Towards fully automated machine learning VMAT planning for oropharyngeal cancer

I.G. van Bruggen

TOWARDS FULLY AUTOMATED MACHINE LEARNING BASED VMAT PLANNING FOR OROPHARYNGEAL CANCER

A thesis submitted to the University of Twente for the degree of

Master of Science

by

I.G. van Bruggen

23-08-2019

Faculty of Science and Technology
Master Technical Medicine
Track 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:	Dr. M. Groenier (UT)
Extra member:	Dr. ir. R.G.J. Kierkels (UMCG)
Extra member:	Dr. ir. E.W. Korevaar (UMCG)
Extern member:	J.K. van Zandwijk, MSc (UT)

UNIVERSITY OF TWENTE.

Preface

Before you lies the thesis: 'Towards fully automated machine learning based VMAT treatment planning for oropharyngeal cancer'. It has been written in partial fulfilment to the graduation requirements of Technical Medicine at the University of Twente (UT). I have worked on this project from November 2018 to August 2019 at the radiotherapy department of the University Medical Center Groningen (UMCG).

This project on automation in clinical application, fits perfectly with my interests. The reason why I choose my study Technical Medicine in the first place, was because of the interest in improving healthcare with technical solutions. I get satisfaction of improving quality or processes or both and I remain motivated by learning something new. This project uses machine learning, a relatively new and highly potential technique to improve health care. In this project, I combined these interests to achieve more efficient and higher quality radiotherapy treatment plans. I aim to continue improving automation in radiotherapy in the future.

I would thank Roel Steenbakkers for the medical supervision during this internship. Roel his confidence and giving me the ability to practice in the clinic, resulted in a high learning rate on consultation and medical decision making. I would like to thank my supervisors of the UT for learning me to write a solid report (Kees Slump) and be a professional technical physician (Marleen Groenier). Furthermore I would like to thank the machine learning team from RaySearch Laboratories (Mats Holmström, Karl Berggren, Hanna Gruselius, Marco Trincavelli and Fredrik Löfman) for their excellent and quick support during the project. Finally, I would like to thank my technical supervisors of the UMCG (Erik Korevaar and Roel Kierkels). Their critical look and enthusiasm in this project have highly improved the results of the project and my research skills.

I hope you enjoy reading.

Ilse van Bruggen

Groningen, August 2, 2019

CONTENT

Abbreviations

AI	Artificial Intelligence
ARF	Atlas Regression Forest
CI	Conformity Index
CLM	Constrained Leaf Motion
CT	Computed Tomography
CTV	Clinical Target Volume
DL	Deep Learning
DVH	Dose Volume Histogram
ENE	Extra-Nodal Extension
GAN	Generative Adversarial Network
HI	Homogeneity Index
HNC	Head and Neck Cancer
HPV	Human Papilloma Virus
IMPT	Intensity Modulated Proton Therapy
IMRT	Intensity Modulated RadioTherapy
KBP	Knowledge Based Planning
ML	Machine Learning
MLO	Machine Learning Optimization
MRI	Magnetic Resonance Imaging
MU	Monitor Units
NTCP	Normal Tissue Complication Probability
OAR	Organ At Risk
PCM	Pharyngeal Constrictor Muscle
PEG	Percutaneous Endoscopic Gastrostomy
PET	Positron Emission Tomography
pRF	prediction Random Forest
PTV	Planning Target Volume
QA	Quality Assurance
ROI	Region Of Interest
SCC	Squamous Cell Carcinoma
UMCG	University Medical Center Groningen
VMAT	Volumetric Modulated Arc Therapy
3D	three Dimensional

CHAPTER 1 BACKGROUND

Head and Neck Cancers (HNC) are well known to be aggressive tumours. The overall 5-year survival rate ranges between 42-77%, in respectively HNC stage IV and I[1]. Furthermore, locoregional control is limited to 50% in patients with locally advanced head and neck Squamous Cell Carcinoma(SCC)[2]. One of the possible treatment options for HNC is radiotherapy. Radiotherapy treatment planning is challenging in HNC, since Organs At Risk (OARs) are located close to targets and dose related side effects in this area have major effect on quality of life[3]–[5]. Currently, manual treatment planning is used to create clinical acceptable radiotherapy treatment plans. Treatment planning is a time consuming process, taking hours up to days for each patient[6]. In clinical practice, time is limited and it is shown that plan quality is correlated with time invested in an individual plan[7]. Furthermore, plan quality is dependent on the skills and experience of dosimetrists[6]. Automation in HNC radiotherapy treatment planning may improve plan quality[6], [8]. This can be achieved by using Machine Learning (ML)[9], [10].

Chapter 1 provides background information about head and neck cancer, radiotherapy and machine learning.

Chapter 2 describes the main content of this study: generating clinical acceptable machine learning based radiotherapy treatment plans for oropharyngeal cancer.

1. Head and neck cancer

1.1 Clinical presentation

The prevalence of HNC is low compared to other cancer types, only 3% of all malignant tumours in the Netherlands are found in the head and neck region. This anatomical region reaches from the skull base to the clavicles[11]. Nowadays, HNC has an incidence of approximately 3000 in the Netherlands[12]. This is a notable increase of 50%, compared to 1990, which can be explained by a still growing and aging population[13], [14]. Therefore, it is expected that the HNC incidence will further increase in the future and more patients will need treatment.

Patients with HNC can suffer from a painful tongue, sore throat, hoarseness, dysphagia and nose bleedings[15]. HNC can be further categorized by the anatomical area of origination of the tumour. An overview of HNC regions is shown in figure 1[16].

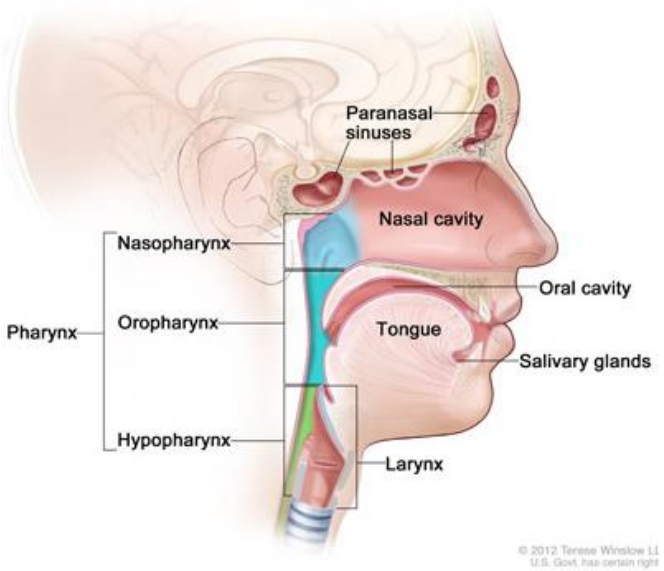

Figure 1 Overview of head and neck cancer regions. The head and neck area can be subdivided into: salivary glands, paranasal sinuses, oral cavity, nasopharynx, oropharynx, hypopharynx and larynx.[16]

HNC can originate at the following sites, mentioned in decreasing incidence: oral cavity, larynx, oropharynx, salivary glands, hypopharynx, paranasal sinuses and nasopharynx tumours[12]. These tumour sites include several anatomical structures [17]:

- Oral cavity, including the upper and lower lips, buccal mucosa, the floor of the mouth, retromolar trigone, anterior two thirds of the tongue, hard palate and upper and lower gingiva
- Larynx, including the glottic, the supraglottic or the subglottic area
- Oropharynx, including the tonsils, tongue base, soft palate and the oropharyngeal wall
- Salivary glands, including the large salivary glands (parotid gland and submandibular gland) and several small salivary glands
- Hypopharynx, including the hypopharyngeal wall, lateral and medial piriform sinuses and postcricoid
- Paranasal sinuses and nasal cavity, including the sinuses in bones of the head or in the nasal vestibule
- Nasopharynx: Tumours originating from epithelial between soft palate, the base of the skull, the lateral and posterior pharyngeal wall

The majority of malignancies in the head and neck region is SCC. Other HNC histology like adenocarcinoma, adenoid cystic carcinoma and mucoepidermoid carcinomas occur less frequent [18]. The risk of development SCC in the head and neck region increases by exposure to tobacco and alcohol[19]. Another risk factor for developing SCC is the Human Papilloma Virus (HPV)[20]. Noticeable, is that patients with oropharyngeal HPV positive tumours have a significant better prognosis compared to patients with HPV negative tumours[21]. Besides the HPV status, the tumour site and overall stage do significantly influence the prognosis [21].

1.2 Diagnostics

The diagnosis of HNC is usually performed by physical examination, nasopharyngoscopy, Computed Tomography (CT) and/or Magnetic Resonance Imaging (MRI). Physical examination is sufficient for superficial oral cavity tumours when depth infiltration is not present or doubted. Nasopharyngoscopy is used to determine superficial extension of tumours in the pharynx and mobility of vocal cords in laryngeal tumours. The choice between MRI and CT is dependent on tumour location and probable contraindications. In general, patients with tumours above the epiglottis undergo MRI. In case a partial or total larynx extirpation is considered for larynx cancer, MRI is indicated to determine cartilage invasion. Furthermore, a diagnostic Positron Emission Tomography (PET) scan is indicated when the patient is at high risk for distant metastasis[22].

1.3. Treatment

Curative treatment options for HNC include surgery, radiotherapy and systemic therapy. For patients with limited or early-stage disease (stage I and II), treatment with only surgery or radiotherapy is sufficient [23]. Patients with locally advanced tumours receive often a combination of these modalities to enlarge the local control and survival prognosis. Decisions for treatment are made in a multidisciplinary team, including surgeons, medical oncologists, radiation oncologists, radiologists and dentists. Both surgery and radiotherapy can be performed as primary treatment. There are no randomized controlled trials available to compare the results of both modalities[17][24]. In general, surgery is performed when resection margins of minimal 5 mm can be achieved without the risk of damaging critical structures. Radiation therapy can be employed concurrent with chemotherapy or additional to surgery[19].

2. Radiotherapy

2.1 Indications

Radiotherapy will be indicated when surgery is limited by anatomical extent of the tumour and function loss is likely. As an example, oropharyngeal cancer surgery may induce dysphagia, which may result in tube feeding dependency and has a major impact on quality of life[17]. Furthermore, primary radiotherapy should be considered for irresectable tumours, with the expectation of small resection margins (<5 mm) or in patients with limited physical performance. Adjuvant radiotherapy will be given if local tumour control is not expected since one or more of the following observations of the pathologist and/or the surgeon are present: [24]

- lymph node metastasis with Extra-Nodal Extension (ENE)
- small resection margins(< 5 mm)
- irradical resection margins (<1 mm)

- macroscopic tumour residual
- multiple lymph node metastasis
- one or more lymph node metastases at more than 3 cm distance from the primary tumour

Furthermore, adjuvant radiotherapy is advised when one or more of the following clinical presentations are observed by the pathologist and/or radiologist:

- perineural extension
- tumour cross section of more than 4 cm (cT3)
- tumour growth in bone, cartilage or intracranial invasion (pT4)
- irregular tumour growth in oral cavity tumours

Radiotherapy of these latter mentioned indications will decrease the risk on a locoregional recurrence, however, patients and/or physicians can decide to renounce the adjuvant radiotherapy[25].

2.2 Toxicity

During and after treatment, patients can suffer from acute mucositis of the throat and mouth, alternation of taste, xerostomia, skin reaction, hoarseness and dysphagia. Some of these side effects will only be present during and a few weeks after treatment. However, some side effects, like xerostomia and dysphagia are common late complications[17]. The risk of developing xerostomia seems highly related with the dose on parotid and submandibular glands[26]. Furthermore, it is known that complication rates will increase when concomitant chemoradiation therapy is given [17]. Grading of xerostomia and dysphagia is based on subjective (dry mouth), functional consequences (impact eating pattern) and objective consequences (saliva production) and can be found in appendix A.

2.3 Treatment preparation

When an indication for radiotherapy is determined and both patient and physician agree on the treatment, treatment preparations start. First a mask is made, to immobilize the patient during the fractionated treatment course. Thereafter a CT scan and if indicated MRI and PET scans are acquired with the mask. The radiation oncologist delineates the target and subsequently planning technicians delineate the OARs. Then, a dosimetrist makes a treatment plan using clinical goals, see table 1[24]. Sufficient target coverage is reached when 98% of the Planning Target Volume (PTV) received 95% of the prescribed dose (D98 ≥95%), according to the ICRU recommendations[27]. During treatment planning, the goal is to minimize the dose to OARs to reduce side effects while maintaining adequate target coverage.

Table 1 Overview of clinical goals for multiple ROIs used in radiotherapy treatment planning[24], [28]

ROI	Clinical goal
PTV 7000	D_{98} ≥ 6650 cGy
	D_{mean} ≥ 6950 cGy
	D_{mean} ≤ 7050 cGy
	D_{2} ≤ 7490 cGy
PTV 5425	D_{98} ≥ 5154 cGy
Brainstem	$D_{0.1}$ ≤ 6300 cGy
Brain	$D_{0.1}$ ≤ 6300 cGy
Optic nerve	$D_{0.1}$ ≤ 6000 cGy
Optic chiasm	$D_{0.1}$ ≤ 6000 cGy
Retina	$D_{0.1}$ ≤ 6000 cGy
Spinal cord	$D_{0.1}$ ≤ 5400 cGy
Other OARs	D_{mean} = ALARA

ROI: Region Of Interest, PTV: Planning Target Volume, OARs: Organs At Risk, D_{98}: dose at 98% of a volume, D_{mean}: mean dose of a volume, D_2: dose at 2% of a volume, $D_{0.1}$: dose at 0.1% of a volume, ALARA: As Low As Reasonably Achievable

2.4 Model-based selection procedure for proton therapy

The standard photon treatment in the UMCG for HNC is Volumetric Modulated Arc Therapy (VMAT). Since January 2018, the UMCG has the possibility to treat HNC with Intensity Modulated Proton Therapy (IMPT). Proton therapy is able to achieve major reduction of dose on OARs. The maximum energy is released when a particle is almost stopped, this phenomenon is known as the Bragg peak[29]. Therefore, the normal tissue behind a beam receives less dose as

compared to photon treatments[30][31]. Dose prescription and clinical goals of proton therapy are similar to photon therapy.

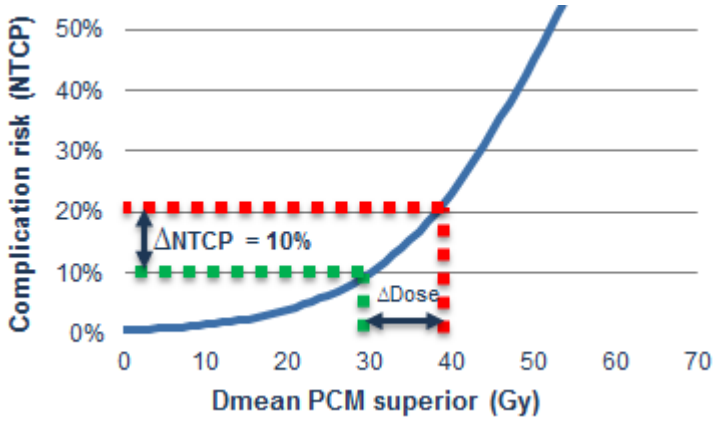

Figure 2 NTCP curve of mean dose on PCM superior. A patient will be selected for proton therapy when the difference in dose, in NTCP involved organs like the PCM superior, between the photon (red) and proton plan (green) results in a significant difference in NTCP (ΔNTCP of any grade 2≥ 10%).*PCM: pharyngeal constrictor muscle, D_{mean}: mean dose of a volume, NTCP: normal tissue complication probability*

Both VMAT and IMPT radiotherapy treatment plans will be made for all HNC patients, when proton therapy might be indicated. A patient will be selected for proton therapy via a model based approach following the 'Dutch National Indication Protocol for Proton Therapy'[32]. Predictions of radiation-induced side effects are compared for both photon and proton treatment plans with Normal Tissue Complication Probability (NTCP) models. Only patients with a significant benefit, calculated with the difference between NTCP (ΔNTCP) values of both plans, are eligible for proton therapy. Table 2 indicates the minimal ΔNTCP to be eligible for proton therapy. An example of sufficient benefit with proton therapy is shown in the NTCP curve of the Pharyngeal Constrictor Muscle (PCM) superior in figure 2. The PCM superior is involved in dysphagia and tube feeding dependence. The patient receives 40 Gy on the PCM superior in the photon plan and 30 Gy in the proton plan. The difference in dose results in a ΔNTCP of 10% that is enough for qualification for proton therapy and as can be seen in table 2.

Table 2 NTCP model and decision criteria for proton therapy selection for head and neck cancer [32]

NTCP model	**Organs involved**
Xerostomia (6 months; grade 2-3)	Contralateral parotid
Dysphagia (6 months; grade 2-3)	Oral cavity; superior PCM
Tube feeding dependence (6 months; grade 3-4)	Contralateral parotid, superior PCM, inferior PCM, cricopharyngeal muscle
Decision criteria	**ΔNTCP**
ΔNTCP of any grade ≥2 complication	≥ 10%
ΔNTCP of any grade ≥3 complication	≥ 5%
∑ ΔNTCP of grade ≥2 complication	≥ 15%

NTCP: Normal Tissue Complication Probability, PCM: Pharyngeal Constrictor Muscle

2.5 Treatment

Patients who will be treated primarily with radiotherapy will receive 70 Gy in PTV7000 and 54.25 in PTV5425. The fractionation scheme is 35 x 2 Gy and 1.55 Gy respectively. In case of T1a and T2b tumours, a total dose of 66Gy is given in 33 fractions of 2 Gy. The total dose for adjuvant radiotherapy is 66 Gy for high risk tumours, i.e. resection margins < 1mm and/or lymph node metastasis with ENE. The dose will be given in 33 fractions of 2Gy. Tumours with intermediate risk (resection margins >1 mm without lymph nodes with ENE) after surgery receive a maximum of 56 Gy in 28 fractions of 2 Gy[17][24].

The 6 MV VMAT plans comprise a dual arc of 360 degrees with a maximum Constrained Leaf Motion (CLM) of 0.25 degrees/cm, or without a maximum CLM. A dose grid of 0.3 x 0.3 x 0.3 cm^3 was used. The proton plans use four beams, two anterior oblique and two posterior oblique beams are used. The beam angles range from 150-160 (beam 1), 40-60

(beam 2) 310-325 (beam 3) and 200-225 (beam 4).The caudal part of the target is irradiated by only the anterior beams, to avoid range uncertainties in shoulder and neck region. A range shifter of 4.0 cm was used for all beams to ensure coverage of superficial targets. A range uncertainty of 3.0 % is used to optimize and test plan robustness and radii of 3.0 and 5.0 mm are used for the setup uncertainties. Initial beam energy ranged between 70 and 225 MeV. The minimum spot weight was 0.01 Monitor Units (MU) per spot, 1 MU is defined as 1 cGy in a field of 10x10 cm^2 at 10 cm depth[33].

3. Machine learning

3.1 What is machine learning?

The field of Artificial Intelligence (AI) tends to develop a functional intelligence system what is able to argue and solve problems without the human brain. According to Arthur Samuel, a pioneer in the field of AI, ML is a field within AI that gives computers the ability to learn and improve themselves without being explicitly programmed[34]. Another field within AI is Deep Learning (DL) is based on multiple layers of neural networks to learn from large amounts of data. A DL network tends to simulate or beats the human brain in specific applications[35]. In general, ML can be used for automation of simple or repeating tasks, recognition of patterns in data, perform predictions and take complex decisions. ML has the ability to compare, analyse and classify large amounts of data fast and reliable. This section provides a concise, global overview of ML categories and applications in radiotherapy.

3.2 Machine learning categories

ML algorithms can roughly be separated into three categories: supervised learning, unsupervised learning and reinforcement learning[35]. In supervised learning, the algorithm learns how input and known, labelled, output are related to make predictions on novel input data. The algorithm compares predicted data with known data and adjusts the weights in the algorithm if necessary to create acceptable results. Supervised learning can either be a classification or a regression problem. In a classification problem, the output variable is a category, for example ‘yes’ or ‘no’. The output in a regression problem is a value like radiation dose. Unsupervised learning algorithms tend to recognize patterns in unlabelled data, this can be used for example to cluster data based on corresponding characteristics or associate rules in large portions of data, for example patients with characteristic X often have complication Y. The final category of ML, reinforcement learning, attempts to learn actions by highly repeated trial and error. Reinforcement learning is one of the methods used to beat the game of Go master Lee Sedol in 2017, a historical moment in the field of AI to outperform humans[36]. An action is either punished or rewarded, based on a time delayed reward. The goal is to maximize the cumulative reward and this type of ML differs from supervised learning by allowing sub optimal actions and does not require labelled data[37]. Further division of ML categories and algorithms can be found in appendix B.

Traditional ML algorithms perform better when a task is more repeated, however, the impact of adding more data reduces as the training data increases[38]. In ML applications should be aimed to reach the plateau of the learning curve to obtain optimal model performance. Neural networks and DL algorithms require a minimum of training data to reach sufficient model performance[39].

3.3 Machine learning in radiotherapy

ML can play a role in the field of radiotherapy by improving efficiency and especially quality of patient care in multiple components within the radiotherapy workflow. Several opportunities of the use of ML in the radiotherapy workflow are described by Feng et al.[40]. A short overview of opportunities in each step of the workflow, see figure 3, are described here.

Patient assessment

In the patient assessment is the patient informed about risks and benefits of treatment and the physician and patient discuss the patient’s goal of care and treatment strategy. ML can play a role in patient assessment by predicting treatment outcome and toxicity based on patient characteristics like treatment stage, viral status, prior and current therapies, eventual resection margins and overall performance status. Furthermore, prediction models can be used to predict the benefit and risks of concomitant chemo radiotherapy, time to pain relief and the risk of undesirable OAR complications blindness, or paraplegia for individual patients[41]. Both physicians and patients can make better decisions about the treatment since they are better informed about the probable impact of their decision[40]. However, the development of adequate prediction models for a heterogeneous patient population remains challenging and needs further investigation.

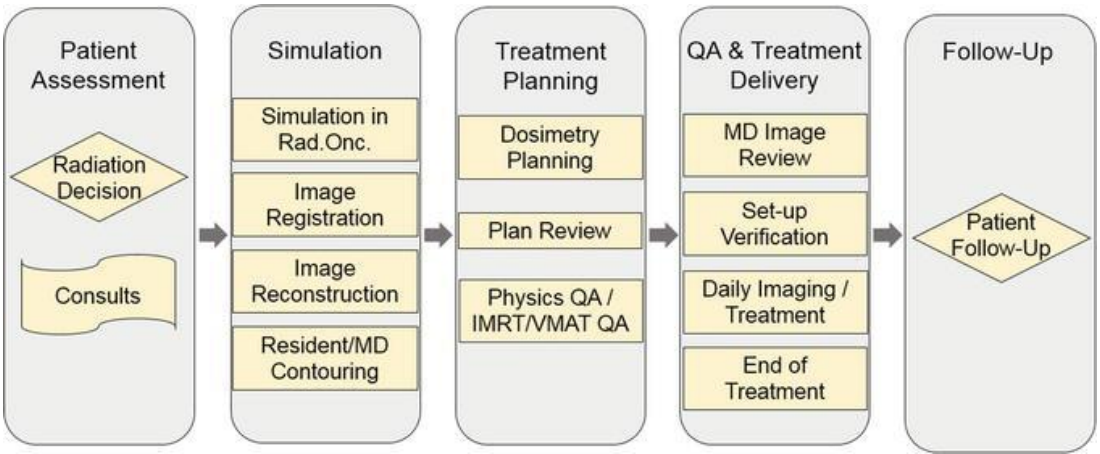

Figure 3 Schematic overview of radiotherapy workflow components; from patient assessment to follow-up. *QA: quality assessment*[40]

Simulation

In the simulation phase, images are acquired and reconstructed and tumour and OARs are delineated. The main challenge in contouring a tumour is to decide if tissue is part of the tumour or not. ML models may help a physician in this decision making[42]. Although the field of target auto-contouring is not widely investigated yet, there are some promising results. For example the study of Cardenas et al. showed close agreement in predicted high risk Clinical Target Volumes (CTVs) in oropharyngeal cancer patients compared with manual observers[43]. Automated OAR contouring using DL shows already superior performance than manual contouring in head and neck OARs[44].

Treatment planning

ML can be of added value in treatment planning by selection of an appropriate treatment technique i.e. VMAT or IMPT or in the iterative treatment planning process. The iterative process can be replaced or supplemented by machine learning optimization planning. ML plans can serve as a good starting point for complex plans or final plans in easier plans. Physicians can be more critical on radiotherapy treatment plans, since improvements in the plan do not require intensive manual labour. These factors are likely to lead to improved plan quality compared with only manual planning[6], [8]. Furthermore, ML can be used to predict patient specific plan outcome based on previous plans[45]. Once treatment plans are created and evaluated automatically, treatment technique selection can be performed easily by comparing the plan quality of multiple plans. Further optimization of IMPT plan quality can be done by predicting optimal beam angles, that is already done for Intensity Modulated RadioTherapy IMRT[46], [47].

QA and treatment delivery

Several studies have shown that ML has potential in Quality Assurance (QA) and treatment delivery: QA passing rates and performance of linear accelerators over time can be predicted[48]–[50]. These applications can help physicists to focus on outliers, the patients that impact workload the most[51]. Furthermore ML can be used to predict the need for re-planning in HNC[52].

Follow up

In the follow-up, post treatment decisions will be made, like the need of additional surgery of primary tumour and/or lymph nodes. ML can correlate patient characteristics and image features with clinical outcome[53], [54]. A drawback is the need of a large amount of patient data, which is not always available in health care.

CHAPTER 2 FULLY AUTOMATED MACHINE LEARNING BASED VMAT PLANNING FOR OROPHARYNGEAL CANCER

Ilse G. van Bruggen[1,3], Roel G.J. Kierkels[1], Roel J.H.M. Steenbakkers[1], Mats Holmström[2], David Lidberg[2], Karl Berggren[2], Cornelis H. Slump[3], Stefan Both[1], Johannes A. Langendijk[1], Fredrik Löfman[2] and Erik W. Korevaar[1]

[1]*University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, the Netherlands.* [2]*RaySearch Laboratories AB, Stockholm, Sweden.* [3]*University of Twente, Enschede, the Netherlands*

ABSTRACT

Objective To demonstrate that fully automated Volumetric Modulated Arc Therapy (VMAT) dose distributions for oropharyngeal cancer patients can be generated, with similar quality as the clinical 'dosimetrists-optimized' dose distributions, further indicated as reference plans. Furthermore, the influence of model size and composition on Machine Learning Optimization (MLO) plan quality is investigated.

Method MLO planning involved training of a model, which was used to predict the voxel dose for novel patients. CT scans, structures and dose distributions of 155 consecutive primary Head and Neck Cancer (HNC) patients, previously treated with dual arc VMAT, were retrieved from our clinical database. In the final step, the predicted dose distribution was input to a mimicking optimization to generate a deliverable dose distribution. The main goal of this study, generating clinical acceptable MLO plans, was investigated with a model containing 60 oropharyngeal cancer plans. In order to assess the effect of model size and composition on treatment plan quality, 3 additional models were trained (30 and 90 oropharyngeal cancer plans and 60 all HNC plans). Validation was performed with 39 oropharyngeal cancer patients to tune prediction and mimicking settings using both target and Organ At Risk (OAR) quality measures on models with 60 oropharyngeal cancer plans. The dose distributions of the validation plans were compared against the reference plans and the additional models.

Results The predicted dose was in accordance with reference dose for all plans. Plan quality was highly dependent on prediction and mimicking settings. In the final settings, the mimicked plans of the model with 60 oropharyngeal cancer plans had adequate target coverage, acceptable OARs dose and sum NTCP lower or within 2% increase compared to reference plans in 26/39 (67%) plans. Clinical acceptable plan quality was reached in 30/39 (77%), 31/39 (80%) and 26/39 (67%) for the models with 30, 90 and 60 all HNC plans, respectively.

Conclusion In this study, we have demonstrated that clinical acceptable MLO VMAT plan quality can be reached in the majority of oropharyngeal cancer patients. Comparable plan quality was reached as reference plans in all models. This study indicates that MLO planning can serve as a step towards automated treatment planning in radiotherapy.

Keywords: *Automated treatment planning, knowledge based planning, machine learning, machine learning optimization, random forests, dose distribution, head and neck cancer*

1. Introduction

Head and Neck Cancer (HNC) treatment planning is challenging. Organs At Risk (OARs) are located close to targets and dose related side effects in this area have a major effect on quality of life[55]–[57]. To minimize the side effects, a balance in dose distribution between target and OARs is to be found. Currently, manual treatment planning is used to create clinically acceptable radiotherapy treatment plans. Treatment planning is a time consuming process, taking hours up to days for each patient[6]. Multiple iterations are required for plan optimisation. In clinical practice, time is limited and it is shown that plan quality is correlated with time invested in an individual plan[7]. Furthermore, plan quality is dependent on the skills and experience of dosimetrists[6]. Radiotherapy treatment plan quality impacts clinical outcome in HNC patients[58]. Hansen et al. and Fogliata et al. have already shown that automation in HNC radiotherapy treatment planning may improve plan quality[6], [8].

Automated radiotherapy treatment planning has recently received much attention from the pertinent research community[6], [8], [59]. The user intervention can be eliminated entirely by using a pre-set dose volume objective list, as implemented in treatment planning system Pinnacle[6]. Other automated planning techniques, such as RapidPlan in Varian Eclipse treatment planning system, use a database of previous plans, also known as Knowledge Based Planning (KBP)[8]. RapidPlan is a Dose-Volume Histogram (DVH) based algorithm to estimate DVHs for OARs using a trained model, an inverse planning optimization algorithm is then applied[60]. Results in HNC treatment planning with RapidPlan show similar plan quality as manual optimized plans. However, a main drawback of DVH-based KBP is the lack of spatial information. Since plan quality is highly associated with target and OAR location[45], [61] using spatial information like three Dimensional (3D) dose distributions, can potentially further improve quality of automated planning. Machine Learning (ML) KBP techniques as random forest and neural networks, U-net and Generative Adversed Network (GAN), are capable to include spatial information in automated 3D dose distribution prediction[9], [59], [62]. However, neural networks and Deep Learning (DL) algorithms, require high amounts of patient data to perform well, especially in complex treatment sites such as head and neck[63]. In clinical studies, patient data can be limited and neural networks and DL may be less suitable for automated treatment planning in HNC.

Machine Learning Optimization (MLO) planning has shown to be able to predict acceptable treatment plan quality with limited amount of patient data[10]. The contextual atlas regression forest planning pipeline of McIntosh et al. is based on the assumption that patients with identical geometry and appearance should be treated in the same way[10]. A model learns how Computed Tomography (CT) data and contour features are related to dose distributions. After feature extraction, a dose distribution can be predicted on a patient and a mimicking optimization can be applied to create a deliverable plan[64].

The study of McIntosh et al. generated 12 clinical acceptable plans using a model with 54 right-sided oropharyngeal cancer treatment plans[10]. The first results are promising, however, the study from McIntosh et al. had small variability in model size and composition. Furthermore, it is unknown if the results can be translated to bilateral elective neck irradiation, the most common treatment in oropharyngeal cancer. The method of McIntosh is fully implemented in the commercial treatment planning software system of RayStation v8 (RaySearch Laboratories AB, Stockholm, Sweden). Plan quality after this implementation needs to be evaluated, which will be done in this study. The required number of training atlases is a function of the variation and complexity of a given tumour site[9]. Boutilier et al. investigated the requirement of a minimum sample size in training prostate cancer OARs models to reach clinical acceptable plans[65]. The minimum required sample size was different for each model and varied between 20 samples for the rectum and 75 to predict the bladder. Another study of McIntosh et al. showed that 40 atlases were sufficient to achieve high conformity in lung, breast and prostate[9]. Since variation in a model may impact the required number of training atlases to reach adequate model performance, model composition (i.e. single tumour sites versus multiple tumour sites in a model) should be investigated as well. Since treatment planning in the head and neck area is highly complicated, it is likely that model size and composition can influence treatment plan quality in HNC[66].

The main goal of this study is to generate clinical acceptable machine learning based VMAT treatment plans for oropharyngeal cancer in RayStation. Furthermore, the influence of the model size and composition on oropharyngeal cancer treatment plan quality are investigated. We used a fully automated machine learning based optimization approach using random forests and conditional random fields to train a model, predict voxel dose and perform a mimicking optimization in novel patients. A brief description of the method is described in section 2, as well as the used study population, plan characteristics and plan evaluation criteria. In this study, further adjustments of prediction and

mimicking settings were made, in close collaboration with the RaySearch ML team, and described in section 2.5. The results are shown in section 3 and the discussion can be found in section 4.

2. Materials and methods

2.1 Study population

In this study, we included a total of 155 HNC patients, who started curative intent (chemo)radiotherapy treatment between January 2016 and May 2019 at the UMCG. Patient selection was restricted to tumours originating in the oropharynx, larynx, oral cavity, hypopharynx and nasopharynx, an overview of patient characteristics is shown in table 3. The treatment for all patients included a dose level prescription of 7000 cGy to the primary Clinical Target Volume (CTV) and involved nodes and 5425 cGy to the bilateral elective lymph node region in 35 fractions. Patients were treated with either VMAT or IMPT depending on the result of the model-based selection procedure for IMPT. All included patients had a clinical approved VMAT plan, further indicated as 'reference plan'.

A repeated random subset validation was applied, to split patient data into training and validation sets, see figure 4. The main goal of this study, generating clinical acceptable MLO plans, was investigated using a model containing 60 oropharyngeal cancer plans. In order to assess the effect of model size and composition on treatment plan quality, three additional models were trained, see figure 4.

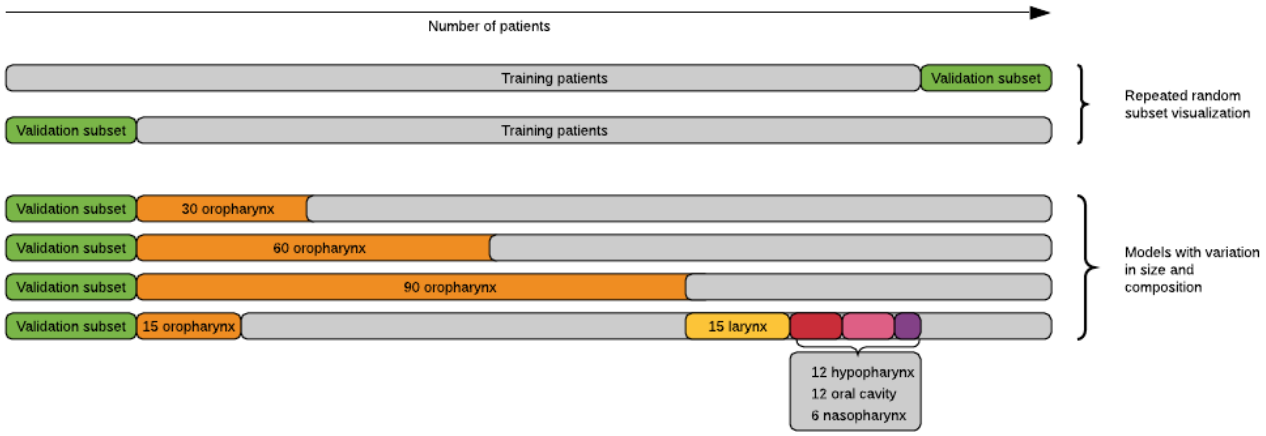

Figure 4 Upper part: schematic overview repeated random subset validation. The study population was split into two random sets of training and validation patients, the results of the two validation subsets were combined. Lower part: schematic overview of training and validation patients of one subset in four models used to investigate influence of model size (models 30, 60 and 90 oropharyngeal cancer plans) and composition (60 all head and neck cancer plans) on machine learning optimization plan quality.

2.2 Plan characteristics

The clinical VMAT plans were optimized in the RayStation treatment planning system (v4.5 and v8b RaySearch Laboratories AB, Stockholm, Sweden). A dose grid of 0.3 x 0.3 x 0.3 cm^3 was used. CTVs were expanded with 3 (planned since 2019) or 5 mm (planned before 2019) to create a planning target volume (PTV). Plans were optimized by achieving D98≥95% and D2<107% in the PTV and minimizing dose on Normal Tissue Complication Probability (NTCP) involved organs[32]. We used an evaluation structure to determine adequate target coverage in case part of the PTV is located outside or within 5 mm of the skin. Plans planned before 2019 had a CLM of 0.25 deg/cm, plans planned since 2019 had no constraint leaf motion. Final dose distributions were calculated by a collapsed cone dose engine. The 6 MV Volumetric Modulated Arc Therapy (VMAT) plans comprised a dual arc of 360 degrees were delivered by an Elekta linear accelerator.

2.3 Machine learning optimization planning

2.3.1 Training

The training of a model in MLO planning consisted of feature extraction, Atlas Regression Forest(ARF) and Prediction Random Forest (pRF)[9], [10], [67]. The features consisted of image and Region Of Interest (ROI) features. Voxel-wise features were extracted by a set of first- and second-order 3D Gaussian filters convolved with the CT image, resulting in

86 image features. The ROIs used in the training were; PTV7000, PTV5425, brain, brainstem, spinal cord, parotid glands, oral cavity, larynx, cricopharyngeal muscle, Pharyngeal Constrictor Muscles (PCMs), cervical oesophagus, mandible and thyroid. Four features per ROI were used, resulting in 64 ROI features, describing the direction and distance to the closest point of the target boundary and a signed distance transform to the target. An ARF was trained to predict dose and feature density. An ARF consisted of a set of decisions trees trained on random voxels and features from the training data. One ARF per patient was trained from the input features against the corresponding clinical dose. The ARF consisted of 96 trees with a maximum depth of 10 node levels in each tree. Selection of features in the nodes was done by maximizing the information gain at each split. The information gain was calculated for multiple features and the highest gain was used in the node. A pRF model was then trained to select the best matching ARFs, given the presence or absence of specific features. In the pRF, first, feature distance was calculated against the other ARFs. Thereafter, a Bhattacharyya distance was calculated at node level 7 for all ARF combinations[68].

The pRF was then trained to predict DVH distances at gamma 20, the clinical dose value at 20% of the prescription and above. The accuracy of an ARF could then be predicted for a patient without knowing the dose distribution, only features were used. The models were trained on an Intel Core tm i9-7940X CPU @ 3.1 GHz.

2.3.2 Prediction

Given the features of a novel patient, a dose estimation at each voxel could be predicted based on the trained model. The predicted accuracy measures were used to select the five best matching ARFs. A joint probability distribution per voxel from the five selected ARFs is used. Then, a conditional random field model was used to find the most likely spatial dose distribution while adhering to the dose prior. A predicted dose distribution from most similar atlases combined with scalar dose estimation resulted in a predicted plan.

2.3.3 Mimicking optimization

The final step in MLO planning was to perform a mimicking optimization on the predicted plan to generate a deliverable dose distribution. The beam setup was copied from the reference plans. A collapsed cone convolution dose engine was used to calculate a mimicked dose, while taken into account the delivery machine parameters, beam geometry, scatter and attenuation. Three sets of 60 iterations were used. The first 20 iterations were only in the fluence domain and during the last 40 iterations the positions of the leafs and Monitor Units (MU) were optimized. There was no CLM. The mimicking optimization used both voxel-based and DVH-based objectives. Voxel-based objectives equal or improve the predicted dose at each voxel[10]. DVH-based objectives consider the dose distribution by the DVH of each ROI[69]. Prediction and mimicking optimization were executed using remote HPE DL380 Gen9 servers with 2 Intel Xeon E5-2698v4-core CPU processors.

2.4 Analysis

The dose distributions (i.e. predicted and mimicked dose) of the plans of the model with 60 oropharyngeal cancer plans were compared against the reference plans. The plans of the models with 30 , 90 and all HNC plans were compared against mimicked plans of the model with 60 plans. Evaluation parameters were determined by simulating physicians's plan observation method by calculating and translating plan approval observations into measurable parameters. All plans were evaluated by means of dosimetric parameters (evaluation structures of PTV7000 and PTV5425: D98, D2; OARs: Dmean or D0.1), NTCP for xerostomia, grade 2-4 dysphagia, and Percutaneous Endoscopic Gastrostomy (PEG) tube dependence and MU. Target structures were contracted by 1mm and evaluated on D99.9 to examine gaps in dose distribution, as a surrogate parameter for visual dose distribution observation.

Furthermore, the Homogeneity Index (HI), HI=(D2-D98)/D50 and Conformity Index (CI), CI = TV_{95}/V_{95} were calculated. TV95 was the volume within the 95% isodose line of the target and V95 the volume in the patient what receives 95% or more of the dose prescription[70]. The OARs evaluated in this study were: brainstem, brain, spinal cord, eyes posterior, eyes anterior, parotid glands, submandibular glands, oral cavity, PCM superior-medius-inferior, cricopharyngeal muscle, supraglottis, glottis and thyroid. All dosimetric parameters were extracted from RayStation with RaySearch analytics software and visualised in Tableau (version 2019.1.0, Tableau Software Inc., Seattle, United States of America).

A plan was considered clinical acceptable when criteria as listed in table 4 are met. The threshold of accepting 2% increase in sum NTCP values was chosen as a non-inferiority margin to show that MLO plans perform not clinically relevant worse than reference plans. Evaluation parameters not included in this table were considered plan quality objectives instead of hard constraints since these are clinically less relevant(anterior eyes, CI, HI, non-NTCP OARs) and/or exact criteria have not yet been determined (hotspots in PTV 5425, D99.9 in 1mm contracted target structures).

Table 3 Overview of patient characteristics for all models. The study population was split into two random sets of training and validation patients, this table shows patient characteristics of each training subset and the validation subsets.

		Model 30		Model 60		Model 90		Model all HNC		Validation	
Characteris-tics		subset 1 (n)	subset 2 (n)	subset 1 (n)	subset 2 (n)	subset 1 (n)	subset 2 (n)	subset 1 (n)	subset 2 (n)	subset 1 (n)	subset 2 (n)
Patients		30	30	60	60	90	90	60	60	23	16
Sex	Male	19	19	39	35	58	58	41	46	19	10
	Female	11	11	21	25	32	32	19	14	4	6
Age	≤59	12	15	23	29	43	47	17	16	10	4
	>60	18	15	37	31	47	43	33	34	13	12
Tumour site	Oropharynx	30	30	60	60	90	90	15	15	39	39
	Larynx	0	0	0	0	0	0	15	15	0	0
	Oral cavity	0	0	0	0	0	0	12	12	0	0
	Hypopharynx	0	0	0	0	0	0	12	12	0	0
	Nasopharynx	0	0	0	0	0	0	6	6	0	0
Pathological T-stage	T1	6	4	14	13	19	18	7	8	1	1
	T2	5	7	8	8	15	18	11	9	7	2
	T3	3	6	8	6	11	12	18	23	3	2
	T4*	16	13	30	33	45	42	14	20	13	11
Pathological /clinical N-stage	N0	2	3	6	8	9	10	21	17	3	1
	N1	5	4	9	10	13	13	10	12	4	3
	N2**	17	18	34	35	53	53	16	28	14	9
	N3	6	5	11	7	15	14	13	3	2	3
Target location	Left	11	11	23	19	32	30	14	16	4	6
	Right	7	8	15	16	25	23	21	19	11	6
	Middle	12	11	22	33	33	37	15	15	8	6

*PTV: Planning Target Volume, *T4 includes T4a, T4b, TNOS, ** N2 includes N2a, N2b, N2c and N2NOS,*

Table 4 Criteria for clinical acceptable plan quality in machine learning optimized plans[24], [28]	
ROI	**Clinical acceptable threshold**
PTV 7000	D_{98} ≥ 6650 cGy
	D_{mean} ≤ 7050 cGy
	D_{mean} ≥ 6950 cGy
	D_2 ≤ 7490 cGy
PTV 5425	D_{98} ≥ 5154 cGy
Brainstem	$D_{0.1}$ ≤ 6300 cGy
Brain	$D_{0.1}$ ≤ 6300 cGy
Spinal cord	$D_{0.1}$ ≤ 5400 cGy
Eyes posterior	$D_{0.1}$ ≤ 6000 cGy
Sum NTCP	(Sum $NTCP_{MLO}$-Sum $NTCP_{ref}$) ≤ 2%
ROI: Region Of Interest, PTV: Planning Target Volume, NTCP: Normal Tissue Complication Probability, MLO: Machine Learning Optimization. Note that the clinical acceptable threshold for sum NTCP was turned when clinical acceptable plan quality was determined in reference plans	

Only mimicked plans were evaluated on significant differences. Two-tailed p-values were calculated by a paired Wilcoxon signed-rank test. A Bonferroni correction was used to correct p-values for multiple experiment-wise testing (multiple parameters), since a large number of independent tests are performed and the results of all tests combined are relevant in this study[71][72]. Family-wise testing (across multiple models) was not corrected by Bonferroni because the study is restricted to four pre-planned comparisons and known hypotheses. Differences were considered statistically significant if $p<0.006$($\alpha=0.05/9$ parameters) for target coverage, $p<0.01$($\alpha=0.05/5$ structures) for maximum OAR dose, $p<0.005$($\alpha=0.05/10$ structures) for average OAR dose, $p<0.013$($\alpha=0.05/4$ NTCPs) for NTCPs and $p<0.016$($\alpha=0.05/3$ parameters) for other plan evaluation parameters. Statistical analyses were performed by SPSS (IBM Corp. Released 2015. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY).

2.5 Tuning MLO plans

Prediction and mimicking settings can be tuned to optimize the MLO plan. Each voxel has a most probable value, which is shown in a predicted plan. In addition to that, each voxel has less likely, but possible values. The dose distribution of a predicted plan can be influenced by adjusting clinical constraints and objectives in the prediction phase. During mimicking optimization, target and OARs weights can be added and adjusted to impact dose value selection of the predicted dose in each voxel. Tuning of prediction and/or mimicking settings will influence the trade-off between target coverage and OARs sparing of the final MLO plan. The optimal trade-off is dependent on institution's and physician's preferences and patient's tumour characteristics.

MLO planning is able to generate multiple plans for each patient, focussing on target coverage, OAR sparing or both, based on specific preferences. A standard strategy can be used to achieve both adequate target coverage and OARs sparing. Other strategies can reach additional target coverage while accepting higher OAR dose or focus on sparing critical OARs like brain, brainstem and spinal cord while allowing less target coverage.

2.5.1 Requirements

In this study, MLO planning was validated by tuning prediction and mimicking optimization settings in 39 patients to reach clinical acceptable MLO plans using the model with 60 oropharyngeal cancer plans. MLO planning will supplement or replace manual treatment planning, therefore we determined that MLO plans has to equal or improve clinical relevant parameters: time (hands-on time and calculation time), ease of use, quality and reliability. These parameters were translated into the following requirements to validate MLO:

- MLO should reach clinical acceptable plan quality in 90% of the plans
- MLO shoud reach clinical acceptable plan quality within three strategies

In these requirements, hands-on time and calculation time will be non-inferior to manual treatment planning. In 90% of the plans the hands-on time and calculation time will be decreased from 4 hours to 15 minutes and 4 hours to at maximum 3 hours, respectively. MLO plans can be generated during the night so that in the 10% failing MLO plans a manual treatment plan can be created in the next day without losing time in the treatment preparation. Ease of use of MLO will be guaranteed in the requirements by demanding only a few clicks to start a strategy. A maximum of three strategies is chosen to keep plan quality review achievable for dosimetrists and physicians. Reliability will be reached by consistency of MLO planning, in at least 90% of the plans, clinical acceptable plan quality is achieved.

2.5.2 Tuning approach

A schematic overview of the tuning approach can be found in figure 5. A subset of 5 patients of the validation set was used as starting point for tuning prediction and mimicking settings. Plan quality of both predicted and mimicked plans were evaluated. In case predicted or mimicked plans were clinical unacceptable, prediction and mimicking optimizations settings, respectively, were adjusted until clinical acceptable plans were reached. Furthermore, plan quality objectives were evaluated and optimized if possible. Thereafter, the settings were applied to a larger subset or total set of 39 validation patients. The settings were iteratively adjusted. The tuning was done on model level, contrary to manual treatment planning where tuning is done per plan. This process was repeated until 90% of all plans in the validation set reached clinical acceptable plan quality. In this study, only one standard strategy is used due to time constraints. No additional tuning was done between models; the prediction and mimicking settings were the same across all 39 patients generating plans using other models.

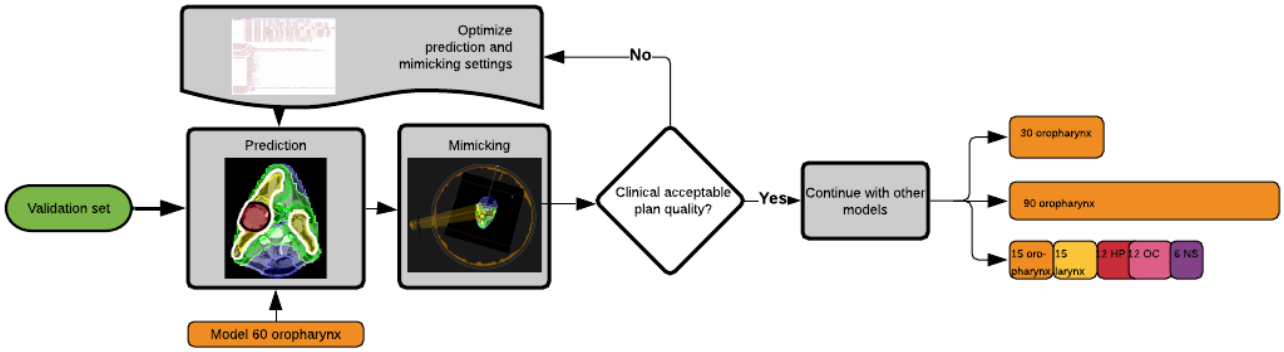

Figure 5 Flowchart of tuning approach prediction and mimicking settings. The input was the validation set containing 39 oropharyngeal cancer patients. Machine learning optimization plans were predicted using the model with 60 oropharyngeal cancer plans and then mimicked. If plan quality was unacceptable, prediction and mimicking settings are optimized, this process was repeated until the plan quality was sufficient. Then, the validation set was input in the other models with variation in size and composition.

3. Results

Training of the model with 60 plans took 48 hours. Prediction was executed in 5-10 minutes and mimicking optimization was finished in 50 minutes, fully automated. Manual treatment planning took approximately 240 minutes hands-on time per plan.

3.1 Tuning results

An overview of the number of clinical acceptable plans and results during the tuning process is shown in table 5 and figure 6. The first prediction and mimicking settings(1.1) resulted in clinical unacceptable plans for the 5 patients of the subset since NTCP values were higher than observed in the reference plans. Both predicted and mimicked plans of 1.1 were clinical unacceptable, therefore prediction and mimicking optimizations settings were adjusted. In further tuning only mimicking optimization settings were adjusted, since predicted plan quality was clinical acceptable. The adjusted settings (1.2) resulted in better OAR sparing, however, unacceptable hotspots were observed in PTV 5425 and the spinal cord. The second round of adjustments (1.3) resulted in inadequate target coverage of PTV 5425. The final round (1.4) showed clinical acceptable plans in 26/39 (66.7%) of the MLO plans. The complete final settings can be found in appendix C.

Table 5 Results of four rounds in tuning approach of prediction and mimicking settings. The table shows the used prediction and mimicking settings, the number of clinical acceptable plans reached, the main reason for unacceptable plan quality and indicates prediction or mimicking problem for each round of tuning.

Round	Prediction and mimicking settings	Clinical acceptable plans/total (%)	Reason unacceptable plan quality (number of unacceptable plans)	Prediction or mimicking problem?
1.1	PTV 7000: 2 PTV 5425: 4 All OARs: 1 DecreaseMaxdose: 0.5	0/5 (0%)	$NTCP_{MLO}$ >$NTCP_{ref}$ (n=5)	Prediction and mimicking
1.2	*CG: overlap parotid glands and PTV: D10<1300cGy* PTV 5425: 8 Parotid glands: 12 PCM superior: 6 Crico: 6 PCM inferior: 6	5/39 (13%)	D2 PTV 5425 >107% (n=35) D0.1 spinal cord >5400 (n=1)	Mimicking
1.3	PTV 7000: 5 PTV 5425: 11 DecreaseMaxdose: 2	5/7 (71%) 28/35 (80%)	D98 PTV 7000 <95% (n=1) D0.1 spinal cord >5400 (n=1) D98 PTV 7000 <95% (n=3) D98 PTV 5425 <95% (n=4) D0.1 spinal cord >5400 (n=1)	Mimicking Mimicking
1.4	PTV 7000: 10 PTV 5425: 15 Spinal cord: 3	6/7 (86%) 26/39 (67%)	D0.1 spinal cord >5400 (n=1) D98 PTV 5425 <95% (n=2) D0.1 spinal cord >5400 (n=1) $NTCP_{MLO}$>$NTCP_{ref}$ (n=12)	Mimicking Mimicking

PTV: Planning Target Volume, PCM: Pharyngeal Constrictor Muscle, NTCP: Normal Tissue Complication Probability, MLO: Machine Learning Optimization, ref: reference plans, *Italic* settings indicate prediction settings.

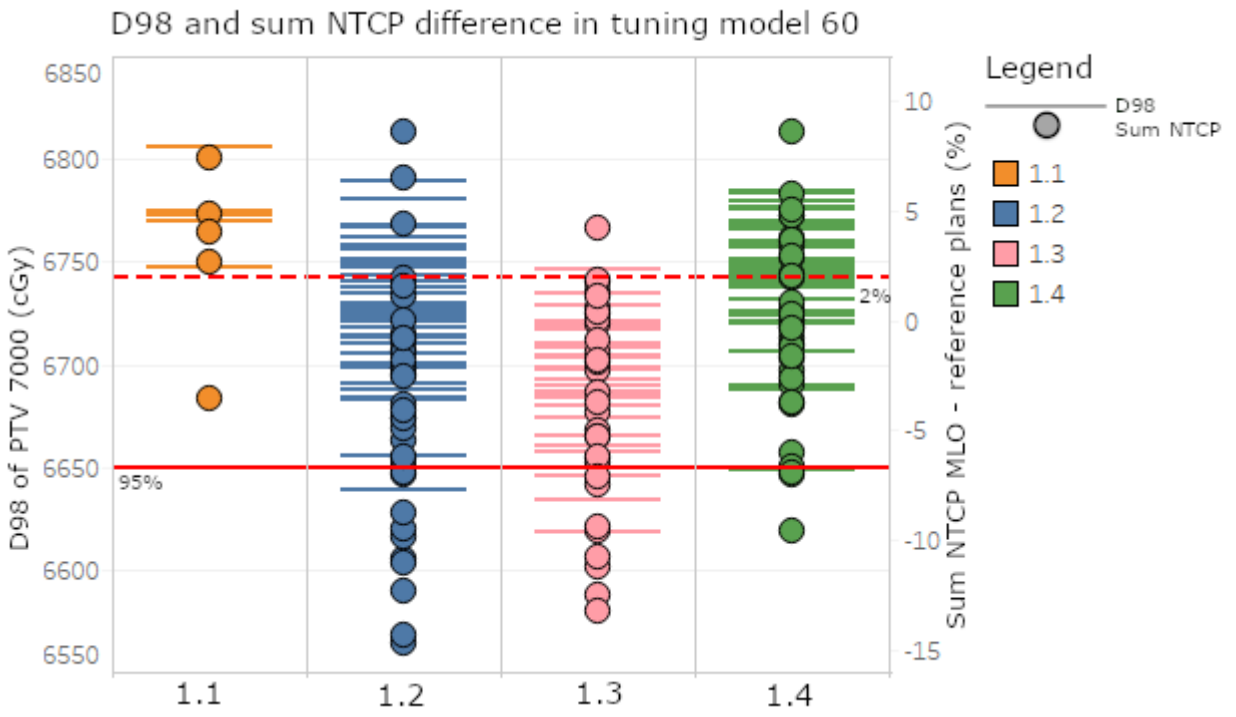

Figure 6 D98 and sum NTCP difference of validation patients for each round in the tuning approach. Each plan is shown by lines (D98) and dots (Sum NTCP). The solid red line indicates the D98 threshold of 95% and the dashed red line indicate the 2% threshold for sum NTCP. Note that the plans of round 1.2 had unacceptable high D2 values in PTV 5425, not visualized in this figure. *NTCP: Normal Tissue Complication Probability, PTV: Planning Target Volume*

3.2 Clinical acceptable plan quality

All predicted plans had higher plan quality regarding all evaluation parameters compared to reference and mimicked plans, see table 6. All MLO and reference plans reached adequate target coverage in terms of D98 for both high risk and elective volumes. The average dose at high risk volume was significant lower (Δ=27 cGy, p=0.000) in MLO plans and within the range (6950-7050 cGy) for 38/39 plans for both MLO and reference plans. One MLO plan had minor underdosage (6938 cGy) and one reference plan had minor overdosage (7052 cGy). Figure 7 shows comparable target DVHs for reference and mimicked plans. The maximum dose on high risk target was not exceeded in all MLAP and reference plans. The maximum dose in the spinal cord was significant higher in MLAP plans (Δ=539 cGy, p=0.000). One MLO plan exceeded the maximum dose on spinal cord (D0.1=6286 cGy). Other maximum doses for brain, brainstem, spinal cord and posterior eyes were not exceeded.

In MLO and reference plans, the contracted high risk volume had adequate target coverage in 36/39 (92%) and 31/39 (80%) respectively. Adequate target coverage in the contracted elective volume was achieved in 5/39 (12.8%) and 28/39 (71.8%) for MLO and reference plans respectively. The average difference was significant (Δ=60 cGy, p=0.000), an example of inadequate elective target coverage is shown in figure 8 by patient B. The D2 was significant higher in MLO in elective target volumes compared to reference plans (Δ=175 cGy, p=0.000, xxx), see figure 8 patient C for an example. High risk and low risk target homogeneity was significant lower in MLO plans compared to reference plans (high risk: Δ=0.013, p=0.001, low risk: Δ=0.010, p=0.001). The sum NTCP value was lower or within 2% increase from the value observed in the reference plan in 27/39 (69.3%) MLO plans. Xerostomia NTCP values of the MLO were lower(Δ=1.6%, p=0.006) and PEG NTCP values were higher (Δ=0.5%, p=0.003) compared to the reference plans. The average dose on OARs is lower in MLO in parotid glands (Δ=156 cGy, p=0.000) and higher in PCM superior (Δ=168 cGy, p=0.000) and PCM medius (Δ=359 cGy, p=0.000) and supraglottic (Δ=311 cGy, p=0.000). The average dose of other OARs is comparable in MLO and reference plans. The maximum dose on anterior eyes was exceeded in 3/39 (7.7%) MLO plans and 2/39(5.1%) reference plans. MU were significant higher in MLO (Δ=31, p=0.010). Clinical acceptable plan quality was reached in 26/39 (66.7%) of the reference plans, the same as in MLO plans.

3.3 Model size

Training of the models with 30 and 90 plans took 22 and 90 hours respectively. Predicted plans generated by models with 30 and 90 plans were comparable to the predicted plans generated by the model with 60 plans. An overview of results of model size can be found in table 6 and figure 9. Adequate target coverage in D98 was reached in all plans in both high and low risk target volumes in plans of model 30. The maximum dose on spinal cord was exceeded in two plans (D0.1 = 6160 cGy and 5612 cGy). Plans generated by the model with 30 plans had significant lower PEG NTCP values (Δ=0.41%, p = 0.004) and lower elective target conformity (Δ=0.006, p=0.012) compared to plans generated by the model with 60 plans. Sum NTCP values were lower or within 2% increase compared to the reference plans in 30/39 (77%) plans. Clinical acceptable plan quality was reached in 30/39 (77%) plans generated by the model with 30 plans.

All target coverage parameters were significant worse for the MLO plans of model 90 compared to MLO plans of model 60, except for D2 for high and low risk target which were significant lower (p = 0.005) in the plans of model 90. Adequate target coverage was reached in 35/39 (90%) and 36/39 (92%) for high risk and low risk target volumes, respectively, in plans of model 90. The maximum dose on the spinal cord was exceeded in one patient (D0.1=6003 cGy). The D0.1 was significant lower in model 90 plans compared to model 60 plans (Δ=325 cGy, p=0.000). Average dose on parotid glands, submandibular glands and PCM superior were lower in the plans of model 90 (Δ=66 cGy, Δ=53 cGy, Δ=72 cGy respectively, all p=0.000). Furthermore, all NTCP values were lower in the model with 90 plans (Xerostomia: Δ=0.8%, Dysphagia: Δ=0.6%, PEG: Δ=0.9%, Sum NTCP: Δ=2.1%, p=0.000-0.001). Plans of model 90 had lower or within 2% increase sum NTCP values compared to reference plans in 35/39 (90%). Clinical acceptable plan quality was reached in 31/39 (80%) plans of model 90.

3.4 Model composition

Predicted plan quality of the plans generated by the model with all HNC plans were comparable with the plans generated by the model with 60 oropharyngeal cancer plans. All model all HNC plans reached adequate target coverage in both high risk and elective target volumes. Only high risk target homogeneity and conformity were lower in the all HNC model plans (HI: Δ=0.003, p=0.002, CI: Δ=0.010, p=0.000). Two plans exceeded the maximum dose on the spinal cord (D0.1 = 6253 cGy and 5555 cGy). In 27/39 (69%) of the all HNC model plans was clinical acceptable plan quality reached.

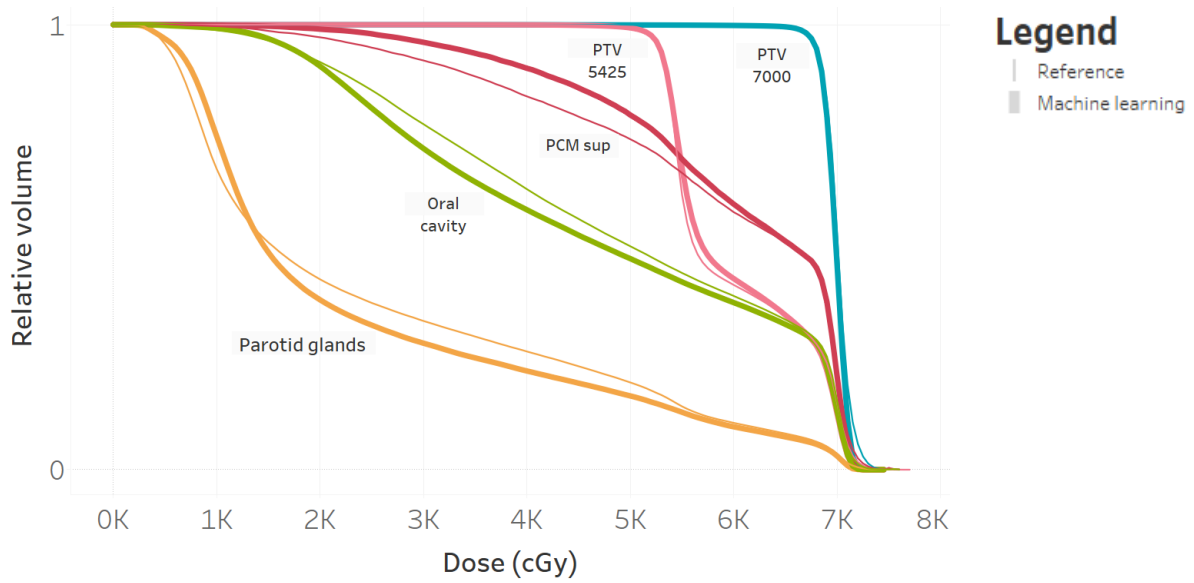

Figure 7 Average DVH of PTV 7000 (blue), PTV 5425 (pink), Pharyngeal Constrictor Muscle (PCM) superior (red), oral cavity (green) and parotid glands (orange) of 39 validation patients using the model with 60 oropharyngeal cancer treatment plans

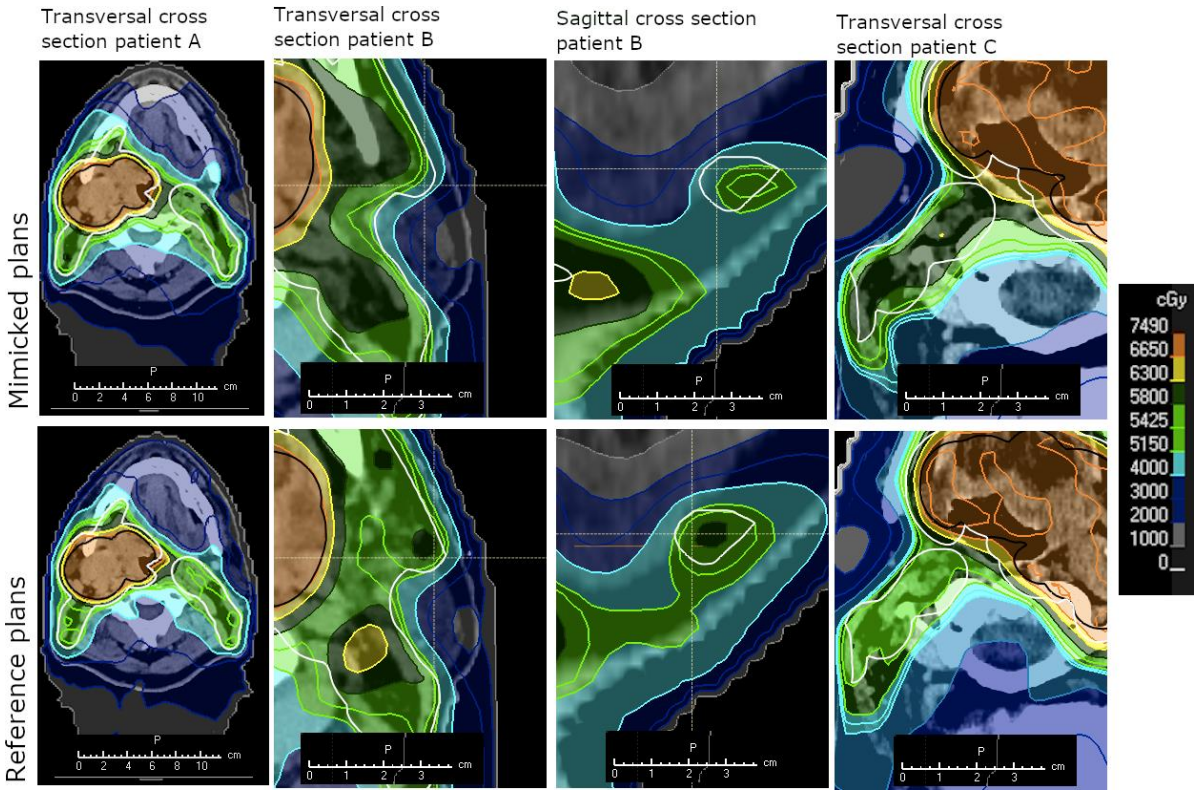

Figure 8 Transversal and sagittal cross sections of three model 60 mimicked and reference dose distributions. PTV 7000 is indicated in black contours and PTV 5425 in white contours. Patient A is a representative patient regarding OARs dose, patient B visualizes isodose line gaps of in this case 3mm in PTV 5425 in two dimensions and patient C shows high dose in elective target volume. D98 and difference in sum NTCP of the mimicked plans are shown in figure 9. *PTV : Planning Target Volume*

Table 6 Average results and standard deviation of evaluation on dosimetric parameters, HI, CI and NTCP values of predicted and mimicked plans of all models

	Parameters	Reference	Prediction model 30	Mimicking model 30	Prediction model 60	Mimicking model 60	Prediction model 90	Mimicking model 90	Prediction model all HNC	Mimicking model all HN
Target coverage	PTV 7000 D_{98} (cGy)	6737 ± 37	6764 ± 30	6744 ± 25	6766 ± 32	6745 ± 28	6756 ± 33	**6697 ± 52****	6741 ± 65	6723 ± 61
	PTV 7000 D_{mean}(cGy)	7010 ± 27	6985 ± 8	6987 ± 12	6983 ± 11	**6983 ± 15***	6981 ± 13	**6965 ± 22****	6945 ± 62	6949 ± 62
	PTV 7000 D_2 (cGy)	7238 ± 89	7212 ± 25	7163 ± 15	7204 ± 33	**7156 ± 23***	7201 ± 33	7154 ± 24	7152 ± 81	7112 ± 67
	PTV 7000-1mm $D_{99.9}$ (cGy)	6646 ± 127	6724 ± 39	6700 ± 29	6723 ± 38	6698 ± 36	6712 ± 40	**6644 ± 59****	6700 ± 67	6676 ± 63
	PTV 7000 HI	0.080 ± 0.03	0.075 ± 0.04	0.068 ± 0.04	0.073 ± 0.1	**0.067 ± 0.04***	0.075 ± 0.04	**0.074 ± 0.04****	0.071± 0.03	**0.064 ± 0.03****
	PTV 5425 D_{98} (cGy)	5243 ± 41	5279 ± 28	5230 ± 30	5276 ± 31	5226 ± 31	5249 ± 34	**5198 ± 41****	5284 ± 28	5228 ± 33
	PTV 5425-PTV 7000 D_2 (cGy)	5922 ±112	5750 ± 113	6098 ± 82	5771 ± 94	**6097 ± 118***	5791 ± 79	**6003 ± 112****	5764 ± 75	6122 ± 105
	PTV 5425-1mm $D_{99.9}$ (cGy)	5144 ± 88	5214 ± 53	5092 ± 57	5209 ± 59	**5084 ± 76***	5136 ± 83	**5030 ± 98****	5198 ± 44	5080 ± 67
	PTV 5425 HI	0.345 ± 0.026	0.339 ± 0.025	0.336 ± 0.025	0.335 ± 0.025	**0.335 ± 0.025***	0.343 ± 0.027	**0.343 ± 0.027****	0.335 ± 0.027	0.333 ± 0.027
$OARD_{0.1}$ (cGy)	Brain	3013 ± 1122	2958 ± 562	2790 ± 1111	2710 ± 751	**2769 ± 1149***	2518 ± 723	2585 ± 1112	3010 ± 632	2815 ± 1147
	Brainstem	3380 ± 1067	3433 ± 395	3262 ± 1174	3146 ± 467	3172 ± 1183	3099 ± 509	2923 ± 1099	3298 ± 482	3271 ± 1251
	Spinal cord	4265 ± 382	4226 ± 200	4911 ± 441	4085 ± 253	**4804 ± 485***	3997 ± 257	**4479 ± 451****	4182 ± 215	4867 ± 494
	Eyes anterior	219 ± 200	128 ± 65	230 ± 266	109 ± 75.0	224 ± 261	101 ± 55	216 ± 241	118 ± 124	228 ± 271
	Eyes posterior	259 ± 213	183 ± 68	309 ± 412	157 ± 98.0	305 ± 417	149 ± 70	289 ± 378	158 ± 116	299 ± 383
OAR mean dose (cGy)	Parotid glands	2584 ± 836	1809 ± 789	2401 ± 819	1817 ±798	**2428 ± 812***	1802 ± 777	**2362 ± 807****	1806 ± 793	2423 ± 815
	Submandibular glands	5785 ± 1239	5803 ± 922	5941 ± 980	5720 ± 1063	5914 ± 1029	5764 ± 955	**5861 ± 989****	5756 ± 931	5908 ± 984
	Oral cavity	4691 ± 992	4514 ± 798	4639 ± 845	4384 ± 90 1	4573 ± 906	4412 ± 859	4548 ± 874	4363 ± 932	4574 ± 917
	PCM superior	5824 ± 904	5661 ± 785	5981 ± 724	5631 ± 816	**5992 ± 716***	5655 ± 808	**5920 ± 744****	5617 ± 811	5954 ± 738
	PCM medius	5509 ± 1008	5477 ± 1007	5872 ± 694	5134 ± 1016	**5868 ± 660***	5231 ± 968	5820 ± 697	5249 ± 905	5904 ± 644
	PCM inferior	4068 ± 1239	4011 ± 1236	4154 ± 1094	3725 ± 1113	4224 ± 1062	3707 ± 1061	4120 ± 1042	3801 ± 1077	4308±1009
	Cricopharyngeal muscle	3142 ± 1047	3151 ± 1048	3247 ± 803	2603 ± 675	3373 ± 791	2398 ± 605	3157 ± 722	2663 ± 869	3430 ± 839
	Supraglottic	4946 ± 1239	4908 ± 1216	5219 ± 1032	4586 ± 1232	**5257 ± 982***	4589 ± 1215	5185 ± 999	4685 ± 1205	5346 ± 939
	Glottic	3662 ± 1340	3635 ± 1340	4092 ± 1019	2909 ± 1126	4160 ± 976	2965 ± 1001	4052 ± 920	3033 ± 1199	4235 ± 971
	Thyroid	5109 ± 480	5100 ± 481	5187 ± 396	4934 ± 492	5152 ± 411	4831 ± 467	5076 ± 425	4949 ± 443	5168 ± 414
NTCP (%)	Xerostomia	39.3 ± 6.2	30.55 ± 4.39	37.30 ± 5.96	30.7 ± 4.2	**37.7 ± 5.7***	30.61 ± 4.08	**36.89±5.65****	30.59 ± 4.28	37.73 ± 6.02
	Dysphagia	32.0 ± 8.5	30.02 ± 7.26	32.31 ± 7.40	29.4 ± 7.7	32.2 ± 8.5	29.53 ± 7.52	**31.55±7.49****	29.29 ± 7.74	32.04 ± 7.62
	PEG	14.5 ± 5.1	10.48 ± 3.40	**14.49±4.62****	10.9 ± 3.7	**14.9 ± 4.6***	10.63 ± 3.32	**13.99±4.35****	11.02 ± 3.64	15.05 ± 4.74
	Sum NTCP	84.9 ± 19.4	71.0 ± 15.0	84.1 ± 18.0	70.7 ± 15.2	84.5 ± 17.8	70.8 ± 14.9	**82.4 ± 17.6****	70.9 ± 15.5	84.8 ± 18.3
Plan evaluation	PTV 7000 CI95%	0.818 ± 0.038	0.854 ± 0.038	0.820 ± 0.038	0.850 ± 0.040	0.818 ± 0.038	0.851 ± 0.037	**0.843 ± 0.040****	0.845 ± 0.040	**0.808 ± 0.040****
	PTV 5425 CI95%	0.826±0.067	0.850±0.056	**0.817 ± 0.063****	0.857 ± 0.054	0.823 ± 0.060	0.832 ± 0.056	0.826 ± 0.062	0.854 ± 0.056	0.718 ± 0.058
	Monitor units (#)	349 ± 98.5	-	384 ± 37.5	-	**380 ± 38.8***	-	387 ± 38.3	-	378 ± 33.8
	Total planning time (min)	~240	10 ± 5	55 ±10	10 ± 5	55 ±10	10 ± 5	55 ±10	10 ± 5	55 ±10

PTV: Planning Target Volume, PCM: Pharyngeal Constrictor Muscle, NTCP: Normal Tissue Complication Probability, PEG: Percutaneous Endoscopic Gastrostomy

*Significant difference compared to reference plans

** Significant difference compare to mimicked model 60 machine learning optimized plans

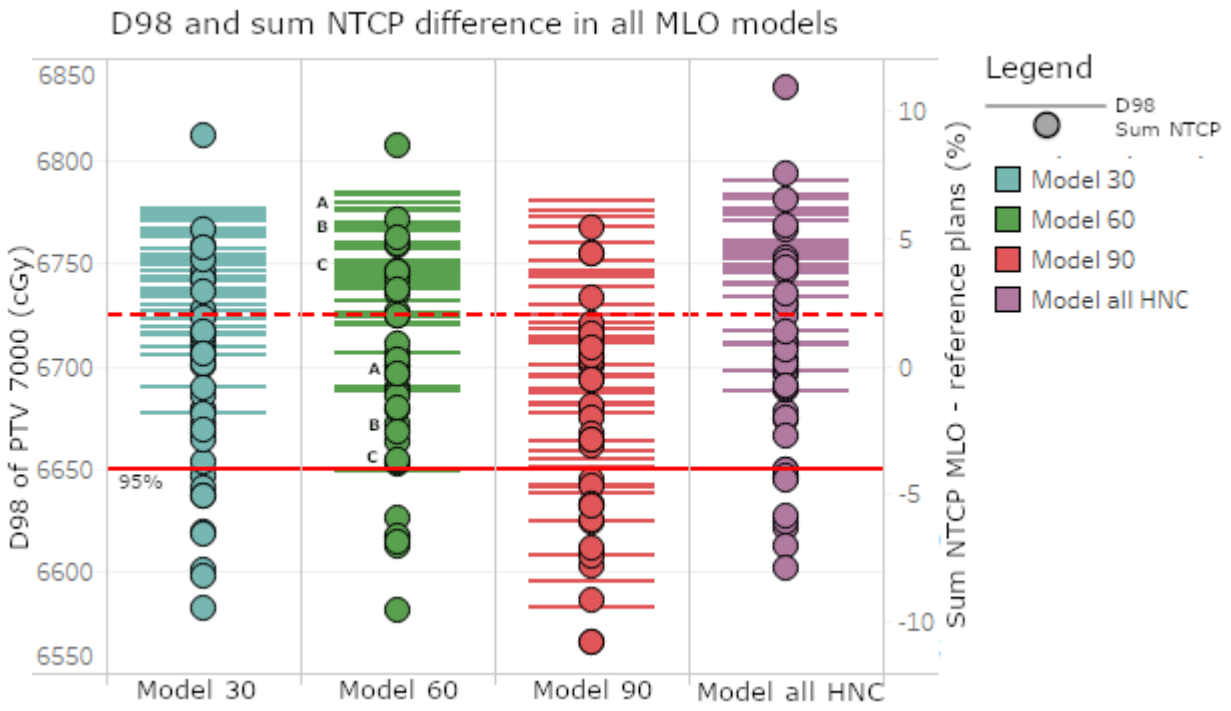

Figure 9 D98 and sum NTCP difference of validation patients for all models. All plans are shown by lines (D98) and dots (Sum NTCP), A,B and C indicate the plans shown in figure 8. The solid red line indicates the D98 threshold of 95% and the dashed red line indicate the 2% threshold for sum NTCP. *NTCP: Normal Tissue Complication Probability, PTV: Planning Target Volume*

4. Discussion

In this study, machine learning optimization planning has shown to be able to achieve clinical acceptable treatment plan quality in the majority of the oropharyngeal cancer plans (26/39 (67%)). We have compared MLO plans of a model with 60 plans with the reference plan and evaluated on dosimetric parameters, CI, HI and NTCP. Tuning of a model using prediction and mimicking settings highly influenced plan quality. Furthermore, an increase in model size showed lower target coverage and NTCP values in the plans generated by a model with 90 plans compared to a plans generated by a model with 60 plans. Furthermore, the influence of model composition on oropharyngeal cancer treatment plan quality was investigated, the plan quality of a model with all HNC plans was similar to the model containing 60 oropharyngeal cancer plans. This study showed that fully automated machine learning optimization planning has potential to achieve similar plan quality as dosimetrists optimized plans.

The predicted plans of our study were all clinical acceptable, unlike the mimicked plans. The mimicking optimization of our study was not able to realize the predicted plan. It is likely that this can be explained by a too optimistic predicted plan since mimicked plan quality is highly comparable with reference plans. This finding is in contradiction with the results in the study of McIntosh et al., they found that almost all mimicked plans had higher plan quality compared to predicted plans[10]. However, the results of adequate target coverage and acceptable OARs dose in the majority of the mimicked plans are in accordance. They found that their automated method can create comparable mimicked dose distributions to clinical. Furthermore, the results of our study are in line with other studies, an overview of results of these studies can be found in appendix E[60], [73]–[75]. For example, the study of Babier et al., who used DVH predictions to generate comparable plans as clinical plans[75]. This study has not been able to generate improved plan quality by decrease in NTCP values, as shown by fogliata et al.[8]. A possible explanation for that may be a relatively high plan quality of reference plans in our institute.

Consistent with the literature, this study found that a model with 30 plans was able to generate clinical acceptable plans[9], [60]. Furthermore, we found that the results of model with 30 plans were comparable with the results from the model with 60 plans. This finding was also reported by the study of Tol et al.[60]. The results of the model with 60

all HNC plans confirm the finding that smaller training atlases are sufficient to reach acceptable plan quality. The plan quality of outliers is not investigated in this study.

A strength of this study is the use of a relatively large validation set compared to some other studies[10]. A larger validation set will result in robust results of a test set and generalizability of the MLO results for novel patients. Especially for outliers, this can be of benefit, since outliers can be recognized and handled in a large validation set. Another strength is that this is the first study, to our knowledge, that compares influence of model size and composition on plan quality in machine learning automated planning studies for HNC.

The study has a few limitations. First, the study lacks an independent test set, which is recommended to verify the results. However, validation plans have shown to be clinical acceptable in already 26/39 plans, similar as in reference plans, and therefore it is expected that the prediction and mimicking settings are robust enough to create clinical acceptable MLO plans in the majority of novel patients. Another limitation of the study is copying the clinical beam orientations of the reference plan. However, beam orientations in oropharyngeal cancer treatment plans are highly standardized. An option to overcome this limitation is to extract the beam orientation from the five selected ARFs in the prediction. A limitation in the plan evaluation method of this study is the lack on hotspot size and location information. The D2 of the elective target volume of MLO plans are significant higher and it is possible that larger hotspots are present and are located in undesired areas like OARs, as shown by patient C in figure 8. It is recommended to implement hotspot size and location in evaluation parameters, for example by using distance to OARs and targets and CT density.

An important finding of this study was that plan quality was highly dependent on prediction and mimicking settings and the model used for tuning. The results of the model containing 90 plans had significant lower target coverage and lower OARs dose, than in the model with 60 plans, shown in figure 9. An explanation for that can be that the mimicking weight on OARs is too high or target weight too low. In a model with 90 plans more similar ARFs are selected and more options for sparing may be available. It is recommended to tune prediction and mimicking settings on each model.

One unanticipated finding was the exceeded spinal cord dose of one patient (MLO plan: 6004 cGy, ref: 5404 cGy). The overlap between the spinal cord and elective target volume can be an explanation for this result, the MLO is not able to create a steep dose fall off near the PTV, since this is currently not included in the objective list. Furthermore, target underdosage in PTV overlap with a critical OAR like the spinal cord is allowed, but not implemented yet in the MLO method. It is recommended to evaluate if and how a steeper dose fall off can be achieved in MLO and adjust the MLO algorithm to allow target underdosage in comparable patients, for example by a critical OAR dose MLO strategy. In case this strategy will not result in clinical acceptable plans, these patients can have manual adjustments after MLO planning or a fully manual plan. It is likely that when only one area has inadequate dose, MLO planning can give a good starting point for manual post processing.

It is interesting to note that elective target volume objectives of D98 were reached comparable in MLO plans (37/39, 95%) and reference plans (38/39, 97%) but the D99.9 of the contracted PTV5425 decreased significant with 60 cGy on average. This implicates that more gaps in isodose line are present in MLO plans, an example is shown in figure 8 by patient B. An explanation for this finding is that a gap larger than 1 mm in the isodose line can be corrected in the relatively large elective target volume to reach the D98 objective. It is recommended to penalize isodose line gaps of more than 1 mm by implementing D99.9 in the model and settings.

The model with 90 plans achieved most clinical acceptable plans (model 30: 30/39, model 60: 26/39, model 90: 31/39, reference: 26/39), however the D99.9 of this model was the lowest (5030 cGy). The latter result is not taken into account in the evaluation of clinical acceptable plan quality. Therefore, it is important to interpret D98 results without additional information about dose distribution with caution.

The requirement of reaching 90% clinical acceptable MLO plans was not met with the only strategy used in this study. As described previously, the exceeded dose on critical OARs is expected to be fixed easily. Furthermore target coverage is likely to be improved by increasing weight on target coverage. However, the remaining failing plans had higher sum NTCP values than the reference plans and require further tuning of target and OAR weights. A possible option for reaching clinical acceptable quality in 90% of the plans is to create multiple strategies and choose dependent on patient characteristics and/or an initial strategy which strategies would fit best for this patients. Ideally mimicking weights will be updated automatically during the optimization, a feature which probable will be released in future RayStation versions. Further work is required to establish a multiple strategy approach in MLO planning.

Contrary to expectations, the mimicking optimization took around 50 minutes per plan in our institute. That was longer than the 29 minutes, reached by RaySearch[76]. This inconsistency may be due to differences in the processor power(UMCG: Intel Xeon E5-2698v4 20-core CPU, RaySearch: Intel i9-7940X CPU)[77]. Furthermore, it is reasonable that computing power can be lower in our institute, since RayStation is shared between users and less random access memory (RAM). High-performance computing processors are recommended if MLO will be used in future in clinical practice in multiple treatment sites to reach fast MLO plans.

Several questions remain unanswered at present. An important issue for future work is to investigate if changes in clinical protocols require new trained models or if tuning a model on the new requirements will be sufficient. Furthermore, it should be investigated if MLO plan quality can be predicted to determine if adequate OARs dose is reached or if manual treatment planning is necessary to improve plan quality in clinical use. Finally, as mentioned before, isodose line gaps and hotspots should be improved.

When MLO planning will be used in clinical practice, several changes will in workflow and decision making will appear. Physicians may ask more frequently for a plan adaptation during the treatment course. It is shown that adaptive planning results in better outcome and less normal tissue complications[78]. Furthermore, the content of daily work of dosimetrists may change. It is likely that dosimetrists will focus on manual treatment planning of outliers or complicated patients. The result for the patient will be a better matching treatment plan and potential shorter treatment preparation.

In conclusion, machine learning optimization planning has shown to be able to generate clinical acceptable VMAT plans for the majority of oropharyngeal cancer patients. Comparable plan quality was reached as reference plans in all models. This indicates that MLO planning can serve as a step towards fully automated treatment planning in radiotherapy.

CHAPTER 3 FUTURE PERSPECTIVES

The main goal of this study was to generate clinical acceptable machine learning based VMAT treatment plans for oropharyngeal cancer. We tuned prediction and mimicking settings in a set of 39 patients on a model of 60 plans. Clinical acceptable plan quality was reached in 67% using one strategy. Results of our study of other model sizes showed that large models are not a requirement in using machine learning optimization planning. However, the plan quality of outliers is not investigated in this research. Since it is expected that larger models especially perform better on outliers it cannot be concluded that large models do not have advantage over smaller models. Furthermore, our results showed that model performance was highly dependent on prediction and mimicking settings. These preliminary conclusions can be used for further research towards clinical implementation of automated treatment planning. In this chapter will be discussed which tasks and questions need to be answered before machine learning optimization planning can be implemented in clinical practice.

Improve machine learning optimization plan quality

As recommended in the discussion section of chapter 2, several characteristics of MLO plans including D2, maximum dose on critical OARs and D99.9, can be improved towards the reference plans. First, dose hotspots in elective target volume should not appear in or near OARs. However, it may be clinically accepted if hotspots appear in fatty tissue or close to high risk targets. Since a sample of the MLO plans already showed unacceptable hotspots, see figure 8 patient C, hotspots should be evaluated properly and reduced if they are present. Secondly, maximum dose on critical OARs should be decreased as discussed before. Each patient should receive at least two plans, the first one with optimal critical OARs sparing and the second one with adequate target coverage. These plans may support patient and physician decision making of sacrificing critical OAR(s) or risk of recurrence. Finally, gaps in the isodose lines in targets, indicated by low D99.9 values, need to be reduced. As discussed before, new structures like contracted PTVs can be added in MLO planning. Furthermore, results showed that some reference plans from 2016 and 2017 had gaps as well. These plans are not accepted in clinical practice nowadays. The study of Witte et al. showed that the PTV margin may be contracted not more than half of an unilateral PTV margin to safely treat the CTV[79]. Therefore, it may be necessary to remove these plans from MLO training and validation sets.

In addition, it is important to investigate the use of strategies in MLO planning to reduce OARs dose and reach clinical acceptable plan quality in 90% of the plans. It is suggested to examine patient and plan characteristics for further tuning of multiple strategies. For example, small and large tumours currently have the same strategy. Furthermore, strategies for other HNC types should be investigated. It may be possible that each HNC type needs different strategies. Finally, results need to be verified with an independent test set, especially when further tuning is performed on smaller sets.

Practical updates of MLO planning

In this study, MLO plans are optimized and evaluated using contracted PTV structure (evaluation structure) in case part of the PTV is located outside the skin. This outdated method is done to avoid high MU in the skin in case of shifts during the treatment course. However, in clinical practice, plans are optimized using a virtual bolus, an option to avoid high MU in the skin without contracting the PTV and risk on lack of coverage. The study of Tyran et al. showed that virtual bolus can be used safely during treatment and did not impact the plan quality during treatment planning[80]. Therefore, we assume that we can keep using contracted PTV structures in this study, but we should switch to the virtual bolus method when MLO planning will be used in clinical practice.

Another practical consideration is a new clinical protocol with higher priority on oral cavity, which will be released in a few months. It should be investigated if MLO plans according the new protocol can be reached with tuning or that new models need to be trained with oral cavity plans. Furthermore, new OARs are added in the clinical protocol and should be added in the training.

Evaluation method

Currently, clinical plan quality evaluation is done by manual scrolling through the slices and additional dosimetric parameters. This process is time consuming when many plans should be evaluated in research or clinical setting, for example when multiple plan strategies are generated. A possible solution is to (partial) automate plan quality evaluation. Automation in plan quality evaluation may lead to objective evaluation and easy comparison between plans for both clinical and research use.

Table 7 Overview of criteria and score for multiple ROIs for automated plan evaluation

	ROI	Criteria	Priority	Score 100	Score 50	Score 0
Constraints	PTV 7000	D_{98} ≥ 6650 cGy	1	≥ 6650	6640	≤ 6630
		6950 ≤ D_{mean} ≤ 7050 cGy	1	6950 ≤ D_{mean} ≤ 7050	6940 ≥ D_{mean} ≥ 7060	6930 ≥ D_{mean} ≥ 7070
		D_2 ≤ 7490 cGy	1	≤ 7490	7500	≥ 7510
	PTV 5425	D_{98} ≥ 5154 cGy	1	≥ 5154	5144	≤ 5134
	Brainstem	$D_{0.1}$ ≤ 6300 cGy	1	≤ 6300	6310	≥ 6320
	Brain	$D_{0.1}$ ≤ 6300 cGy	1	≤ 6300	6310	≥ 6320
	Spinal cord	$D_{0.1}$ ≤ 5400 cGy	1	≤ 5400	5410	≥ 5420
	Eyes posterior	$D_{0.1}$ ≤ 6000 cGy	1	≤ 6000	6010	≥ 6020
	Cochlea	$D_{0.1}$ ≤ 5250 cGy	1	≤ 5250	5260	≥ 5270
Objectives	PTV 7000– 1mm	$D_{99.9}$ ≥ 6650 cGy	2	≥ 6650	6600	≤ 6550
	PTV 5425– 1mm	$D_{99.9}$ ≥ 5154 cGy	3	≥ 5154	5054	≤ 4954
	PTV 5425 – PTV 7000	V5800 = ALARA	3	?	?	?
	NTCP OARs	D_{mean} = ALARA	4	Plan score = 100 - NTCP		
	Other OARs	D_{mean} = ALARA	5	0	35	≥ 70

ROI: Region Of Interest, PTV: Planning Target Volume, OARs: Organs At Risk, NTCP OARs: parotid glands, pharyngeal constrictor muscle superior (PCM), PCM inferior, oral cavity and cricopharyngeal muscle, D_{98}: dose at 98% of a volume, D_{mean}: mean dose of a volume, D_2: dose at 2% of a volume, $D_{0.1}$: dose at 0.1% of a volume, ALARA: As Low As Reasonably Achievable

The process of creating automated plan approval can be divided in the following steps:

Step 1: Determine criteria with physicians
Step 2: Translate scrolling criteria into measurable parameters
Step 3: Score criteria
Step 4: Implement criteria in automated rating method
Step 5: Visualize plan quality

The first two steps are already included in the evaluation criteria of this thesis, except for D2 evaluation, see an criteria overview in table 7. A concept of scores to the criteria is also shown in table 7 (step 3). The following considerations are taken into account in rewarding or punishing the criteria:

- All scores are linear descending, the constraints steeper than objectives, since constraints have higher priority[24].
- Reached criteria are rewarded with 100 points, it is not higher rewarded when a criteria is reached easy since clinical advantages seem minimal.

All scores are summed and divided by the amount of criteria scored. The final scores can be implemented in the RaySearch Analytics pipeline (step 4) and visualized in Tableau (step 5).

When the automated method is finished, it should be evaluated with physicians to verify adequate reflection of plan quality in the plan score. If the method is confirmed, plan quality can be used to select strategies or recognize worse plan quality. In addition, it should be investigated what plan quality is reachable for a specific patient, for example by combining patient and plan characteristics like target size, target location, target and OARs overlap and distance to target[81]. It is recommended to manually evaluate the final plan since currently not the total dose distribution is reflected by the plan score, however, automatic plan evaluation may a good starting point in selecting and evaluation MLO plans.

REFERENCES

[1] "Head and neck cancer survival rates | Seattle Cancer Care Alliance." [Online]. Available: https://www.seattlecca.org/diseases/head-neck-cancers/head-neck-cancers-overview/survival-rates. [Accessed: 10-Jul-2019].

[2] F. Duprez *et al.*, "Distant metastases in head and neck cancer," *Head Neck*, vol. 39, no. 9, pp. 1733–1743, Sep. 2017.

[3] C. B. Simone *et al.*, "Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer," *Radiother. Oncol.*, vol. 101, no. 3, pp. 376–382, Dec. 2011.

[4] H. P. van der Laan *et al.*, "The potential of intensity-modulated proton radiotherapy to reduce swallowing dysfunction in the treatment of head and neck cancer: A planning comparative study," *Acta Oncol. (Madr).*, vol. 52, no. 3, pp. 561–569, Apr. 2013.

[5] E. Wissinger, I. Griebsch, J. Lungershausen, T. Foster, and C. L. Pashos, "The Economic Burden of Head and Neck Cancer: A Systematic Literature Review," *Pharmacoeconomics*, vol. 32, no. 9, pp. 865–882, Sep. 2014.

[6] 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.

[7] F. Pezzuto *et al.*, "Update on Head and Neck Cancer: Current Knowledge on Epidemiology, Risk Factors, Molecular Features and Novel Therapies.," *Oncology*, vol. 89, no. 3, pp. 125–36, 2015.

[8] A. Fogliata *et al.*, "RapidPlan head and neck model: the objectives and possible clinical benefit.," *Radiat. Oncol.*, vol. 12, no. 1, p. 73, Apr. 2017.

[9] 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.

[10] 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.

[11] "Hoofd-halstumoren - Startpagina - Richtlijn - Richtlijnendatabase." [Online]. Available: https://richtlijnendatabase.nl/richtlijn/hoofd-halstumoren/hoofd-halstumoren_-_korte_beschrijving.html. [Accessed: 21-Dec-2018].

[12] "hoofd-halskanker." [Online]. Available: https://www.iknl.nl/oncologische-zorg/tumorteams/hoofd-halskanker. [Accessed: 29-Nov-2018].

[13] "Weet u symptomen van hoofd-halskanker? - Stichting Nationaal Fonds tegen Kanker." [Online]. Available: https://www.tegenkanker.nl/2017/09/28/weet-symptomen-hoofd-halskanker/. [Accessed: 19-Dec-2018].

[14] CBS, "Bevolkingsprognose in 2060," no. December, pp. 1–19, 2017.

[15] A. Degboe *et al.*, "Patients' experience of recurrent/metastatic head and neck squamous cell carcinoma and their perspective on the EORTC QLQ-C30 and QLQ-H&N35 questionnaires: a qualitative study.," *J. patient-reported outcomes*, vol. 2, p. 33, 2017.

[16] "Head and Neck Cancers - National Cancer Institute." [Online]. Available: https://www.cancer.gov/types/head-and-neck/head-neck-fact-sheet. [Accessed: 29-Nov-2018].

[17] S.-A. Yeh, "Radiotherapy for head and neck cancer.," *Semin. Plast. Surg.*, vol. 24, no. 2, pp. 127–36, May 2010.

[18] J. Taxy, "Pathology of head and neck neoplasms - UpToDate." [Online]. Available: https://www.uptodate.com/contents/pathology-of-head-and-neck-neoplasms?sectionName=SQUAMOUS CELL CARCINOMA&topicRef=3393&anchor=H84270723&source=see_link#H84270723. [Accessed: 19-Dec-2018].

[19] S. Marur and A. A. Forastiere, "Head and Neck Squamous Cell Carcinoma: Update on Epidemiology, Diagnosis, and Treatment."

[20] E. M. Sturgis and P. M. Cinciripini, "Trends in head and neck cancer incidence in relation to smoking

prevalence," *Cancer*, vol. 110, no. 7, pp. 1429–1435, Oct. 2007.

[21] C. Fakhry *et al.*, "Improved Survival of Patients With Human Papillomavirus-Positive Head and Neck Squamous Cell Carcinoma in a Prospective Clinical Trial," *JNCI J. Natl. Cancer Inst.*, vol. 100, no. 4, pp. 261–269, Feb. 2008.

[22] "Hoofd-halstumoren - Startpagina - Richtlijn - Richtlijnendatabase." [Online]. Available: https://richtlijnendatabase.nl/richtlijn/hoofd-halstumoren/hoofd-halstumoren_-_korte_beschrijving.html. [Accessed: 04-Jan-2019].

[23] D. G. Pfister *et al.*, "Head and Neck Cancers, Version 1.2015.," *J. Natl. Compr. Canc. Netw.*, vol. 13, no. 7, pp. 847–55; quiz 856, Jul. 2015.

[24] J. G. . van den Hoek *et al.*, "HH-01 Radiotherapie bij plaveiselcelcarcinoom primair Medisch Inhoudelijk," pp. 1–19, 2019.

[25] J. Brierley, "TNM Classification of Malignant tumours eight edition."

[26] A. B. Miah *et al.*, "Recovery of Salivary Function: Contralateral Parotid-sparing Intensity-modulated Radiotherapy versus Bilateral Superficial Lobe Parotid-sparing Intensity-modulated Radiotherapy.," *Clin. Oncol. (R. Coll. Radiol).*, vol. 28, no. 9, pp. e69-76, 2016.

[27] "ICRU Report 83 Prescribing, Recording, and Reporting Photon-Beam Intensity-Modulated Radiation Therapy (IMRT)," *J. ICRU*, vol. 10, no. 1, p. NP.1-NP, Apr. 2010.

[28] N. Hodapp, "Der ICRU-Report 83: Verordnung, Dokumentation und Kommunikation der fluenzmodulierten Photonenstrahlentherapie (IMRT)," *Strahlentherapie und Onkol.*, vol. 188, no. 1, pp. 97–100, Jan. 2012.

[29] A. J. J. Bos, F. S. Draaisma, and W. J. C. Okx, *Inleiding tot de stralingshygiëne*. Sdu Uitgevers, 2009.

[30] C. Kurz *et al.*, "Feasibility of automated proton therapy plan adaptation for head and neck tumors using cone beam CT images," *Radiat. Oncol.*, vol. 11, no. 1, p. 64, Dec. 2016.

[31] M. Steneker, A. Lomax, and U. Schneider, "Intensity modulated photon and proton therapy for the treatment of head and neck tumors," *Radiother. Oncol.*, vol. 80, no. 2, pp. 263–267, Aug. 2006.

[32] Langendijk et al., "Dutch National Indication Protocol for Proton Therapy." [Online]. Available: http://www.nvro.nl/landelijkindicatieprotocol. [Accessed: 26-Apr-2018].

[33] P. Mayles, A. E. Nahum, and J.-C. Rosenwald, *Chapter 20: from Measurements to Calculations, Handbook of radiotherapy physics : theory and practice*. Taylor & Francis, 2007.

[34] A. Pant, "Introduction to Machine Learning for Beginners - Towards Data Science." [Online]. Available: https://towardsdatascience.com/introduction-to-machine-learning-for-beginners-eed6024fdb08. [Accessed: 08-Jul-2019].

[35] "A.I, Machine Learning en Deep Learning, wat is het verschil?" [Online]. Available: https://www.globalorange.nl/artificial-intelligence-machine-learning-en-deep-learning. [Accessed: 08-Jul-2019].

[36] D. Silver *et al.*, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, no. 7676, pp. 354–359, Oct. 2017.

[37] S. Wagner, "Reinforcement Learning and Supervised Learning: A brief comparison," 2018. [Online]. Available: https://hackernoon.com/reinforcement-learning-and-supervised-learning-a-brief-comparison-1b6d68c45ffa. [Accessed: 08-Jul-2019].

[38] X. Zhu, C. Vondrick, C. C. Fowlkes, and D. Ramanan, "Do We Need More Training Data?," *Int. J. Comput. Vis.*, vol. 119, no. 1, pp. 76–92, 2016.

[39] "Artificial Intelligence vs. Machine Learning vs. Deep Learning: What's the Difference? | Sumo Logic." [Online]. Available: https://www.sumologic.com/blog/machine-learning-deep-learning/. [Accessed: 29-Jul-2019].

[40] M. Feng, G. Valdes, N. Dixit, and T. D. Solberg, "Machine Learning in Radiation Oncology: Opportunities, Requirements, and Needs.," *Front. Oncol.*, vol. 8, p. 110, 2018.

[41] G. Valdes, J. M. Luna, E. Eaton, C. B. Simone, L. H. Ungar, and T. D. Solberg, "MediBoost: a Patient Stratification Tool for Interpretable Decision Making in the Era of Precision Medicine," *Sci. Rep.*, vol. 6, no. 1, p. 37854, Dec. 2016.

[42] I. Boon, T. Au Yong, and C. Boon, “Assessing the Role of Artificial Intelligence (AI) in Clinical Oncology: Utility of Machine Learning in Radiotherapy Target Volume Delineation,” *Medicines*, vol. 5, no. 4, p. 131, Dec. 2018.

[43] C. E. Cardenas *et al.*, “Deep Learning Algorithm for Auto-Delineation of High-Risk Oropharyngeal Clinical Target Volumes With Built-In Dice Similarity Coefficient Parameter Optimization Function.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 101, no. 2, pp. 468–478, Jun. 2018.

[44] X. Wu *et al.*, “AAR-RT – A system for auto-contouring organs at risk on CT images for radiation therapy planning: Principles, design, and large-scale evaluation on head-and-neck and thoracic cancer cases,” *Med. Image Anal.*, vol. 54, pp. 45–62, May 2019.

[45] L. M. Appenzoller, J. M. Michalski, W. L. Thorstad, S. Mutic, and K. L. Moore, “Predicting dose-volume histograms for organs-at-risk in IMRT planning,” *Med. Phys.*, vol. 39, no. 12, pp. 7446–7461, Nov. 2012.

[46] C. G. Rowbottom, S. Webb, and M. Oldham, “Beam-orientation customization using an artificial neural network.,” *Phys. Med. Biol.*, vol. 44, no. 9, pp. 2251–62, Sep. 1999.

[47] A. Sadeghnejad, S. Jiang, and D. Nguyen, “A Fast Deep Learning Approach for Beam Orientation Optimization for Prostate Cancer IMRT Treatments | Azar Sadeghnejad Barkousaraie | Request PDF,” 2019.

[48] G. Valdes, R. Scheuermann, C. Y. Hung, A. Olszanski, M. Bellerive, and T. D. Solberg, “A mathematical framework for virtual IMRT QA using machine learning,” *Med. Phys.*, vol. 43, no. 7, pp. 4323–4334, Jun. 2016.

[49] G. Valdes, M. F. Chan, S. B. Lim, R. Scheuermann, J. O. Deasy, and T. D. Solberg, “IMRT QA using machine learning: A multi-institutional validation,” *J. Appl. Clin. Med. Phys.*, vol. 18, no. 5, pp. 279–284, Sep. 2017.

[50] Q. Li and M. F. Chan, “Predictive time-series modeling using artificial neural networks for Linac beam symmetry: an empirical study,” *Ann. N. Y. Acad. Sci.*, vol. 1387, no. 1, pp. 84–94, Jan. 2017.

[51] M. S. Huq *et al.*, “The report of Task Group 100 of the AAPM: Application of risk analysis methods to radiation therapy quality management,” *Med. Phys.*, vol. 43, no. 7, pp. 4209–4262, Jun. 2016.

[52] G. Guidi *et al.*, “A machine learning tool for re-planning and adaptive RT: A multicenter cohort investigation,” *Phys. Medica*, vol. 32, no. 12, pp. 1659–1666, Dec. 2016.

[53] L. Oakden-Rayner, G. Carneiro, T. Bessen, J. C. Nascimento, A. P. Bradley, and L. J. Palmer, “Precision Radiology: Predicting longevity using feature engineering and deep learning methods in a radiomics framework,” *Sci. Rep.*, vol. 7, no. 1, p. 1648, Dec. 2017.

[54] K. H. Cha *et al.*, “Bladder Cancer Treatment Response Assessment in CT using Radiomics with Deep-Learning,” *Sci. Rep.*, vol. 7, no. 1, p. 8738, Dec. 2017.

[55] C. B. Simone *et al.*, “Comparison of intensity-modulated radiotherapy, adaptive radiotherapy, proton radiotherapy, and adaptive proton radiotherapy for treatment of locally advanced head and neck cancer,” *Radiother. Oncol.*, vol. 101, no. 3, pp. 376–382, Dec. 2011.

[56] H. P. van der Laan *et al.*, “The potential of intensity-modulated proton radiotherapy to reduce swallowing dysfunction in the treatment of head and neck cancer: A planning comparative study,” *Acta Oncol. (Madr).*, vol. 52, no. 3, pp. 561–569, Apr. 2013.

[57] E. Wissinger, I. Griebsch, J. Lungershausen, T. Foster, and C. L. Pashos, “The economic burden of head and neck cancer: a systematic literature review.,” *Pharmacoeconomics*, vol. 32, no. 9, pp. 865–82, Sep. 2014.

[58] L. J. Peters *et al.*, “Critical impact of radiotherapy protocol compliance and quality in the treatment of advanced head and neck cancer: results from TROG 02.02.,” *J. Clin. Oncol.*, vol. 28, no. 18, pp. 2996–3001, Jun. 2010.

[59] R. Mahmood, A. Babier, A. McNiven, A. Diamant, and T. C. Y. Chan, “Automated Treatment Planning in Radiation Therapy using Generative Adversarial Networks.” pp. 484–499, 29-Nov-2018.

[60] J. P. Tol, A. R. Delaney, M. Dahele, B. J. Slotman, and W. F. A. R. Verbakel, “Evaluation of a Knowledge-Based Planning Solution for Head and Neck Cancer,” *Int. J. Radiat. Oncol.*, vol. 91, no. 3, pp. 612–620, Mar. 2015.

[61] B. Wu *et al.*, “Patient geometry-driven information retrieval for IMRT treatment plan quality control.,” *Med. Phys.*, vol. 36, no. 12, pp. 5497–505, Dec. 2009.

[62] D. Nguyen *et al.*, “3D radiotherapy dose prediction on head and neck cancer patients with a hierarchically densely connected U-net deep learning architecture,” *Phys. Med. Biol.*, vol. 64, no. 6, p. 065020, Mar. 2019.

[63] H.-H. Tseng, Y. Luo, R. K. Ten Haken, and I. El Naqa, “The Role of Machine Learning in Knowledge-Based Response-Adapted Radiotherapy.,” *Front. Oncol.*, vol. 8, p. 266, 2018.

[64] C. Gui and V. Chan, “Machine learning in medicine,” *Med. Technol.*, pp. 76–78, 2017.

[65] J. J. Boutilier, T. Craig, M. B. Sharpe, and T. C. Y. Chan, “Sample size requirements for knowledge-based treatment planning,” *Med. Phys.*, vol. 43, no. 3, pp. 1212–1221, Feb. 2016.

[66] “How Do You Know You Have Enough Training Data? – Towards Data Science.” [Online]. Available: https://towardsdatascience.com/how-do-you-know-you-have-enough-training-data-ad9b1fd679ee?gi=1fc2ee5f1bf. [Accessed: 29-Jun-2019].

[67] 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, Aug. 2017.

[68] A. Bhattacharyya, “On a Measure of Divergence between Two Multinomial Populations,” *Indian J. Stat.*, vol. 7, no. 4, pp. 401–406, 1946.

[69] A. Fredriksson, “Automated improvement of radiation therapy treatment plans by optimization under reference dose constraints,” *Phys. Med. Biol.*, vol. 57, no. 23, pp. 7799–7811, Dec. 2012.

[70] N. J. Lomax and S. G. Scheib, “Quantifying the degree of conformity in radiosurgery treatment planning.,” *Int. J. Radiat. Oncol. Biol. Phys.*, vol. 55, no. 5, pp. 1409–19, Apr. 2003.

[71] C. Bonferroni, “Teoria statistica delle classi e calcolo delle probabilit,” *del R Ist. Super. di Sci. Econ. e Commer. di*, vol. 8, no. 3, p. 62, 1936.

[72] R. A. Armstrong, “When to use the Bonferroni correction,” *Ophthalmic Physiol. Opt.*, vol. 34, no. 5, pp. 502–508, Sep. 2014.

[73] B. Wu *et al.*, “Data-Driven Approach to Generating Achievable Dose–Volume Histogram Objectives in Intensity-Modulated Radiotherapy Planning,” *Int. J. Radiat. Oncol.*, vol. 79, no. 4, pp. 1241–1247, Mar. 2011.

[74] J. Lian *et al.*, “Modeling the dosimetry of organ-at-risk in head and neck IMRT planning: An intertechnique and interinstitutional study,” *Med. Phys.*, vol. 40, no. 12, p. 121704, Nov. 2013.

[75] A. Babier, J. J. Boutilier, A. L. McNiven, and T. C. Y. Chan, “Knowledge-based automated planning for oropharyngeal cancer,” *Med. Phys.*, vol. 45, no. 7, pp. 2875–2883, Jul. 2018.

[76] X. RaySearch Laboratories AB, “Machine Learning Automated Treatment Planning White Paper,” 2019.

[77] “UserBenchmark: Intel Core i9-7940X vs Xeon E5-2690 v2.” [Online]. Available: https://cpu.userbenchmark.com/Compare/Intel-Xeon-E5-2690-v2-vs-Intel-Core-i9-7940X/m13436vsm353078. [Accessed: 12-Jul-2019].

[78] J. Heukelom and C. D. Fuller, “Head and Neck Cancer Adaptive Radiation Therapy (ART): Conceptual Considerations for the Informed Clinician,” *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 258–273, Jul. 2019.

[79] M. G. Witte, J.-J. Sonke, J. Siebers, J. O. Deasy, and M. van Herk, “Beyond the margin recipe: the probability of correct target dosage and tumor control in the presence of a dose limiting structure,” *Phys. Med. Biol.*, vol. 62, no. 19, pp. 7874–7888, Sep. 2017.

[80] M. Tyran *et al.*, “Safety and benefit of using a virtual bolus during treatment planning for breast cancer treated with arc therapy,” *J. Appl. Clin. Med. Phys.*, vol. 19, no. 5, pp. 463–472, Sep. 2018.

[81] K. L. Moore, “Automated Radiotherapy Treatment Planning,” *Semin. Radiat. Oncol.*, vol. 29, no. 3, pp. 209–218, Jul. 2019.

[82] “Weblet Importer.” [Online]. Available: https://s3.amazonaws.com/MLMastery/MachineLearningAlgorithms.png?__s=tshondqd1xsoaqb7zsis. [Accessed: 12-Jul-2019].

APPENDIX

CONTENT

APPENDIX A GRADING XEROSTOMIA AND DYSPHAGIA

Table I Grading xerostomia according to CTCAEv4.0[32]

Grading	Description
Grade I	Mild symptoms: oral intake is altered
	Unstimulated saliva production: 0.1-0.2 ml/min
Grade II	Adequate oral intake not possible
	Unstimulated saliva production: <0.1 ml/min

Table II Grading dysphagia according to CTCAEv4.0[32]

Grading	Description
Grade I	Symptomatic, but normal diet possible
Grade II	Symptomatic and altered eating and swallowing pattern
Grade III	Highly altered eating and swallowing pattern
	Tube feeding dependence
	Hospitalization required
Grade IV	Life-threatening consequences
	Acute intervention indicated

APPENDIX B OVERVIEW MACHINE LEARNING ALGORITHMS

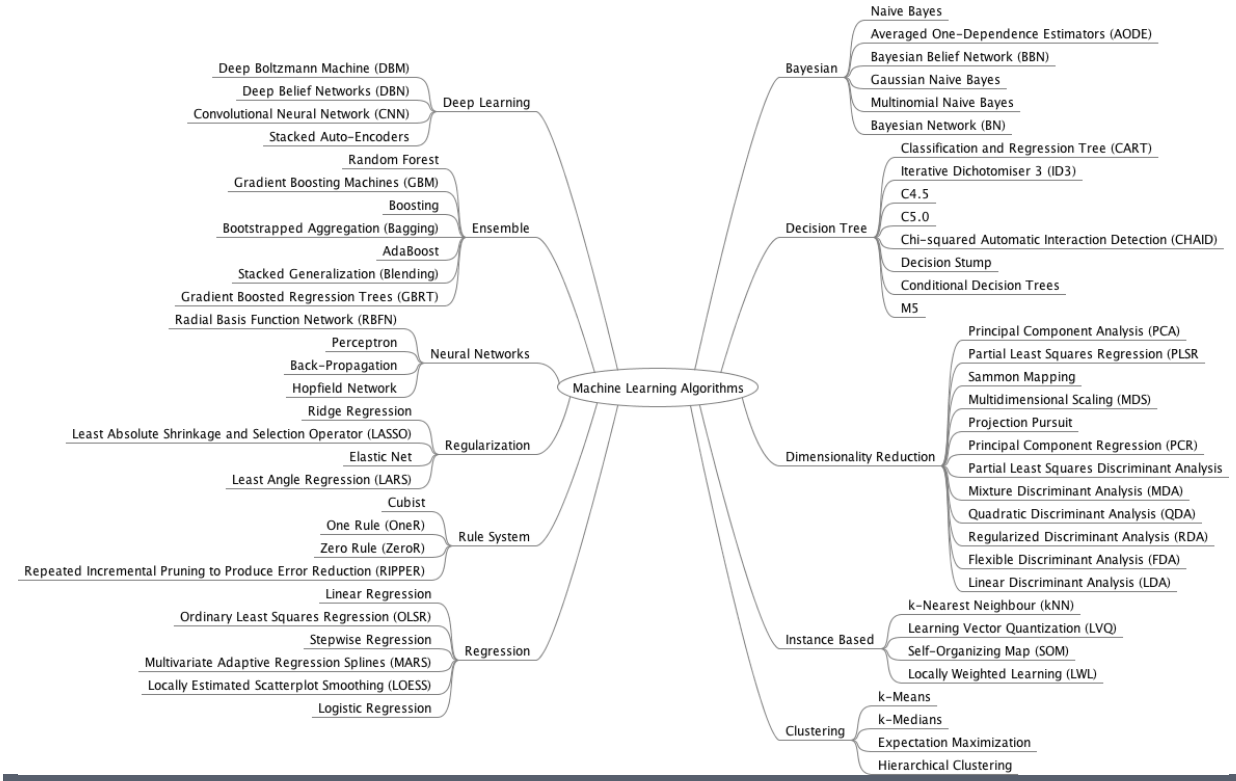

Figure i Overview machine learning algorithms[82]

APPENDIX C FINAL PREDICTION AND MIMICKING SETTINGS


```
"StrategyName" : "Standard",
"MLRoiGoals" : [{ "GoalType" : "max_dose",  "ROI" : "PTV70", "Volume" : 0.50, "Value" : 6950, "ValueType" : "abs"},
                { "GoalType" : "min_dose",  "ROI" : "PTV70", "Volume" : 0.98, "Value" : 6680, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ROI" : "PTV54.25", "ExcludeRoi" : "PTV70", "Volume" : 0.50, "Value" : 5475, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ROI" : "PTV54.25", "ExcludeRoi" : "PTV70", "Volume" : 0.10, "Value" : 5700, "ValueType" : "abs"},
                { "GoalType" : "min_dose",  "ROI" : "PTV54.25", "ExcludeRoi" : "PTV70", "Volume" : 0.98, "Value" : 5180, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoi" : "PTV70", "ROI" : "PTV54.25-Parotid_L", "Volume" : 0.00, "Value" : 5425, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoi" : "PTV70", "ROI" : "PTV54.25-Parotid_R", "Volume" : 0.00, "Value" : 5425, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_R", "Volume" : 0.10, "Value" : 1300, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_L", "Volume" : 0.10, "Value" : 1300, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_R", "Volume" : 0.20, "Value" : 1000, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_L", "Volume" : 0.20, "Value" : 1000, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_R", "Volume" : 0.50, "Value" : 600, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ExcludeRoiType" : "PTV", "ROI" : "Parotid_L", "Volume" : 0.50, "Value" : 600, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ROI" : "SpinalCord", "Volume" : 0.01, "Value" : 5000, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ROI" : "Brain", "Volume" : 0.01, "Value" : 6000, "ValueType" : "abs"},
                { "GoalType" : "max_dose",  "ROI" : "BrainStem", "Volume" : 0.01, "Value" : 6000, "ValueType" : "abs"}
],
"MLReduceOar" : false,
"MLPriorMethodOAR" : "min",
"MLPriorMethodTarget" : "max",
"MLPriorMethodTarget_OAR" : "avg",
"MimickOarList" : ["Brain","BrainStem","SpinalCord","Parotid_L","Parotid_R","Submandibular_L","Submandibular_R",
           "Mandible","OralCavity_Ext","GlotticArea", "Supraglottic","Thyroid","PCM_Inf","PCM_Med","PCM_Sup","Crico","Esophagus_Cerv"],
"MimickingStrategy" : "RayLearner",
"MimickAddClinicalGoals" : true,
"MimickTargetWeights": {"PTV70" : 10,
                        "PTV54.25" : 15},
"MimickOarWeights": {"Brain" : 0.5,
                     "BrainStem" : 0.5,
                     "SpinalCord" : 3,
                     "Parotid_L" : 12,
                     "Parotid_R" : 12,
                     "Submandibular_L" : 6,
                     "Submandibular_R" : 6,
                     "Mandible" : 1,
                     "OralCavity_Ext" : 6,
                     "GlotticArea" : 1,
                     "Supraglottic" : 1,
                     "Thyroid" : 1,
                     "PCM_Inf" : 6,
                     "PCM_Med" : 1,
                     "PCM_Sup" : 6,
                     "Crico" : 6,
                     "Esophagus_Cerv" : 1,
                     "Eye_Ant_L" : 0.5,
                     "Eye_Ant_R" : 0.5,
                     "Eye_Post_L" : 0.5,
                     "Eye_Post_R" : 0.5},
"MimickWeightForMaxDose" : 2.0,
"MimickOarIsoFunctions" : {"Parotid_L" : "avg", "Parotid_R" : "avg", "Submandibular_L" : "avg", "Submandibular_R" : "avg", "OralCavity_Ext" : "avg",
                  "PCM_Sup" : "avg", "PCM_Inf" : "avg", "Crico" : "avg","SpinalCord" : "max71"},
"MimickFinalTargetCheck" : true,
"MimickMaxOarConstraint" : false,
"MimickExternalWeight" : 25,
"MimickOarExpand" : 0
```


Figure ii Final prediction and mimicking settings

APPENDIX D SUPPLEMENTARY MATERIAL MODEL 60 RESULTS

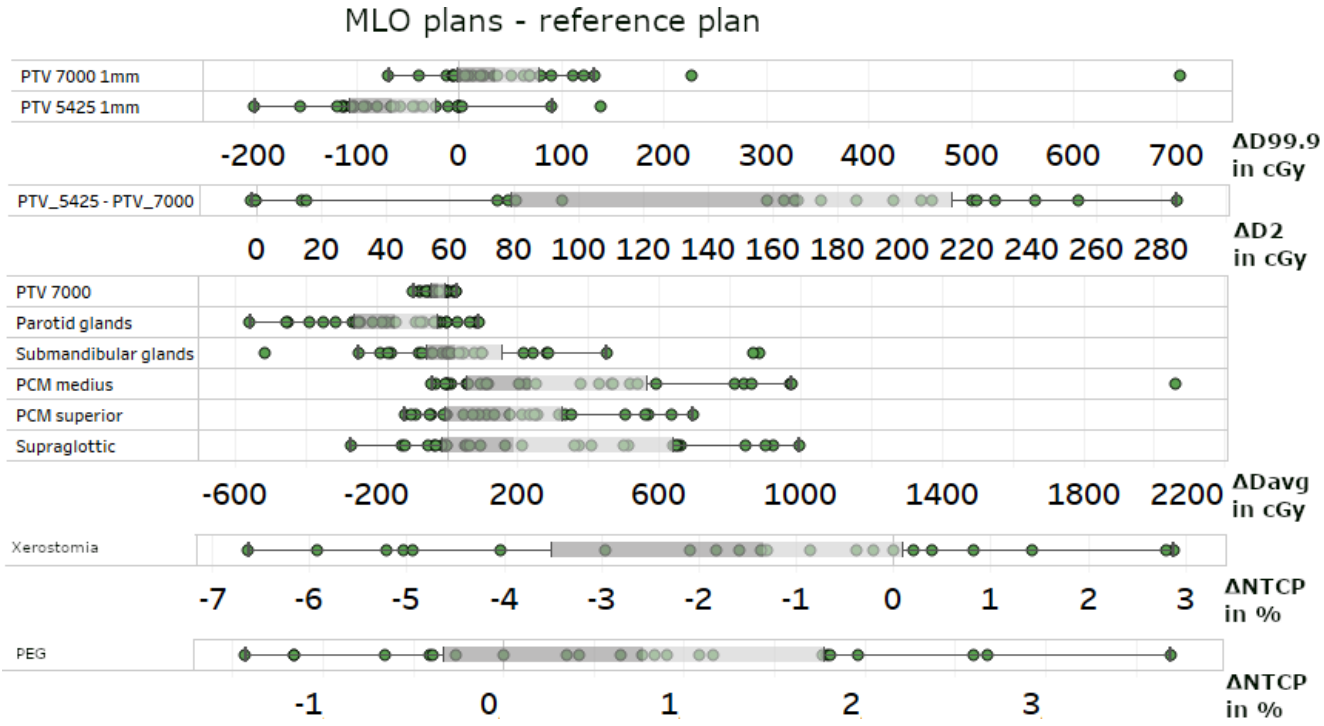

Figure iii Boxplot of dosimetric and evaluation parameters. Reference plans are subtracted from the Machine learning optimization (MLO) plans of model 60.

APPENDIX E OVERVIEW OF KNOWLEDGE BASED HEAD AND NECK CANCER STUDIES

Table III Overview of results knowledge based automated treatment planning head and neck cancer studies

Articles	Method type	Model size	Validation size	Test size	Target comparison		Metrics comparison	Target	Max OARs	Mean OARs	NTCP
Wu 2011	Case/DVH	91	not mentioned	15	Re-planned	vs clinical	Difference reference	comparable	lower	lower	
Wu 2012	Case/DVH		not mentioned	40	Re-planned	vs clinical	Difference reference	comparable	lower	lower	
Yuan 2012	Model/DVH	82	not mentioned	24	Predicted	vs clinical	error bound	comparable	comparable	comparable	
Lian 2013	Model/DVH		not mentioned	53	Predicted	vs clinical	error bound	comparable	lower	lower	
Wu 2013	Case/DVH		not mentioned	12	Re-planned	vs clinical	Difference reference	comparable	lower	lower	
Tol 2015	RapidPlan	30 and 60	not mentioned	30	Re-planned	vs clinical	Difference reference	comparable	comparable	lower	
Schmidt 2015	Case/DVH	103	not mentioned	10	Re-planned	vs clinical	Difference reference	comparable	lower	lower	
Zhang 2018	Model/DVH	80	148		Predicted	vs clinical	Weighted root mean square			higher	
Van Bruggen 2019	Voxel based	63	0	32	Goals		Difference reference	27/32 (84%)	higher	lower	
McIntosh 2017	Voxel based	54	12	12	Difference reference		Difference reference	0.6% higher		2.4% lower	
Fogliata 2017		83	not mentioned	20			Difference reference	comparable		lower	7% decrease
Hansen 2016			not mentioned	30				comparable		lower	
Mahmood 2018	Gan, deliverable	130	not mentioned	87				comparable	lower	lower	
Nguyen 2019	U-net, prediction	80	80	20	Difference reference		Difference reference		lower	lower	

APPENDIX F SHORT PAPER ICCR

Fully automated treatment planning of deliverable VMAT by machine learning dose prediction and mimicking optimization in HNC

Ilse G. van Bruggen[1], Roel G.J. Kierkels[1], Mats Holmström[2], David Lidberg[2], Karl Berggren[2], Stefan Both[1], Johannes A. Langendijk[1], Fredrik Löfman[2] and Erik W. Korevaar[†1]

[1]*University of Groningen, University Medical Center Groningen, Department of Radiation Oncology, Groningen, the Netherlands.* [2]*RaySearch Laboratories AB, Stockholm, Sweden.*

Introduction

Traditional intensity modulated treatment planning requires a manual and iterative loop of changing planning objectives. Therefore, plan quality remains subject to the experience of the dosimetrist and may greatly benefit from automation. Currently, fully and semi-automated methods have been proposed in literature, such as lexicographic-based methods and multi-criteria optimization, amongst others. We hypothesize that a (photon) dose distribution for head and neck cancer (HNC) patients can be predicted within minutes, which can then be mimicked to create a deliverable volumetric modulated arc therapy (VMAT) dose distribution with similar quality as the clinical 'dosimetrist-optimized' dose distributions.

Materials & Methods

The machine learning based automated planning (MLAP) involves training of a model, which is used to predict the voxel dose in novel patients. The CT scan, structures and dose distributions of 71 consecutive primary HNC patients, previously treated with dual arc VMAT and two dose levels (70 Gy and 54.25 Gy in 35 fractions), were retrieved from our clinical database. Patient selection was restricted to tumors originating in the oropharynx, larynx, oral cavity, nasopharynx and paranasal sinuses. As part of the evaluation of the MLAP, we applied a repeated random subset validation approach where patient data was split into four sets; using 8 patients for testing and the remaining 63 patients for training in each set.

The patient data was trained using an Atlas Regression Forest (ARF) model. During training, the ARF creates a set of decisions trees based from random voxels and features from the training data. The features are extracted imaging features, information about distances to target and OARs contours and the dose distributions of the patients in the training set. A prediction random forest (PRF)model is then trained to learn to select the best matching forests from the model using accuracy measures (gamma criteria)from the decision tree information. Hence, the five ARFs with the highest accuracy were used to predict a probability distribution of potential dose values per voxel in the test patients by using their input features. A conditional random field was then used, in combination with the clinical goals, to boil this distribution down to the most probable dose in each voxel. The authors refer to the work of McIntosh and Purdie for more details on the dose prediction step [1]. In the final step, the predicted dose distribution was input to a mimicking optimization algorithm to generate a deliverable dose distribution.

The dose distributions (i.e. predicted and mimicked dose) of the patients in the test set were compared against the dosimetrist-optimized clinical plans, further indicated as reference plans. Plans were compared by means of target conformity (CI95%),dosimetric parameters (targets: D98 (D98≥95%), D2; OARs: Dmean, Dmax) and normal tissue complication probabilities (NTCP) for xerostomia, grade 2-4 dysphagia, and percutaneous endoscopic gastrostomy (PEG) tube dependence.

Results

The predicted dose was in accordance with reference dose for all plans (table 1). The mimicked plans had adequate target coverage for high risk and elective volumes according to clinical goals in 27/32 (84%) and 29/32(91%) of the cases, respectively. For reference plans, 30 of 32 (94%) had adequate target coverage for both high risk and elective volumes. Target conformity was better in mimicked plans compared to reference plans. The NTCP values of mimicked

plans were lower or did not increase with more than 2.0% compared to the reference plans in 23 of the 32 cases (72%). The Dmax on the posterior eye (580cGy) exceeded the maximum eye dose (500cGy) in one mimicked plan. Other maximum doses for brain, brainstem, spinal cord and left eye were not exceeded in both reference plans and mimicked plans. Dmean for oral cavity, pharynx constrictor muscle superior and inferior, cricopharyngeal muscle and supraglottic larynx were lower in mimicked plans than reference plans, on average for these OARs: 36.7 ± 11.6 Gy vs. 39.2 ± 13.0 Gy respectively. Figure 1 shows the median dose volume histogram (DVH) of all mimicked and reference plans.

Table 1: *Overview of results reference, predicted and mimicked plans. SD: standard deviation, PTV: planning target volume, CI: conformity index, NTCP: normal tissue complication probability, PEG: percutaneous endoscopic gastrostomy*

	Reference(SD)	Prediction(SD)	Mimicked (SD)
PTV_7000 D98 (Gy)	67.1 (0.7)	67.3 (0.5)	66.9 (0.5)
PTV_7000 D2 (Gy)	72.3 (0.8)	71.1 (0.0)	72.1 (0.5)
PTV_7000 CI95%	1.3 (0.2)	1.3 (0.1)	1.2 (0.1)
PTV_5425 D98 (Gy)	52.6 (2.0)	52.5 (0.4)	51.9 (1.0)
PTV_5425 D2 (Gy)	71.7 (0.9)	70.9 (0.2)	71.3 (0.7)
PTV_5425 CI95%	1.4 (0.1)	1.2 (0.1)	1.2 (0.1)
NTCPXerostomia (%)	38.2 (11.8)	38.4 (8.5)	39.8 (10.8)
NTCP Dysphagia (%)	20.8 (11.7)	19.3 (9.6)	19.7 (10.2)
NTCP PEG (%)	12.9 (9.0)	10.9 (6.6)	11.9 (7.6)
NTCP Sum (%)	72.3 (30.4)	68.8 (23.4)	71.6 (27.0)

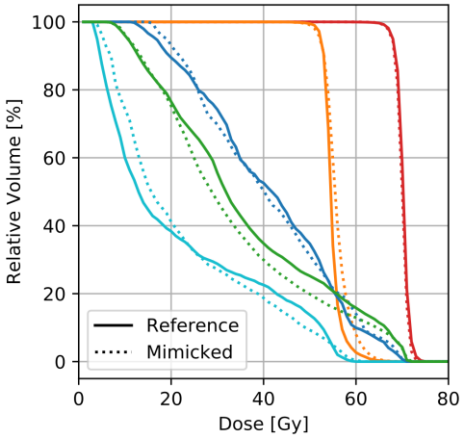

Figure 1: *Median dose volume histogram(DVH) of all mimicked and reference plans; PTV_7000 (red), PTV_5424 (excluding PTV_7000 orange), pharynx constrictor muscle superior (navy), oral cavity (green), combined parotid glands (turquoise)*

Discussion & Conclusions

In this study, we demonstrated that adequate target coverage can be reached using MLAP in the majority of HNC VMAT treatment plans. Similar or lower NTCP values were reached in 72% of the plans This indicates that MLAP can serve as a promising tool for automated treatment planning to overcome lower plan quality due to inexperienced dosimetrists. Further improvements in OARs dose is likely to be reached by optimizing machine learning settings and increasing the trainings set size, as more similar ARFs can be used to predict dose.

References

[1] McIntosh C, Purdie TG, Contextual Atlas Regression Forests: Multiple-Atlas-Based Automated Dose Prediction in Radiation Therapy, IEEE Trans Med Imaging 2016 Apr;35(4):1000-12

†Corresponding Author: e.w.korevaar@umcg.nl

APPENDIX G MANUAL MACHINE LEARNING IN RAYSTATION

Document created by: Maaike Dotinga
Edited by: Ilse van Bruggen
Date: 22-08-2019

TRAINING

After deciding which patients you would like to include in your ML model, please follow the next:

1. Go to \\zkh\appdata\Raystation\Research\ML\trained_models\RayStation8b and create your own folder
    a. Create a txt file comprising a patient list with ID and name of plan, an example can be found here: \\zkh\appdata\Raystation\Research\ML\trained_models\RayStation8b\Modeltest0702\\training_patient_test.txt
    b. Define your model_meta.JSON file, for examples go to: \\zkh\appdata\Raystation\Research\ML\trained_models\model_config\Maaike\JSON_files
    c. Change the txt and JSON file in your start_training file: \\zkh\appdata\Raystation\Research\ML\trained_models\RayStation8b\Modeltest0702\\start_training_modeltest0702.py

       further instructions about parameters in this file are found in: \\zkh\appdata\Raystation\Research\ML\Instructions_ML-models\RaySearch_instruction.docx

       NB: when using the GUI, keep in mind that 'modelname' should correspond with one of the defined folders in \\zkh\appdata\Raystation\Research\ML\trained_models
2. Open RayStation Development (RayStation 8B), open each new patient
    a. Verify if a patient is imported in RayStation Development, if not, import the patient from RayStation Klinisch
    b. Copy and rename the clinical plan to PhotonPlan
    c. Go to 'Plan Design', verify if dual arc is used, if not, create dual arc
    d. Change machine to latest Agility machine

3. In RayStation Development (RayStation 8B) >> Scripting >> Script creation
    a. Open: \\zkh\appdata\Raystation\Research\ML\trained_models\RayStation8b\Modeltest0702\\start_training_modeltest0702.py
    b. Change 'Phyton interpreter' to RayLearnerGron
    c. Start training and wait until it is completed

4. Output is created in the same folder as where your txt and json files are located. A new folder called 'Photon' will appear that includes a log file, pkl and npz files. Check your log file after training to detect empty structures and errors to find out whether adjustments have to be made or if training was successful.

TESTING

To generate predicted and mimicked plans, consider the following steps. For all actions keep in mind that you have opened RayStation in patient data management, that will speed up the process.

5. Copy your 'Photon' folder and JSON file into the folder with corresponding model name found here: \\zkh\appdata\Raystation\Research\ML\trained_models

6. Make sure ‘autoplanning_raystation_script.py’ and ‘autoplanning_service_script.py’ are present in the folder. If they are not there, copy them from: \\zkh\appdata\Raystation\Research\ML\trained_models\model_config\scripts

7. Have a look at the model_settings file. Do you want to change something in the settings?
8. Make sure the plans you want to use for prediction have to following settings:
    a. Current machine: Go to Plan design → double click on machine→ change to latest agility machine. There is no script available for make changes in a batch, so you have to change the machine for each patient manually.
    b. Prescription set to site: go to patient data management → start script ‘change to site’ \\zkh\appdata\Raystation\Research\ML\raylearner\AutomaticPlanning 1.1\Model_oro30_1\batch_change_to_site.py
    and specify the plans you want to change in the txt file \\zkh\appdata\Raystation\Research\ML\raylearner\AutomaticPlanning 1.1\Model_oro30_1\patients_change_to_site.txt

9. The prediction and mimicking consists of three steps. It is recommended to run the script only for one action at the time in new patients. Comment out (#) the other actions. \\zkh\appdata\Raystation\Research\ML\raylearner\AutomaticPlanning 1.1\Model_oro30_1\Model_copy_predict_mimick_expected.py
    a. Copy plans: batch_copy_plans(COPY_PATIENT_PLAN_FILE, NEW_PLAN_NAME). Specify the new plan name in the script. Copy plans takes around 1-2 minutes per plan.
    b. Prediction: batch_autoplanning_predict(PATIENT_PLAN_FILE, MODEL_NAME, MODEL_DIR, [STRAT_NAME]). Choose the strategy you want for prediction in ‘STRAT_NAME’. Prediction takes around 15 minutes per plan.
    c. Mimicking: batch_autoplan_mimick_plans(PATIENT_PLAN_FILE, MODEL_DIR, STRAT_NAME). Mimicking takes around 45 minutes per plan.