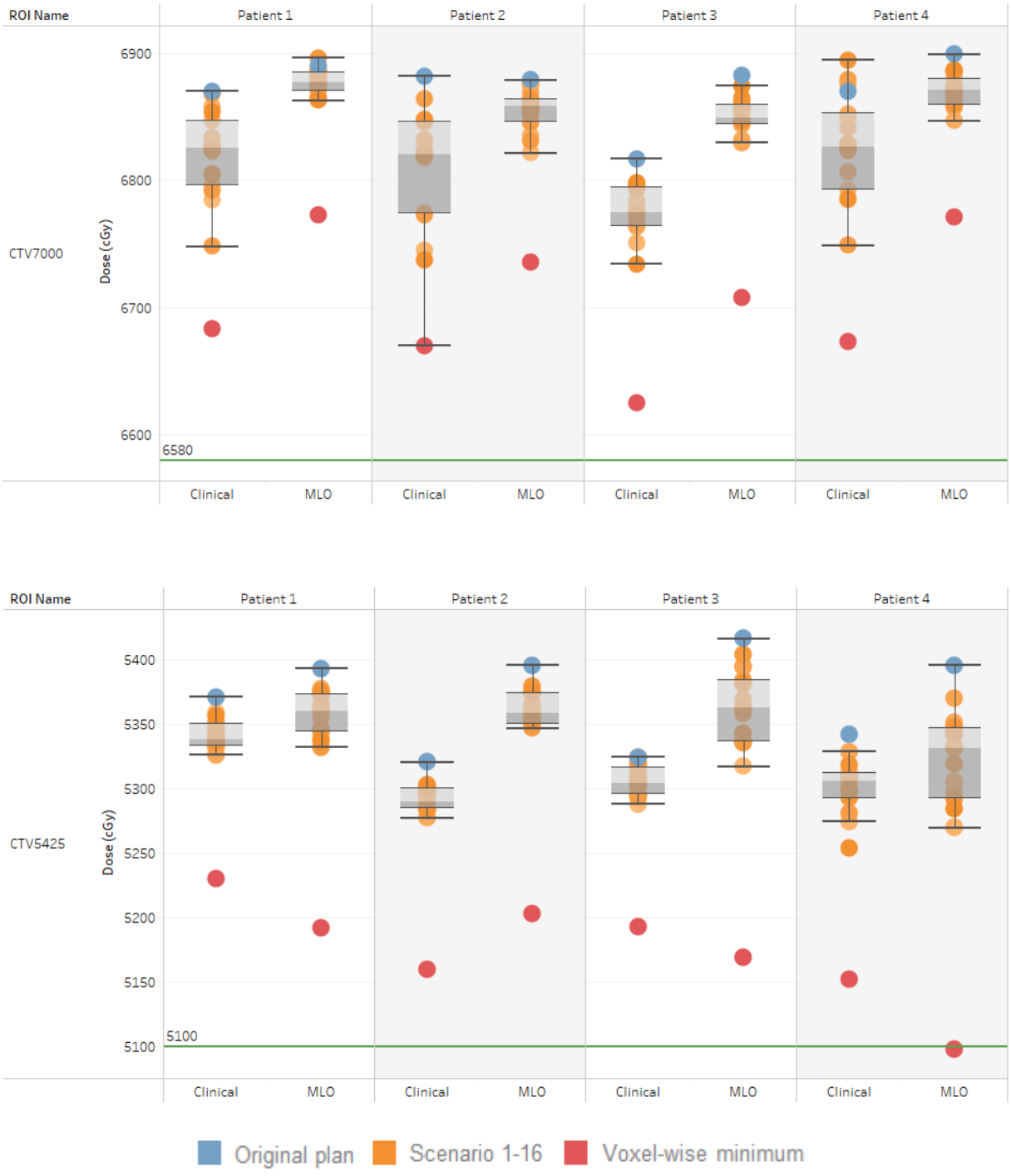


Figure 10 Boxplot of the 16 scenario doses from the robustness evaluation (orange), nominal dose (blue) and vw-min (red) of 4 cross-validation patients of the CTV7000 (above) and CTV5425 (below). The green line indicates the threshold for 94% of the prescribed dose.

Test set

Since the cross-validation test showed robust target coverage, this strategy was tested on an independent test set consisting of 9 treatment plans. In table 3, the results of D98 in the CTVs from the vw-min of the robustness evaluation are shown. Also, the target evaluation criteria CI and HI are shown, as well as the sum NTCPs for xerostomia and dysphagia. The Dmean and Dmax to the OARs can be found in Appendix A, table 10 and 11 respectively.

In the MLO plans, 8/9 generated plans fulfilled the D98>94% of the prescribed dose criterion in the vw-min. The vw-min D98 in CTV7000 in the MLO plans was on average 0.7 Gy higher in comparison to the clinical plans, this difference was statistically significant (p=0.015). Conformity in the PTV7000 was in the MLO plans statistically significant lower (p=0.013) compared to the clinical plan.

The sum NTCP for xerostomia was acceptable, i.e. was lower or did not increase >2% compared to the clinical plans, in 8 of the 9 mimicked MLO plans (89%) and for dysphagia in 5 of the 9 mimicked MLO plans (56%). The sum NTCP for dysphagia in the MLO plans was statistically significant higher as compared to the clinical plans. The Dmax constraints to the brain, brainstem, spinal cord and eyes were not exceeded in the MLO plans and no statistically significant differences were found compared to the clinical plans. The mean dose to the parotids, right submandibular and PCMs were on average comparable to the clinical plans, i.e. differences were <1 Gy. The average Dmean in the oral cavity was statistically significant higher (3 Gy) in the MLO plans compared to the clinical plans.

In figure 11, the mean DVHs of the 9 test patients is shown for the targets and several OARs. It can be seen that the DVH of the CTV7000 from the MLO plans was very similar to the clinical plans. For the CTV5425, the DVH of the MLO plan was somewhat higher compared to the clinical plan, e.g. a higher dose was given to a higher amount of volume in the MLO plans (D50 of 5775 cGy in the MLO plans and 5596 cGy in the clinical plans). For the OARs it can be seen that the dose was lower in the clinical plans compared to the MLO plans.

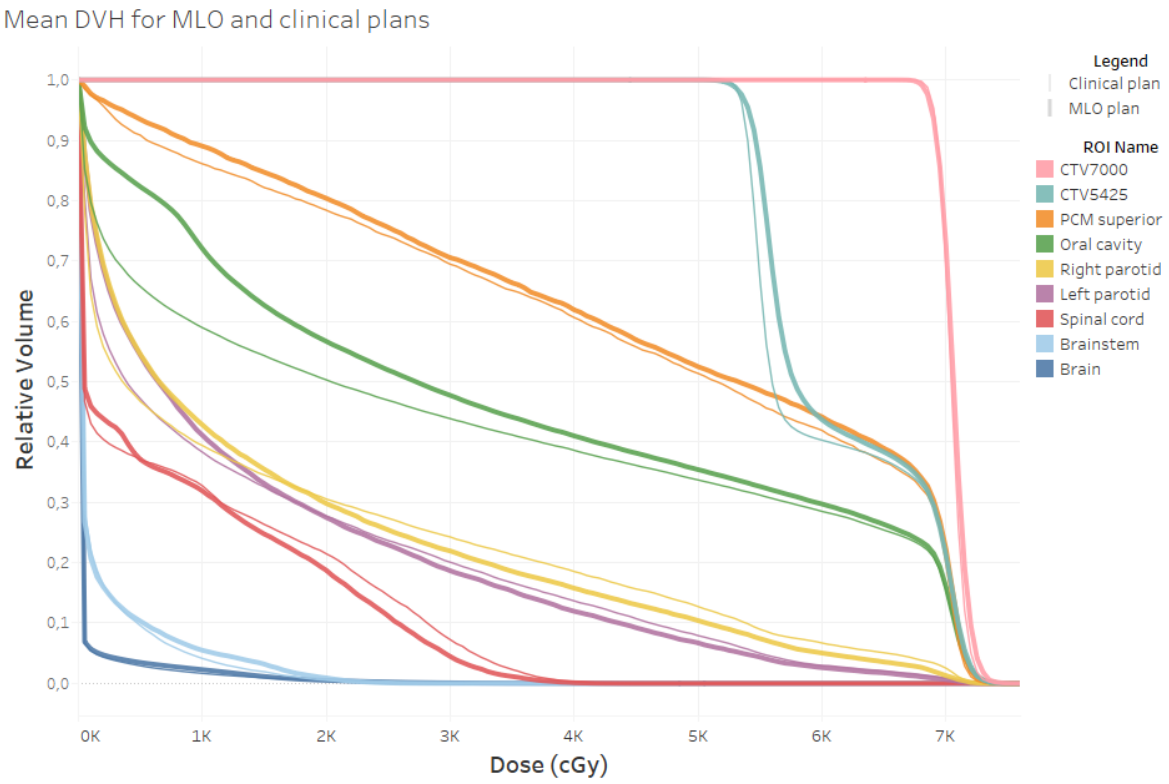


Figure 11 The mean DVH of 9 patients of the clinical plan and the MLO plan used for independent testing.

Table 3 The D98 in the voxel-wise minimum, conformity index and homogeneity index of the targets and sum NTCP of the clinical and MLO plans of the 9 independent test patients, and mean ± SD. The result of the Wilcoxon signed rank test is shown in the last row.

	D98 vw-min (cGy)				Conformity index				Homogeneity index				Sum NTCP (%)			
	Clinical plan		**MLO plan**		**Clinical plan**		**MLO plan**		**Clinical plan**		**MLO plan**		**Clinical plan**		**MLO plan**	
Test Patient	**CTV 7000**	**CTV 5425**	**CTV 7000**	**CTV 5425**	**PTV 7000**	**PTV 5425**	**PTV 7000**	**PTV 5425**	**PTV 7000**	**PTV 5425**	**PTV 7000**	**PTV 5425**	**Xero-stomia**	**Dys-phagia**	**Xero-stomia**	**Dys-phagia**
1	6668	5125	6644	5045^	0.819	0.801	0.800	0.749	0.076	0.351	0.071	0.351	49.7	14.9	46.0	14.7
2	6663	5135	6760	5155	0.804	0.824	0.776	0.822	0.071	0.357	0.067	0.365	47.1	19.3	44.5	24.5
3	6708	5204	6726	5194	0.739	0.692	0.748	0.708	0.080	0.362	0.084	0.367	37.1	5.1	37.9	5.5
4	6650	5153	6763	5128	0.651	0.717	0.642	0.803	0.065	0.349	0.069	0.374	32.0	14.4	32.2	16.9
5	6737	5160	6790	5128	0.733	0.871	0.683	0.825	0.074	0.364	0.071	0.362	39.0	5.4	38.5	6.1
6	6673	5251	6776	5136	0.756	0.832	0.692	0.857	0.072	0.355	0.070	0.374	60.9	12.9	63.6	14.8
7	6720	5147	6778	5236	0.776	0.769	0.739	0.675	0.068	0.358	0.068	0.338	56.7	12.8	52.2	12.3
8	6700	5189	6725	5263	0.857	0.942	0.830	0.928	0.086	0.321	0.075	0.296	70.8	38.3	70.5	43.5
9	6623	5177	6841	5160	0.827	0.925	0.790	0.868	0.092	0.355	0.077	0.348	49.4	18.4	46.5	22.8
Mean ± SD	**6682 ± 36**	**5171 ± 39**	**6756* ± 54**	**5161 ± 65**	**0.774 ± 0.062**	**0.819 ± 0.086**	**0.744* ± 0.062**	**0.804 ± 0.081**	**0.076 ± 0.009**	**0.352 ± 0.013**	**0.072 ± 0.005**	**0.353 ± 0.025**	**49.2 ± 12.3**	**15.7 ± 9.8**	**48.0 ± 12.4**	**17.9* ± 11.6**
p-value			0.015	0.515			0.013	0.374			0.159	0.944			0.173	0.028

***=statistically significant difference (p<0.05) compared to clinical plan**

^= lower than the threshold of 94% of the prescribed dose, but will be clinical acceptable

2.5 Discussion

In this study, we described a MLO method that automatically can generate robust IMPT plans for oropharyngeal patients with clinically acceptable robust target coverage, with reduced planning time up to 1.5 hours. As reflected in the test set, the dosimetric parameters between the MLO and manually generated clinical plans were overall comparable. No statistically significant differences were observed between the MLO plans and clinical plans, except for the D98 vw-min in the CTV7000 and oral cavity mean dose. The dose constraints to the serial organs were never exceeded in the MLO generated plans and were comparable to the clinical plans.

The mean DVH of the CTV7000 in the nominal MLO plan was very similar to the clinical plan. However, the vw-min D98 was statistically significant higher as compared to the clinical plans. This can be explained by the smaller distribution of the scenario doses in the MLO plans, as was seen in the validation set. The statistically significant lower CI of this target in the MLO plans compared to the clinical plans suggests that a higher V95 resulted in a more robust CTV7000.

In the nominal plan, the dose to the CTV5425 was higher compared to the clinical plans, as can be seen in the mean DVHs, but did not always result in a higher D98 in the vw-min. An explanation for this is that the robustness evaluations showed a wider distribution range of the scenario doses, as shown in the boxplots of the robustness evaluation of the initial- and cross-validation plans. However, when lower doses were achieved in the nominal plan, this may result in target coverage below the minimal threshold for robust plans.

The sum NTCP for dysphagia showed a statistically significant higher dose in the MLO plans compared to the clinical plans. This is in accordance with the statistically significant higher mean dose to the oral cavity, which has an essential contribution to the NTCPs for dysphagia. Therefore, the dose to the oral cavity has to be reduced further in the MLO plans to reach clinical plan quality. This was as expected since most time investment during this study was given to target coverage robustness, and then some first attempts to reduce the dose to the OARs were performed.

In the four initial tune patients, the dose to the OARs in the MLO plans was lower or very similar to the clinical plans. However, the cross-validation and independent test showed increased mean dose compared to the initial tuning, for example in the oral cavity. Also, the mean dose to the PCM superior was lower compared to the clinical plans in the validation set, but in the test set this was not the case anymore. Our MLO approach is based on the method used in the study of McIntosh, which is based on the assumption that patient with identical geometry and appearance should be treated in the same way.[13][19][20] However, the mentioned differences between validation and test set indicates that it may be necessary to develop different strategies for specific patient(groups) regarding sparing the dose to the OARs.

To our knowledge, this was the first study which combined a robust optimization method with a machine learning optimization to automatically generate deliverable IMPT plans. Despite the limited institutional amount of available IMPT plans for oropharyngeal cancer, we were able to train a model and could perform initial- and cross-validation and testing subsequent on an independent test set. In addition, a strength of our study is that a voxel-based prediction and mimicking method was applied, which means that no spatial information was lost, which would be the case in DVH-based approaches.

Delaney et al. have investigated also an automatic planning method for IMPT plans, which resulted in comparable plan quality to clinical plans. However, in contrast to our study, they performed a non-robust DVH-based optimization method.[52] Later, they extended their approach with a robust optimization method, and have shown preliminary potential of automated robustly optimized IMPT head-neck plans.[53] However, this was only tested in three head-neck plans and comprised a DVH-based prediction and mimicking. In accordance with our study, they also showed automated plans with comparable OAR doses as in the clinical plans, except for the oral cavity where increased mean dose was found.

In our study, it was necessary to tune settings in the prediction and mimicking part to achieve MLO plans with adequate robust target coverage and comparable plan quality to clinical plans. Bijman et al. used an automatic planning approach to investigate differences between model selection for IMPT in oropharyngeal cancer based on manually or automatically planned IMPT.[54] In the study of Hennings et al., an automatic tool to pre-calculate planning solutions was developed for uveal melanomas.[55] In both studies such tuning as used in our study was incorporated by a predefined list of constraints and objectives. While in the study of Delaney et al., giving additional objectives was not necessary to achieve adequate plan quality.

In our study, the field set up of the clinical plans was copied and used to perform mimicking of the MLO prediction. However, an automated robust beam orientation optimization method of Gu et al. has shown better target coverage under simulated uncertainties while reduced dose to the OARs compared to plans with manual beam selection.[56] In future research, an automatic robust beam orientation optimization method could be combined with a MLO method to compare alternative beam arrangements.

It is often seen that even each physician has its own preferences regarding the evaluation of an IMPT plan, which makes it difficult to have one MLO strategy which will be accepted by all the assessing doctors. A study with knowledge based planning in pancreatic cancer has shown that training separate models for each dedicated physician resulted in the dose distribution desired by a given physician.[57] Also, the implementation in other radiotherapy institutes has to be investigated, since they have different planning aims. Hopefully, with different strategies MLO is also able to generate IMPT plans with adequate clinical plan quality in other centers.

The automated MLO method in our study is a promising method which has numerous potential advantages in clinical practice. It is able to generate robust IMPT dose distributions with reduced workload. When our method is further developed, and consistent plan quality is achieved, MLO will be suitable for fast radiotherapy planning in daily practice. Next to this, MLO can be used in photon to proton plan comparison, to automatically select which patients are suitable for proton therapy based on NTCPs. Also, in adaptive radiotherapy, where fast planning is required to not delay any adaptations to the day-to-day variation inaccuracies seen in the dose distribution, MLO will have great potency, since manual plan adaptations are time consuming.

2.6 Conclusion

The MLO method presented in this study is able to automatically generate robust IMPT plans for oropharyngeal cancer patients with higher target coverage robustness in the high CTV and similar target coverage robustness in the low CTV compared to the clinical plans. Furthermore, the MLO method is also able to achieve OAR doses similar to the OAR doses in the clinical plans. Further reduction of the dose to specific OARs, i.e. oral cavity, and improvement of the plan consistency between patients is part of future investigation.

Chapter 3 Development

3.1 Introduction

The results described in Chapter 2 were the final product of this thesis. Before these results were achieved, different development steps were involved. The main goal of these development steps was to achieve robust target coverage. When this goal was achieved, also some first attempts to reduce the dose to the organs at risk (OARs), while maintaining robust target coverage were performed.

The resulting machine learning optimization (MLO) dose distribution is highly dependent of the various steps prior in the training, prediction and mimicking phase. Adjustments to the steps prior will lead to different outcomes and can be made in four different manners. First, the trained model can be changed by making adjustments in the training settings, such as on which target regions of interest (ROIs) is trained. Also, it is possible to change the model size, e.g. train on a larger set of input data. Next to this, the settings in de prediction and mimicking optimization can be adjusted to create a new strategy. Furthermore, it is possible to make small changes in in the MLO function by change of some functions in the scripting environment. And finally, RaySearch can add whole new functionalities by updating a newer version of the RayStation treatment planning system.

During this study, various changes in different ways are made, which resulted in 6 different models, and a total of 70 strategies which are used to find a strategy which fulfills the plan requirements. This chapter will elaborate on the main steps involved (summarized in 15 strategies) which have led to the final results described in Chapter 2. First the requirements for a clinical acceptable plan will be described in section 3.2. After that, the parameters which could be adjusted are described in section 3.3.

In section 3.4, a selection of the main strategies per model are shown and discussed. Strategies ‘Standard’ and ‘Strategy 1’ were prediction strategies, and therefore were evaluated on the predicted dose distribution. Since ‘Strategy 2’, the mimicking functionality in the RayStation treatment planning system was available. From ‘Strategy 2’ till ‘Strategy 14’ tuning was focused on achieving robust target coverage. Until ‘Strategy 8’, target coverage evaluation was done on the planning target volumes (PTVs). From Strategy 8 until Strategy 15, target robustness evaluation was performed in the clinical target volumes (CTVs). Finally, in ‘Strategy 15’, a first attempt to reduce the dose to the OARs was performed. At the end of this chapter, a general discussion and recommendations for further improvement are given.

3.2 Requirements

The first goal in this study was to achieve enough target coverage which is robust for a certain range- and setup uncertainty. For the mimicked plans, this was evaluated by the voxel wise minimum (vw-min) for each plan, where the D98 in the CTVs has to be at least 94% of the clinically prescribed dose. Since robustness evaluation can only be performed on a mimicked plan, for the prediction, target coverage was evaluated by the D98>95% in the PTV, i.e. 6650 cGy for the PTV7000 and 5154 cGy for the PTV5425. For the mimicked plans, when a robustness evaluation is performed, the targets must receive at least 94% of the prescribed dose in the D98 of the vw-min, see table 4.

Table 4 Clinical goals for different regions of interest (ROIs)

ROI	Clinical goal (cGy)
PTV7000*	D98 ≥ 6650
PTV5425*	D98 ≥ 5154
CTV7000	D98 vw-min ≥ 6580
CTV5425	D98 vw-min ≥ 5100
Spinal cord	Dmax ≤ 5425
Parotid glands	Dmean = ALARA^
Submandibular glands	Dmean = ALARA^
Brain	Dmax ≤ 6000
Brainstem	Dmax ≤ 6300
Eyes	Dmax ≤ 500

***PTV goals are not applied in robust IMPT planning, but is in this development part used as initial evaluation criteria.**
^ As low as reasonably achievable

The second goal in this study was to reduce the dose to the OARs, while maintaining robust target coverage. An overview of the constraints to these OARs are also given in table 4 and has to be as low as reasonably achievable. Dose to the other normal tissue complication probability (NTCP) related organs with no specific constraints (oral cavity, PCMs) had the objective to be as low as reasonably achievable.

3.3 Parameters

A short explanation of the main prediction and mimicking settings which were adjusted during this study is given in this section. The functionalities are shown in italics, but not all options were available yet since the start of this study. An overview of these settings and when these settings became available is given in Appendix C, table 13. An example of a settings file is shown in Appendix B, where the final settings are presented.

For the prediction, with *EnsembleSize* the number of atlas regression forests (ARFs) could be selected. In *ROIGoals,* it was possible to give goals to steer to predicted dose for specific ROIs, such as a minimal dose to a certain volume within that ROI. Within this setting, different sub-settings could be adjusted, i.e. choose minimal or maximal dose, to which ROI it concerns, and which dose has to be achieved by a chosen percentage of volume.

In *ReduceOAR*, when stated to *true*, the function would reduce the dose to all OARs, or per separate OAR, with a certain reduction level between 1-10. With *PriorMethod*, the predicted dose would adhere to the overall dose of the chosen ensemble plans. When “avg” was set, the dose would be similar to the average dose of the ensemble patients in that ROI. If “min” or “max” was set, the dose would adhere to the minimum or maximum dose level in that ROI of the ensemble plans respectively. The same setting was applicable for areas with overlap between target and OAR with function *PriorMethodInsideTarget*.[36][49]

In the mimicking settings, with *VoxelCTVMim* could be chosen for a voxel-based mimicking of the CTVs when stated to “true” or for a dose volume histogram (DVH)-based mimicking optimization when stated “false”. The same setting was available for the OARs; *VoxelOARMim*. In *RobustMimickCTVs* the CTVs which have to be robustly optimized had to be listed. Ideally, with additional minimal and maximal dose goals, but unfortunately this functionality was not yet

available. With *AddTargetsNtimes* it was possible to add the CTVs to the optimization function a certain number of times. In *Strategy_ROIs* it was possible to create new ROIs, for example by the expansion or subtraction of two ROIs. The *NonRobustMimickOARs* had to be a list with all the OARs which did not have to be robustly optimized. Within this function, the objective of a respective OAR could be added a certain number of times in the optimization function with *add_ntimes*. With *reduce_dose* stated on true the dose would be further reduced if possible, and when stated on false, the predicted dose would be maintained. At the end of the mimicker optimization function, the *RobustnessParameters* had to be defined, with position and density uncertainty parameters.[36][49]

3.4 Iterations of adjustments

In this part, the different adjustments which were involved to achieve the final model and settings are described. An overview of the different trained models used during these steps can be found in Appendix C, table 12. Furthermore, an overview of the different strategies and their according settings are given in Appendix C, table 14.

3.4.1 Tune prediction: Model-60

At the start of this study, no mimicking functionality for protons was available in RayStation development version 8b. Therefore, the first goal was generating predicted plans which fulfill the target coverage criteria. A model of 60 patients, with 40 clinical planned patients and 20 plans which were planned for research purposes was trained: Model-60. This model was trained based on PTV target structures, in order to already have a margin around the CTV in the prediction, and to be able to evaluate PTV coverage in the predicted plan. See Appendix C, table 12 for further characteristics of the included patients in this trained model.

Standard Strategy

Initial testing was performed on 5 patients. First, a strategy for prediction was tried without any additional settings to see if and where tuning in the settings is necessary. In table 5, the D98 of this Standard MLO strategy is given. It can be seen that the D98 requirements for both PTVs were not fulfilled for this strategy, and the sum of the NTCP for xerostomia and dysphagia was higher compared to the clinical plan. Therefore, different *ROIGoals* settings for the PTVs in the prediction were added.

Strategy 1

Also, it was investigated what the impact was of the *ReduceOar* functionality, and the function *ReduceOar* was set to true in order to reduce the sum NTCP. The results of the D98 and SUM NTCP of these adjustments, Strategy 1, are shown in Table 6. It seemed promising that in the prediction it was relatively easy to fulfill the target coverage requirements by adding *ROIGoals* for the PTVs and to reduce the NTCP by set *ReduceOar* to true.

3.4.2 Tune mimicking: Model-60

With the update to RayStation development version 9A in September 2019, the MLO mimicking functionality for protons became available. In this version of the scripting environment (2.2), the dose to the OARs would be reduced in the mimicker by default. The first goal of testing this mimicker functionality, was to achieve robust target coverage. To have a quick initial evaluation method, target coverage of each plan was accessed by the D98 criteria in the PTVs, until this goal was fulfilled.

Strategy 2

Initial testing was performed on 5 plans which were clinical robustly optimized for 0.5 cm position- and 3% density uncertainty. The same prediction settings as for the 'Strategy 1' were applied, but now also ROI goals for the CTV were added in order to be included in the robust mimicker optimization. In the mimicking settings, DVH based mimicking was used as default. Robustness uncertainty settings according to the clinical plans were used.

In table 5, the result to the target coverage of this strategy ('Strategy 2') is shown and in table 6, the sum NTCP can be found. It can be seen that the MLO mimicked plans did not fulfill the target coverage criteria in all five patients. Unfortunately, it was found out that setting 'weights' for a specific 'dose_level' in ROI goals for the targets in the mimicker had zero effect; this functionality was not available in RayStation version 9A. So, there were less options in the mimicker than in the prediction settings to tune to preserve enough target coverage in the mimicked plans.

Strategy 3, Scripting environment version 2.3

Because of the low target coverage in the MLO mimicked plans, the scripting environment functionality was adjusted. In the previous version (2.2), the dose to the OARs was reduced in the MLO mimicker by default. Now, this functionality was adjusted to maintain the predicted dose to the OARs. Also, in the prediction, the function *ReduceOar* was set to false, since the first priority was to achieve enough target coverage. After this has been achieved in a later stadium, we could focus on reducing the dose to the OARs. Also, in the prediction settings, only CTVs were included instead of both CTVs and PTV, because in manual intensity modulated proton therapy (IMPT) planning there is optimized on CTVs and the intention is that they will be robustly optimized by the mimicker. This has led to 'Strategy 3', and the D98 to the PTVs of this strategy are shown in table 5.

On average, the predicted D98 dose to the PTV7000 was slightly increased and the D98 to the PTV5425 was slightly reduced by excluding the PTVs from the ROI goals and setting the function *ReduceOar* to false. With excluding the PTVs from the prediction settings, as in 'Strategy 3', the D98 in the PTVs still got enough target coverage. Despite the lower predicted dose in the PTV5425, the target coverage in the MLO mimicked plans was improved. However, for 4 out the 5 patients the PTV7000 did not achieve enough target coverage. Also, for the PTV5425, not enough target coverage was achieved. In table 6 the result of the sum of the NTCPs is shown. It can be seen that with the new functionality in the scripting environment, the sum NTCP was increased. This was as expected since now the predicted dose to the OARs was set to maintain instead of further reduce.

Table 5 The resulting target coverage per patient of the different strategies compared to the predicted and clinical dose.

		Clinical plan	Standard MLO	Strategy 1 MLO	Strategy 2 MLO		Strategy 3 MLO	
	D98 (cGy)		Predicted	Predicted	Predicted	Mimicked	Predicted	Mimicked
Patient 1	PTV7000	6804	6560	6673	6694	6123	6739	6637
	PTV5425	5260	5168	5234	5236	4548	5219	5323
Patient 2	PTV7000	6390	6530	6625	6699	5869	6716	6577
	PTV5425	5089	5164	5236	5234	4718	5237	5146
Patient 3	PTV7000	6593	6549	6608	6690	6328	6643	6592
	PTV5425	5113	4983	5158	5197	4570	4943	4837
Patient 4	PTV7000	6662	6676	6678	6701	6116	6763	6594
	PTV5425	5240	5122	5141	5221	4749	5162	5194
Patient 5	PTV7000	6655	6599	6623	6714	6317	6722	6641
	PTV5425	5091	5082	5178	5244	4593	4884	4994
Mean ± SD	**PTV7000**	**6621 ± 150**	**6583 ± 58**	**6641 ± 32**	**6700 ± 9**	**6151 ± 187**	**6716 ± 45**	**6608 ± 29**
Mean ± SD	**PTV5425**	**5159 ± 84**	**5104 ± 76**	**5189 ± 44**	**5222 ± 16**	**4636 ± 92**	**5089 ± 164**	**5099 ± 188**

Table 6 The resulting sum NTCP of per patient of the different strategies compared to the predicted and clinical plan.

Sum NTCP (%)	Clinical plan	Standard MLO	Strategy 1 MLO	Strategy 2 MLO		Strategy 3 MLO	
		Predicted	Predicted	Predicted	Mimicked	Predicted	Mimicked
Patient 1	97.6	94.3	79.5	81.2	78.5	105.2	123.6
Patient 2	102.3	97.4	80.4	83.7	75.5	109.0	119.1
Patient 3	66.6	75.5	63.3	64.0	62.3	89.1	84.8
Patient 4	88.2	92.5	78.6	80.3	78.0	104.0	96.2
Patient 5	54.9	59.5	47.0	48.4	47.7	76.6	73.0
Mean ± SD	**81.9 ± 20.4**	**83.9 ± 16.1**	**69.7 ± 14.5**	**71.0 ± 15.1**	**67.9 ± 13.3**	**96.0 ± 12.2**	**98.3 ± 21.8**

Strategy 3; AddTargetsNtimes

As a next step, it was tried to include the targets several times in the optimizer. However, when we included the function *add_ntimes* in the *RobustMimickCTVs*, no effect was seen in the mimicked plans, and it worked out that this functionality only could be applied to the *NonRobustMimickOars*. Therefore, a function which could add the targets multiple times to the optimizer, *AddTargetsNtimes*, was created. However, if this was applied to the CTVs with their minimal and maximal dose as well, this probably will not improve the minimal dose since the minimal and maximal dose will counteract each other. Therefore, a copy of the CTVs was made and only their minimal goals were included. Then, *AddTargetsNtimes* was applied to the minimal dose to the CTVs, repetitively with increasing number of times when using Strategy 3. Values between 1 and 10000 were tried. It was investigated whether there was a number of times the CTVs had to be included to achieve enough target coverage, evaluated by the D98 in the PTV7000 and PTV5425.

The time required to complete the MLO plans was increased by the number of times the CTVs were included in the optimizer, e.g. it took several hours to generate a plan with the CTVs added 250 times. For adding the CTVs 10.000 times, generating one plan was not finished within a day, and therefore the MLO was stopped manually. In figure 12, the D98 in the PTV7000 (above) and PTV5425 (below) of each patient and the average is shown. It can be seen that for the PTV7000 the D98 was increased by adding this CTV multiple times. Also, the distribution of the D98 in this PTV became smaller and was comparable to the clinical plan when this CTV is added ≥100 times.

For the PTV5425, adding the CTV multiple times did not have as much effect on the D98 as for the PTV7000; the differences were all <0.5 Gy. Adding the CTVs more than 100 times, led to a larger distribution of the D98, which was not as expected. Also, on average, a slightly increasing effect on the target coverage was seen when the target was added ≥100 times, but then the time to finish a plan was increased. So, we decided that when setting the *addTargetsNtimes* to 100, the agreement between time and target coverage was acceptable.

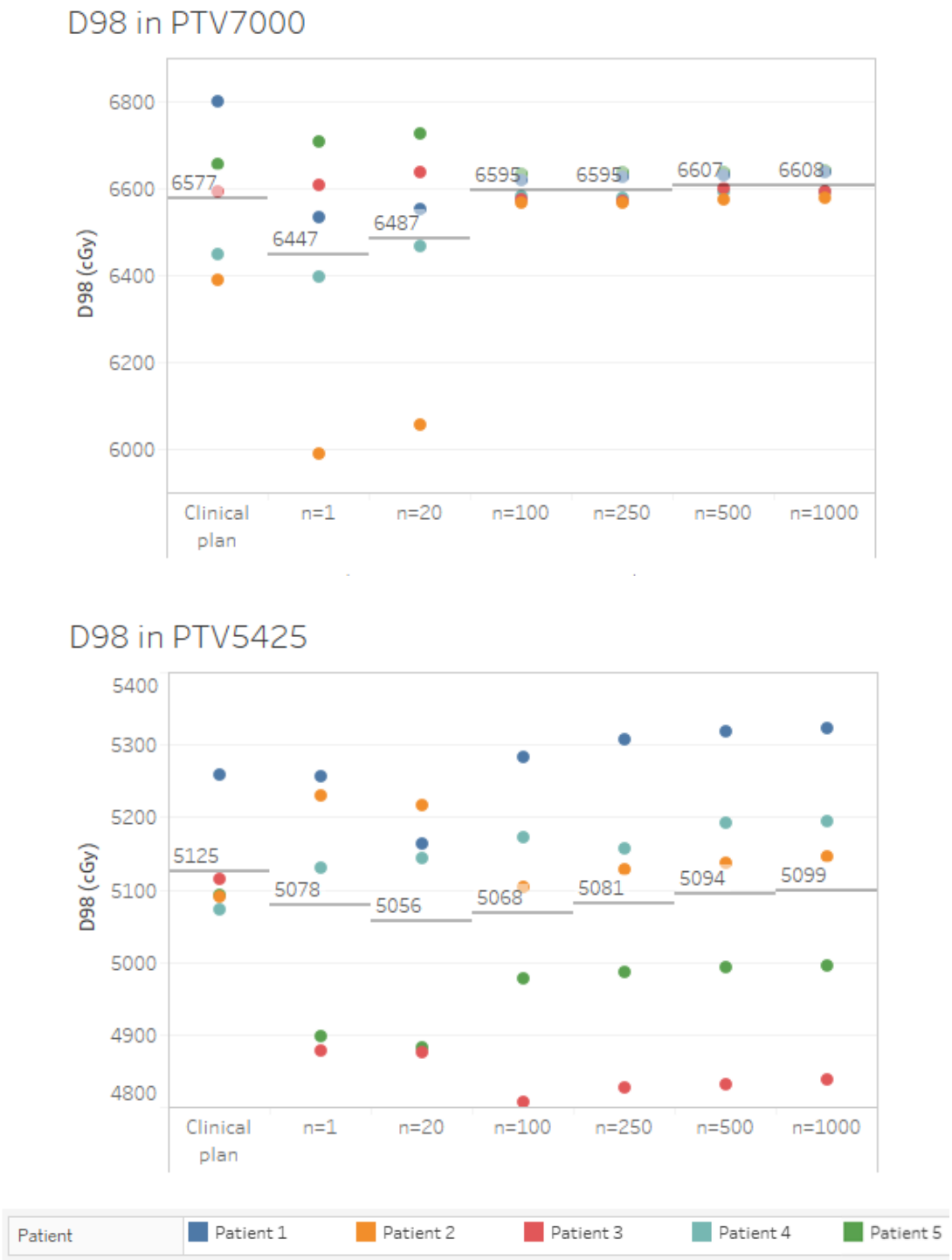


Figure 12 D98 in the PTV7000 (above) and PTV5425 (below) per patient of strategies when adding the CTVs different times compared to the clinical plan. The grey lines and displayed value represent the average D98 of the 5 patients.

Strategy 4 and 5; DVH- vs voxel-based mimicking

Next to this, also the effect of DVH-based mimicking instead of voxel-based mimicking on the target coverage was investigated. This was performed on the strategy where he CTVs were added 500 times in the optimizer. In table 8, the results of both mimicking strategies can be found, where 'Strategy 4' represents voxel-based mimicking and in 'Strategy 5' the mimicker is based on DVHs. It can be seen that on average, the DVH-based mimicking D98 in the PTV7000 and PTV5425 was slightly higher, but this difference is almost negligible since the difference is <0.5 Gy. In patient 3 a drop down was seen in the D98 for PTV7000 of nearly 3 Gy, which resulted in a non-acceptable target coverage. Therefore, at this point it was chosen to continue with voxel-based mimicking.

Strategy 6 and 7; robustness uncertainty settings

Furthermore, it was investigated if increasing the robustness settings also would lead to a higher D98 in the PTVs. Therefore, 'Strategy 7' with robustness position uncertainty of 0.8 was compared to 'Strategy 6' with a position uncertainty of 0.5, both strategies were planned with the CTVs included 1000 times. It was expected that this adjustment would improve the mimicked dose to at least the 95% criterion of the prescribed dose. The results are shown in table 8, where it can be seen that on average the D98 in the PTVs did improve, but not as much as expected. In the PTV5425 the 95% criterion was achieved on average, but in the PTV7000 this was not the case.

Since the 'Strategy 3' was applied, it can be seen that for patient 3 and 5 the predicted and mimicked dose was lower compared to the clinical plan. Therefore, a strategy with a higher predicted dose would maybe result in a higher mimicked dose, so this was investigated in a next step (Strategy 8).

3.4.3 Mixed vs. homogenous model

Plans included in the training which were made in 2018, were robustly optimized with 0.5 cm position uncertainty and 3% density uncertainty, and differed from plans made since the beginning of 2019, where 0.3 cm position uncertainty was taken into consideration.

Until December 2019, a total of 35 dose distributions planned with 0.3 cm position uncertainty were available. To investigate whether there was a difference in performance regarding target coverage and sum NTCP due to the different robustness settings in the trained models, a mixed model (both 0.3 and 0.5 cm position uncertainty optimized plans included) and a homogeneous model (only 0.3 cm position uncertainty optimized plans included) were compared. Therefore, a separate model with exclusively plans which were robustly optimized for 0.3 cm position uncertainty was trained: Model-0.3. This model was trained with 32 plans and subsequently tested on 3 novel patients, see Appendix C, table 12. The same 3 patients were also tested on Model-66 with the same settings. An ensemble size of 3 ARFs was used for both predictions. To assess target coverage, robustness evaluation was performed with 0.3 cm set-up and ±3 % range uncertainty and assessed by the vw-min D98>94% criterion.

In table 7, the results of the D98 in the vw-min in the CTVs are shown. It can be seen that for the three patients tested on, two of them showed robust D98 in the CTVs in both the mixed as well as the homogeneous model. On average, the homogeneous model showed a higher D98 in the vw-min, of approximately 0.2 Gy for the CTV5425 and 0.3 Gy for the CTV7000. Next to this, the sum NTCP between the two models was similar. Since the homogeneous model had very small effect on the target coverage and sum NTCP, we have chosen to continue with the mixed model because this consists of more patients.

Table 7 The result of the voxel-wise minimum in D98 in the targets per patient of the mixed and homogeneous model. Also, the sum NTCP is shown.

	Mixed model			Homogeneous model		
	Vw-min D98 CTV5425 (cGy)	**Vw-min D98 CTV7000 (cGy)**	**Sum NTCP (%)**	**Vw-min D98 CTV5425 (cGy)**	**Vw-min D98 CTV7000 (cGy)**	**Sum NTCP (%)**
Patient 1	5161	6685	60.4	5198	6761	58.6
Patient 2	5127	6559*	57.9	5136	6564*	58.7
Patient 3	5239	6775	63.9	5259	6785	65.1
Mean ± SD	**5176 ± 47**	**6673 ± 89**	**60.7 ± 3.0**	**5198 ± 50**	**6703 ± 99**	**60.8 ± 3.7**

*** Below the threshold of 94% of the prescribed dose**

Table 8 The D98 in the PTVs per patient of the different strategies compared to the prediction and clinical plan.

		Clinical plan	**Strategies 4-7 MLO**	**Strategy 4 MLO**	**Strategy 5 MLO**	**Strategy 6 MLO**	**Strategy 7 MLO**	**Strategy 8 MLO**	
	D98 (cGy)		Predicted	Mimicked	Mimicked	Mimicked	Mimicked	Predicted	Mimicked
Patient 1	PTV7000	6804	6736	6631	6675	6639	6653	6767	6694
	PTV5425	5260	5259	5319	5350	5323	5351	5001	5062
Patient 2	PTV7000	6390	6716	6576	6285	6577	6596	6750	6690
	PTV5425	5089	5237	5136	5266	5146	5236	5240	5292
Patient 3	PTV7000	6593	6643	6600	6696	6592	6649	6801	6640
	PTV5425	5113	4943	4832	4949	4837	4934	5176	5069
Patient 4	PTV7000	6622	6763	6592	6641	6534	6649	6808	6716
	PTV5425	5240	5162	5191	5192	5194	5246	5213	5083
Patient 5	PTV7000	6655	6722	6638	6789	6641	6663	6731	6634
	PTV5425	5092	4884	4992	4899	4994	5058	5022	4847
Mean ± SD	**PTV7000**	**6621 ± 150**	**6716 ± 45**	**6607 ± 26**	**6617 ± 194**	**6597 ± 45**	**6642 ± 26**	**6771 ± 33**	**6675 ± 36**
Mean ± SD	**PTV5425**	**5159 ± 84**	**5089 ± 164**	**5094 ± 188**	**5131 ± 198**	**5099 ± 188**	**5165 ± 167**	**5130 ± 111**	**5071 ± 158**

3.4.4 Tune mimicking: Model-40 and Model-40.1

The research patients included in Model-60 had plans which were not clinically reviewed. In the test patients, it was noticed that the research plans were planned in a different way than conventionally in clinical patients. The goal of this study was to generate plans with clinical accepted plan quality. Therefore, we continued with a model which consists of clinically accepted plans, and thus excluded the plans which were planned for research purposes. For model characteristics, see table 12, Appendix C.

This model was tested again on 5 patients, but not the same patients where previous tests were performed on, due to some test patients were research patients. All 5 patients were clinical robustly optimized for 0.5 cm position uncertainty and 3% density uncertainty, so these settings were also used in the robust mimicking optimization. Evaluation of the MLO plans was done by assessment of the D98 in PTV7000 and PTV5425 until the goal of 95% of the prescribed dose was achieved. From that point, we continued by assessing the robustness by the D98 of the vw-min in the CTV7000 and CTV5425.

Strategy 8

Testing this model was started with a new strategy; 'Strategy 8', where higher ROI goals were given to the CTV5425 compared to the 'Strategy 3'. With Strategy 8, for the PTV7000 the D98 goal of 95% of the prescribed dose, i.e. 6650 cGy, was achieved, see table 8. Therefore, from this moment, a robustness evaluation for 16 perturbed dose scenarios was performed to access robust target coverage.

The results of the D98 of the vw-min in the CTV5425 and CTV7000 for plans made with Strategy 8, are shown in table 9. It can be seen that robust target coverage was achieved in 2 out the 5 patients. Three plans were not robust in the D98 of the CTV5425.

Strategy 9

In the settings of 'Strategy 8', the *EnsembleSize*; the number of ARFs that were selected for the prediction, was set at 5. However, it is advised to include maximal 10% of model size, so in the model size of 40, an ensemble size of 5 was a somewhat large. Therefore, the same strategy was applied, but now with an ensemble size of 3; represented by 'Strategy 9'. This resulted in robust target coverage in all patients, see table 9.

Model-40.1: Strategy 10

As a next step, a new model was trained with the same patients included as previous, but now the training was also performed on CTVs next to PTVs, Model-40.1. It is expected that training the model on the CTVs too, will improve target coverage in the CTVs because it is learnt what dose value the voxels inside the CTVs and around the CTVs have. The same settings as in Strategy 9 were applied again, now named 'Strategy 10'. In table 9 it can be seen that now the plans were not robust anymore for 2 out the 5 patients; the vw-min of the D98 in the CTV 5425 was below the goal of 94%, and in one plan the CTV7000 did not fulfill the robustness criterion. So, with the current strategy, it was more difficult to achieve robust target coverage when CTVs are included in the training.

Table 9 Results of the voxel-wise minimum D98 per patient of the different strategies compared to the clinical plan.

Vw-min D98 (cGy)	Clinical plan		Strategy 8		Strategy 9		Strategy 10	
			Mimicked		Mimicked		Mimicked	
	CTV 7000	CTV 5425	CTV 7000	CTV 5425	CTV 7000	CTV 5425	CTV 7000	CTV 5425
Patient 1	6674	5172	6716	5022	6683	5204	6515	5154
Patient 2	6691	5146	6769	5230	6770	5210	6730	5234
Patient 3	6457	5122	6737	5093	6715	5124	6623	5082
Patient 4	6675	5154	6789	5139	6756	5103	6744	5155
Patient 5	6664	5139	6733	4976	6688	5107	6643	5089
Mean ± SD	**6632 ± 98**	**5147 ± 18**	**6749 ± 30**	**5092 ± 100**	**6722 ± 39**	**5150 ± 53**	**6651 ± 93**	**5143 ± 62**

3.4.5 Tune mimicking: Model-66

A new model was trained with all the clinical IMPT oropharyngeal plans available at that moment. Also, in this model plans planned according to the Dutch National Indication Protocol for Proton therapy (LIPP) 2 were included. The model was tested on 4 patients, all with robustness parameters of 0.3 cm set up and 3% range uncertainty, and thus differ from the patients were tests were performed at earlier. Two of them had clinically LIPP 1 planned plans and the other two patients had LIPP 2 planned plans. Evaluation was done by assessment of the D98 in the vw-min in the CTV7000 and CTV5425, which has to be at least 94% of the prescribed dose, i.e. 6580 cGy for CTV7000 and 5100 cGy for CTV5425.

Strategy 11

The same strategy as used in Model-40.1 ('Strategy 10') was applied to this new trained model, but with an ensemble size of 5 since the new trained model consists of more patients and names 'Stratgey 11'. The results of the robustness evaluation are shown in figure 13, as n=100. It can be seen that for patient 3, the CTV7000 was not robust and in patient 4, the D98 in the vw-min for the CTV5425 was not robust. Therefore, again the *addTargetNtimes* setting was set to a higher value, 250 and 500 respectively, to see if this would result in a higher robustness. The results of these adjustment on the vw-min are shown in figure 13. It can be seen that adding the targets multiple times to the optimizer did not result in achieving robust target coverage in the D98 of the vw-min; in patient 3 and 4 the vw-min is below the clinical goal.

Strategy 12 and 13

In a next step, some higher *ROIGoals* in prediction settings were added; 'Strategy 12'. This resulted in robust target coverage in all 4 test patients, see figure 14. Therefore, cross-validation was performed in a new set of 4 test patients, and a new model was trained with the same number of patients and settings. The same strategy was applied, and subsequently robustness evaluation was performed. However, as can be seen in figure 14, for tune patient 1 and 2, this strategy did not result in robust target coverage.

As a next step, a DVH-based mimicking strategy was applied ('Strategy 13') to see if this would result in robust target coverage. However, in figure 14 it can be seen that this resulted in no robust target coverage in cross-tune patient 1 and 2 regarding the CTV5425.

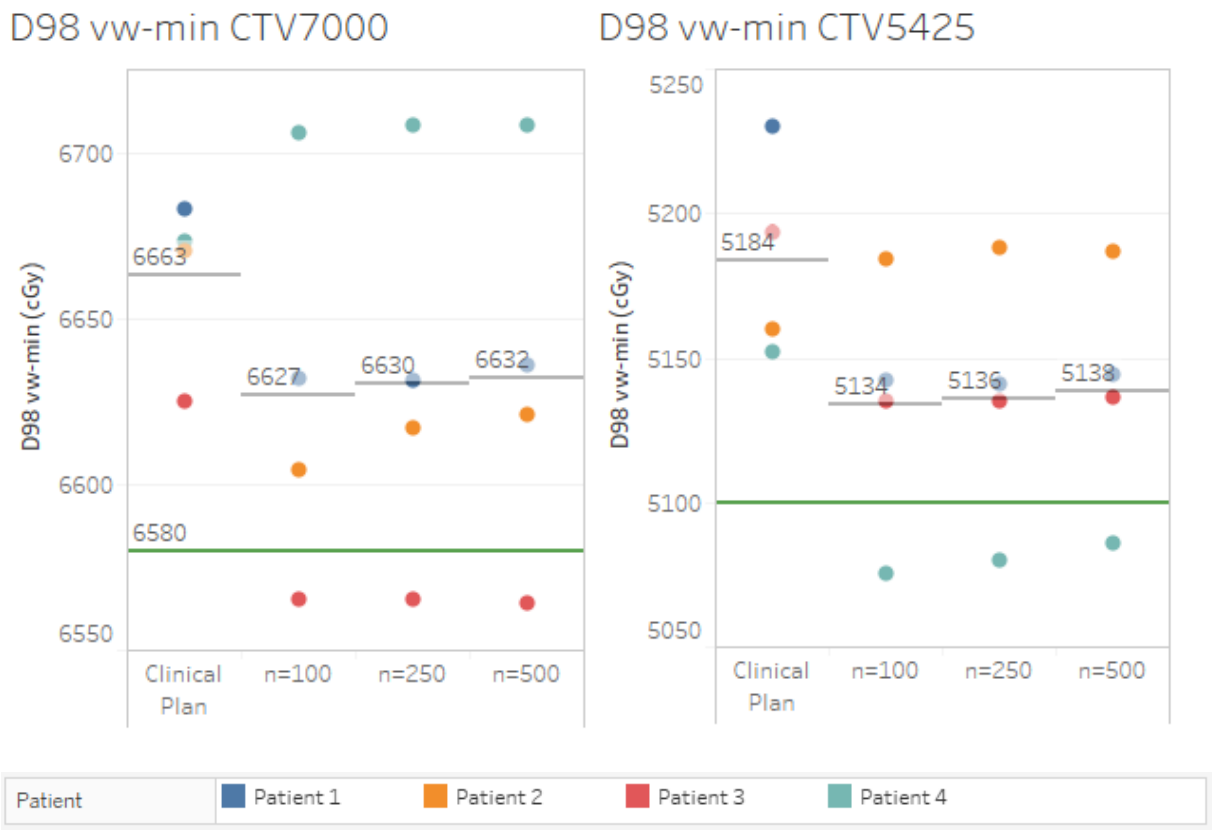


Figure 13 Results of the voxel-wise minimum D98 per patient in the CTVs of Strategy 11 (n=100) and when the CTVs are added more times compared to the clinical plan. The grey lines and displayed value represent the average D98 of the 5 patients. The green line indicates the threshold for 94% of the prescribed dose.

Strategy 14

Then, the same settings as in Strategy 12, but again with somewhat higher *ROIGoals* were given to the CTV7000 and CTV5425, named 'Strategy 14'. Higher *ROIGoals* were given because in Strategy 12 this had demonstrated an improvement in the D98 of the vw-min in the CTVs.

Applying Strategy 14 resulted in robust target coverage in all 4 patients, see figure 14. As a next step, cross-validation with Strategy 14 was performed on the initial tune patients. Also, in these 4 patients, this strategy resulted in robust target coverage, see figure 14. The results of the perturbed scenarios of robustness evaluation in the cross-validation patients is shown in figure 15. It can be seen, that for the CTV7000 the vw-min in the MLO plans were higher compared to the clinical plans and were all above the 94% threshold. Also, the distribution of the of the perturbed scenarios doses was smaller compared to the clinical plans. For the CTV5425 all plans were above the threshold of 94% of the prescribed dose. In figure 16, the mean DVHs of the targets and OARs of this strategy in the cross-validation patients are shown compared to the clinical plan. It can be seen that on average the DVHs of the CTVs were nearly the same, the DVH of the CTV5425 is somewhat higher. The dose to the OARs was higher in the MLO plans compared to the clinical plans, except for the spinal cord which dose was on average lower in the MLO plans.

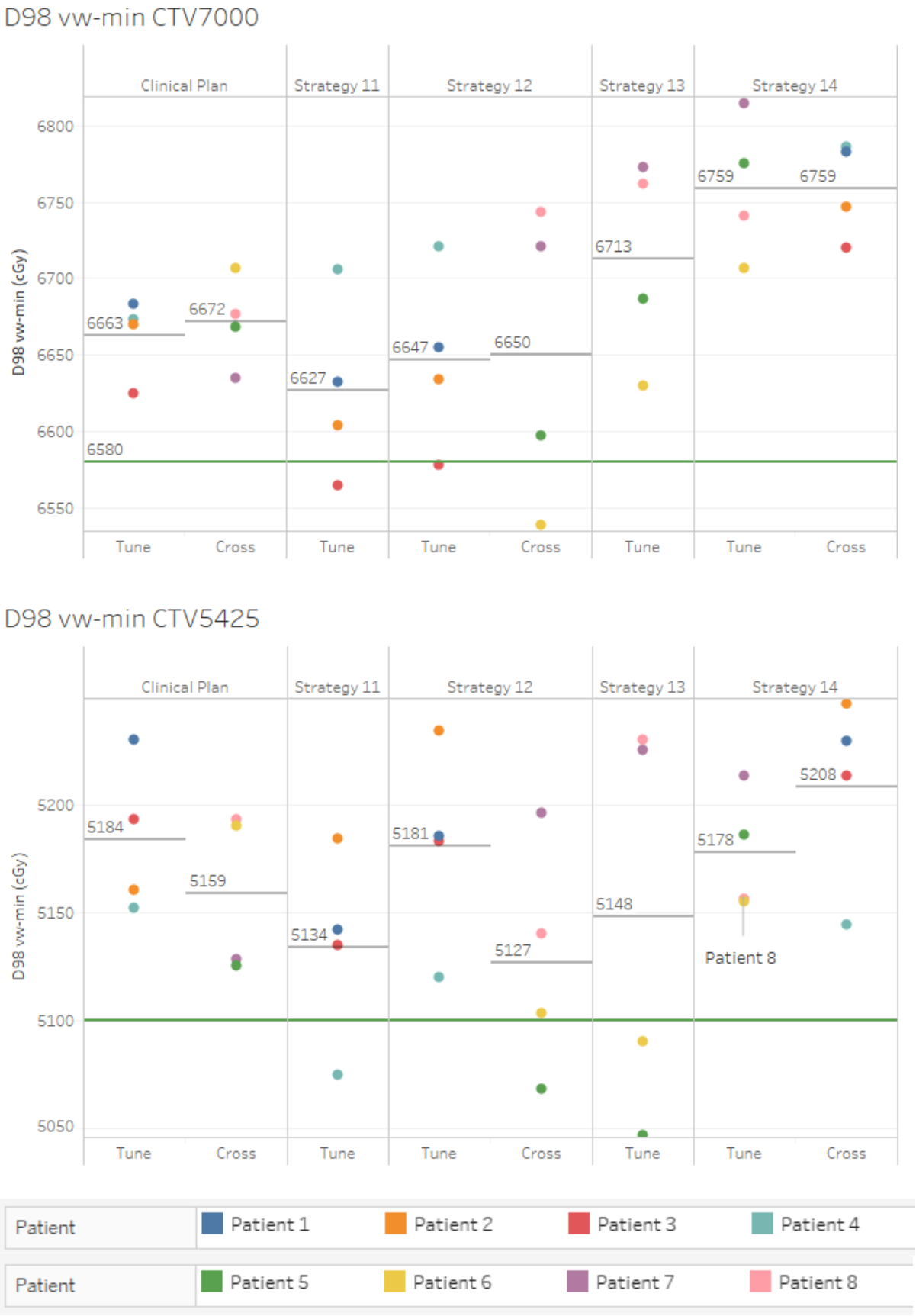


Figure 14 Results of the voxel-wise minimum D98 per tune- and cross-patient of strategies 11-14 compared to the clinical plan. The grey lines and displayed value represent the average D98 of the 5 patients. The green line indicates the threshold for 94% of the prescribed dose.

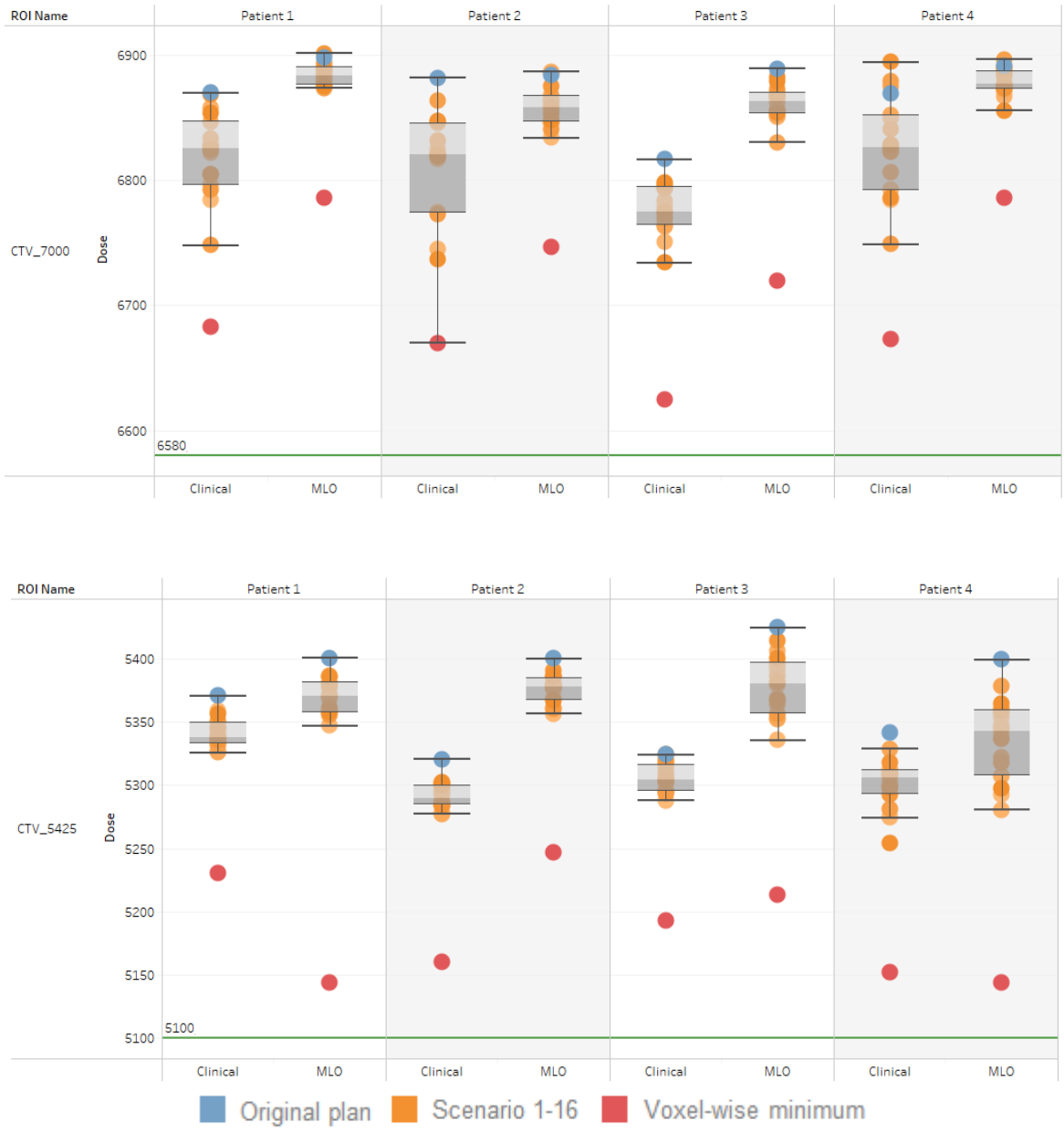


Figure 15 Boxplot of the 16 scenario doses for Strategy 14 from the robustness evaluation (orange), nominal dose (blue) and vw-min (red) of 4 cross-validation patients of the CTV7000 (above) and CTV5425 (below). The green line indicates the threshold for 94% of the prescribed dose.

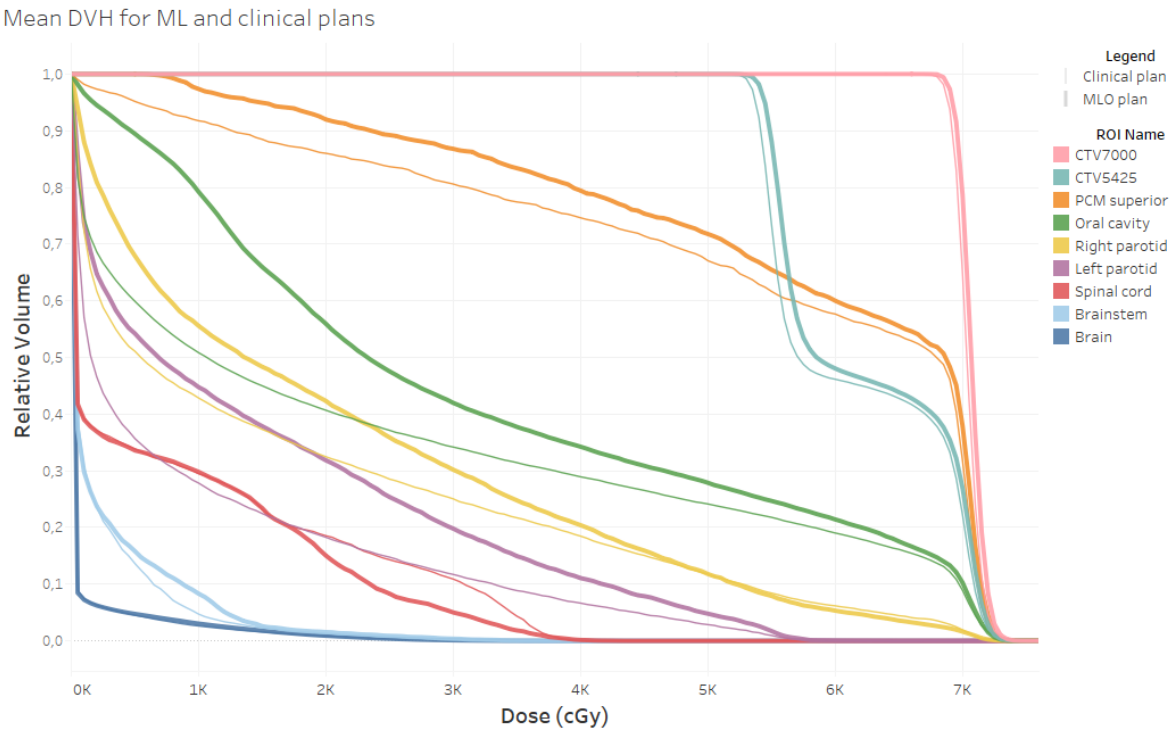


Figure 16 The mean DVH of 4 patients of the clinical plan and the MLO plan of Strategy 14 where cross-validation was performed on.

Strategy 15, Reduce OAR dose

Since the goal of robust target coverage was achieved now, a strategy to reduce the dose to the OARs was applied. In this strategy, 'Strategy 15', the *ReduceOAR* was set to true per OAR with a *ReduceOARLevel* of 0.25 in *ROIGoals.* Also, the *PriorMethod* was set to 'avg' instead of 'max'. This resulted in robust target coverage and less dose to the OARs as compared to Strategy 14. The same results were seen in the cross-validation which was performed subsequently. After that, this strategy was tested on an independent test set of 9 patients. Next to this, also the setting *reduce_dose* in the *NonRobustMimickOARs* function was set to true, to test if this would even more reduce the dose to the OARs. However, with this setting, the MLO plan generation was not competed within 24 hours, and thus was stopped manually. Due to time, Strategy 15 was the endpoint of the development part for this study. The final results were presented and discussed in detail in Chapter 2.

3.5 Discussion and future recommendations

In this first application of MLO for robust mimicking of IMPT plans, we found out that achieving robust target coverage was hard, and a lot of tuning in the settings was necessary to find a strategy which fulfills the clinical goals regarding target coverage. Therefore, initially we have chosen to exclude the goals for the OARs. When a strategy was found which could achieve robust target coverage, some first attempts to reduce to dose to the OARs while maintain the robust target coverage were performed.

The difference between DVH and voxel-based mimicking was not always clear, and had in some patients improving effect on target coverage, while decreasing effect in others. In the final strategy of our study, a voxel-based mimicking method was used, which has the advantage of preserving spatial information. It may be that DVH-based mimicking could lead to less target homogeneity, however this was not investigated in our study since the main goal was to achieve robust target coverage.

The functionality *addTargetsNtimes* was a cumbersome way to influence the weight of the CTV goals, and increased the time to generate a MLO plan. It has been shown that adding the CTVs 100 times was a good agreement between target coverage and the time it took to optimize the plan. However, only adjustments of this function were made for both CTVs simultaneous. When treating the CTV7000 and CTV5425 separately, so that different 'weights' are given for each CTV, this could influence the target coverage in both CTVs. So, this functionality could be further investigated in the future. Also, in a future version of RayStation, it is advised to include the *'weights'* functionality for the *'RobustMimickCTVs'* because then it probably will be easier to steer the dose in the targets.

Next to this, in the mimicking optimization there is optimized for different scenario doses regarding the robustness for a certain setup- and density error. However, the evaluation criteria are based on the vw-min D98. It may be an improvement to include this evaluation criterion also as a goal in the robustness optimizer.

In testing the difference in target coverage between the mixed and the homogeneous model, testing was only performed in three patients. To be able to train a model of a certain size, as much as possible plans which were planned since 2019 were included in the training of the homogeneous model. Therefore, only 3 patients were left to perform testing on. However, since more plans were available to train the mixed model, the model size of the mixed model was larger than the homogeneous model. Therefore, an ensemble size of 3 was used in the test strategy for both the mixed and the homogeneous model, to exclude the eventual benefit of more similar ARFs available in the larger model size.

There were only small differences found between the homogeneous and mixed model regarding the target coverage vw-min D98 and sum NTCP. This indicates that for the number of patients which were available, using a mixed model instead of a homogeneous model would not result in much difference in target robustness. Therefore, in this study we continued the use of a heterogeneous model. However, when a larger number of patients planned with 0.3 cm robustness uncertainty is available, it will be advised to investigate if a homogeneous model with a similar model size as the mixed model will result in more difference between these modes. Next to this, it will be interesting to see if the same ARFs are selected in both models, this can be performed in the future.

Since August 2019, clinical plans were optimized according to LIPP2. In this study, only 13 plans were available which were planned according to this protocol. When more patients are available, it has to be investigated if it is necessary to train a new model for LIPP2 planned plans or that tuning the settings will be able to generate plans with comparable plan quality to the clinical plans.

The final strategy found in this study provides robust target coverage with similar target coverage to most OARs as compared to the clinical plans. However, the mean dose to the oral cavity was higher in the MLO plans compared to the clinical plans. So, in a next step, the first focus in tuning has to be to reduce the dose to the oral cavity. Also, it has to be investigated if the dose to all OARs can be further reduced, while maintaining robust target coverage. Since the *reduce_dose* mimicking setting was very time consuming to generate a MLO plan, it is advised to tune the ReduceOAR settings in the prediction to reduce the dose to the OARs.

Chapter 4 References

[1] "Cijfers over kanker; hoofd hals," 2018. [Online]. Available: https://www.cijfersoverkanker.nl/selecties/Dataset_1/img5d09435fdce39. [Accessed: 18-Jun-2019].

[2] J. A. van Spil, H. A. M. van Muilekom, and B. F. H. van de Walle-van de Geijn, *Oncologie: Handboek voor verpleegkundigen en andere hulpverleners*, 2nd ed. Bohn Stafleu van Loghum, 2013.

[3] P. Economopoulou and A. Psyrri, "ESMO Essentials for Clinicians Head and Neck Cancers Chapter 1," in *Essentials for Clinicians: Head & Neck Cancers*, 2017.

[4] M. D. Stubblefield and M. W. O'Dell, *Synopsis of clinical oncology*. Demos Medical, 2010.

[5] Z. J. Cappello and A. B. Dublin, *Anatomy, Head and Neck, Nose Paranasal Sinuses*. StatPearls Publishing, 2019.

[6] A. Christopoulos, "Mouth Anatomy," *Medscape*, 2013. [Online]. Available: http://emedicine.medscape.com/article/1899122-overview#showall. [Accessed: 18-Jun-2019].

[7] F. H. Martini, E. F. Bartholomew, W. C. Ober, H. Isselée, and J. E. Bruijn, "Chapter 15: Het ademhalingsstelsel," in *Anatomie en fysiologie: een inleiding*, Pearson Education, 2008.

[8] N. Vigneswaran and M. D. Williams, "Epidemiologic trends in head and neck cancer and aids in diagnosis.," *Oral Maxillofac. Surg. Clin. North Am.*, vol. 26, no. 2, pp. 123–41, May 2014, doi: 10.1016/j.coms.2014.01.001.

[9] J. D. Klein and J. R. Grandis, "The molecular pathogenesis of head and neck cancer.," *Cancer Biol. Ther.*, vol. 9, no. 1, pp. 1–7, Jan. 2010, doi: 10.4161/cbt.9.1.10905.

[10] N. Denaro, M. C. Merlano, and E. G. Russi, "Follow-up in Head and Neck Cancer: Do More Does It Mean Do Better? A Systematic Review and Our Proposal Based on Our Experience.," *Clin. Exp. Otorhinolaryngol.*, vol. 9, no. 4, pp. 287–297, Dec. 2016, doi: 10.21053/ceo.2015.00976.

[11] D. G. Deschler, M. G. Moore, and R. V. Smith, "Quick Reference Guide to TNM Staging of Head and Neck Cancer and Neck Dissection Classification," American Academy of Otolaryngology–Head and Neck Surgery Foundation, 2014.

[12] C. Nutting, "Radiotherapy in head and neck cancer management: United Kingdom National Multidisciplinary Guidelines.," *J. Laryngol. Otol.*, vol. 130, no. S2, pp. 66–67, May 2016, doi: 10.1017/S0022215116000463.

[13] John M. Eisenberg Center for Clinical Decisions Science, *Understanding Radiotherapy for Head and Neck Cancer: A Guide for Adults and Their Caregivers*. Agency for Healthcare Research and Quality (US), 2005.

[14] M. Baumann and C. Petersen, "TCP and NTCP: a basic introduction.," *Int. J. Radiol. Sci.*, vol. 30, no. 2, pp. 99–104, 2005.

[15] J. R. Hubenak, Q. Zhang, C. D. Branch, and S. J. Kronowitz, "Mechanisms of injury to normal tissue after radiotherapy: a review.," *Plast. Reconstr. Surg.*, vol. 133, no. 1, pp. 49e-56e, Jan.

2014, doi: 10.1097/01.prs.0000440818.23647.0b.

[16] S.-A. Yeh, “Radiotherapy for head and neck cancer.,” *Semin. Plast. Surg.*, vol. 24, no. 2, pp. 127–36, May 2010, doi: 10.1055/s-0030-1255330.

[17] S. Nuyts, “Defining the target for radiotherapy of head and neck cancer.,” *Cancer imaging Off. Publ. Int. Cancer Imaging Soc.*, vol. 7 Spec No, no. Special issue A, pp. S50-5, Oct. 2007, doi: 10.1102/1470-7330.2007.9009.

[18] H. Lewis-Jones, S. Colley, and D. Gibson, “Imaging in head and neck cancer: United Kingdom National Multidisciplinary Guidelines.,” *J. Laryngol. Otol.*, vol. 130, no. S2, pp. S28–S31, May 2016, doi: 10.1017/S0022215116000396.

[19] A. K. Berthelsen *et al.*, “What’s new in target volume definition for radiologists in ICRU Report 71? How can the ICRU volume definitions be integrated in clinical practice?,” *Cancer Imaging*, vol. 7, no. 1, pp. 104–16, Jun. 2007, doi: 10.1102/1470-7330.2007.0013.

[20] P. Maingon, V. Mammar, K. Peignaux, G. Truc, and I. Barillot, “Les contraintes aux organes à risque en radiothérapie par modulation d’intensité des cancers ORL,” *Cancer/Radiothérapie*, vol. 8, no. 4, pp. 234–247, Aug. 2004, doi: 10.1016/J.CANRAD.2004.06.002.

[21] P. Klinker, H. Credoe, P. Kalk, and M. Landeweerd, “Techniekbeschrijving Protonen Raystation (Versie 7),” 2019. [Online]. Available: https://docportal.umcg.nl/iDocument/Viewers/Frameworks/ViewDocument.aspx?DocumentID=cdc73fb3-3ef7-4eb6-ab3f-b1488c67e33a&NavigationHistoryID=22339491&PortalID=160&Query=hoofd+hals+constraints&FromLogin=1. [Accessed: 18-Jun-2019].

[22] E. Evans and J. Staffurth, “Principles of cancer treatment by radiotherapy,” *Surg.*, vol. 36, no. 3, pp. 111–116, Mar. 2018, doi: 10.1016/J.MPSUR.2017.12.006.

[23] C. Garibaldi *et al.*, “Recent advances in radiation oncology.,” *Ecancermedicalscience*, vol. 11, p. 785, 2017, doi: 10.3332/ecancer.2017.785.

[24] E. B. van Dieren *et al.*, “Beam intensity modulation using tissue compensators or dynamic multileaf collimation in three-dimensional conformal radiotherapy of primary cancers of the oropharynx and larynx, including the elective neck.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 47, no. 5, pp. 1299–309, Jul. 2000.

[25] W. P. Levin and T. F. DeLaney, “Charged Particle Radiotherapy,” *Clin. Radiat. Oncol.*, pp. 358-372.e2, Jan. 2016, doi: 10.1016/B978-0-323-24098-7.00019-8.

[26] J. Widder *et al.*, “The Quest for Evidence for Proton Therapy: Model-Based Approach and Precision Medicine,” *Int. J. Radiat. Oncol.*, vol. 95, no. 1, pp. 30–36, May 2016, doi: 10.1016/j.ijrobp.2015.10.004.

[27] J. A. Langendijk, M. S. Hoogeman, R. Monshouwer, and M. Verheij, “Landelijk Indicatie Protocol Protonen Therapie HOOFD-HALSTUMOREN,” 2017.

[28] J. A. Langendijk, M. S. Hoogeman, R. Monshouwer, and M. Verheij, “Landelijk Indicatie Protocol Protonen Therapie (versie 2.2) HOOFD-HALSTUMOREN,” 2019.

[29] J. E. Leeman *et al.*, “Proton therapy for head and neck cancer: expanding the therapeutic window,” *Lancet Oncol.*, vol. 18, no. 5, pp. 254–265, May 2017, doi: 10.1016/S1470-2045(17)30179-1.

[30] J. Unkelbach *et al.*, “Robust radiotherapy planning,” *Phys. Med. Biol.*, vol. 63, no. 22, p. 22TR02, Nov. 2018, doi: 10.1088/1361-6560/aae659.

[31] E. W. Korevaar *et al.*, “Multi-scenario robustness evaluation; transition to a ‘proton proof’ alternative to PTV evaluation,” *Radiother. Oncol.*, vol. 127, pp. S1073–S1074, 2018.

[32] R. G. J. Kierkels, A. Fredriksson, S. Both, J. A. Langendijk, D. Scandurra, and E. W. Korevaar, “Automated Robust Proton Planning Using Dose-Volume Histogram-Based Mimicking of the Photon Reference Dose and Reducing Organ at Risk Dose Optimization,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 103, no. 1, pp. 251–258, Jan. 2019, doi: 10.1016/j.ijrobp.2018.08.023.

[33] C. M. Bishop, *Pattern recognition and machine learning*. Springer, 2006.

[34] P. Giraud *et al.*, “Radiomics and Machine Learning for Radiotherapy in Head and Neck Cancers.,” *Front. Oncol.*, vol. 9, p. 174, 2019, doi: 10.3389/fonc.2019.00174.

[35] C. McIntosh, M. Welch, A. McNiven, D. A. Jaffray, and T. G. Purdie, “Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method,” *Phys. Med. Biol.*, vol. 62, no. 15, pp. 5926–5944, Jul. 2017, doi: 10.1088/1361-6560/aa71f8.

[36] M. Holmström, “Internal document RaySearch Laboratories AB - PMH Autoplanning overview,” 2017.

[37] T. A. van de Water, H. P. Bijl, C. Schilstra, M. Pijls-Johannesma, and J. A. Langendijk, “The Potential Benefit of Radiotherapy with Protons in Head and Neck Cancer with Respect to Normal Tissue Sparing: A Systematic Review of Literature,” *Oncologist*, vol. 16, no. 3, pp. 366–377, Mar. 2011, doi: 10.1634/theoncologist.2010-0171.

[38] C. R. Hansen *et al.*, “Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans,” *Clin. Transl. Radiat. Oncol.*, vol. 1, pp. 2–8, Dec. 2016, doi: 10.1016/j.ctro.2016.08.001.

[39] J. Krayenbuehl, I. Norton, G. Studer, and M. Guckenberger, “Evaluation of an automated knowledge based treatment planning system for head and neck,” *Radiat. Oncol.*, vol. 10, no. 1, Nov. 2015, doi: 10.1186/s13014-015-0533-2.

[40] I. El Naqa, R. Li, and M. J. Murphy, *Machine Learning in Radiation Oncology: Theory and Applications*. Springer International Publishing, 2015.

[41] I. Hazell *et al.*, “Automatic planning of head and neck treatment plans,” *J. Appl. Clin. Med. Phys.*, vol. 17, no. 1, pp. 272–282, 2016, doi: 10.1120/jacmp.v17i1.5901.

[42] A. Fogliata *et al.*, “Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer,” *Radiat. Oncol.*, vol. 9, p. 236, Oct. 2014, doi: 10.1186/s13014-014-0236-0.

[43] T. G. Purdie, R. E. Dinniwell, A. Fyles, and M. B. Sharpe, “Automation and intensity modulated radiation therapy for individualized high-quality tangent breast treatment plans,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 90, no. 3, pp. 688–695, Nov. 2014, doi: 10.1016/j.ijrobp.2014.06.056.

[44] C. McIntosh and T. G. Purdie, “Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy,” *IEEE Trans. Med. Imaging*, vol. 35, no. 4, pp. 1000–1012, Apr. 2016, doi: 10.1109/TMI.2015.2505188.

[45] C. McIntosh and T. G. Purdie, “Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning,” *Phys. Med. Biol.*, vol. 62, no. 2, pp. 415–431, Jan. 2017, doi: 10.1088/1361-6560/62/2/415.

[46] “RaySearch white paper: Machine learning automated treatment planning,” 2019.

[47] D. Wagenaar *et al.*, “Head and Neck IMPT probabilistic dose accumulation: feasibility of a 2 mm setup,” *Int. J. Radiat. Oncol.*, 2020.

[48] B. K. Sasidharan *et al.*, “Clinical Monte Carlo versus Pencil Beam Treatment Planning in Nasopharyngeal Patients Receiving IMPT,” *Int. J. Part. Ther.*, vol. 5, no. 4, pp. 32–40, Mar. 2019, doi: 10.14338/ijpt-18-00039.1.

[49] RaySearch Laboratories AB, “Raystation 9B Scripting Environmens for Machine Learning,” 2019.

[50] Z. Khan *et al.*, “Ensemble of optimal trees, random forest and random projection ensemble classification,” *Adv. Data Anal. Classif.*, 2019, doi: 10.1007/s11634-019-00364-9.

[51] E. W. Korevaar *et al.*, “Practical robustness evaluation in radiotherapy – A photon and proton-proof alternative to PTV-based plan evaluation,” *Radiother. Oncol.*, vol. 141, pp. 267–274, Dec. 2019, doi: 10.1016/j.radonc.2019.08.005.

[52] A. R. Delaney, W. F. Verbakel, J. Lindberg, T. K. Koponen, B. J. Slotman, and M. Dahele, “Evaluation of an Automated Proton Planning Solution,” *Cureus*, Dec. 2018, doi: 10.7759/cureus.3696.

[53] A. R. Delaney *et al.*, “Automated knowledge-based intensity-modulated proton planning: An international multicenter benchmarking study,” *Cancers (Basel).*, vol. 10, no. 11, Nov. 2018, doi: 10.3390/cancers10110420.

[54] R. G. Bijman *et al.*, “Impact of model and dose uncertainty on model-based selection of oropharyngeal cancer patients for proton therapy,” *Acta Oncol. (Madr).*, vol. 56, no. 11, pp. 1444–1450, Nov. 2017, doi: 10.1080/0284186X.2017.1355113.

[55] F. Hennings, A. Lomax, A. Pica, D. C. Weber, and J. Hrbacek, “Automated Treatment Planning System for Uveal Melanomas Treated With Proton Therapy: A Proof-of-Concept Analysis,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 101, no. 3, pp. 724–731, Jul. 2018, doi: 10.1016/j.ijrobp.2018.02.008.

[56] W. Gu, R. Neph, D. Ruan, W. Zou, L. Dong, and K. Sheng, “Robust Beam Orientation Optimization for Intensity-Modulated Proton Therapy,” *Med. Phys.*, p. mp.13641, Jun. 2019, doi: 10.1002/mp.13641.

[57] W. G. Campbell *et al.*, “Neural network dose models for knowledge-based planning in pancreatic SBRT:,” *Med. Phys.*, vol. 44, no. 12, pp. 6148–6158, Dec. 2017, doi: 10.1002/mp.12621.

Appendix A

Table 10 Maximal doses to the serial OARs per patient in the MLO plans and clinical plans.

Dmax (cGy)	Brain		Brainstem		Spinal cord		Left lens		Right lens		Left eye		Right eye	
Test patient	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan
1	5025	5057	2739	2003	4312	4304	13	21	19	23	16	23	28	29
2	2998	3731	3216	3127	4286	4392	12	16	12	16	12	17	20	24
3	540	780	21	17	4274	3359	4	5	11	8	8	8	11	10
4	3060	2914	2601	3045	4538	373	14	13	14	19	17	19	22	21
5	2464	2048	1250	1092	1907	1626	10	16	18	20	17	24	30	28
6	2098	3091	1623	2598	3158	3222	22	22	21	15	35	31	24	24
7	2995	4461	1611	1765	3975	3742	16	18	17	13	25	19	17	19
8	4951	4678	1964	1625	3193	3532	27	25	19	23	27	28	26	30
9	1984	2751	1996	2494	4355	4502	28	23	13	17	32	25	16	16
Mean ± SD	**2902 ± 1415**	**3201 ± 1536**	**1891 ± 939**	**1974 ± 994**	**3778 ± 864**	**3601 ± 871**	**16 ± 8**	**18 ± 6**	**16 ± 4**	**17 ± 5**	**21 ± 9**	**22 ± 7**	**22 ± 6**	**22 ± 7**
p-value		0.314		0.767		0.441		0.326		0.433		0.673		0.347

Table 11 Mean doses to the parallel OARs per patient in the MLO plans and clinical plans.

Dmean (cGy)	Oral cavity		Left parotid		Right parotid		Left submandibular		Right submandibular		PCM inferior		PCM medial		PCM superior	
Test patient	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan	Clinical plan	MLO plan
1	3018	3372	1257	1255	3657	3524	2946	2392	6713	6621	1775	1598	4532	4145	5062	4896
2	4348	4814	963	1322	1720	1619	3032	2632	6302	6210	877	1474	3527	3964	5817	5865
3	1069	1244	620	738	859	945	1722	2117	5816	5558	1789	1969	4249	4113	2720	2782
4	2077	2376	1090	1165	1113	1240	4312	4083	2070	2324	2718	2767	3711	3925	5876	6120
5	1551	2107	1211	1197	1490	1568	2140	1891	5676	5815	2390	2071	2893	2550	2653	2847
6	3559	3968	1404	1459	851	872	5996	6138	5661	5941	1772	1998	2681	2513	4671	4724
7	2533	2630	1641	1522	2153	2059	4001	3308	6884	6849	2360	2249	5441	5238	4020	3896
8	5969	6133	2321	2395	1220	1640	6456	6463	6633	6539	1792	2214	6011	6116	5474	5616
9	3790	3976	2090	2009	1207	1196	6445	6304	3261	2891	1652	2251	4997	5368	4960	5171
Mean ±SD	**3102 ± 1515**	**3402* ± 1504**	**1400 ± 540**	**1451 ± 490**	**1586 ± 881**	**1629 ± 803**	**4117 ± 1826**	**3925 ± 1898**	**5446 ± 1667**	**5416 ± 1650**	**1902 ± 534**	**2066 ± 382**	**4227 ± 1133**	**4215 ± 1216**	**4584 ± 1219**	**4657 ± 1236**
p-value		0.008		0.515		0.767		0.139		0.678		0.173		0.953		0.139

***=statistically significant difference (p<0.05) compared to clinical plan**

PCM = Pharyngeal constrictor muscle

Appendix B


```
{
    "ModelName": "HN-Proton-New_66",
    "BeamSetsDefaults": [
        {
            "SpatialMetric" : "p-prior&&gamma20",
            "EnsembleSize": 5,
            "PriorMetric" : "simplified-prior",
            "TreatmentMachine": "RSL_VersaHD",
            "BeamListTemplate": "1 Arc",
      "IsocenterType": "CenterOfRoi",
            "IsocenterRegionAlias": ["PTV5425","PTV54.25","PTV5425total","ptv5425",
                                     "ptv54.25", "PTV_5425"],
       "TreatmentTechnique": "ProtonPencilBeamScanning"
        }
    ],
  "RoiNameAlias" :  {"PTV70" : ["A1 PTV_7000", "PTV_7000-A1", "PTV_7000", "PTV7000"],
                     "PTV54.25" : ["A1 PTV_5425", "PTV_5425-A1", "PTV_5425", "PTV5425"],
                     "CTV70": ["A1 CTV_7000", "CTV_7000_A1", "CTV_7000-A1", "CTV7000",
                               "CTV70", "CTV_7000"],
                     "CTV54.25": ["A1 CTV_5425", "CTV_5425_A1", "CTV_5425-A1", "CTV5425",
                                  "CTV54.25", "CTV_5425"],
                    "Brain" : ["Brain"],
                    "BrainStem" : ["BrainStem"],
                    "SpinalCord" : ["SpinalCord", "myelum"],
                    "Parotid_L" : ["Parotid_L"],
                    "Parotid_R" : ["Parotid_R"],
                    "Submandibular_L" : ["Submandibular_L"],
                    "Submandibular_R" : ["Submandibular_R"],
                    "Mandible" : ["Mandible", "Mandibula"],
                    "OralCavity_Ext" : ["OralCavity_Ext","OralCavityExtended"],
                    "GlotticArea" :["GlotticArea", "GloticArea"],
                    "Supraglottic" : ["Supraglottic", "SupraglotticLarynx"],
                    "Thyroid" : ["Thyroid"],
                    "PCM_Inf" : ["PCM_Inf", "PharynxConst_Inf"],
                    "PCM_Med" : ["PCM_Med", "PharynxConst_Med"],
                    "PCM_Sup" : ["PCM_Sup", "PharynxConst_Sup"],
                    "Crico" : ["Crico","Cricopharyngeus", "CricoPharyngeal_Inlet"],
                    "Esophagus_Cerv" : ["Esophagus_Cerv", "CervicalEsophagus"],
                    "Eye_Ant_L": ["Eye_Ant_L","Lens_L"],
                    "Eye_Ant_R": ["Eye_Ant_R","Lens_R"],
                    "Eye_Post_L": ["Eye_Post_L","Eye_L"],
                    "Eye_Post_R": ["Eye_Post_R","Eye_R"]
            },
    "MultiPlanSettings" : [
{
  "StrategyName" : "Strategy 15",
      "PredictSettings": {
            "RoiGoals": [
                    { "GoalType" : "max_dose",  "ROI" : "CTV70", "Volume" : 50, "Value" :
                       7000, "ValueType" : "abs"},
                    { "GoalType" : "min_dose",  "ROI" : "CTV70", "Volume" : 98, "Value" :
                       6950, "ValueType" : "abs"},
                    { "GoalType" : "max_dose_serial",  "ROI" : "CTV70", "Volume" : 0.0,
                      "Value" : 104, "ValueType" : "prc"},

                    { "GoalType" : "min_dose",  "ROI" : "CTV54.25", "Volume" : 100,
                      "Value" : 5425, "ValueType" : "abs"},
                    { "GoalType" : "min_dose",  "ROI" : "CTV54.25", "Volume" : 100,
                      "Value" : 5289, "ValueType" : "abs"},
                    { "GoalType" : "min_dose",  "ROI" : "CTV54.25", "ExcludeRoi" :
                      "CTV70", "Volume" : 100, "Value" : 5425, "ValueType" : "abs"},

                    { "GoalType" : "max_dose_serial",  "ExcludeRoi" : "PTV70", "ROI" :
                      "PTV54.25&&Parotid_L", "Volume" : 0.00, "Value" : 5425, "ValueType"
                      : "abs"},

                    { "GoalType" : "max_dose_serial",  "ExcludeRoi" : "PTV70", "ROI" :
                      "PTV54.25&&Parotid_R", "Volume" : 0.00, "Value" : 5425, "ValueType"
                      : "abs"},
                    { "GoalType" : "max_dose_serial",  "ROI" : "SpinalCord","Volume" : 1,
```

```
                "Value" : 5000, "ValueType" : "abs"},
              { "GoalType" : "max_dose_serial",  "ROI" : "BrainStem", "Volume" : 1,
                "Value" : 6000, "ValueType" : "abs"}
              ],
    "Functions":[
              {"ROI": "ANY", "ReduceOar": false, "PriorMethod": "avg",
                      "PriorMethodInsideTarget": "avg"},
              {"ROI": "CTV70", "ReduceOar": false, "ReduceOarLevel": 0.25,
                      "PriorMethod": "max", "PriorMethodInsideTarget": "avg"},
              {"ROI": "CTV54.25", "ReduceOar": false, "ReduceOarLevel": 0.25,
                      "PriorMethod": "max", "PriorMethodInsideTarget": "avg"},

              {"ROI": "OralCavity_Ext", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Parotid_R", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Parotid_L", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "PCM_Sup", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "PCM_Med", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "PCM_Inf", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Submandibular_R", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Submandibular_L", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Supraglottic", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "GlotticArea", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Crico", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"},
              {"ROI": "Esophagus_Cerv", "ReduceOar": true, "ReduceOarLevel": 0.25,
                      "PriorMethod": "avg", "PriorMethodInsideTarget": "avg"}
              ]
},
    "MimickSettings":{
    "VoxelCTVMim": true,
    "VoxelOARMim": true,
    "MimicNominalScenarioDoseInAllScenarios": true,
    "RobustMimickCTVs" : {"CTV70": [{"FunctionType": "MinDose"},
                                    {"FunctionType": "MaxDose"}],
                          "CTV_7000_copy": [{"FunctionType": "MinDose"}],
                        "CTV54.25": [{"FunctionType": "MinDose"},
                                     {"FunctionType": "MaxDose"}],
                          "CTV_5425_copy": [{"FunctionType": "MinDose"}],
                            "CTV_L-CTV_H": [{"FunctionType": "MinDose"},
                                            {"FunctionType": "MaxDose"}]},

    "AddTargetsNtimes": {"CTV70": 1, "CTV_7000_copy": 100, "CTV54.25": 1,
                         "CTV_5425_copy": 100, "CTV_L-CTV_H": 1},

    "Strategy_ROIs": [{"roi_name": "CTV_7000_copy", "algebra_type": "Expand",
                       "expand_from": "CTV70", "expand_cm": 0, "roi_type": "CTV"},
                          {"roi_name": "CTV_5425_copy", "algebra_type": "Expand",
                       "expand_from": "CTV54.25", "expand_cm": 0, "roi_type":
                                      "CTV"},
                         {"roi_name": "CTV_L-CTV_H", "algebra_type": "Subtract",
                       "outer_roi": "CTV54.25", "subtract_roi": "CTV70",
                          "roi_type": "CTV"}],

    "NonRobustMimickOARs": {"Brain": {"FunctionType": "MaxDose", "add_ntimes": 1,
                            "reduce_dose": false},
                            "BrainStem": {"FunctionType": "MaxDose", "add_ntimes":
                                          1, "reduce_dose": false},
                            "SpinalCord": {"FunctionType": "MaxDose", add_ntimes":
                                                1, "reduce_dose": false},
                            "Parotid_L": {"FunctionType": "MaxDose", "add_ntimes":
                                          1, "reduce_dose": false},
                            "Parotid_R": {"FunctionType": "MaxDose", "add_ntimes":
                                         1, "reduce_dose": false},
                            "Submandibular_L": {"FunctionType": "MaxDose",
```

```
                                        "add_ntimes": 1,"reduce_dose":
                                        false},
                                "Submandibular_R": {"FunctionType": "MaxDose",
                                        "add_ntimes": 1, "reduce_dose":
                                         false},
                                "Mandible": {"FunctionType": "MaxDose", "add_ntimes":
                                         1, "reduce_dose": false},
                                "OralCavity_Ext": {"FunctionType": "MaxDose",
                                        "add_ntimes": 1, "reduce_dose":
                                        false},
                                "Thyroid": {"FunctionType": "MaxDose", "add_ntimes": 1,
                                        "reduce_dose": false},
                                "Esophagus_Cerv": {"FunctionType": "MaxDose",
                                        "add_ntimes": 1, "reduce_dose":
                                         false},
                                "PCM_Inf": {"FunctionType": "MaxDose", "add_ntimes": 1,
                                        "reduce_dose": false},
                                "PCM_Med": {"FunctionType": "MaxDose", "add_ntimes": 1,
                                        "reduce_dose": false},
                                "PCM_Sup": {"FunctionType": "MaxDose", "add_ntimes": 1,
                                        "reduce_dose": false},
                                "Crico": {"FunctionType": "MaxDose", "add_ntimes": 1,
                                        "reduce_dose": false},
                                "Eye_Ant_L": {"FunctionType": "MaxDose", "add_ntimes":
                                         1, "reduce_dose": false},
                                "Eye_Ant_R": {"FunctionType": "MaxDose", "add_ntimes":
                                         1, "reduce_dose": false},
                                "Eye_Post_L": {"FunctionType": "MaxDose", "add_ntimes":
                                         1, "reduce_dose": false},
                                "Eye_Post_R": {"FunctionType": "MaxDose","add_ntimes":
                                         1, "reduce_dose": false}
                                },

            "RobustenessParameters": {"PositionUncertaintyAnterior": 0.3,
                                      "PositionUncertaintyPosterior": 0.3,
                                      "PositionUncertaintySuperior": 0.3,
                                      "PositionUncertaintyInferior": 0.3,
                                      "PositionUncertaintyLeft": 0.3,
                                      "PositionUncertaintyRight": 0.3,
                                      "DensityUncertainty": 0.03
                                      },
            "NumberOfMimicOptimizations": 3,
            "MachineSetup" : []
            }
         }
        ],

"OptimzationDefaults" : {
        "Runs": [{ "Iterations" : 60, "FinalDose" : true, "OptimalityTolerance" : 1e-6 ,
                  "FinalCheck" : false},
                 { "Iterations" : 60, "FinalDose" : true, "OptimalityTolerance" : 1e-6 ,
                   "FinalCheck" : true},
                 { "Iterations" : 60, "FinalDose" : true, "OptimalityTolerance" : 1e-6 ,
                   "FinalCheck" : true}
                  ],
        "SegmentationSettingsDefaults":{
            "ArcDefaults":{
                "UseMaxLeafTravelDistancePerDegree" : false,
                "FinalArcGantrySpacing" : 2
            }
        },
        "OptimizationsSettingsDefaults":{
            "IterationsBeforeConversion" : 20
        }
    }
}
```

Appendix C

Table 12 Overview of the patient and plan characteristics per model.

		Model-60 (n=60)	Model-40 (n=40)	Model-40.1 (n=40)	Model-66 (n=66)	Model-66 Cross (n=66)	Model-0.3 (n=32)
Tumor Location	Oro-pharynx	31	23	23	42	43	22
	Base of tongue	17	10	10	13	12	4
	Tonsil	12	7	7	11	11	6
T-stage	T1	7	5	5	12	12	3
	T2	9	8	8	11	11	1
	T3	9	6	6	11	12	7
	T4*	35	21	21	32	31	21
N-stage	N0	7	8	8	9	8	7
	N1	11	10	10	15	16	4
	N2**	31	13	13	28	28	20
	N3***	11	9	9	14	14	1
Plan	Clinical	35	40	40	66	66	32
	Research	25	-	-	-	-	-
Targets trained on		PTVs	PTVs	PTVs & CTVs	PTVs & CTVs	PTVs & CTVs	PTVs & CTVs
LIPP version	1	60	40	40	60	60	22
	2	-	-	-	6	6	10
Robustness position uncertainty (cm)	0.3	4	8	8	23	23	32
	0.5	56	32	32	43	43	-

***T4a, T4b, T4NOS**

**** N2a, N2b, N2c, N2NOS**

*****N3a, N3b**

Table 13 Description of the prediction and mimicking settings and availability.

	Parameter	Desciption	Additional settings	Availability
Prediction	*EnsembleSize*	Selection of a number of ARFs	-	Since start
	ROIGoals	Give goals to specific ROIs	*Goaltype:* minimal/maximal dose; *ROI* or *ExcludeROI:* a specified ROI; *Volume*: an amount of the ROI volume; *Value:* a dose value *ValueType:* absolute or percentage	Since start
	ReduceOAR	If stated to *true*, it will reduce the OAR dose	*ROI:* a specified ROI; *ReduceOARlevel:* a reduction level between 1-10; *PriorMethod:* choose if dose in OAR will be based on the minimal/average/maximal dose of the ensemble patients; *PriorMethodInsideTarget:* same as *PiorMethod* but for ROIs which have target and OAR overlap	Since start
Mimicking	*VoxelCTVMim*	If stated to *true*, it will perform voxel-wise mimicking	-	Since start
	VoxelOARMim	If stated to *true*, it will perform voxel-wise mimicking	-	Since start
	RobustMimickCTVs	Robustly optimizes the CTVs specified	*FunctionType:* minimal/maximal dose; *Weight:* *Dose_level:* specify the desired dose level	*Weight* and *Dose_level* not available in RayStation version 9A
	AddTargetsNtimes	The number of times *RobustMimickCTVs* can be added to the optimizer		Since scripting environment v2.3
	Strategy_ROIs	Can create new ROIs	*ROI_name:* new ROI name; *Algebra_type:* expand or substract; *Expand_cm:* value in cm	Since scripting environment v2.3
	NonRobustMimick OARs	Non-robustly optimizes the OARs specified	*add_ntimes:* number of times the OAR is added to the optimized; *reduce_dose:* choose true for further dose reduction, otherwise false.	*Reduce_dose* since scripting environment v2.3 available to set to false.

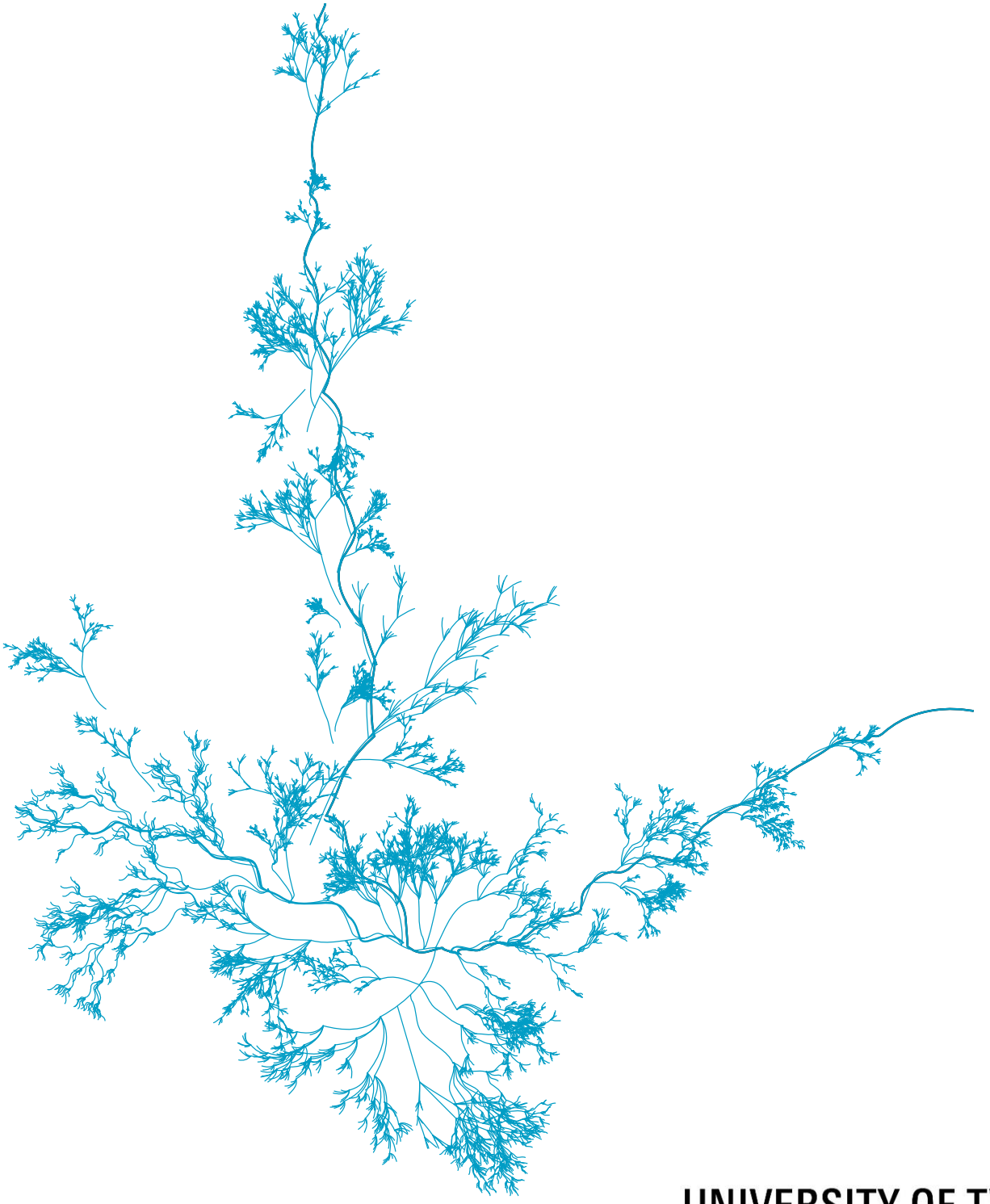
Table 14 Overview of the different strategies and their according prediction and mimicking settings, and in which model and scripting environment they were applied.

		Prediction settings					Mimicking settings				
Strategy	**Model**	***Ensemble size***	***MLRoi-Goals***	***Reduce-Oar***	***Reduce-OarLevel***	***Prior-Method***	***Voxel-CTV-Mim***	***Voxel-OAR-Mim***	***AddTargets-Ntimes***	***Robustness parameters***	**Scripting Environ-ment**
Standard	60	5	-	False	-	-	-	-	-	-	1.1
Strategy 1	60	5	PTVs	True	-	-	-	-	-	-	1.1
Strategy 2	60	5	PTVs & CTVs	Per organ true	1.5	min	False	False	-	0.5 cm/3%	2.2
Strategy 3	60	5	CTV	False	-	max	True	True	1000	0.8 cm/3%	2.3
Strategy 4	60	5	CTV	False	-	max	True	True	500	0.5 cm/3%	"
Strategy 5	60	5	CTV	False	-	max	False	False	500	0.5 cm/3%	"
Strategy 6	60	5	CTV	False	-	max	False	False	1	0.5 cm/3%	"
Strategy 7	60	5	CTV	False	-	max	True	True	100	0.5 cm/3%	"
Strategy 8	40	5	CTV	False	-	max	True	True	100	0.5 cm/3%	"
Strategy 9	40	3	CTV	False	-	max	True	True	100	0.5 cm/3%	"
Strategy 10	40.1	3	CTV	False	-	max	True	True	100	0.5 cm/3%	"
Strategy 11	66	5	CTV	False	-	max	True	True	100	0.3 cm/3%	"
Strategy 12	66	5	CTV*	False	-	max	True	True	100	0.3 cm/3%	"
Strategy 13	66	5	CTV	False	-	max	False	False	100	0.3 cm/3%	"
Strategy 14	66	5	CTV*	False	-	max	True	True	100	0.3 cm/3%	"
Strategy 15	66	5	CTV	Per organ true	0.25	avg	True	True	100	0.3 cm/3%	"

***Within this setting, different ROI goals were adjusted**

Appendix D

```
import logging
from raylearner.data.patient.raystation.patient_data import read_patient_data_file
from raylearner.data.patient.raystation.patient_data import get_plan, get_examination
from raylearner.data.patient.raystation.patient_data import check_patient_plan_presence
from raylearner.data.patient.raystation.patient_data import read_raystation_patient_from_db
import os
import sys
logger = logging.getLogger("Extracting scenario doses for robust evaluation")
import connect

# Configure paths to meta file and patient-plan file
patient_plan_file = r"\\zkh\appdata\Raystation\Research\ML\raylearner\ML Planning 9A
(v2.2)\RobustnessEvaluation\patients_plans_to_auto_RE_0.3.txt"
patient_db = connect.get_current("PatientDB")
patient_plan_list = read_patient_data_file(patient_plan_file)
patient_plan_list = check_patient_plan_presence(patient_db, patient_plan_list)

# define robust eval parameters
robust_eval_name = 'auto_RE_0.3/3'
uncentainty_cm = 0.3
uncertainty_density_perc = 3
num_dens_discret_points = 2

for patient_id, baseplan_name in patient_plan_list:
    try:
        patient = read_raystation_patient_from_db(patient_db, patient_id)
        case = patient.Cases[0]
        case.SetCurrent()
        plan = get_plan(case, baseplan_name)
        plan.SetCurrent()
        treatment_delivery = case.TreatmentDelivery
        patient_name = patient.Name
        examination = get_examination(plan)
        beam_set = connect.get_current("BeamSet")

        retval_0 = beam_set.CreateRadiationSetScenarioGroup(Name=robust_eval_name,
                            UseIsotropicPositionUncertainty=False,
                            PositionUncertaintySuperior=uncentainty_cm,
                            PositionUncertaintyInferior=uncentainty_cm,
                            PositionUncertaintyPosterior=uncentainty_cm,
                            PositionUncertaintyAnterior=uncentainty_cm,
                            PositionUncertaintyLeft=uncentainty_cm,
                            PositionUncertaintyRight=uncentainty_cm,
                            PositionUncertaintyFormation="DiagonalEndPoints",
                            PositionUncertaintyList=None,
                            DensityUncertainty=uncertainty_density_perc,
                            NumberOfDensityDiscretizationPoints=num_dens_discret_points,
                            ComputeScenarioDosesAfterGroupCreation=False)

        retval_0.ComputeScenarioGroupDoseValues()


plan.TreatmentCourse.EvaluationSetup.ApplyClinicalGoalTemplate(Template=root.TemplateTreat
mentOptimizations['HN_ML_Proton_Merle'])

    except:
        logger.exception('uncaught exception for patient {:}'.format(patient_id))

        patient.Save()
```



UNIVERSITY OF TWENTE.