TAMPEREEN TEKNILLINEN YLIOPISTO
TAMPERE UNIVERSITY OF TECHNOLOGY

TOMPPA PAKARINEN

IMPLEMENTATION OF A KNOWLEDGE-BASED TREATMENT PLANNING SYSTEM AT TAMPERE UNIVERSITY HOSPITAL

Master of Science Thesis

Examiner: Prof. Hannu Eskola
Examiner and topic approved by the Faculty Council of the Faculty of Biomedical Sciences and Engineering on 6th June 2018

ABSTRACT

PAKARINEN, TOMPPA: Implementation of a Knowledge-Based Treatment Planning System at Tampere University Hospital
Tampere University of Technology
Master of Science Thesis, 71 pages
June 2018
Master's program of Electrical Engineering
Major: Biomeasurements and bioimaging
Supervisor: PhD Antti Aula
Examiner: Professor Hannu Eskola

Keywords: Rapidplan, knowledge-based treatment planning, intensity modulated radiation therapy, volumetric arc therapy, model building, model training, model validation, optimization.

External beam radiotherapy is the most often used radiation therapy method in curative and palliative cancer treatment. Since the discovery of X-rays, radiotherapy has developed to highly sophisticated treatment system consisting of multiple phases and challenges. Successful cancer treatment requires expertise and continuous co-operation across different professions. Today's radiotherapy methods aim for optimal dose delivery with dynamically conformed field shapes, minimizing the harmful dose effects in surrounding normal tissue.

In this Master of Science thesis the radiotherapy plans were constructed for intensity-modulated radiotherapy (IMRT) and volumetric arc therapy (VMAT) using Rapidplan (RP), a knowledge-based treatment planning (KBTP) system. Without KBTP, the planner must interactively guide the plan optimization. This is time consuming and may produce lower plan coherence between different planners. Previous studies have shown that RP generated plans shorten the planning time, increase planning coherence within hospitals and can generate clinically acceptable plans with proper organs at risk (OAR) sparing.

In this thesis two head and neck cancer- (HNC) and a prostate model were built in RP. In addition, a previously built robust prostate model was modified for further validation. Prostate models were trained using 126 and 38 plans and HNC models were trained with 156 plans. Model evaluation statistics were used as guiding indicators and most OAR structures yielded good model fit statistics ($R^2 > 0.7, X^2 < 1.1$). Only the robust prostate model had large deviations ($\Delta R^2 > 0.1$) from the guidelines.

The model validation against clinical plans showed similar results to previous research. All RP models could create individual plans meeting the clinical dose-volume constraints and were mainly comparable with the clinical validation plans with no statistically significant deviations ($p < 0.05$). Differences were found in higher PTV doses for prostate and for those OAR structures, which have high sparing priority in clinical planning. This thesis shows that RP models can produce clinically acceptable plans with proper OAR sparing and conformal PTV dose distributions. As a conclusion, RP-generated plans can be used in treatment planning directly or as a starting point for manual optimization.

TIIVISTELMÄ

PAKARINEN, TOMPPA: Automatisoidun sädehoitosuunnittelu-ohjelmiston käyttö Tampereen yliopistollisessa sairaalassa
Tampereen teknillinen yliopisto
Diplomityö, 71 sivua
Kesäkuu 2018
Sähkötekniikan diplomi-insinöörin tutkinto-ohjelma
Pääaine: Biomittaukset ja biokuvantaminen
Ohjaaja: Tohtori Antti Aula
Tarkastaja: Professori Hannu Eskola

Avainsanat: Rapidplan, dataan perustuva hoitosuunnittelu, intensiteetti moduloitu sädehoito, kaari-sädehoito, sädehoitomalli, sädehoitomallin harjoittaminen, validointi, optimointi.

Ulkoinen sädehoito on yleisimmin käytetty sädehoidon tekniikka sekä parantumiseen tähtäävässä, että palliatiivisessa syövän hoidossa. Menetelmänä sädehoito on kehittynyt aina röntgensäteilyn löytymisestä asti yhdeksi teknologialtaan hienostuneimmista hoitomenetelmistä. Sädehoito koostuu useasta eri vaiheesta sekä haasteista, joiden toteuttamiseen ja ratkaisemiseen tarvitaan usean eri ammattiryhmän yhteistyötä. Tänä päivänä sädehoito tähtää optimaaliseen annoksen kohdistamiseen, jossa hyödynnetään dynaamista sädehoitokentän modulointia, tarkoituksena minimoida haitallinen tervekudoksen annos.

Tässä opinnäytetyössä sädehoitosuunnitelmat luodaan intensiteetti modulaatio- (IMRT) ja kaari-menetelmille hyödyntäen Rapidplan (RP) sädehoitosuunnittelu-ohjelmaa. RP luo automaattisesti hoitosuunnitelmaan vaadittavat optimointiparametrit ilman hoitosuunnittelijan interaktiivista ohjausta. Optimoinnin manuaalinen ohjaus on aikaa vievää ja voi johtaa alentuneeseen yhtenäisyyteen hoitosuunnitelmien välillä. Aikaisemmat tutkimukset osoittavat, että RP:n avulla luodut hoitosuunnitelmat lyhentävät suunnitteluaikaa ja lisäävät koherenssia, luoden myös kliinisesti hyväksyttyjä hoitosuunnitelmia riittävällä tervekudossäästöllä.

Työssä luotiin kaksi RP-mallia pään- ja kaulan alueen syöville (HNC) sekä yksi prostata syövälle. Myös yhtä aiemmin luotua prostata-mallia muokattiin validointia varten. Prostata mallit harjoitettiin 126 ja 38 kliinisellä suunnitelmalla. HNC-mallit harjoitettiin käyttäen 156 suunnitelmaa. Lopullisten regressiomallien ennusteet kuvasivat hyvin harjoitusdataa ($R^2 > 0.7, X^2 < 1.1$). Vain yleisen prostata-mallin tulokset poikkesivat suuresti tavoitteesta ($\Delta R^2 > 0.1$).

Mallin luomien suunnitelmien vertailu kliinisiin suunnitelmiin tuotti aikaisempiin tutkimuksiin verrattavia tuloksia. Kaikki RP-mallit pystyivät luomaan kliinisesti hyväksyttäviä sädehoitosuunnitelmia, jotka eivät suurilta osin eronneet ($p < 0.05$) kliinisistä suunnitelmista. Eroja esiintyi osittain korkean annoksen kohderakenteissa sekä IMRT suunnitelmissa prostatalle. Myös riskielimet, joiden matala annos priorisoitiin korkealle, poikkesivat tilastollisesti kliinisistä suunnitelmista pääasiassa RP-suunnitelmien korkeammilla annoksella prostatelle ja HNC:lle. Tämän opinnäytetyön perusteella voidaan kuitenkin todeta, että RP-pohjaiset suunnitelmat voivat luoda kliinisesti hyväksyttyjä suunnitelmia riittävällä tervekudossäästöllä.

PREFACE

This master's thesis was done at Tampere University Hospital in radiotherapy ward. At first, I want to thank my thesis supervisor PhD. Antti Aula for his great input and excellent guidance throughout the whole project. I also want to thank my examiner Professor Hannu Eskola for his advices and for introducing this possibility to me.

Special thanks belong to Department Chief Physicist Mika Kapanen for giving me the opportunity to join their radiotherapy group, and to all physicists in TAYS radiotherapy ward for their help and support.

Last but not least, I would like to thank my family and grandparents for their support and especially my spouse, who have been there for me over all these years.

At Tampere, 13.5.2018

-Tomppa Pakarinen

CONTENTS

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVATIONS

A	Total area of multi-leaf collimator's aperture
a	Parameter describing the dose-induced volume effect to a specific tissue
α	Level of statistical significance
α_{cor}	Bonferroni-method corrected level of significance
b_i	Optimization constraint function bounds
CI_{RTOG}	Conformity index defined by Radiation Therapy Oncology Group
D	Dose
$D_{OAR}(x)$	Dose in OAR volume
$D_{target}(x)$	Dose delivery to the target volume
D_{50}	Dose which causes certain tissue response with 50 % probability
E	Regression model's expected result
f_o	Optimization objective function
$l = 1, \dots, m$	Optimization constraint functions
k	Steepness of the dose response curve
m	Number of comparisons in multiple comparison test
O	Observation
$P(D)$	Dose response probability
RF	Repeat factor
R^2	Coefficient of determination
τ	Leaf thickness
$\|t\|$	Test value for t- test
TMR	Tissue maximum ratio
TMR_{ave}	Average tissue maximum ratio
TV	Target volume
v_i	$i{:}\,th$ fractional organ volume
V_{RI}	Reference isodose volume
WF	Weight factor
$\boldsymbol{x}$	Vector including MLC leaf positions
$\boldsymbol{x}_{opt}$	Vector for optimal solution
X	Geometric feature's value
X_{max}	Maximum value in the training set
X_{med}	Median of the training set
X_{min}	Minimum value in the training set
X_{10}	10 % percentile of the values in the training set
X_{90}	90 % percentile of the values in the training set
χ^2	Chi-squared model fit statistic
Y	Predicted or observed value

y_x	Observation
3D CRT	3-dimensional conformal therapy
aA	Areal difference of estimate
AAA	Anisotropic Analytical Algorithm
Acuros XB	Dose calculation algorithm in Eclipse software
AIO	Automatic interactive optimization
CAT	Conformal arc therapy
CI	Conformity index
CT	Computed tomography
Df	Degrees of freedom
DVH	Dose volume histogram
EUD	Equivalent uniform dose
GED	Geometry-based expected dose
gEUD	Generalized equivalent uniform dose
HNC	Head and neck cancer
HU	Hounsfield unit
ICRU	International Commission on Radiation Units and Measurements
IMRT	Intensity modulated radiation therapy
KBTP	Knowledge Based Treatment Planning
LINAC	Linear accelerator
MLC	Multi-leaf collimator
MRI	Magnetic resonance imaging
MSE	Mean squared error
MU	Monitor unit
NTCP	Normal tissue complication probability
OAR	Organ at risk
PDD	Percentage depth dose
PO	Photon optimizer
PTV	Planning target volume
RP	Rapidplan
RT	Radiation therapy
RTOG	Radiation Therapy Oncology Group
SD	Standard deviation
SIB	Simultaneous integrated boost
TAYS	Tampere University Hospital
TAYS coh	Coherent Rapidplan model for head and neck cancer
TAYS tot	Robust Rapidplan model for prostate cancer
TCP	Tumor control probability
UI	User interface
VMAT	Volumetric arc therapy

1. INTRODUCTION

During the past years, Knowledge Based Treatment Planning (KBTP) has attracted interest among clinicians, physicists, and other medical professionals in the field of radiotherapy (RT). The term knowledge-based treatment planning derives from the core idea behind the method, which is to use tens or even hundreds of clinical radiotherapy plans to create and train a single estimation model. The model holds and uses the wide range of clinical planning experience to create individually defined optimization objectives. The objectives are then used for plan optimization generating plans, which reflect the hospitals planning conventions.

There are several steps before a patient with radiotherapy prescription can be treated. Often the most time-consuming step is to create the treatment plan, which must meet the prescribed dose requirements and yet spare surrounding tissues and organs from. Different cancers require different levels of complexity from the treatment plans, e.g. in general a treatment plan for prostate cancer may be considered relatively simple compared to head and neck area. Creating a complex plan with multiple organs at risk (OAR) may consume hours of work and several re-planning iterations, which is one of the main motivators behind development of automated treatment planning software.

Multiple large medical companies, including e.g. Varian (Palo Alto, USA) and Philips (Amsterdam, The Netherlands) have released their automated planning systems. (Schubert 2017) Also, a variety of research have been published by companies and research groups around the world including several different types of cancer, such as breast-, head and neck- and cervical cancer. (Wang 2017) (Tol 2015) (Wu 2016) (Fogliata 2015, 2017) Developers of the automated KBTP algorithms have proposed that use of KBTP in RT - treatment planning would increase plan coherence between different planners, within and between hospitals. Additionally, the developers propose decreased treatment planning time. (Krauenbuehl 2015)

Shortly described, KBTP algorithm takes organ structures and further features from the training plans and uses them to create a model which can reconstruct the plan's dose volume histograms (DVH) with a small error margin. The model is then later used to create DVH estimations and optimization objectives for completely new patients. In currently used automatic interactive optimization (AIO) the user must dynamically change the optimization objectives for optimal optimization result. When KBTP software's automatically generated planning objectives are used, the planner is no longer required to interactively guide the optimization process. The plan can be validated or used as a starting point for a manual optimization. Even though a substantial portion of the previous

research have demonstrated better planning target coverage and normal tissue sparing, KBTP is not claimed to produce superior plans in comparison to individual manual plans, but rather increase the planning quality in wider scale. (Eclipse IOU 2015) (Fogliata 2015) (Tol 2015)

In this thesis work the Varian's KBTP system named Rapidplan (RP) is studied for its capabilities in treatment planning. RP generated plans are compared to their clinical counterparts and to each other and the plan quality is measured with different parameters, such as target volume's conformity index (CI) and monitor unit (MU) delivery. Plans are also tested against the Tampere University Hospital's (TAYS) planning constraints and guidelines for further confirmation whether RP can create clinically acceptable RT plans.

The first objective of this thesis work is to test several rebuilt RP models for prostate cancer and head and neck cancer (HNC). The models will be ranked by preliminary DVH comparisons using small validation data sets. The second objective is to choose and improve the best performing model for later validation against the clinical plan. The third and the main goal of this thesis work is to build and validate completely new prostate- and HNC models. Additional objectives are to build an analysis program with Matlab, which can be used in model testing and in validation stages. Finally, the second round, iteratively trained RP model is studied for its capabilities of producing either higher OAR sparing or target coverage compared to the clinical plans and to the original HNC model.

The necessary theoretical background of this thesis work is presented in chapter 2. Chapter 2 includes general information about RT and RT-planning, the used programs, algorithms and their functions, and a review of previous research. The materials and methods of this thesis work are presented in chapter 3, which roughly divides in three sections; the methods used in model building, model estimation and model validation. Chapter 4 consists of the results acquired with the methods mentioned above. The results are divided in 2 main parts, prostate model and HNC model, which include several subsections considering model evaluation, dose- and statistical analysis. In chapter 5, the results are analyzed with more depth together with observations from previous research and possible reasons are further discussed. Finally, conclusions are presented in chapter 6.

2. THEORETICAL BACKGROUND

In this chapter the necessary theoretical background for this thesis is presented. The chapter is an overview of radiotherapy care path and of the two treatment methods used in this work. Chapter also provides insight for Eclipse treatment planning software and for the Rapidplan software, including an algorithm overview, dose analysis and optimization sections. For optimization, an illustrative example for radiotherapy optimization problem construction is presented. The theoretical background for this work focuses strongly towards the computerized treatment planning, rather than in the physics or biological effects behind radiotherapy.

2.1 Radiotherapy

Radiotherapy is a cancer treatment method including several steps and sub-methods. The physical radiotherapy treatment is based on the biological tissue response for ionizing radiation. Radiotherapy may be divided in two main types: internal radiotherapy and external radiotherapy. (Tenhunen 2007) In this thesis, we focus only on the latter.

In early 20th century, after the discovery of X-rays, scientists found that the negative normal tissue responses to radiation decreased when the total dose was delivered in fractions rather than all at once. Later it was also discovered that different tissues have characteristic dose-time responses to radiation and the treatment result with number of unwanted effects of radiation can be controlled by careful treatment planning. (Moonen 1994) The number of fractions is usually between 15 to 35 and the single fraction dose is typically 1.8 Gy – 2.0 Gy (Tenhunen 2007).

The macroscopic effect of radiation therapy is measured with normal tissue complication probability (NTCP) and with tumor control probability (TCP) as function on dose. There are several methods to define mathematically the dose response, where the two parameters are considered. One way to represent the dose response probability is presented in equation 1.

$$P(D) = \frac{1}{1+\left(\frac{D_{50}}{D}\right)^{k}}. \quad (1)$$

Here D_{50} is the dose, which causes the chosen response (NTCP or TCP) with 50 % probability, D is the dose and k describes the steepness of the curve. (Tenhunen 2007) (Baumann 2005) The most important point of equation 1 may not be the exactness, but rather the general characteristic of increased small probability with small doses and the rapid increase in probability after a certain tissue dependent threshold. (Tenhunen 2007)

A definition of a good radiotherapy treatment plan is case dependent, but number of general metrics apply in every treatment plan. It is important that the delivered (prescribed) dose is conformal for the target volume and the dose distribution inside the planning target volume (PTV) is homogenous. Also, the radiation dose in surrounding tissues and OARs should be minimized (Yoon 2007) (Oh 2007). The radiation doses are prescribed by a physician and the dose constraints for different tissues and organs are defined either by physician, guidelines based on research, or both.

2.1.1 Radiotherapy care path

Radiotherapy care path can be considered as an iterative loop, which starts from the diagnosis of the disease and ends to the treatment delivery. The status of the patient is followed though-out the treatment and the treatment planning (prescription, structure contouring, computed plan etc.) is modified according to the progress if needed.

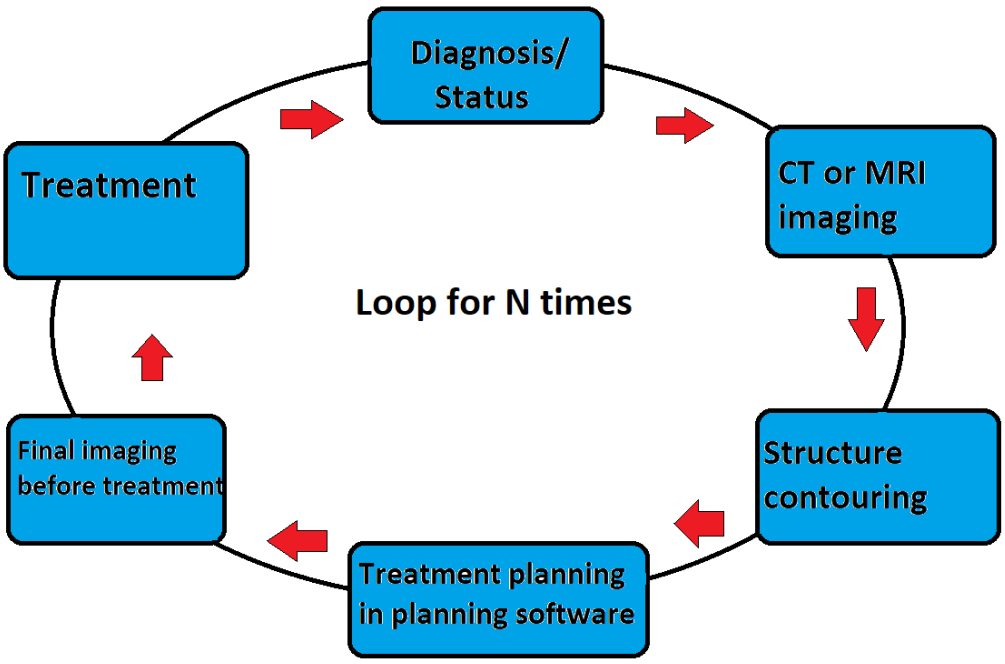

***Figure 1**. Author's perception of the radiotherapy care path.*

As mentioned above, the first stage is the diagnosis of the cancer. This stage is often done by medical imaging or possibly biopsy from the tumor (Galimberti 2002). The next stage is to use computed tomography (CT) imaging while the patient is in planned treatment position. The acquired CT stack is then used to locate and contour the target tissue/tissues, normal tissue, and organs at risk. Physician who also prescribes the radiation dose performs contouring. Next, the radiotherapy treatment technique is chosen prior to dose-volume optimization and dose calculation. In this thesis, these steps are concluded in the Eclipse Software (see chapter 2.3). When the treatment plan is ready, the patient is imaged in the treatment position at the linear accelerator or LINAC (treatment machine). The

final step in the treatment chain loop is treatment session itself, where the fractionated dose is delivered to the target volume. After the final step the loop return to status checking and starts over. (Mayles 2007)

2.1.2 Intensity Modulated Radiation Therapy

In this thesis the treatment plans and the treatment model were constructed for 2 radiotherapy treatment methods, intensity modulated radiation therapy (IMRT) and volumetric arc therapy (VMAT). IMRT is based on multi leaf collimator (MLC) controlled beam intensity modulation, where the collimator leaves absorb the excess radiation surrounding the PTV (planning target volume), shaping the incoming radiation field. Unlike in traditional radiation therapy such as 3-dimensional conformal therapy (3D CRT) or in conformal arc therapy (CAT), IMRT offers possibility to simultaneous field shape modulation and intensity modulation. (Bortfield 2006) 3D CRT and CAT techniques use fields conformed for the PTV shape, but the shapes are not dynamically modulated but manually chosen (in CRT) by the treatment planner along with other beam parameters. IMRT has been shown to have superior OAR sparing compared to conformal methods, but also to have increased monitor units during the treatments. (Bakiu 2003) (Maier 2016) (Palma 2008) Another advantageous feature connected to intensity modulating techniques is the simultaneous integrated boost (SIB)-technique, which allows the treatment of multiple target volumes with different dose prescriptions during the same treatment fraction (Orlandi 2016).

The leaves positions are changed by multiple motors to achieve the approximate shape of the radiated target volume and the intensity of beams is increased for the rays that do not penetrate any OARs and lowered to those which do. (Bortfield 2006) Beam intensities, number of the beams and the radiation angles depend on the location of the tumor and OAR’s. The goal is to minimize the dose around the PTV and maximize the dose inside the PTV. In figure 2 is illustrated an example of MLC system.

Figure 2. *Multi-leaf collimator (MLC). (Bortfield 2006)*

As seen in the figure 2. MLC resolution is limited by the leaf width. Increasing the number of beams will have a positive effect on the dose distributions at the target volume, but only until a certain level. According to Bortfield (2010) by dismissing the variance of scattering by depth and by assuming flat dose-depth profile the mathematical formulation for IMRT fields yields a result which supports this claim.

Usually, IMRT planning is done as inverse planning, where the leaf positions are optimized according to the prescribed dose to target areas. This kind of problem is commonly treated as an optimization problem. Inverse IMRT optimization result depends largely on the prescription plan and without a consistent plan optimization result may be far away from optimal or not meeting the minimum optimization goals. This leads to a technique called as iterative planning, which is used dynamically together with the inverse optimization. In iterative planning the prescription is modified to improve the optimization algorithms performance. In Eclipse's DVH optimization similar iterative loop is used to help the optimization, but instead of tweaking the prescription, the output parameter objectives such as mean dose, generalized equivalent uniform dose (gEUD) and maximum dose of the optimization are modified. *(Bortfield 2006)* (Eclipse Algorithm 2015)

In this thesis work IMRT is the one of the two techniques used in treatment planning. Approximately half of the cases for both, prostate and head and neck models were IMRT plans and the built models can produce plans for both treatment techniques. In future VMAT is becoming more common in TAYS treatment planning, but IMRT, CAT and 3D CRT remain as an option.

2.1.3 Volumetric Arc Therapy

Volumetric Arc Therapy is an advanced form of IMRT. Similar to IMRT, VMAT delivers the prescribed dose by modulating the intensity of the fields from different angles, but instead of multiple fixed fields, VMAT operates dynamically over gantry rotation (360 degrees) around the patient. The final dose can be delivered completely during the rotation, which also decreases the delivery time. (Ghandour 2014) Varian's Eclipse uses the direct aperture optimization in VMAT planning. Instead of multiple fixed fields, VMAT plan consists of arc or arcs, which cover the prescribed dose using several field shapes inside the arc. Typically, single arc is used e.g. for prostate plan, but as the complexity and target size increase the number of arcs needed increases because several arcs allow more shapes for the same gantry angle and thus, for large targets multiple arcs instead of single arc may increase the target coverage. (Cao) (Nithya 2014) (Teoh 2011) Figure 3 presents the comparison between 1 and 2 arcs used for the same HNC case.

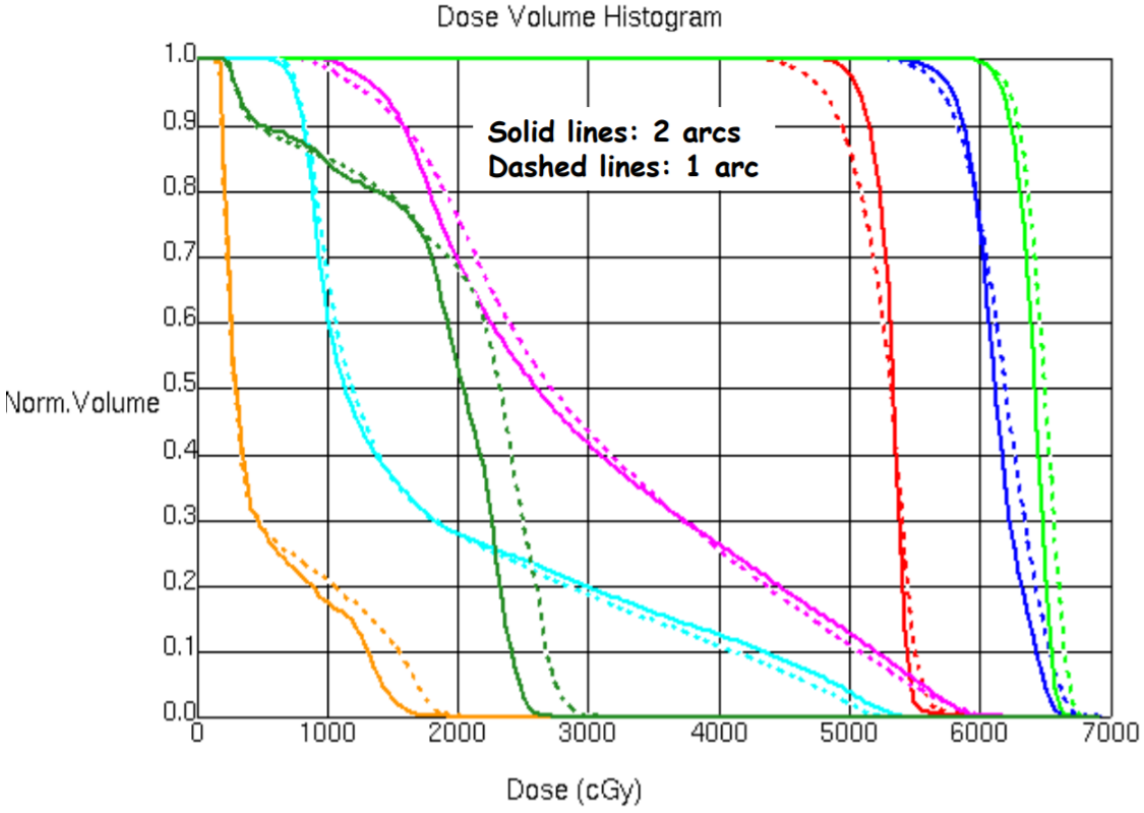

Figure 3. *Comparison between single and 2 arc VMAT DVH plan. (Cao)*

The figure includes 3 target volumes (green, blue and red) and 4 OARs. The higher target coverage for multiple arcs is visible especially in 2 lower dose PTVs (blue and red) and increased uniformity. The effect of adding yet another arc for this HNC case is presented in figure 4.

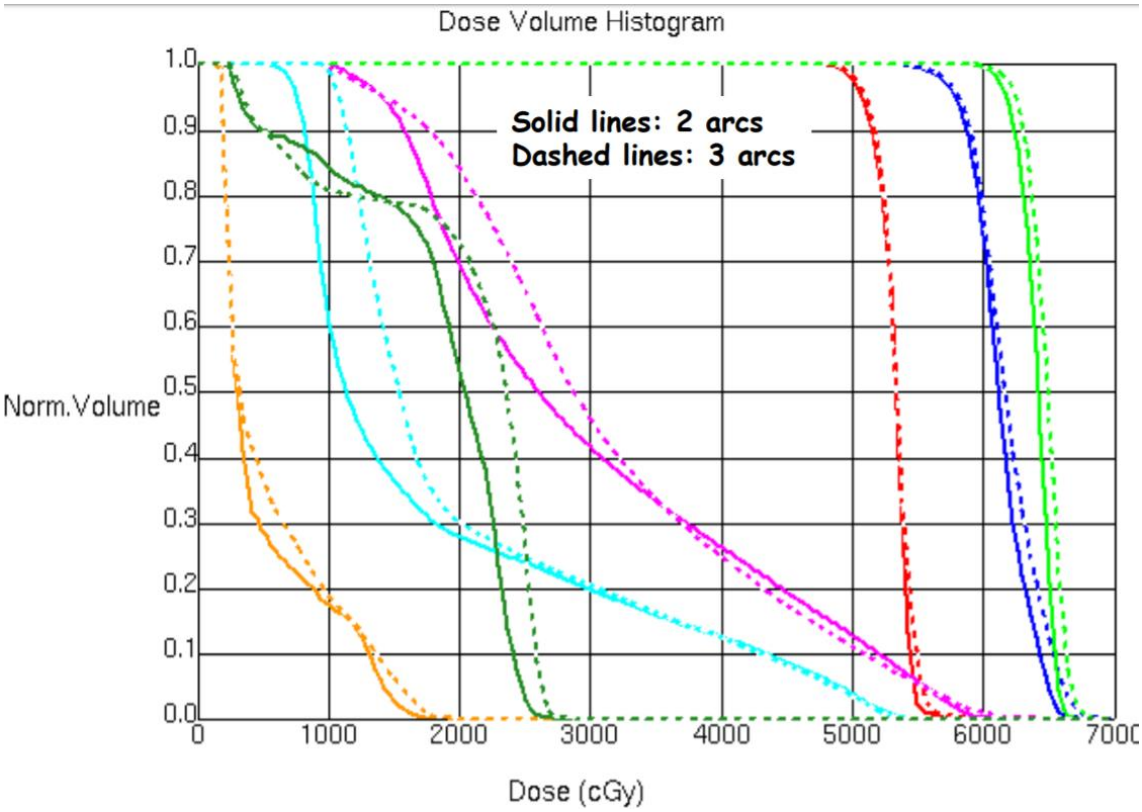

Figure 4. *Comparison between 2 and 3 arc VMAT DVH plan. (Cao)*

As seen in figure 4, the third arc does not bring additional benefit to the target coverage in this case and the OAR sparing is ambiguous for higher doses. One should also take in account the increased delivery time for each added arc, (Cao) or more trivial aspects such as, increasing number of arcs to increase the delivery time so that larger shape modulations are allowed by the optimization algorithm (Tol 2014).

2.1.4 Dose-volume histograms

Understanding dose-volume histograms is an essential requirement in understanding the results of this thesis work. DVH can be defined as a quantitative tool in evaluating radiotherapy treatment plans (Ting 1997). The figure 5. Presents an example of a dose volume histogram produced from HNC treatment plan.

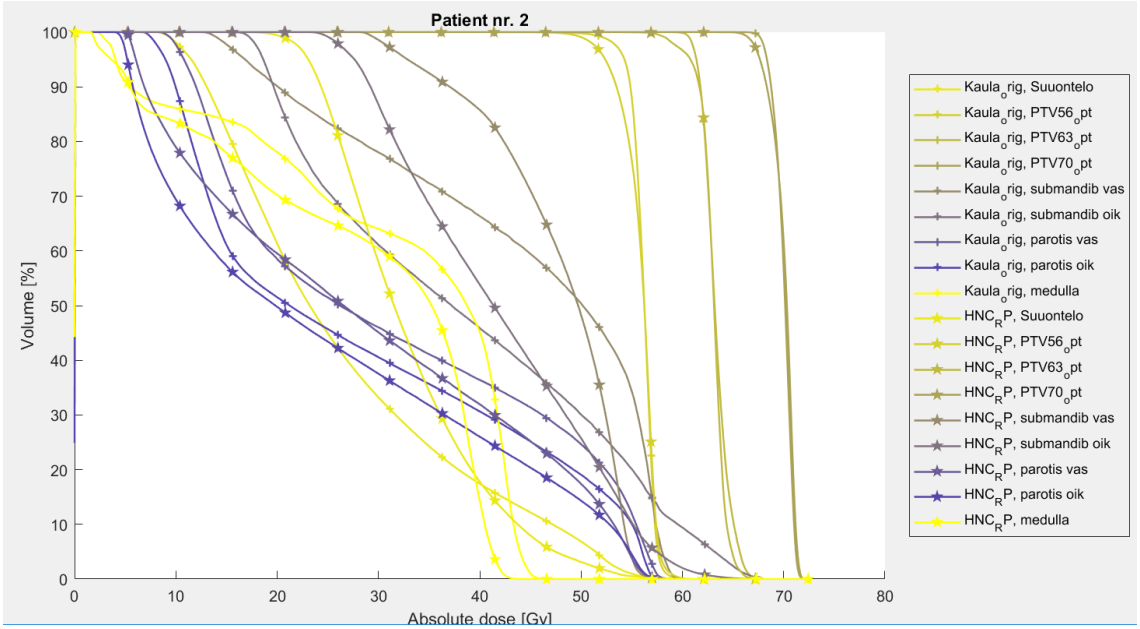

Figure 5. *Example of DVH plot from HNC treatment plan.*

As seen in figure 5, covered volume percentage is plotted as a function of the absolute dose in grays (Gy). The different curves in the figure represent different structures, which can be divided in this thesis as target volumes and OAR. Targets are structures that have a prescribed dose, which must be delivered in certain volume. In this work the PTV structures in DVH analysis were used. PTV is defined as the true volume of the tumor with excluded margins covering the uncertainty and variability in patient positioning and beam adjustment (Grosu 2006) (Burned 2004) (Radiotherapy Board 2015)

OARs are structures that are aimed to be spared from radiation but are located near the target-volume (Grosu 2006) Thus, an ideal DVH curve would show 100 % volume fill in the target-volume at the prescribed dose, and no dose to the OARs.

The colors in figure 5 present the different structures and the line-symbol presents the model used to produce the line, but in this example different models may be omitted. The three targets in this case are located at higher doses ($56 - 70\ Gy$). The steep angle in the PTV curve indicated high conformity and dose coverage in the volume. For example, 95 $vol - \%$ of the highest dose PTVs receive approximately 68 Gy doses, which is 97 % from the prescribed 70 Gy dose.

Unlike for PTVs, it is important that OARs doses are minimized in radiotherapy plan (Mao 2015). Because the OARs (by definition) are organs which lay near the high dose structures, it is impossible with current technology to achieve complete OAR sparing. Thus, some prioritization must be made during treatment planning. In general, organs may be separated in two groups, serial and parallel organs. Serial organs are prone to complications when even a small volume inside the organ receives high dosage. Parallel organ complication is more dependent on the total volume that receives the radiation.

(Fiorino 2009) Dose objectives for DVH curves are usually set with complication types in mind, especially if the planner uses gEUD objectives (chapter 2.3.3). For the resulting DVH curve this is seen as serial organs having higher mean dose, with steeper decrease before the maximum (e.g. medulla in figure 5). Parallel organs may show higher dose maximums because usually the mean dose is minimized (e.g. parotid gland).

As mentioned above, the OAR structures in figure 5 are represented by the curves under PTVs (high dose curves). If the OAR is located near the target volume the dose is increased e.g. by the dose spillover and penumbra (Sasane 1981) and the OARs DVH curve is expanded to the higher doses. In figure 6 this effect is shown clearly for submandibular glands (purple and gray curves).

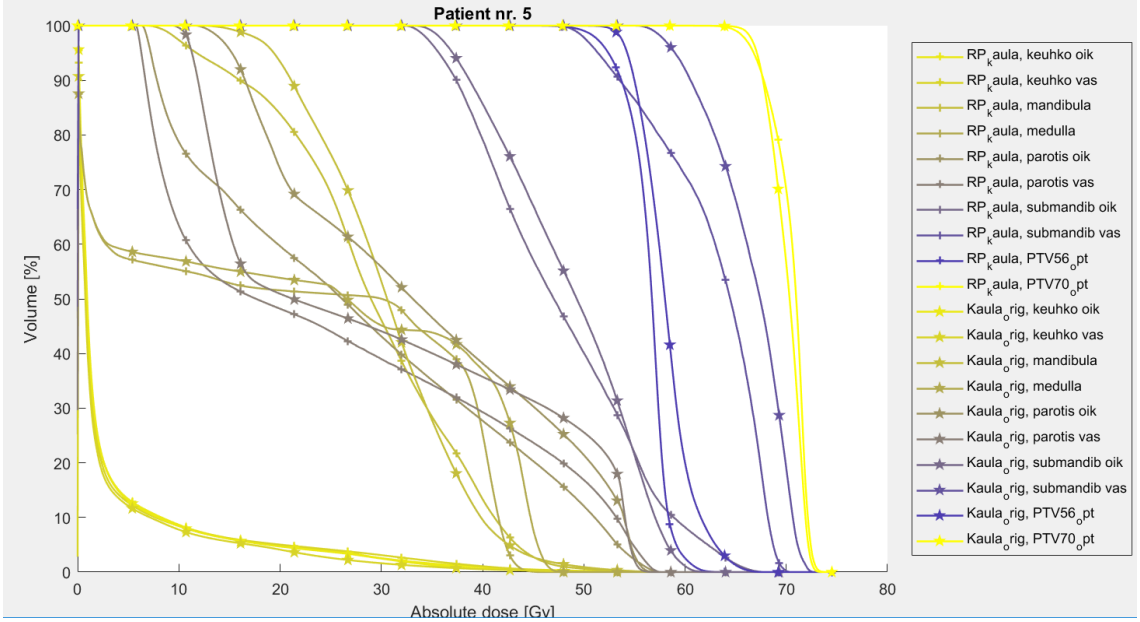

Figure 6. *DVH curve example for HNC case.*

The opposite effect for OAR locating far from PTV/PTVs is clear in case of lung DVH curves in figure 6 (located in lower left corner). Note also that the figures represent the total cumulative dose over the whole treatment instead of dose per fraction.

Even though the DVH curve representation is generally considered as a good evaluation tool and has a major role in this work, there has been some criticism about the method. For example, Ting's group in their study Dose-Volume histograms for bladder and rectum concluded that DVH concept may not be accurate representation of the dose distribution, since the DVH treats the structures as solid volumetric objects instead of the shell-like nature of some organs. (Ting 1997) But the DVH accuracy deviations depend largely on the system and algorithm in use and the deviations were found to be low in Linac-based treatment planning system (max error 1.7 %) by Grossmann's group when using a proper quality assurance method (Grossmann 2010).

2.1.5 Dose analysis parameters

In this thesis work, the dose analysis is performed using statistical metrics treating the DVH constrains, dose-volume point values, delivered monitor units and conformity indices. This chapter reviews only the necessary theory of the latter two. Theory considering DVH curves and dose-volume objectives are reviewed in chapters 2.1.4 and 2.3.3 respectively.

Conformity index is a measure of dose coverage in the target volume and it generally presents the ratio between the isodose volume and the target volume. Or in other words, it indicates as one numerical value how well the actual dose meeting a chosen dose-level is covering the total target volume. (Feuvret 2006) CI has multiple slightly different definitions, but in this work the Radiation Therapy Oncology Group's (RTOG) definition is used.

$$CI_{RTOG} = \frac{V_{RI}}{TV}, \tag{2}$$

where V_{RI} is the reference volume/isodose and TV is the planned target volume. (Petkovska 2010) (Feuvret 2006) The V_{RI} was defined in this work as the isodose covering 95 % of the prescribed dose for each PTV separately. The detailed explanation of the procedure is presented in chapter 3.4.5.

The second parameter used in dose analysis apart from DVH statistics is the Monitor unit. For sake of simplicity, MU can be considered as a treatment machine -specific calibrated unit, which controls accurately the dose output of each field or arc. (Bourdland 2016) The maximum error has been stated by International Commission on Radiation Units and Measurements (ICRU) to be 5 % for any dose delivery system (Gibbons 2001). The monitor unit for photons is formulated in Eclipse as

$$MU = \frac{RF \times WF}{TMR_{ave} \times OF_{TMR_{max}}(S) \times DR_{ref}}. \tag{3}$$

Here RF is the repeat factor, WF is the weight factor, TMR the tissue maximum ratio and $OF_{TMR_{max}}(S) \times DR_{ref}$ contributes as calibration factor. The denominator of equation (3) corresponds to the absolute dose from and arc to isocenter and TMR_{ave} is the average tissue maximum ratio from the edge of body contour to the isocenter. (Eclipse Algorithms) In TAYS the monitor unit is defined so that 100 MUs correspond to 1 Gy open field dose. Monitor units are calculated by the Eclipse Software and the resulting values from individual IMRT fields or VMAT arcs can be added together to achieve the total monitor units.

2.2 Knowledge-Based Treatment Planning

Knowledge-Based Treatment Planning is a relatively new method to produce radiotherapeutic treatment plans. Even though RT treatment planning is rapidly developing towards automated planning, still multiple steps are done manually, e.g. optimization objectives and structure contouring. (Wang 2017) Because the conventions and experiences vary between different planners, hospitals and regions, manual treatment planning may decrease coherence between treatment plans. (Schubert 2017)

The idea behind Rapidplan or in any KBTP system is to create a trained model, which automatically produces DVH estimations and optimization constrains for a given (new) patient. These constraints and estimations are further used in optimization process to produce the final dose-volume distributions. Typical size for a model's training set is between 20-150 plans depending on the cancer type and expected plan feature variation and complexity. The training set's planning parameters are then used to create a general model, which is further can generate individualized DVH optimization objectives. According to Varian Medical, optimized plans based on RP can increase the plan coherence and shorten the planning time (Eclipse Algorithms 2015)(Eclipse IOU 2015)

2.3 The Eclipse Software

Varian's Eclipse is a radiotherapy treatment planning software, which allows physicians and physicist to create, visualize and optimize treatment plans in clinical use, based on CT- or magnetic resonance image (MRI) data. Eclipse uses the CT stack, or the pseudo CT stack created from MRI data to calculate the Hounsfield unit (HU) values (see chapter 2.3.2). Eclipse includes a wide range of functionality, including 3D CRT, IMRT, VMAT, conformal arc, proton planning, several optimization algorithms, dose calculation, knowledge-based planning (Rapidplan) and plan evaluation. (Eclipse Algorithms 2015) (Eclipse IOU 2015) In this thesis work KBTP is used to create IMRT and VMAT treatment plans and the dose-volume histograms are further analyzed. All the plans and models are built with external beam, 6 MV photon energy and thus, other algorithms considering other methods, such as electron treatment are excluded.

2.3.1 Optimization

Optimization is a mathematical procedure where the goal is to find the most suitable (optimal) solution from a group of possible solutions. An optimization problem can be generally formulized as

$$\begin{aligned} &minimize\ f_o(x) \\ &subject\ to\ f_i(x) \leq b_i, l = 1, \dots, m. \end{aligned} \tag{4}$$

In equation 4, x is a vector containing n variables and is defined as the optimization variable, and $f_o(x)$ is the objective function (or cost function) of the problem. The problem is constrained by the constraint functions, $l = 1, \dots, m$ and the constraint functions are limited by the bounds $b_i = b_1, b_2, \dots, b_m$. The vector x is the solution for the optimization problem when it satisfies all the constrains. Thus, the optimal solution is vector x_{opt}, which yields the minimal value for the objective function amongst all the solutions. (Boyd 2004) The optimization problems in this work are divided in linear and nonlinear optimization problems. Linear optimization problem is defined by the linear nature of the objective and constraint functions. Linear optimization problems are usually simpler than nonlinear problems (Boyd 2014). An example optimization problem derived from these rules is presented in figure 7.

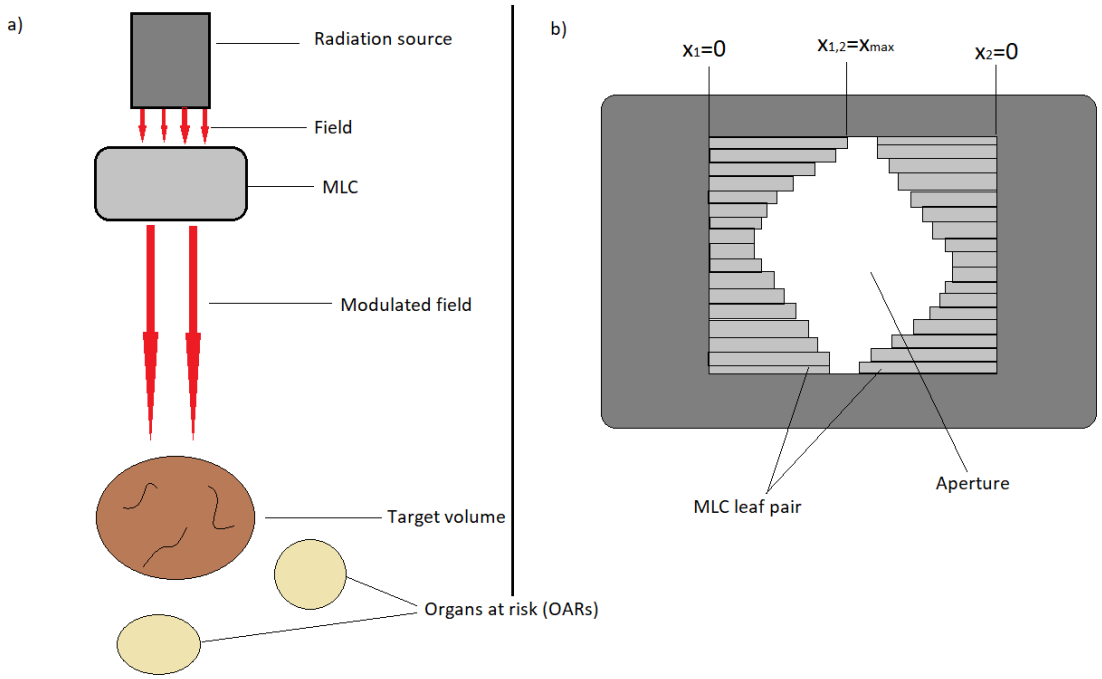

Figure 7. *Schematic of the simplified optimization problem arrangement (a) and MLC bottom view schematic (b).*

As an imaginary example we can consider simplified dose-optimization problem. Visualization of the arrangement (side view) is presented in figure 7a and bottom view of the MLC alone is presented in figure 7 b. For the optimization problem the constrains could be e.g. the possible MLC leaf positions and required dose-volume delivery to the target volume, so that the solution vector becomes

$$x = [x_1\ x_2\ \dots x_n].$$

Here x_i is the leaf position for the ith MLC leaf, when the leaves are counted from top left (i) to right ($i + 1$) and then left and down ($i + 2$). The leaf positions are limited by

$$0 \leq x_i \leq x_{\max},$$

which is the first constraint. The second constraint could then be the prescribed dose delivery to the target volume,

$$D_{target}(x) \geq prescribed\ dose$$

And the objective function would be the dose in OAR volume (to be minimized).

$$f_o(x) = D_{OAR}(x),$$

Where variable x was again the vector containing the MLC leaf positions and since x holds all the leaf positions, the total area of the aperture would then be

$$A = \sum_{i=1}^{n}(2x_{max} - x_i)\tau, \quad (6)$$

where τ is the leaf thickness and n is the total number of leaves. The modulated field size can then be assumed to be linear with respect to A and thus x. If all organs in figure 6 (a) would have constant thickness, absorption properties and uniform shapes, and all the position depended parameters would be omitted, such as field profile and scattering; the objective function and the constrain function would also then be linear and the whole optimization problem would become linear (Tenhunen 2007).

However, if these parameters are not omitted, the constraint functions and objective function become nonlinear with respect to x, which makes the problem also to a nonlinear optimization problem (by definition) (Boyd 2004) (Tenhunen 2007). When considering these two examples, the linear version is much simpler than the nonlinear version.

Linear optimization does not have an analytical solution, but effective algorithms have been developed for finding the global optimal solution. Most common are the Simplex method and interior point methods. The same is not true for nonlinear optimization, and instead of trying to find the global minimum for the objective function, the algorithms try to find the best solution to locally minimize the objective function. Or in other words, the algorithm tries to find the optimal solution from a smaller group of possible solutions and after a while of trying, is satisfied by the best solution inside the limited group. (Boyd 2004)

The eclipse software uses Simplex method to solve a linear optimization problem for example in fluence optimization. The Photon optimizer (PO) algorithm uses iterative optimization to find the local minima of the objective function. In PO algorithm the objective function is defined as the sum of dose-volume objectives and other user defined objectives. The algorithm also calculates intermediate dose during the optimization and uses the result to calculate the difference to the first optimization result. This can be used to compensate for the next optimization iteration round. (Eclipse Algorithms 2015)(Eclipse IOU 2015) In figure 8 is presented the PO view of Eclipse software.

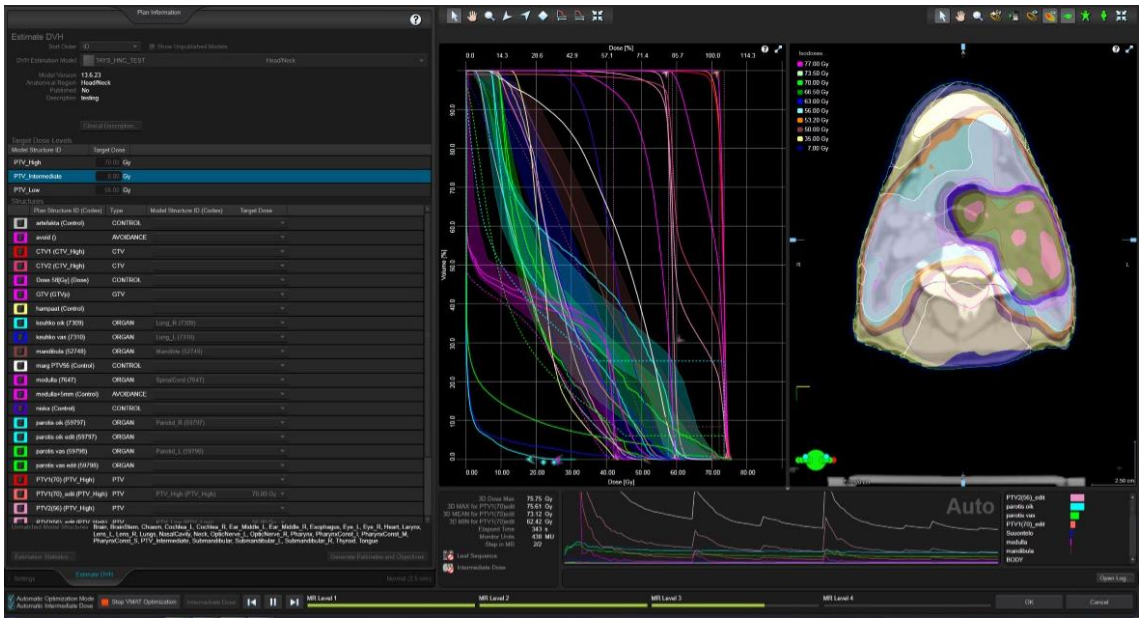

Figure 8. *PO optimizer screen view in Eclipse Software. (Eclipse v.13.6 2015)*

In the left side of the figure are presented the organs matched to the Rapidplan model, which defines the DVH objectives. In the middle of the image are the DVH curves, which present the optimization process in real time. The curves confining the colored areas in the DVH plot are DVH estimations generated by Rapidplan. Convergence of the optimization is presented below the DVH curves. Next to the convergence plot (right side) are presented the objective function values for each structure represented as bars. Finally, above the objective functions are CT slice images of from the patient with colored isodose areas. (Eclipse Algorithms 2015)

In this work, Eclipse Software was set to use photon optimization in DVH optimization, Anisotropic Analytical Algorithm (AAA) algorithm for final dose calculation and DVH estimation algorithm for Rapidplan. Note that the example presented above was not an optimization problem that is used in eclipse, but merely a simplified example for comprehensive demonstration of optimization problem construction.

2.3.2 Dose calculation

Eclipse has several alternative dose calculation algorithms. As described in the validation workflow, dose calculation algorithm must be chosen before calculating the final doses. Suitable algorithms for external beam planning for photons are Acuros XB and AAA. (Eclipse Algorithms 2015) (Herman 2011)

The AAA algorithm uses the CT calibration curve to transform the HU values from CT images to electron densities and truncates the exceeding HU values to the maximum HU values in the CT calibration data. Acuros XB algorithm does the same procedure but the

calculation is halted if CT image's HU values exceed the maximum HU from the calibration curves. Also, the information is transformed to mass density in Acuros XB instead of electron density. (Eclipse Algorithms 2015) (Eclipse Instructions 2015) In this work the AAA algorithm was used for dose calculation. The accuracy of AAA algorithm is presented in figure 9.

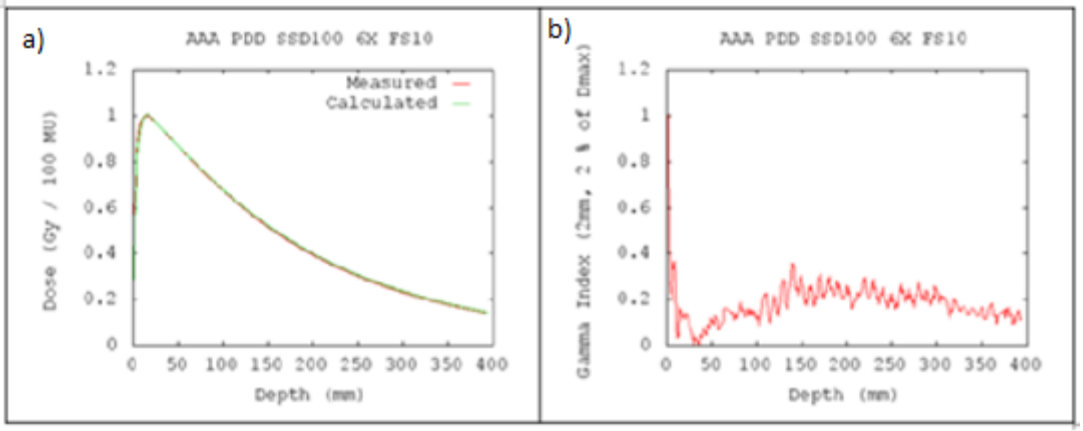

Figure 9. *Depth dose curve comparison with calculated and measured for AAA, where a) shows the depth-dose difference and b) presents the corresponding gamma error. Modified from (Eclipse Algorithms 2015)*

The accuracy of the dose calculation algorithms in Eclipse are measured by comparing the measured depth dose in water phantom to calculated depth dose using gamma error, which is the shortest distance in 4-dimension space where percentage depth dose (PDD) represents the 4th dimension (Eclipse Algorithms 2015). The AAA algorithm is divided in two main parts, configuration algorithm and final dose calculation algorithm. The configuration phase configures different parameters for further fluence, energy and scattering computations. The properties are defined in water equivalent medium. (Sievinen)

The actual dose calculation algorithm calculation is based on beam superposition-convolution using 3-dimensional pencil beams. The algorithm models the primary beam photons and the secondary (scattering) photons. The final dose is the calculated as dose-superpositions from convoluted photons and electrons. (Eclipse Algorithms 2015) (Sievinen)(Herman 2011)

2.3.3 Optimization objectives

In Eclipse Software the optimization objective function is defined by dose objectives and user generated objectives. PO algorithm supports upper, lower, mean and three different biological optimization objectives. (Fogliata, 2017) The upper objective is used to define the desired upper dose limit for a certain volume. Similarly, the lower objective is used to set the lowest dose for chosen volume and the mean objective to define the maximum mean dose for the structure. The objectives are shown in the DVH curve figure 10 and

may be manually changed during the optimization to achieve better optimization results. Objectives and DVH plot in Eclipse are presented in figure 10. (Eclipse Algorithms 2015)(Eclipse instructions 2015)

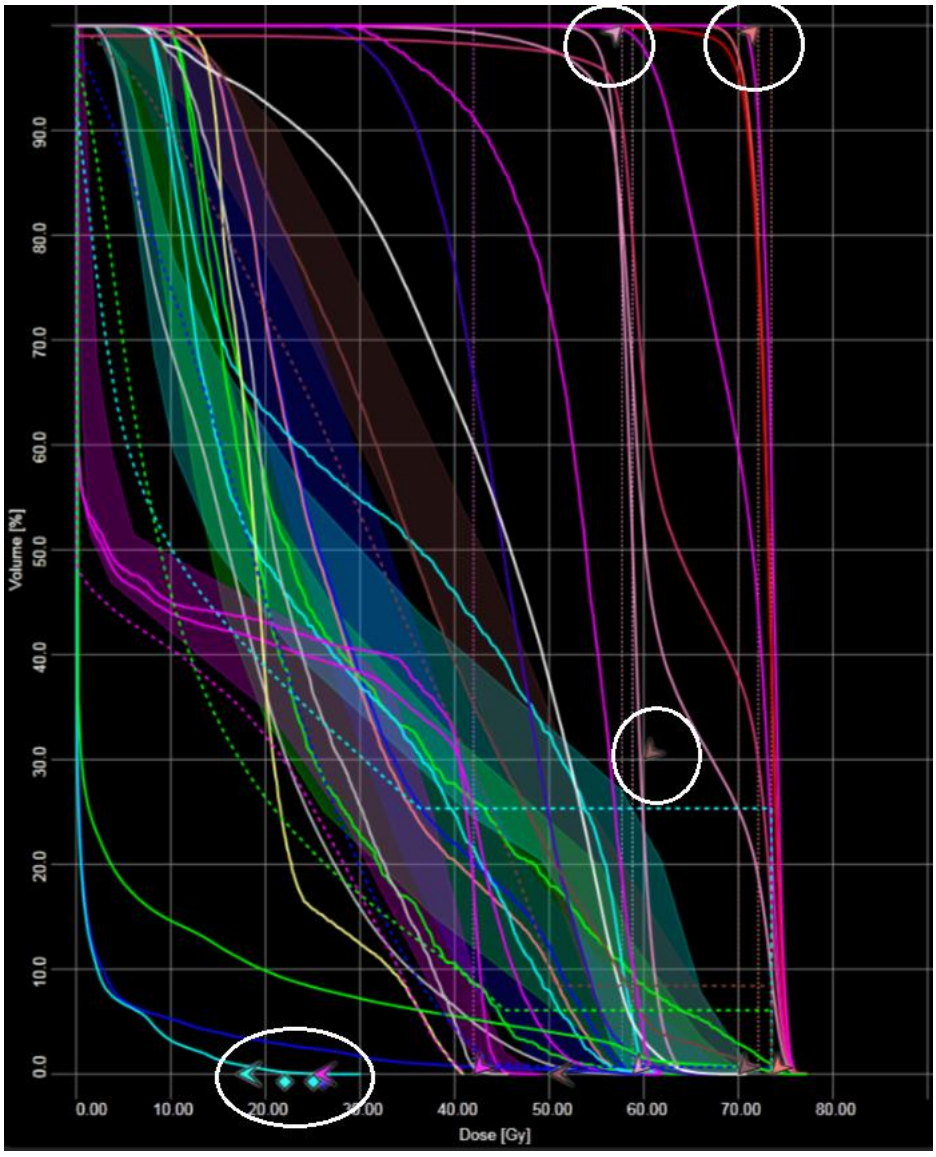

Figure 10. *PO optimizer's DVH with dose objectives. The dose objectives are highlighted with white circles. The upper objectives are diagonal arrows pointing down, lower objectives diagonal arrows pointing up, mean objectives are represented by the diamond shape object and generalized equivalent uniform dose (gEUD) objectives are presented with left pointing arrows. Modified from (Eclipse Software)*

Rapidplan also generates its own dose-volume line objectives, which are limiting the dose across the whole structure's volume. The line objectives are chosen during the RP model construction for desired OARs. Objectives are then generated prior to optimization, if the given structure exists in the patient's structure set. (Eclipse Algorithms 2015)

As mentioned above, the PO algorithm has the possibility to use gEUD objectives. The mean objective is a trivial case of gEUD. The formulation of EUD is presented in equation 7 (Apinorasethkul 2017) (Eclipse Algorithms 2015) (Eclipse instructions 2015)

$$EUD = (\sum_i v_i D_i^a)^{\frac{1}{a}}, \quad (7)$$

Where v_i is the $i{:}\,th$ fractional organ volume, D_i the corresponding dose to the fraction i and a is the parameter describing the dose-induced volume effect to a specific tissue. GEUD objective can be added to the plan optimization by defining the target dose, a parameter and the priority relative to the other objectives. Parameter a roughly defines the section of the DVH curve where the objective weights, so that a low (absolute) value of a contributes to the mean DVH ($a = 1$) and high absolute values of a ($|a| \leq 40$) contribute to high doses. It is meaningful to treat the target dose as a function of a, since increasing the parameter will increase the affected dose-section. (Apinorasethkul 2017) (Eclipse photon 2015) (Sovik 2008)

Higher values of a shift the weight towards higher doses. The values of a -parameter range from -40 to 40 with 1 corresponding to the mean target. In chapter 2.1.4 OAR structures were generally divided in two groups, serial and parallel. Serial organ complication was more dependent on the maximal dose and more invariant from volume. From perspective of gEUD, the a -parameter should be >1 for all the serial organs to affect the DVH curve around the maximal dose-area and thus, reduce the risk of complication (Claudio 2009). For parallel organs, as described in chapter 2.1.4. the complication risk is proportional to the volume receiving dose, leading to conclusion that gEUD for parallel organs should be set to affect the mean DVH. (Claudio 2009) (Fogliata 2018) (Zinchenko 2008) (Åste 2008) In this work no lower gEUD objectives were used.

2.4 Rapidplan algorithm

The Rapidplan algorithm is an integrated part of Eclipse software and is used to generate DVH estimations, which are further transferred to plan objectives in the optimization stage. The motivation for the algorithm is to speed up and increase the coherence in the planning process between different planners and hospitals. The RP algorithm is divided in two parts: Model configuration- and DVH estimation part. (Eclipse Algorithms 2015) (Tol 2015)

2.4.1 RP model configuration

The first part of the RP algorithm is the model configuration part, which consists of two phases and prepares or permits further steps in DVH estimation part. The configuration is further divided in data extraction phase and model training phase. The model configuration is also part of the model building. First, the plans are extracted to the RP model.

Plan information consists of structure sets, dose prescription and dose matrices, which are saved as binary information in the model. (Eclipse Algorithms 2015)

The structures in structure sets are divided to voxels with 2.5 mm resolution. The voxels for matched OAR structures are then classified belonging in one of the four regions which together cover the whole organ volume. The regions are: out of field region, leaf transmission region, in-field region and overlap region. The first two are defined by the visibility from the jaw aperture, the third by the targets field projection over the OAR and the last by the structure overlapping. Next, relative volume, cumulative volume histograms are calculated, and the geometric distributions of the regions are evaluated as cumulative volume histograms of geometry-base expected dose (GED). (Eclipse Algorithms 2015)

After the data extraction the DVH model can be trained. The histogram information is used to create a set of DVH estimation models for the matched OARs. The GEDs of the OARs are processed with Principal component analysis (PCA) to create number of PCA scores from which the DVH can be reconstructed with error less than 5 %. Finally, the PCA scores are combined with OAR voxel and anatomical feature information. The results are then used to create a regression model through forward and backward iterations until convergence. The final PCA-regression model consists of the mean and principal components of GED and DVHs for each OAR, regression model, statistical parameters for further outlier detection and the range of target coverages from the training set. (Eclipse Algorithms 2015)

2.4.2 RP DVH estimation

DVH estimation part consists of two phases, estimation- and objective generation phases that are performed before the plan optimization. The estimation algorithm takes the RP model's previously generated estimation models, structures from the plan to be optimized and target dose levels. The same metrics are then calculated in estimation generation phase as in model configuration part, excluding the DVH metrics. Anatomical features are considered to have major deviations from the training set and flagged if the geometric feature's value (X) is either smaller than the minimum value in the training set (X_{min}) or larger than the maximum value in the training set (X_{max}), or one of the following equations is satisfied

$$\frac{X-X_{med}}{X_{90}-X_{med}} > 1.56 \tag{8}$$

$$\frac{X-X_{med}}{X_{med}-X_{10}} > 1.56 \tag{9}$$

Where X_{90} is the 90 % percentile of the values in the training set, X_{10} the 10 % percentile and X_{med} is the median of the training set. Next, the PCA-regression model described in chapter 2.4.1 is used to generate the parametrized GED histogram and to calculate the

DVH principal component scores. The most probable DVH is reconstructed from this information and the upper and lower bounds are generated from the most probable estimate. Finally, the final variation curve is computed using the regression model and point wise square root. In the end different regions are summed together weighted by the relative volume (Eclipse Algorithms 2015)

The last phase is the optimization objective generation phase, which uses the upper and lower bounds of the DVH estimation as objectives, i.e. the optimization goal is to achieve the estimated DVH. (Eclipse Algorithms 2015)

2.5 Validation workflow with Rapidplan in Eclipse Software

In this chapter the validation workflow when using RP model is shortly described based on the planning procedure used in this thesis work. The main steps, depending on the RP model type, are importing the patient data, plan duplication and modification, structure addition and cropping, plan optimization, dose calculation and data exportation and analysis. The workflow here is described in the case where the clinical plan (manual) has been made and can be re-planned with RP.

Planning workflow in Eclipse system starts from the 4th step in the radiotherapy care path described in chapter 2.1.1 figure 1, when the patient's CT stack has been imported to the system together with the structures and dose prescription. The patient data is loaded in external beam planning program integrated in Eclipse. The desired clinical plan is chosen and duplicated together with the CT stack and structure set.

After the new duplicated plan is renamed, the structures may be modified to meet the requirements of the RP model. For example, if the RP model training set's PTV structures are drawn with a certain margin to the skin surface, as in the HNC model in this thesis, the patient's structures must be cropped with same guidelines. Structure modification is done with a cropping tool in Eclipse.

When the structures match the RP model, the plan can be optimized. In this thesis work PO algorithm was used in optimization. RP model includes the predefined dose objectives and generated the possible additional objectives and line objectives for chosen OARs. Before the optimization DVH estimation model is calculated for the plan. Next, the plan is optimized by starting either the IMRT or VMAT optimization. Also, an Intermediate dose computation can be chosen to help the optimization process. (Zacarias 2009)

When the optimization has converged, Eclipse returns to the External beam planning program and calculates the dose using the AAA algorithm. The dose normalization point is chosen by the user and the final DVH curves can then be collected to the same DVH plot. Other plans DVH data also may be added for DVH comparison and exportation.

Finally, the data is exported as a text file, which includes the DVH data (as absolute dose) for chosen organs, conformity index, homogeneity index, etc. The MU values are added together from Eclipse and added to the text file, which is then ready to be analyzed.

2.6 Previous research

Rapidplan's implementation and validation in clinical use has been lately the focus of several studies. Previous research included here consists of prostate, HNC, breast and lung cancer models. The model types vary from robust models with large (>150) training sets to coherent models with small carefully chosen training sets (20).

In 2017 Fogliata's head and neck cancer group built 2 HNC models using 83 clinical treatment plans as training set. The clinical plans were narrowed only to VMAT SIB technique with 3 PTV levels. The PTVs included user defined upper and lower limits with high priorities, all OAR optimization objectives were generated using RP. The model validation was concluded with validation set of 20 patients using DVH analysis. 10 of patients were chosen from the training set and 10 plans outside of the training set, yet all the plans resembled the 83 plans used in training. (Fogliata 2017) Similar to the HNC group, Fogliata's bilateral breast cancer group (2015) used only VMAT SIB plans in training. Now the training set size larger with 150 clinical plans with no plans from the training set. The validation data size was 70 patients with 50 single- and 20 bilateral cases, chosen from different hospitals. The validation methods were also similar using dosimetric validation. Both models could create acceptable clinical plans, meeting the DVH constraint set for the studies. HNC plans generated by the model improved the planning quality with significantly increased OAR sparing. Breast cancer model was also able to create plans meeting the dose constraints but not with superior or comparable dose distributions to the clinical plans. (Fogliata 2015) (Fogliata 2017)

Wu's group (2016) studied the RPs capabilities to generate DVH estimation and optimization objectives for rectal cancer, when the RP model is trained with 81 clinical VMAT plans but validated with (30) IMRT plans. The model performance was first validated with 10 VMAT plans similar to the training set. The results showed that VMAT plan generation with the RP model created sufficient target dose distributions and improved significantly OAR sparing. IMRT validation group showed also improvement in OAR sparing for bladder and femoral heads. Both techniques changed target dose conformity, either slightly (≤ 0.01) for VMAT or significantly for IMRT ($\Delta = 0.17$, $P < 0.01$). Wu concludes that RP can be used for IMRT planning even though the model is trained completely with VMAT plans. (Wu 2016)

Tol's HNC group (2015) used 90 AIO optimized clinical HNC treatment plans to train an RP model for quality assurance purposes. Unlike in Fogliata's or Wu's groups, the training set was chosen arbitrarily from HNC patient pool. Model fit and model goodness statistics (R^2 and χ^2) were used for outlier detection, validation and exclusion. Validation plans were re-planned with the model and analyzed with DVH analysis. The study shows that Rapidplan can accurately predict the achievable doses, even though generally the estimated OAR sparing was higher than actual. (Tol 2015)

Similar to Tol's group, Wang's (2017) breast cancer group used 80 largely varying patient plans in model training. All the patients were previously treated with IMRT SIB technique, but the training set included large variance in geometrical features such as different breast sizes. The final model was constructed for left side breast cancer treatment. The validation set was constructed by 6 planners with different experiences (beginner, junior and senior planners). The results showed that all the RP generated plans could fulfill the dose prescription requirements and could improve OAR sparing compared to beginner and junior planners with statistical significance. OAR sparing was found either inferior or similar to the plans optimized by senior planners. (Wang 2017)

Berry's group (2016) studied the inter-campus treatment plan consistency with based RP based models. The research used 58 esophagus RT treatment plans for model training and was validated using 172 clinical plans. The analysis was concluded by DVH band comparison between RP based and clinical plans. The validation group was divided in 4 regional (RS) sites. The results showed that the first RS group's RP plans had the most deviations from the RP model and less OAR sparing and variation especially for liver doses compared to clinical plans. Second RS group generated best matching plans compared to the RP model. 3rd RS group had high variance between RP plan and model, but the OAR sparing for final plans was comparable to the clinical version. The final RS group presented no statistical significance because of small sample size. Berry concludes that based on the results RP can be used to identify increased need for planning coherence in institutions. (Berry 2016)

Finally, Ma's group (2017) used small and coherent set of cervical cancer plans used clinically to train an IMRT RP model. The model was validated using 19 clinical patients. The RP model generated plans highly comparable (statistically) to the clinical plans with proper OAR sparing. Ma emphasizes that RP can produce high quality treatment plans even with small training data. (Ma 2017)

3. MATERIALS AND METHODS

3.1 Patient data

All patient data used in this thesis work was collected from TAYS database and all the patients in both, prostate and HNC models were previously planned and treated in TAYS. Based on previous research and Varian's preference, none of the data used in model training was used in model validation stage. Note that the HNC model's training set re-planning data (chapter 4.2.4) was not performed as part of the model validation.

3.1.1 Prostate cancer model training and validation data

The initial goal was to test and build a prostate cancer model that is robust but still can produce plans comparable to the clinical plans. It was also preferred that the model could create DVH predictions with minimal number of additional structures to complete OARs and PTVs. The reason for this was that generation of extra structures increase the planning time.

Number of training plans must be much higher than the minimum of 20 plans to build a robust model that could create DVH predictions within the dose-volume constrains. Initially 5 different previous built models were compared and further several model modifications were made and competed against each other to find the most applicable option. One of the 5 models provided with the software, 2 were built by a physicist from TAYS hospital and 1 was built as part of this thesis work.

First of the TAYS-made models (TAYS tot) included 104 training plans, randomly chosen from treated prostate cases. Only excluded patient cases included either prosthesis or significant outliers (see chapter 3.4.3). The plans included different sized (empty and full) bladders and rectums, varying target to organ distances, plans with and without seminal vesicles and plans with and without 46 Gy PTV-structure. The guideline in TAYS is to have 'fairly full' bladder and empty rectum during imaging and treatment, but the actualization is patient dependent. Majority of the training data consisted of 7 field IMRT plans, rest were 1 or 2 arc VMAT plans. The model was trained by using the total OAR structures from the plans. The second TAYS model (TAYS crop) included the same training set as the first model, but now the model was trained by using modified OAR structures, where the OAR and PTV overlapping was subtracted from the final OARs.

The model built as part of this thesis included training data of 38 plans, chosen coherently compared to random selection in TAYS tot. The model was then analyzed with the regression- and residual plots and with a set of model fit statistics (see chapter 3.4). The training set was further cleaned from verified outliers to achieve higher coherence.

Model comparison in this stage was performed to a set of 9 patients (4 IMRT and 5 VMAT) chosen by random selection (excluding prostheses). Overall best performing model (TAYS tot) was then chosen for further optimization and training. Outlier detection, further training and DVH objectives modification resulted in several versions of the original TAYS tot -model, which were again competed against each other and compared to the original treatment plans. Comparison method was DVH value comparisons based on TAYS conventions and DVH limits using DVH analysis -Matlab program made for this thesis (see chapter 3.3). Final model version was trained with 126 out from 150, mostly randomly selected plans with 10 outliers for rectum and 2 outliers for bladder. Experience from previous validation rounds indicated that the model-fit and model goodness statistics cannot predict accurately the model performance, especially when the training and validation sets have high variance. Thus, fit and goodness statistics were used to only exclude the most significant outliers. Therefore, the final model remained robust with highly variating, yet numerous training set.

The second tested model was the coherent model built as part of this thesis work. The model was trained with 38 seemingly coherent plans and cleaned from significant outliers. All the plans were 7 field IMRT plans with 6 MeV photon energy with similar dose-volume histograms. The main goal for this model was to find out the effect of higher coherence in training, thus only IMRT plans were tested in validation stage. Initial testing with VMAT plans indicated that the model cannot compete in OAR sparing nor in PTV filling with the TAYS tot or with the clinical plan. The IMRT validation set was the same as for the TAYS tot -model.

3.1.2 Prostate model optimization objectives

Prostate model's DVH optimization objectives were primarily set according to TAYS's dose-volume limits. All the prostate plans were divided in 20 fractions having 3 Gy dose per fraction. The limits for prostate model are presented in table 1.

Table 1. *Prostate model's optimization objectives for both prostate models*

Structure	Limit type	Vol [%]	Dose [Gy] from prescribed	Priority
PTV high	Upper	0	60.0	100
	Lower	100	59.4	135
PTV intermediate	Upper	15	46.2	95
	Upper	0	57.0	105
	Lower	100	45.6	105
Bladder	Upper	0	60.0	80
	Upper	10	30.0	36
	Upper	30	18.0	36
	Line	Generated	Generated	50
Rectum	Upper	10	30.0	44
	Upper	0	60.0	80
	Upper	30	9.0	44
	Upper	50	1.8	36
	Line	Generated	Generated	50
Prostate	Upper	0	60.1	115
	Lower	100	60	135

The dose is presented as grays in table 1. but both, dose and volume were originally defined as percentages from the prescribed dose, so that the model could be used in principle to produce plans also with different target dose prescription for single target dose-level.

3.1.3 Head and neck cancer training and validation data

Head and neck model's training data included 156 patient patients with HNC located in neck and mouth area. All the training set's patients were treated with VMAT technique and only bilateral (PTV area) treatment plans were chosen with individually selected arc geometries. Training plans included either 2, 3 or 4 PTVs with combinations of $70/56\ Gy, 66/54\ Gy$, $70/63/56\ Gy, 66/60/54\ Gy$ and $70/66/63/56\ Gy$. Other combinations and dose levels were excluded from the training set. This decreased the number of validation plans, since most of suitable clinical cases treated in TAYS hospital were used in training stage and as mentioned before, validation data and training data is recommended to be kept separated. 19 IMRT and 13 VMAT cases were used in validation stage. Number is small compared to the training set but is not significantly different in comparison with previous research.

In model iteration stage, the whole training set was re-planned with the HNC model and the re-planned data was then used to retrain the model. This left large amount of model planning data from the original HNC model, which may have valuable information of the model performance. 148 of re-planned DVHs were exported and analyzed with paired sample testing (Wilcoxon). Because the cases are part of the training set the results are treated independently from the validation results.

3.1.4 Head and neck cancer model building and objectives

Rapidplan is in theory capable to separate overlapping structures, e.g. between high and low dose-level PTVs (Fogliata 2017), but pre-contouring is also possible and may support the optimization process (Eclipse Algorithms). Hence, the PTV structures were modified before the data extraction to the model. PTV structures are usually cropped with certain margin to the skin surface to prevent the optimization algorithm filling the skin dose, which is not desired nor possible. Margins are also constructed between PTVs with different dose levels, so the dose boundaries will not become too steep for optimization. TAYS conventions are not defining the exact values for these margins, which leads to small differences between different planners. Thus, all the PTVs in each plan were re-cropped for the HNC model with following guidelines:

- High dose PTV is cropped so that a 4mm margin is left between PTV and the skin surface.
- Intermediate dose PTV is cropped 4mm from the skin surface and 3 mm from high dose PTV.
- Low dose PTV is cropped 4mm from the skin surface, 3 mm from intermediate dose PTV- and 6mm from high dose PTV structure.

PTVs were cropped and matched to the model's PTV high-, PTV intermediate- and PTV Low structures, so that PTVs with dose prescriptions of 70 and 66 Gy corresponded to PTV High, 63 and 60 Gy corresponded to PTV intermediate and 56 and 54 Gy dose levels corresponded to PTV Low. Plans including 4 PTVs were excluded from the validation set, because the optimization algorithm supports maximum of 3 PTVs. It is also important to note that the validation set's PTV structures are cropped by same guidelines as the training set (Fogliata 2017).

Unlike PTVs, the OAR structures were chosen to be extracted with the original contouring without cropping. The reason for this is that the HNC model includes several OARs $(5-12)$ compared e.g. to the prostate model and in future generation of additional structures slows down the planning process. Also, in prostate model testing there was no indication of benefits in additional OAR cropping. OAR structures used in model training with model objectives and priorities are presented in table 2.

Table 2. *HNC model optimization objectives for PTVs*

Structure	Limit type	Vol [%]	Dose [%]	priority
PTV high	Upper	0	105	120
	Lower	100	103	143
PTV intermediate	Upper	0	105	120
	Lower	100	103	143
PTV lower	Upper	0	105	120
	Lower	0	103	143

Again, as in prostate model the PTV dose objective is defined as percentage of the prescribed dose. This means, that the model can be used for different dose prescriptions. (Eclipse Algorithms 2015) The optimization objectives for OAR structure are presented in table 3.

Table 3. *HNC model optimization objectives for OAR structures. Note that the dose is now defined in grays and* a *-parameter is used for gEUD objectives.*

Structure	Limit type	Vol [%]	Dose [Gy]	Priority	a parameter
Brain	Upper	0	48	55	-
	Line	Generated	Generated	30	-
Brainstem	Upper	0	50	120	-
	Line	Generated	Generated	30	-
Chiasm	Upper	0	54	105	-
Inner ears	Mean	N/A	45	55	-
Larynx	Upper	25	60	55	-
	Upper	65	50	50	-
	Mean	N/A	35	50	-
	Line	Generated	Generated	40	-
Lungs	Mean	N/A	25	25	-
Mandible	Upper	0	70	85	-
	Upper	30	60	50	-
	gEUD	N/A	50	35	10.0
	Line	Generated	Generated	35	-
Neck	Mean	N/A	25	30	-
Optic nerves	Line	Generated	Generated	120	-
Oral cavity	Mean	N/A	26	37	-
	Line	Generated	Generated	40	-
Parotids	Mean	N/A	22	35	-
	gEUD	N/A	17	35	1.0
	Line	Generated	Generated	45	-
Pharynx	Upper	70	25	40	-
	Upper	60	50	50	-
	Mean	N/A	38	40	-
	gEUD	N/A	30	35	2.0

	Line	Generated	Generated	35	-
Spinal cord	Upper	0.0	42	120	-
	gEUD	N/A	25	35	1.0
	Line	Generated	Generated	35	-
Submandibular glands	Mean	N/A	35	35	-
	gEUD	N/A	15	35	1.0
	Line	Generated	Generated	Generated	-

Model's upper objectives were chosen based on TAYS dose-volume limits, previous research and planning experience (Fogliata 2017) (Snyder 2016) (Wang 2017). Other objectives were chosen for maximizing OAR sparing, based on previous research and experience. OARs such as submandibular- and parotid glands were optimized with upper gEUD objectives in addition to mean objectives. GEUD objectives are used to affect different parts of the DVH curve by changing the constant a in gEUD (equation 7). GEUD objectives allow weighted optimization of the DVH curves, e.g. by lowering the dose-volume levels for OARs only in higher doses. Note that objectives in table 3 are valid only if the corresponding structure is matched with the model's structure before optimization. The HNC model covers several types of cancer in head and neck area, hence not all the structures are included in all structure sets and thus can't be matched with the model.

3.2 Iterative training of the HNC model

After HNC model training and validation a duplicate HNC model was created. Varian's consultation suggested that iterative training of the existing model with the already trained plans could increase the planning uniformity. All the objectives and existing training data were kept unchanged for the duplicate plan.

The optimized HNC plans were reoptimized with the original model and the resulted plans were used to train the duplicate model. The matched dataset was the same as with the original model, except a neck structure was included in retraining to avoid excess dose spilling to the normal tissue. After the iteration process the duplicate model included 312 HNC plans from which the old training set and outliers were excluded. The final model was then trained with 126 plans and was used to re-optimize the validation set. The main goal for this stage was to inspect if reiteration increases the coherence in planning quality. All the results were exported to text files and analyzed with Matlab DVH analyzing program (see chapter 3.3) SPSS 23 (Armonk, NY) and R 3.5.0 (Vienna, Austria).

3.3 RP data-analysis program

As part of this thesis work a Matlab analysis-program was written for the analyzation of the validation data. Eclipse has the possibility to extract patient's DVH data for each

model in a single text file. The file includes the conformity indices, gradient measures, absolute doses, relative doses, and the corresponding volume values for each organ and for each plan included in the exported DVH plot. Additionally, each plan's MU values were summed together and included in the text file. The data can be thousands of lines long and dozens of patients were analyzed in the verification stage and over 150 patients in iterated HNC model evaluation stage thus, it was essential to use an automated algorithm for data analysis.

The program includes a user interface (UI), where the user can choose the source folder including the text file or files. The model is chosen from a dropdown-menu (prostate or head and neck) and dose or volume points and limits can be manually inputted. Otherwise they are automatically set according to the TAYS conventions. The data can then be analyzed. The program finds and arranges the data from all the text files in the source folder, so that the chosen DVH-values, computed mean values, conformity indices and MU-values are sorted for each model and structure under each patient. The program also compares and displays, if the computed DVH values meet the dose-volume limits. In the end of the resulted file the mean values for each structure in every model are presented as a summary. Additionally, the program computes the DVH plots for each patient which can be browsed and analyzed in the UI. The UI with an example plot is presented in figure 11.

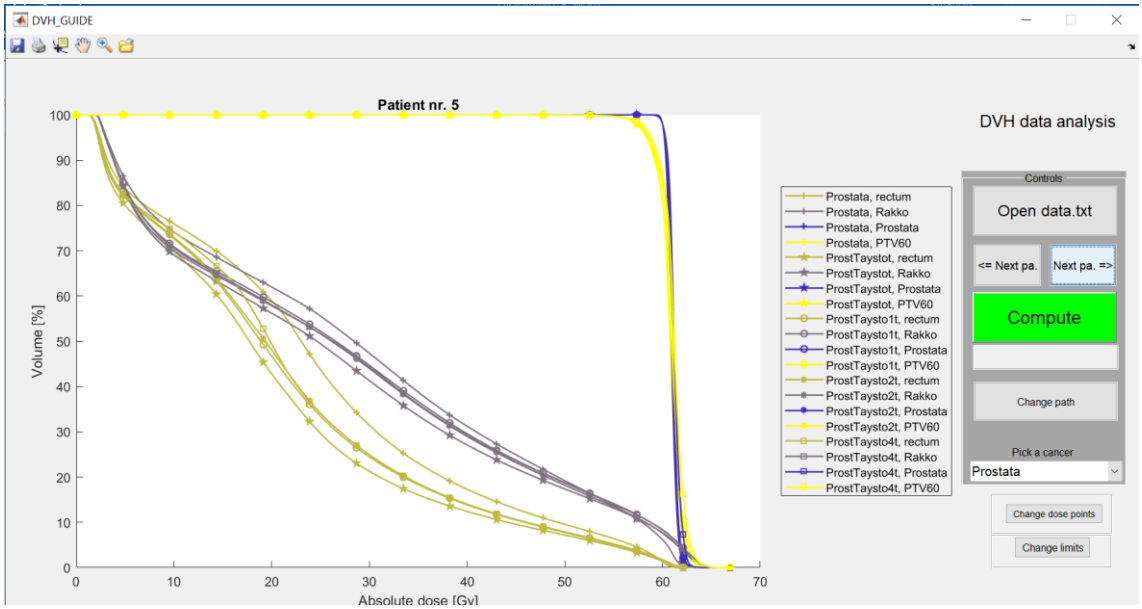

Figure 11. *DVH analysis program UI.*

In figure 11 the structures volume percentage is plotted as a function of absolute dose [Gy]. The color code of the curves indicates the structure, e.g. in the figure, gray color corresponds to bladder and the line-symbol identifies the model, so that a gray line with "+"-symbol presents the bladder's DVH data from manually generated treatment plan named "prostate". Figure 12 presents a section from the resulted text file.

Structure:	rectum					rakko		ptv60		lonkkaV	lonkkaO	
Dose[Gy]:	60	58	54	46	38.5	49	38.5	98	2	mean	mean	
Model name: Pros_to5t												Conformity Index:
Vol [%]:	1.368	4.196	7.409	12.292	17.358	1.955	3.592	58.003	60.901	18.869	21.386	1.58
Model name: Pros_coherent												Conformity Index:
Vol [%]:	3.078*	4.792	7.111	10.3	13.668	2.41	4.146	58.633	61.161	19.71	21.343	1.99
Model name: Prostata												Conformity Index:
Vol [%]:	1.14	4.918	8.202	13.663	20.282	1.329	2.587	57.624	60.886	18.162	19.518	1.49
----------						Patient nr.		1		----------		

Figure 12. *DVH analysis program generated result table for a single patient.*

The patients are sorted successively in the text file. In the end of the file, the mean results are computed and sorted for each model similarly as for single patient in figure 12. The figure presents the patient number (personal information is deleted in the analysis) and the structures. In figure 12, the structures are rectum, bladder, 60 Gy planned target volume (PTV60) and the femoral heads. Only the structures that have defined limits are presented in the results, but all exported structures are displayed in the DVH plot. Below the structure is the OARs dose (volume for targets), for which the corresponding volume (or dose) -value is desired. The volume-% values are presented in the columns under the model names. The program informs the user with a star (*) -symbol for values which do not meet the constrains. In figure 12 this can be seen under model "Pros_coherent" for 60 Gy rectum dose.

3.4 Model evaluation and validation methods

The prostate and HNC model construction and optimization required several training iterations and modified number of patient data. Potential outliers were identified by using the Rapidplan's data-analysis tools including DVH-, regression- and residual plots, and outlier statistics; cook's distance, studentized residual, differential area and modified Z-score. Outlier statistics are further explained in chapter 3.4.3. The Outliers were verified by excluding the potential outlier structures from the training set and retraining the model. After retraining the R^2 and χ^2-values were compared to the previous training set's values and mean squared error (MSE) was evaluated for model goodness. The prediction power (model goodness) of the model had tendency to decrease while the model fit increased (when outliers were excluded). Thus, R^2 and χ^2 were used only as directional indicators.

The model evaluation consisted of 5 stages, evaluation of model goodness and model fit, outlier analysis, dose- and statistical analysis in the validation stage. Model goodness, model fit, and outlier analysis were conducted with the Eclipse's model configuration -program. Matlab, SPSS and R were used for dose- and statistical analysis.

3.4.1 Model fit

Model fit is a statistical concept, included in the Rapidplan's model evaluation which consists of multiple parameters. The idea of model fit statistics is to indicate numerically how well the trained model represents the training data. The metrics used for the model fit were the χ^2 and R^2 values. Both values represent the model-goodness-of-fit. χ^2 (or chi square) measures the squared distance between the clinical plan and the model's estimation. Chi square is the sum over squared differences between observation and expected result relative to the expected result. The definition of χ^2 is presented in equation 10

$$\chi^2 = \sum_{i=1}^{N} \frac{(O_i - E_i)^2}{E_i} \quad (10)$$

Where O_i is the ith observation and E_i is the ith expected result. (Humpfrey 2016) (Liu 2016)

According to Varian Medical, the goodness of fit -statistics are determined by the distance from unity. R^2, or the coefficient of determination measures the data representability of the regression line, or in other words it describes the model's ability to explain the variance in the data set. The value ranges [0,1], where values closer to 1 (or $R^2 > 0.7$) are considered as close to optimal (Henseler 2009) (Moore 2013). Data overfitting was considered by inspecting the combination of R^2 and χ^2. For example, high R^2 and low χ^2 values may indicate overfitting (Legates 1999).

The values were analyzed during every step of model training and modification. For TAYS tot-model, the weight of model fit -statistics was considered small, since the model was built as robust by randomly chosen plans. Highly robust models require large datasets, and even though the data set was relatively large compared to other models, good model fit could have not been expected. Also, χ^2 and R^2 reflect only the model fit for the training set, not the estimation power of the model for new plans. (Eclipse Algorithms 2015)

Unlike for TAYS tot-model, the coherent model was trained with high priority of R^2 and by moderate priority of χ^2. The values were used as part of the coherence measure with aim of $R^2 \geq 0.7$ and $1.0 < \chi^2 < 1.1$. Although, the training set was relatively small, the coherence is high which compensates for the model fit and good results may be expected. Again, the model fit alone does not predict the estimation power, but now a good reflection of the planning conventions was desired.

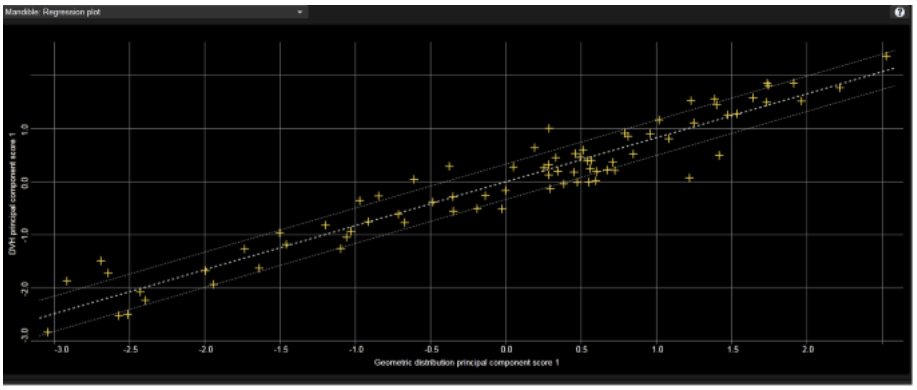

Figure 13. *Regression plot of HNC model's mandible structure with* $R^2 > 0.7$ *and* $\chi^2 < 1.1$. *(Eclipse v.13.6 2015)*

Note that the regression plot in figure 13. is presented through DVH and geometric principal component scores as explained in chapter 2.4.

3.4.2 Model goodness

The model goodness or predictive power is evaluated by MSE between the plan and the model's estimation. Rapidplan algorithm uses the training set to estimate the MSE by fractionating the training set into plans used in training and to plans to evaluation. (Eclipse Algorithms 2015) The mean squared error is calculated to evaluate the regression model, so that values closer to zero indicate higher predictive power. (Ivanescu 2005) In practice this is done by calculating the error of DVH principal component score 1 between the regression model's predictions and every observation (see figure 13). The total error is then the squared sum over the number of observations (n).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}\left(Y_{i_{predicted}} - Y_{i_{observed}}\right)^2 \quad (11)$$

Where Y_{i_X} represents either the predicted value or the observed value. (Wackerly 2008)

3.4.3 Outlier detection and verification

In this work statistical outliers were identified mainly by following the Varian's guidelines for outlier detection and identification. First, the model fit was robustly evaluated by R^2 and χ^2 parameters, defined in chapter 3.4.1. Next, Rapidplan's outlier statistics and the regression plots were evaluated for each structure with line objective. Varian has suggested that Cook's distance larger than 10.0 and studentized residual larger than 3.0 are potential outliers. This suggestion was used as a guideline for outlier detection. Also 2 other outlier parameters were used in evaluation, Areal Difference of Estimate (aA) and modified Z-statistic, but only as support for the main parameters. In figure 13, the blue "+"-sign in the top left corner is considered as a potential outlier by Cook's (CD) distance > 10. (Hao 2016) (Eclipse IOU 2015)

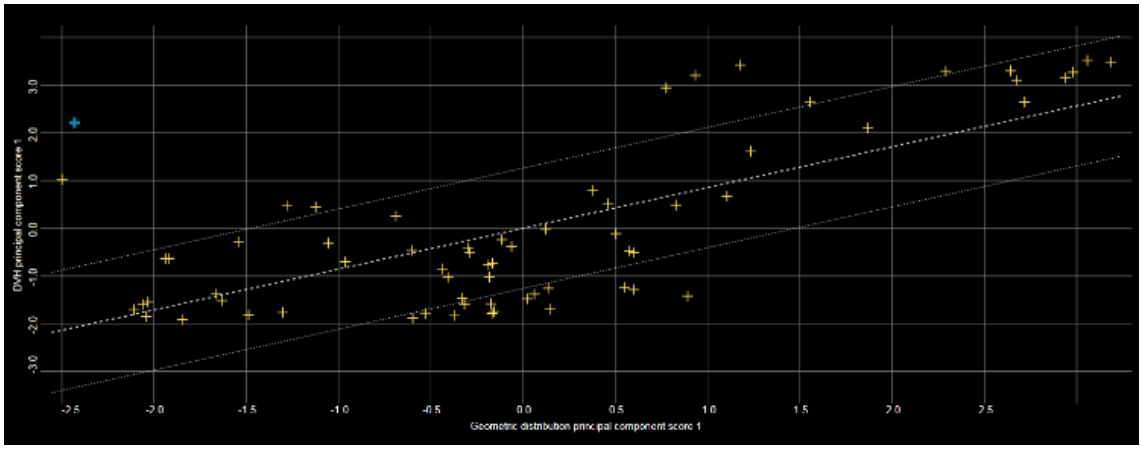

Figure 14. *Regression plot for an example HNC structure with potential outlier marked as blue "+". (Eclipse v.13.6 2015)*

Cook's distance is defined as the sum over the changes in the regression model when the value/observation is removed from the data set. High value of Cook's distance corresponds to large change of the model when the point is removed or added, thus CD can be considered more intuitively as leverage or amount of influence of the point. High CD values may help in outlier detection since outliers usually deviate considerably from the regression line and simultaneously increase the leverage. It is also clear that a single data point should not alone contribute to the model excessively. Points fluctuating >2 standard deviations from the regression line were considered as potential outliers.

After detecting the potential outlier from elevated CD and (or) from regression plot, residual plot, in field DVHs, other outlier metrics like standard deviation (SD) or dA and geometrical statistics were used to verify the outlier. In figure 15 is presented the residual plot from the same case as in figure 15. The potential outlier is again marked as blue "+" sign.

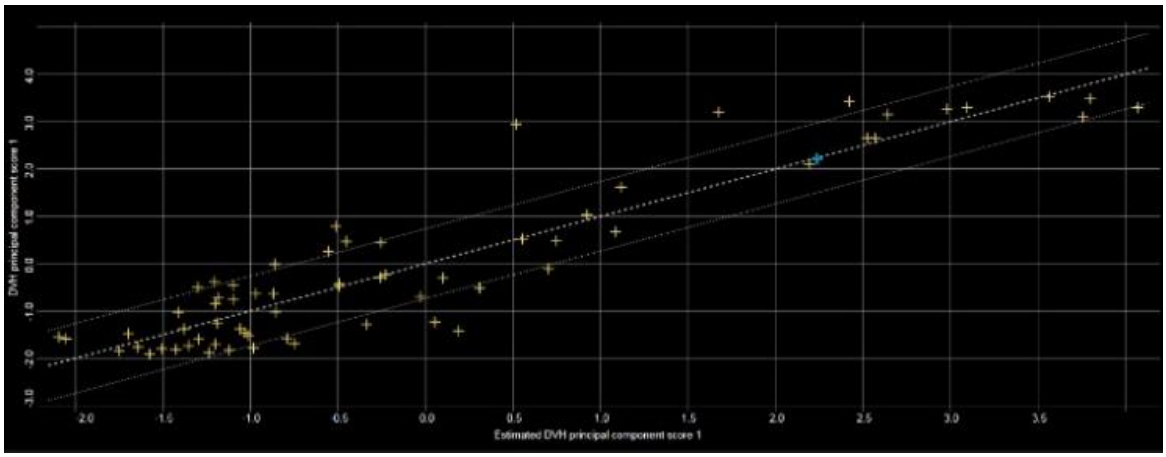

Figure 15. *The structure's (as in figure 14) residual plot with the potential outlier marked as blue "+". Modified from (Eclipse v.13.6 2015)*

After the evaluation of the outlier statistics, the case was evaluated from the in field - DVH plots and geometrical plots. Next, the point was excluded from the data set and the model was retrained without the potential outlier. R^2 and χ^2 values were inspected for change so that $R_1^2 > R_2^2$ and $\chi_1^2 < \chi_2^2$ indicated better model fit and thus, supported the outlier assumption. In figures 14 and 15, the assumption of potential outlier was refuted.

3.4.4 Statistical methods

The final number of validation patients for the prostate model was 16 patients for IMRT and 31 patients for VMAT treatment. VMAT validation was emphasized since, in future the treatment convention in TAYS is expected to incline towards VMAT. It was also interesting to see how mostly IMRT trained RP model handles VMAT plans. Statistical analysis was made for the PTVs and OARs, for which TAYS's radiotherapeutic unit has predefined minimum dose-volume constraint recommendations. PTV ($60\ Gy$) was analyzed by the doses that cover 98 and 2 −percent of the PTV's volume. OAR structures were statistically tested by comparing the volume-values at given dose points, excluding femoral heads, for which mean doses were compared. Reason for this was, that the volumes at the TAYS limit doses were generally zero for femoral heads, which makes numerical comparison irrelevant.

Validation set size for HNC model was 19 IMRT plans and 13 VMAT plans. The robustness of the validation set must be close to the robustness of the training set, so the validation set was narrowed to the neck and lower jaw area, excluding single sided PTVs and

including only full rotation or near full rotation arc geometry for VMAT. This resulted to shortage of applicable patient data in validation stage. Hence, the small sample size.

The small number of samples ($n < 30$) resulted in use of paired sample- and single tailed t-tests and Wilcoxon signed rank test in model comparison (RP vs manual) and evaluation. Three different null hypotheses were defined for the statistical analysis:

1. H_0: RP generated DVHs are equivalent to the manual plans ($mean\ difference = 0$).
   H_1: RP generated DVHs are not equivalent to manual plans.
2. H_0: TAYS dose constrains are not met by RP generated DVH-values.
   H_1: RP generated DVH-values are higher for PTVs and lower for OARs than TAYS constrains (plans meet the dose constrains).
3. H_0: RP generated conformity indices are comparable to clinical plans ($mean\ difference = 0$).
   H_1: RP generated conformity indices differ from clinical plans.

The first null-hypothesis tests the robustness and planning convention representability of the model. All the clinical plans were validated and used in curative treatments. Thus, if the null hypothesis cannot be rejected, it implies that RP can statistically generate similar models from the manual plans and hence, the model represents the planning conventions of the hospital within the statistical constraints. The 2nd hypothesis tests the suitability of RP algorithm in clinical use by testing if the generated plans meet statistically the DVH constrains set by the hospital. Finally, the 3rd hypothesis defines whether the model can create acceptable dose conformalities for the target volumes in the test groups. $\alpha < 0.05$ was considered as a proper indicator of statistically significant difference and $\alpha < 0.01$ as highly significant difference resulting to null-hypothesis rejection (Dahiru 2008).

Paired sample t-test or Wilcoxon signed rank test were used for the model comparison (1st hypothesis) depending on the normality of the data. These tests are applicable for this purpose, since paired samples test is used directly from the RP generated and manual plan DVHs i.e. the same patient, giving the significance of the mean difference between the models. (Kim 2015). The t-test was chosen for normally distributed data because of the small sample size ($n < 30$). Normality of the data was determined by Shapiro-Wilk test in SPSS and in R. For non-normal data distributions Wilcoxon signed rank test was used. Also, additional Bland-Altman plots were created in SPSS from the prostate sample pairs which did not indicate statistically significant difference. The trend around the mean difference was evaluated visually and with data regression analysis.

The 2nd hypothesis used single tailed t-test. The t test was again chosen because of the small sample size. It is sufficient to use single tailed t-test since, only one directional values are important. The test was performed and analyzed in R.

Similar to the 1st hypothesis, paired samples t-test or Wilcoxon signed rank test was used for the 3rd hypothesis (data was tested to be normally distributed). $\alpha = 0.05$ was again considered as statistically significant and $\alpha = 0.01$ as highly significant.

HNC models were also tested with Friedman's non-parametric test. The goal of this test was to determine if the original RP model or the iteration model produced means deviate from original plans or from each other. Friedman's test tests the null hypothesis of groups having the same means (comparable to hypothesis no. 1 above) and indicates if there is significant statistical difference between the plans. The test does not provide any additional information about the deviating plan nor identify the plan. For this purpose, complementary post hoc test was carried out with Bonferroni method. Friedman and post hoc tests were computed and analyzed in Matlab, producing the p-value for Friedman's test and p-values with visual presentation for post hoc. Post hoc results were acquired by using the multi-compare -Matlab function included in the Statistics and Machine Learning -Toolbox. The critical value for the multiple comparisons test is defined by Bonferroni method presented in equation 12.

$$|t| = \frac{|\overline{y_i} - \overline{y_j}|}{\sqrt{MSE\left(\frac{1}{n_i} - \frac{1}{n_j}\right)}} > t_{\frac{\alpha}{2}\binom{m}{2},N-m} \tag{12}$$

Here $N - m$ is the difference between the number of total observations and groups, MSE is the mean squared error and y_x is the observation (MathWorks 2018). Bonferroni method uses the Student's t-distribution to find the critical values. Method also compensates for type-I error caused by multiple comparisons by correcting the significance level as presented in equation 13

$$\alpha_{cor} = \frac{\alpha}{2}\binom{m}{2} \tag{13}$$

Here α is the original significance level and m is the number of comparisons. (MathWorks 2018)

3.4.5 Additional dose analysis

Statistical analysis focuses in model specific mean values over the whole validation set, but the interest concentrates also in the patient specific and individual DVH point-results. Hence, an additional dose analysis was concluded. The analysis compares the DVH results, conformity indices and MU values between patients. Also, different plans were tested with Friedman's test for differences between groups and further with multiple comparison Post hoc test similarly as described in chapter 3.4.4. For conformity indices paired sample t-test was concluded after the data was confirmed as normally distributed with Shapiro-Wilk test. The RTOG protocol defines CI to be in acceptable region if $1 \leq CI \leq 2$. (Stanley 2011) (Petkovska 2010)

MU values were computed from the eclipse planning software by adding the MU values from each field (or arc). The values were included in the exported text file for each patient. For prostate model, the CI values were exported for the PTV high structure directly from the Eclipse Software. However, the HNC plans included at least 2 target volumes. The Eclipse Software calculates the CI values based on the prescribed dose (PTV high) from equation 7, so the conformity indices of the lower dose level PTV cannot be directly calculated. To achieve this, additional 95 % isodose structures were created for each patient, which were used as the reference volume in equation 7.

4. RESULTS

In this chapter the results of this thesis work are presented. The results are divided in two main sections; to prostate model and to HNC model, which are further divided to statistical analysis and to additional dose analysis. HNC model includes also the comparison results of the re-planned training set of the original HNC RP model against the training set consisted of clinical plans.

4.1 Prostate model

One of the prior goals in this thesis work was to find and improve, and to create a Rapidplan prostate model, which can generate acceptable DVH predictions for varying group of prostate cancer patients. The results considering the final prostate models are presented in this chapter.

4.1.1 Prostate model evaluation results

The prostate models were evaluated mainly with model fit and model good statistics as described in chapters 3.4. The model fit, and model goodness results are presented in table 4 for the Tays tot -model.

Table 4. *Model evaluation results for the TAYS tot model. Both bladder and rectum showed relatively low level of model fit considering values for* R^2.

Structure	Training set's size	MSE	R^2	X^2	Outliers
Bladder	126	0.00184	0.444	1.057	10
Rectum	126	0.00300	0.379	1.033	2

Total number of excluded plans was 23 (126 from 149). The model was trained for 126 PTV high and prostate targets and for 40 PTV intermediate targets. In table 5 are presented the model fit and model goodness statistics for the coherent prostate model (TAYS coh).

Table 5. *Model evaluation results for the TAYS coh model. Coherence of the training set data had improving effect to model fit.*

Structure	Training set's size	MSE	R^2	X^2	Outliers
Bladder	34	0.00186	0.497	1.020	0
Rectum	34	0.00348	0.827	1.214	2
Femoral head L	25	0.00666	0.811	1.009	9
Femoral head R	25	0.00457	0.883	1.141	11

TAYS coh model was trained with 36 plans, from which 36 PTV high and prostate structures were matched with the model. PTV intermediate structures were matched with 10 plans. Coherent model was trained with 36 out of 36 extracted plans, which means that no outliers were excluded.

4.1.2 Statistical analysis of the prostate model

Table 6 presents the prostate model results with IMRT validation set for the 1st and 2nd null hypothesis (chapter 3.4.4) concerning model equivalence and DVH constraints respectively.

Table 6. *Paired samples t-test results for the prostate model with IMRT validation set. PTV results are presented as volume-doses and OAR results as dose-volumes, excluding femoral heads for which mean doses were computed.*

Structure	Dose constraint	Original plan $(\boldsymbol{mean \pm SD})$	TAYS tot $(\boldsymbol{mean \pm SD})$	TAYS coh $(\boldsymbol{mean \pm SD})$	Df
PTV D98 [Gy]	$D_{98} \geq 57$	57.79 ± 0.34	57.55 ± 0.47	57.88 ± 0.12	15
PTV D2 [Gy]	$D_2 \geq 60$	61.07 ± 0.21	61.09 ± 0.36	61.28 ± 0.17	15
Rectum $V60_w$ [%]	$V_{60} < 3$	0.87 ± 0.72	$1.35 \pm 0.69^{**}$	$1{,}36 \pm 0.65^{*}$	15
Rectum $V58$ [%]	$V_{58} < 5$	3.46 ± 1.42	3.64 ± 1.8	3.47 ± 1.47	15
Rectum $V54$ [%]	$V_{54} < 20$	6.40 ± 2.02	6.66 ± 1.70	$5.34 \pm 1.59^{**}$	15
Rectum $V46$ [%]	$V_{46} < 35$	11.67 ± 3.51	12.14 ± 2.46	$9.13 \pm 2.29^{**}$	15
Rectum $V38.5$ [%]	$V_{38.5} < 50$	19.03 ± 6.32	19.86 ± 5.00	$14.34 \pm 3.98^{**}$	15
Bladder $V49_w$ [%]	$V_{49} < 25$	6.52 ± 4.08	6.76 ± 3.28	$8.37 \pm 3.64^{**}$	15
Bladder $V38.5$ [%]	$V_{38.5} < 50$	11.30 ± 6.44	$12.35 \pm 5.81^{*}$	$14.31 \pm 6.66^{**}$	15
Femoral head (L) D_{mean_w} [Gy]	-	10.14 ± 4.56	10.55 ± 4.87	$10.88 \pm 4.74^{*}$	10
Femoral head (R) D_{mean} [Gy]	-	10.78 ± 5.06	10.62 ± 5.37	10.78 ± 5.06	9

*: $p < 0.05$ **: $p < 0.01$ †: Dose constraint is not met

Table 6 includes the mean (over the validation group) of the OAR structures and the dose values for PTVs with SD and degrees of freedom (Df). Wilcoxon signed rank tested structures are marked with "lowercase w".

Similar to table 6, table 7 presents the statistical results for the VMAT validation plans.

Table 7. *Paired samples t-test results for the prostate model with VMAT validation set. PTV results are presented as volume-doses and OAR results as dose-volumes, excluding femoral heads for which mean doses were computed.*

Structure	Dose-volume constraint	Original plan $(\boldsymbol{mean \pm SD})$	RP model (TAYS tot) $(\boldsymbol{mean \pm SD})$	Df
PTV $D98_w$ [Gy]	$D_{98} \geq 57$	58.16 ± 0.50	$58.40 \pm 0.49^{*}$	29
PTV $D2_w$ [Gy]	$D_2 \geq 60$	61.65 ± 0.39	$62.27 \pm 0.47^{*}$	30
Rectum $V60_w$ [%]	$V_{60} < 3$	1.68 ± 1.13	$2.33 \pm 1.09^{*}$	30
Rectum $V58$ [%]	$V_{58} < 5$	4.05 ± 1.56	4.32 ± 1.39	30
Rectum $V54$ [%]	$V_{54} < 20$	7.01 ± 2.31	7.00 ± 1.87	30
Rectum $V46$ [%]	$V_{46} < 35$	11.56 ± 3.73	11.63 ± 2.75	30
Rectum $V38.5$ [%]	$V_{38.5} < 5$	17.79 ± 5.86	17.63 ± 4.17	30
Bladder $V49$ [%]	$V_{49} < 2$	9.27 ± 6.90	8.93 ± 6.26	30
Bladder $V38.5$ [%]	$V_{38.5} < 50$	15.31 ± 10.99	14.93 ± 9.99	30
Femoral head (L) D_{mean} [Gy]	-	15.45 ± 5.45	14.71 ± 4.62	12
Femoral head (R) D_{mean} [Gy]	-	14.31 ± 5.30	10.62 ± 5.37	12

*: $p < 0.05$ **: $p < 0.01$ †: Dose constraint is not met

Original number of validation plans was 31, but one plan had to be discarded for PTV60 as outlier. Wilcoxon signed rank tested structures are again marked with "lowercase w". Note that for femoral heads the mean doses are presented in tables 6 and 7 instead of volumes corresponding to the DVH constraint. Femoral heads DVHs met the TAYS constraints clearly in every case, hence mean dose was measured for further comparison.

Results, which had no statistical difference in t-test or Wilcoxon rank tests, were further analyzed by creating Bland-Altman plots from the difference data. Plots were primarily analyzed visually and unclear cases additionally by linear regression. Figure 16a shows an example case of a Bland-Altman plot for PTV60 structure (VMAT). Figure 16b shows an example from IMRT PTV60 structure.

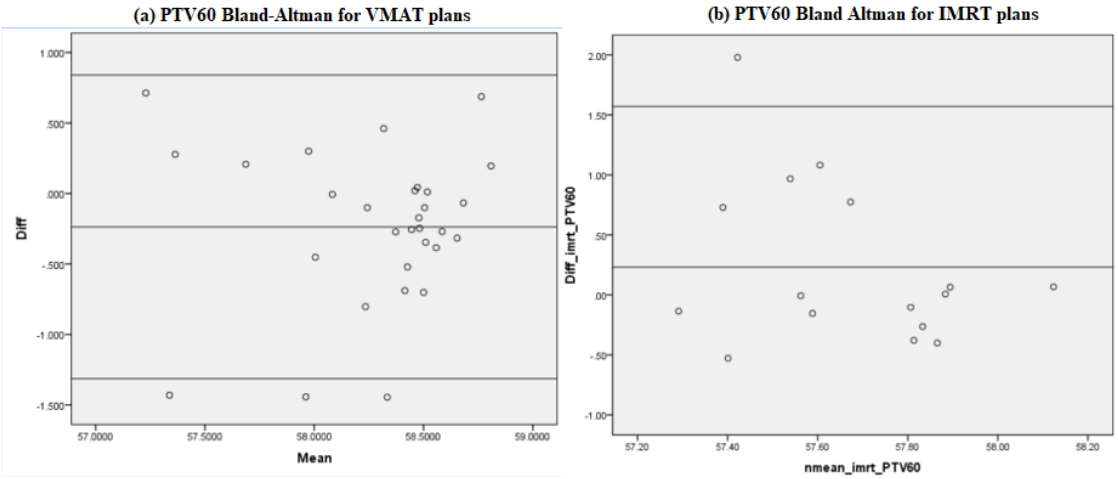

Figure 16. *Bland-Altman plots for PTV60-structures for VMAT (a) and IMRT (b) plans. Neither of the plots show bias.*

Bland-Altman plot includes the scatter point differences between the two plans, where y-axis is the difference value and x-axis the mean. The middle line represents the mean difference of the data and the upper and lower lines represent the 95% confidence interval. Figure 16a was analyzed as having no indication of a trend, since the data scatters approximately evenly around the mean difference. This was confirmed by linear regression in SPSS, having $p = 0.851$ for mean, which strongly indicated that there is no proportional bias ($\alpha = 0.05$). Unlike figure 16, figure 17a indicates visually a trend in the results, having 11 data points below the mean and only 5 data points above the mean with one potential outlier.

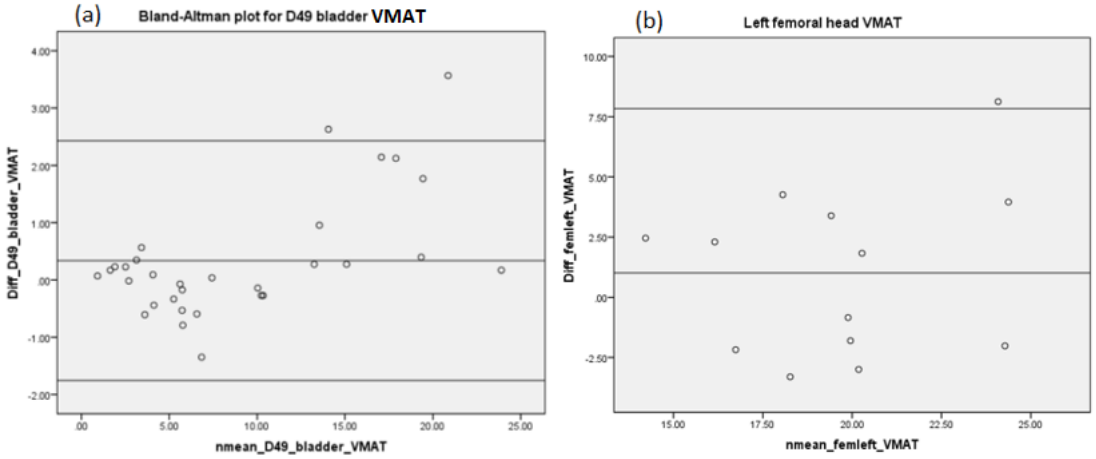

Figure 17. *Bland-Altman plots for bladder (a) and left femoral head (b). The bladder plot (a) shows signs of bias.*

This may be a result e.g. from proportional bias, outliers or from too small data set. Plot's scatter characteristics (large in diff-axis and semi-random) may suggest that the data set is too small for such a large variating data set.

4.1.3 MU and CI values for the prostate models

The CI values inside the region $(1-2)$ have no deviation from the RTOG's protocol. Values $0.9-1.0$ and $2.0-2.5$ are classified having minor deviations and values out from this range are considered as majorly deviating. (Stanley 2011) Table 8 presents the results for the CI single sample- and paired sample t-tests respectively.

Table 8. *Paired samples t-test results for TAYS tot and TAYS coh model's conformity indices.*

Target	$\mathbf{Original_{imrt}}$ $\mathbf{(mean \pm SD)}$	$\mathbf{TAYS\ tot_{imrt}}$ $\mathbf{(mean \pm SD)}$	$\mathbf{TAYS\ coh_{imrt}}$ $\mathbf{(mean \pm SD)}$	$\mathbf{Original_{vmat}}$ $\mathbf{(mean \pm SD)}$	$\mathbf{TAYS\ tot_{vmat}}$ $\mathbf{(mean \pm SD)}$
PTV high	1.52 ± 0.19	1.66 ± 1.17	$1.73 \pm 0.17^{*}$	1.70 ± 0.05	$1.49 \pm 0.04^{**}$

$^{*}: p < 0.05$ $^{**}: p < 0.01$

For IMRT plans $p > \alpha$, where $\alpha = 0.05$ is again considered as the threshold for statistically significant deviation.

MU values were compared with Friedman's test and multiple comparisons post hoc in Matlab. The resulting box plots are presented in figure 18.

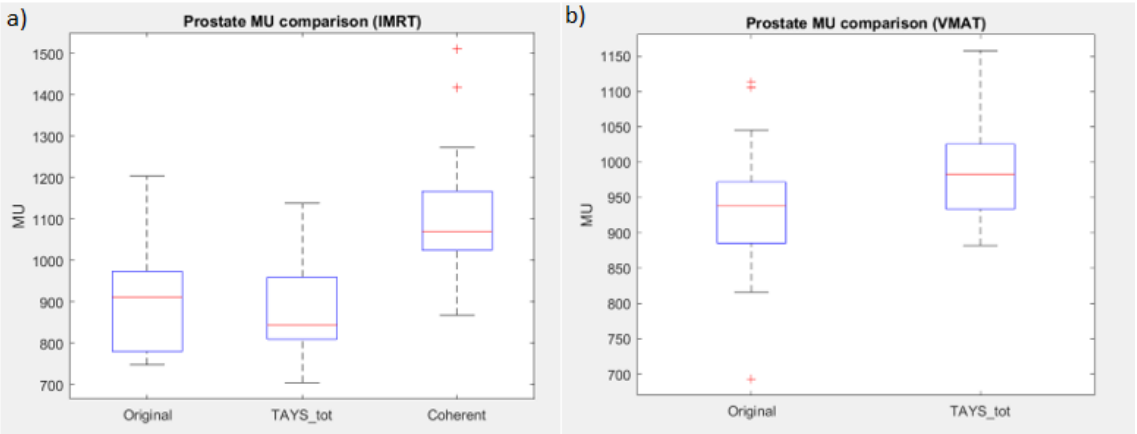

Figure 18. *MU comparison box plots for prostate IMRT and VMAT validation.*

In comparison between IMRT plans, the coherent model's produced MU values exceeded both, the original plan and the TAYS_TOT (RP) plans MU values in all except one case. Comparison statistics are presented in table 9.

Table 9. *MU value comparison results for IMRT plans.*

Model	Mean MU	SD	$\Delta mean$ from original
Original	916.53	141.35	0
TAYS tot	878.3	117.66	-38.20
TAYS coh	1113.9	175.99	197.40

Results for Friedman's test and multiple comparison post hoc with Bonferroni correction are presented in table 10.

Table 10. *Friedman tests results for IMRT plan MU comparisons.*

Friedman test's p-value	Orig vs RP p-value	Orig vs coh. p-value	RP vs coh. p-value
0.0000	0.820	0.003	0.000

In table 11. is presented the MU comparison results for the VMAT plans and the result for the paired sample t-test.

Table 11. *MU value comparison results for VMAT plans.*

Model	Mean MU	SD	$\Delta mean$ from original	p-value (t-test)
Original	932.66	85.76	0	
TAYS tot	987.71	67.08	55.05	0.001

VMAT validation included 30 prostate plans and the IMRT validation 15 plans. Statistical significance for the statistical testing was again chosen as 0.05. In 90 % of the cases the original (manual made) plan produced lower MU values than the RP model.

4.2 HNC models

The final goal in this work was to build, test and evaluate a head and neck cancer model. Additional goal was to iteratively re-plan and re-train the model. The model evaluation results are presented in chapter 4.2.1. The model performance was evaluated by several statistical methods presented in chapter 4.2.2. and for individual patients presented in chapter 4.2.3. The results for the iteratively built model are presented together with the original model and original clinical plans. Finally, additional data from the iterative re-planning (not used in validation) is presented in chapter 4.2.4.

4.2.1 HNC model evaluation results

Model evaluation results for model fit (R^2 and X^2) and model goodness (MSE) statistics are presented in table 12.

Table 12. *Model evaluation results for the first RP HNC model.*

Structure	Training set's size	MSE	R^2	X^2	Possible outliers
Parotid (R)	149	0.00830	0.606	1.047	9
Parotid (L)	146	0.00856	0.606	1.036	9
Submandibular gland (R)	64	0.02604	0.776	1.076	4
Submandibular gland (L)	70	0.02202	0.777	1.146	1
Medulla	150	0.01050	0.309	1.034	8
Mandible	85	0.00620	0.838	1.065	13
Larynx	107	0.01583	0.306	1.030	9
Brain	38	0.00964	0.806	1.144	8
Brainstem	37	0.01411	0.618	1.046	4
Inner ear (R)	35	0.05803	0.786	1.092	6
Inner ear (L)	37	0.02146	0.630	1.015	2
Oral cavity	124	0.00930	0.660	1.018	22

Table 13 shows the results for model fit and model goodness statistics for the iteratively trained HNC model.

Table 13. *Results for the iteratively trained HNC model. The statistics reflect better model fit than with the original model and smaller MSE values.*

Structure	Training set's size	MSE	R^2	X^2	Possible outliers
Parotid (R)	138	0.00383	0.830	1.062	23
Parotid (L)	135	0.00370	0.820	1.034	22
Submandibular gland (R)	59	0.01483	0.912	1.176	8
Submandibular gland (L)	63	0.01333	0.869	1.121	7
Medulla	133	0.00245	0.479	1.045	29
Mandible	78	0.00526	0.927	1.085	18
Larynx	100	0.00661	0.623	1.061	17
Brain	34	0.01519	0.911	1.210	17
Brainstem	32	0.01301	0.848	1.177	8
Inner ear (R)	31	0.01987	0.782	1.148	18
Inner ear (L)	33	0.00900	0.806	1.056	21
Oral cavity	113	0.00311	0.923	1.074	36
Neck	132	0.00322	0.725	1.033	11

4.2.2 Statistical analysis of the HNC model

All the HNC paired sample tests where thus made by non-parametric Wilcoxon signed rank tests. This also makes the $p-$ value comparison between structures more convenient than between parametric and non-parametric tests. The paired sample test results for IMRT plans are presented in table 14.

Table 14. *The non-parametric paired sample test for IMRT plans. Table presents the mean$\pm SD$ values for each model and structure and the tested constraint guideline.*

Structure	Constraint [Gy]	Clinical plans ($\boldsymbol{mean \pm SD}$ [Gy])	RP model ($\boldsymbol{mean \pm SD}$ [Gy])	Iterated model ($\boldsymbol{mean \pm SD}$ [Gy])	Df
PTV 70	$D_{95} \geq 66.5$	67.64 ± 0.74	68.97 ± 0.64	67.78 ± 0.70	9
PTV 66	$D_{95} \geq 62.7$	61.07 ± 0.21	61.09 ± 0.36	61.28 ± 0.17	15
PTV 63	$D_{95} \geq 59.9$	60.18 ± 0.91	61.03 ± 0.65	61.06 ± 0.65	4
PTV 60	$D_{95} \geq 57.0$	58.07 ± 0.37	58.56 ± 0.38	58.73 ± 0.28	4
PTV 56	$D_{95} \geq 53.2$	53.78 ± 0.53	54.47 ± 0.67	54.46 ± 0.67	9
PTV 54	$D_{95} \geq 51.3$	52.38 ± 0.38	52.79 ± 0.35	52.81 ± 38	4
Submandibular gland (R)	$D_{mean} < 45$	53.82 ± 9.85	$46.99 \pm 9.98^{*}$	56.79 ± 9.47	6
Submandibular gland (L)	$D_{mean} < 45$	52.96 ± 15.40	48.97 ± 14.89	48.64 ± 14.64	4
Parotid (R)	$D_{mean} < 26$	32.37 ± 9.99	31.48 ± 8.36	32.44 ± 8.03	14
Parotid (L)	$D_{mean} < 26$	33.05 ± 8.63	33.15 ± 8.63	33.05 ± 10.08	13
Medulla	$D_{max} < 50$	46.78 ± 6.41	46.89 ± 3.80	48.15 ± 3.79	14
Brain	-	18.93 ± 3.15	$28.05 \pm 2.15^{*}$	$28.36 \pm 1.52^{*}$	4
Brainstem	$D_{max} < 55$	50.80 ± 3.22	50.24 ± 1.42	50.04 ± 1.52	4
Oral Cavity	$D_{mean} < 26$	33.29 ± 7.41	30.67 ± 8.53	30.94 ± 9.00	5
Mandibula	$D_{max} < 70$	53.08 ± 5.32	53.64 ± 5.12	53.53 ± 4.79	9
Larynx	$D_{mean} < 45$	49.69 ± 6.91	$39.29 \pm 4.11^{**}$	$39.93 \pm 2.97^{*}$	8

*: $p < 0.05$ **: $p < 0.01$ †: Dose constraint is met

Table 14 presents the results for Wilcoxon ranked sign test for the IMRT treatment plans for related samples. Structures including less than 5 samples were excluded from statistical testing. Note that the degrees of freedom (Df) in table 14. is defined as $N - 1$, where N is the total sample size. The results for paired sample statistical testing for VMAT plans are presented in table 15.

Table 15. *The non-parametric paired sample test for VMAT plans. Table presents the mean±SD values for each model and structure and the constraint guideline.*

Structure	Constraint [Gy]	Clinical plans ($\boldsymbol{mean \pm SD}$ [Gy])	RP model ($\boldsymbol{mean \pm SD}$ [Gy])	Iterated RP model ($\boldsymbol{mean \pm SD}$ [Gy])	Df
PTV 70	$D_{95} \geq 66.5$	67.82 ± 0.50	68.05 ± 0.68	67.98 ± 0.70	9
PTV 66	$D_{95} \geq 62.7$	61.07 ± 0.21	61.01 ± 0.36	61.28 ± 0.17	4
PTV 63	$D_{95} \geq 59.9$	60.94 ± 0.38	60.83 ± 0.69	60.79 ± 0.57	6
PTV 60	$D_{95} \geq 57.0$	58.07 ± 0.37	58.56 ± 0.38	58.73 ± 0.28	4
PTV 56	$D_{95} \geq 53.2$	54.28 ± 0.40	54.05 ± 0.52	53.82 ± 0.84	9
PTV 54	$D_{95} \geq 51.3$	52.38 ± 0.38	52.79 ± 0.35	52.81 ± 38	4
Submandibular gland (R)	$D_{mean} < 45$	43.82 ± 18.08	41.93 ± 15.99	41.25 ± 16.73	8
Submandibular gland (L)	$D_{mean} < 45$	42.45 ± 14.22	43.92 ± 13.62	40.97 ± 13.09	9
Parotid (R)	$D_{mean} < 26$	30.21 ± 10.78	29.95 ± 9.22	30.85 ± 9.37	12
Parotid (L)	$D_{mean} < 26$	24.46 ± 6.62	25.95 ± 6.17	$30.84 \pm 5.91^{*}$	12
Medulla	$D_{max} < 50$	44.59 ± 3.36	$43.93 \pm 1.60^{\dagger}$	$43.83 \pm 1.71^{\dagger}$	12
Brain	-	18.93 ± 3.15	$28.05 \pm 2.15^{*}$	$28.36 \pm 1.52^{*}$	4
Brainstem	$D_{max} < 55$	$50.80 \pm 3.22^{\dagger}$	$50.24 \pm 1.42^{\dagger}$	$50.04 \pm 1.52^{\dagger}$	4
Oral Cavity	$D_{mean} < 26$	29.41 ± 7.22	30.29 ± 6.66	30.29 ± 6.79	7
Mandibula	$D_{max} < 70$	40.74 ± 14.83	40.74 ± 14.42	39.97 ± 15.02	5
Larynx	$D_{mean} < 45$	44.18 ± 8.18	44.95 ± 6.91	45.47 ± 5.97	4

*: $p < 0.05$ **: $p < 0.01$ †: Dose constraint is met

Again, structures with less than 5 samples were excluded from the analysis. Due to small sample sizes, the IMRT and VMAT data was combined and tested once more with non-parametric Wilcoxon test. The results are presented in table 16.

Table 16. *The non-parametric paired sample test for all the HNC plans combined. Table presents the mean±SD values for each model and structure.*

Structure	Dose constraint [Gy]	Original plan ($mean \pm SD$ [Gy])	RP model ($mean \pm SD$ [Gy])	Iterated RP model ($mean \pm SD$ [Gy])	Df
PTV 70	$D_{95} \geq 66.5$	67.73 ± 0.62	68.01 ± 0.64	67.88 ± 0.69	19
PTV 66	$D_{95} \geq 62.7$	63.15 ± 0.86	64.34 ± 1.26	64.38 ± 1.27	7
PTV 63	$D_{95} \geq 59.9$	60.63 ± 0.73	$60.91 \pm 0.65^{*}$	$60.90 \pm 0.59^{*}$	11
PTV 60	$D_{95} \geq 57.0$	57.61 ± 0.87	58.32 ± 1.83	58.40 ± 1.94	6
PTV 56	$D_{95} \geq 53.2$	53.82 ± 0.62	54.26 ± 0.62	54.139 ± 0.8	19
PTV 54	$D_{95} \geq 51.3$	52.02 ± 0.86	$52.79 \pm 1.13^{*}$	52.522 ± 1.71	7
Submandibular gland (R)	$D_{mean} < 45$	48.20 ± 15.47	$44.14 \pm 13.52^{**}$	$43.71 \pm 13.89^{**}$	15
Submandibular gland (L)	$D_{mean} < 45$	45.96 ± 14.97	45.60 ± 13.74	45.53 ± 13.78	14
Parotid (R)	$D_{mean} < 26$	31.37 ± 8.55	30.77 ± 8.63	31.70 ± 10.22	27
Parotid (L)	$D_{mean} < 26$	28.91 ± 9.50	29.68 ± 8.27	$30.66 \pm 7.93^{*}$	26
Medulla	$D_{max} < 50$	45.77 ± 3.68	45.52 ± 3.30	46.15 ± 5.24	26
Brain	-	18.93 ± 3.15	$28.05 \pm 2.15^{*}$	$28.36 \pm 1.52^{*}$	6
Brainstem	$D_{max} < 55$	50.80 ± 3.22	50.24 ± 1.42	50.04 ± 1.52	4
Oral Cavity	$D_{mean} < 26$	31.01 ± 7.27	30.45 ± 7.20	30.57 ± 7.49	13
Mandibula	$D_{max} < 70$	48.45 ± 10.90	48.80 ± 11.25	48.44 ± 11.62	15
Larynx	$D_{mean} < 45$	47.72 ± 7.58	$41.31 \pm 5.74^{*}$	$41.91 \pm 4.90^{*}$	13

*: $p < 0.05$ **: $p < 0.01$ †: Dose constraint is met

Finally, a small script was written for the test to run one-way Friedman and post hoc tests for the three models. The resulted box plots are presented in figures 19, 20 and 21.

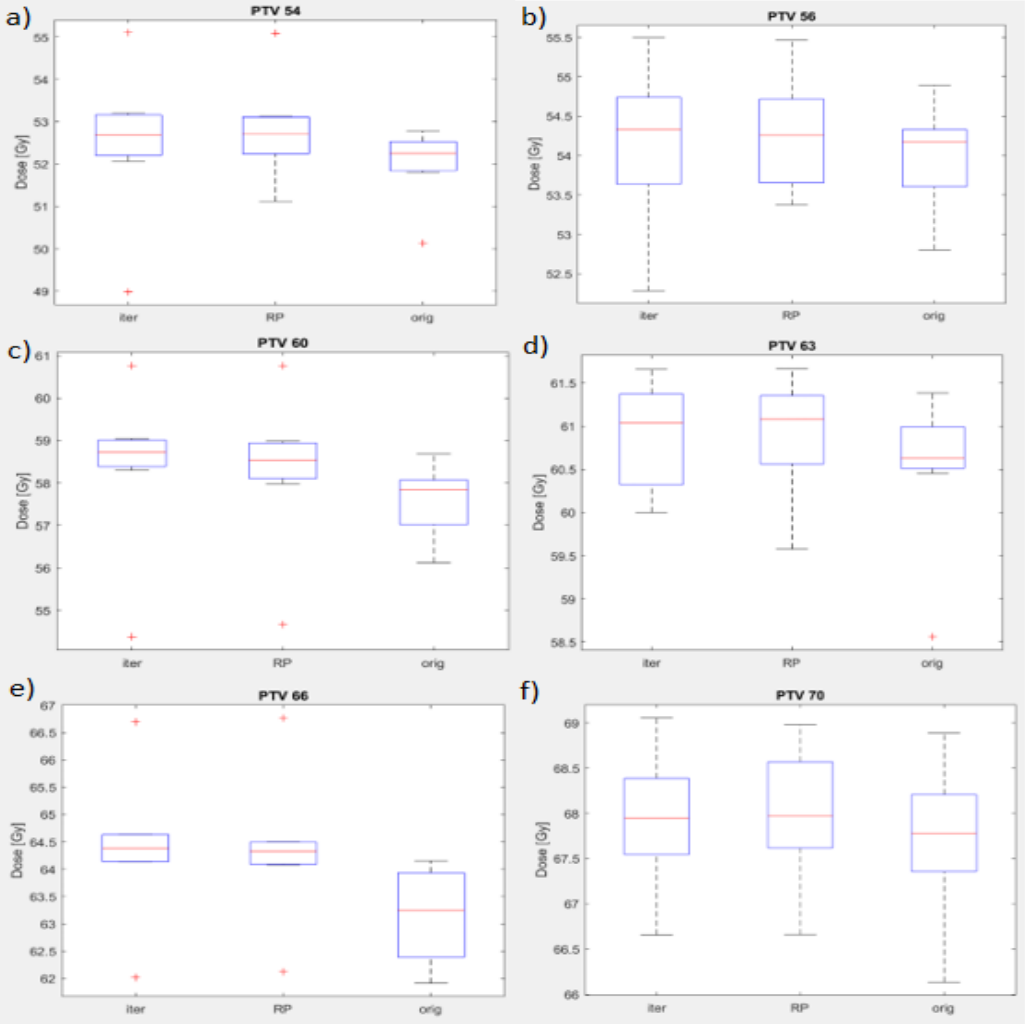

Figure 19. *Box plots for PTVs. The Friedman test showed no statistically significant difference between the plans (*: p<0.05, **- p<0.01). IMRT and VMAT data was combined for Friedman and Post hoc tests.*

Figure 19 presents the results for planned target volumes between the clinical plans (orig), first HNC RP model and the iterative HNC model (iter). Box plots were created with Matlab, and the plots consists of the central red line which is the median, box bottom and top edge represent the 25 and 75 percentile values and the whiskers are the lowest and highest values in the group. (MathWorks 2018) Outliers are marked with red "+" -symbol.

In figure 20 are presented the box plots for combined VMAT and IMRT data for submandibular- and parotid glands, medulla and mandible.

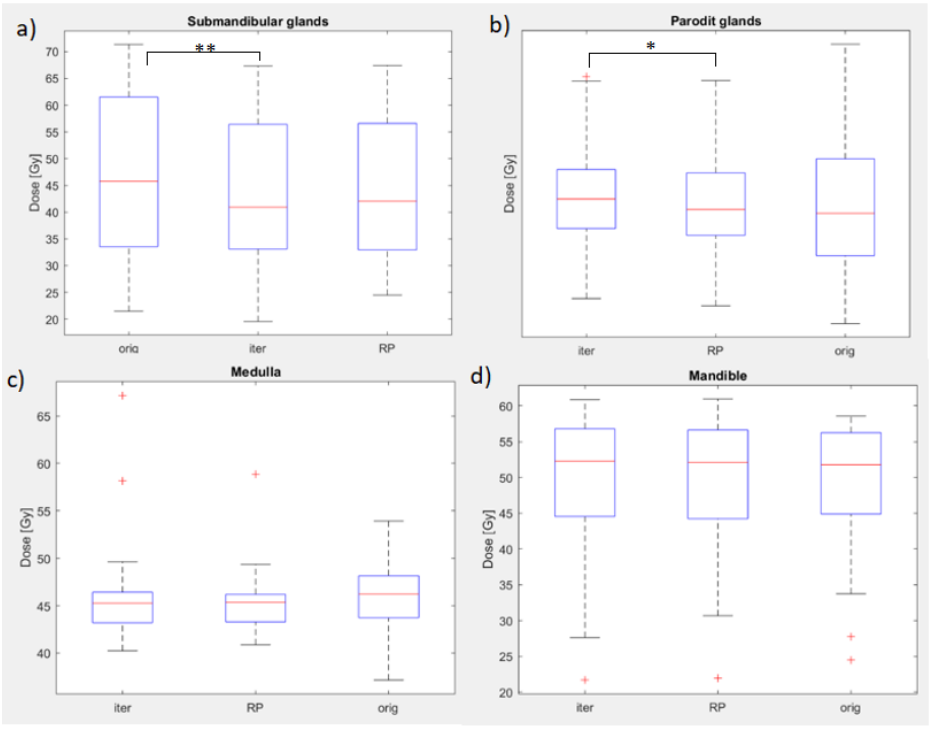

Figure 20. *Box plots for submandibular (a) and parotid glands (b), medulla (c) and mandible (d). The post hoc test showed statistically significant difference for submandibular and parotid glands (* p<0.05, ** p<0.01).*

The last remaining OAR structures analyzed with multiple comparisons are presented in figure 20.

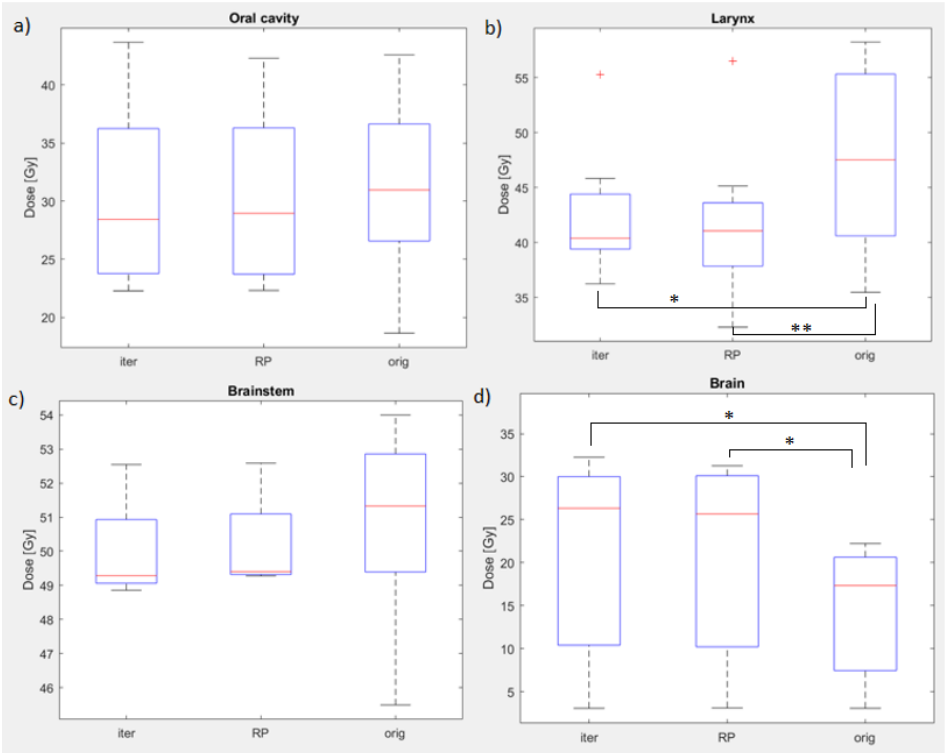

Figure 21. *Box plots for oral cavity (a), larynx (b), brainstem (c) and brain (d). The post hoc test showed statistically significant difference for brain (* p<0.05, ** p<0.01)*

The numeric statistical results for the Friedman's test and post hoc tests are presented in table 17.

Table 17. *Statistical results for the Friedman and post hoc tests for all patients.*

Structure	Friedman (p-value)	Post hoc Iter-RP (p-value)	Post hoc Iter-Orig (p-value)	Post hoc RP-Orig (p-value)	Plan/plans differ (p-value)
Parotids	0.002	0.002	0.033	1.000	Yes
Submandibu-lar glands	0.000	0.067	0.000	0.170	Yes
Medulla	0.156	1.000	0.544	0.184	No
Brainstem	0.247	0.342	0.618	1.000	No
Brain	0.021	1.000	0.049	0.049	Yes
Mandible	1.000	1.000	1.000	1.000	No
Larynx	0.010	0.392	0.392	0.008	Yes
Oral Cavity	0.947	1.000	1.000	1.000	No
PTV54	0.197	1.000	0.240	0.634	No
PTV56	0.951	1.000	1.000	1.000	No
PTV60	0.277	1.000	0.326	1.000	No
PTV63	0.339	1.000	0.459	0.922	No
PTV66	0.197	1.000	0.24	0.634	No
PTV70	0.047	1.000	0.081	0.120	No

Table 17. presents the p-value for Friedman's test, with $\alpha = 0.05$. Probability $p < 0.05$ indicates that at least one group's mean deviates significantly from the other groups. Post-hoc results show the $p - value$ for each comparison between individual groups. The critical value for this comparison was defined by the Bonferroni method defined by equation 13.

4.2.3 CI and MU values for the HNC models

Results for VMAT conformity indices are presented in table 18.

Table 18. *Paired sample comparison for conformity indices. (VMAT)*

Target	Original plan CI $(\boldsymbol{mean\ CI \pm SD})$	RP model CI $(\boldsymbol{mean\ CI \pm SD})$	Iterated RP model CI $(\boldsymbol{mean\ CI \pm SD})$	Df
PTV high	1.19±0.14	1.24±0.17	1.24±0.17	15
PTV intermediate	1.38±0.19	1.51±0.19*	1.46±0.19	8
PTV low	1.37±0.14	1.46±0.21*	1.44±0.22	15

Table 19 presents the same CI comparison for IMRT plans.

Table 19. *Paired sample comparison for conformity indices. (IMRT)*

Target	Original plan ($mean\ CI \pm SD$)	RP model ($mean\ CI \pm SD$)	Iterated RP model ($mean\ CI \pm SD$)	Df
PTV high	1.23±0.34	1.27±0.10	1.26±0.08	15
PTV intermediate	1.33±0.09	1.52±0.10**	1.52±0.11**	8
PTV low	1.44±0.13	1.67±0.17**	1.67±0.17**	15

MU values were compared with same methods as for prostate models. The resulted box plots are presented in figure 22.

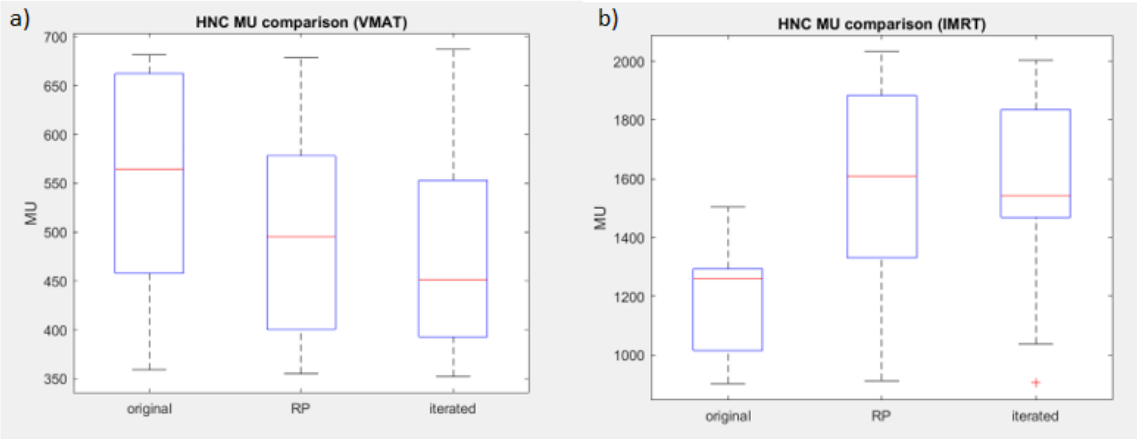

Figure 22. *Box plots for HNC model MU value comparison.*

Comparison statistics for VMAT validation plans are presented in table 20.

Table 20. *MU value comparison results for VMAT plans.*

Model	Mean MU	SD	$\Delta mean$ from original
Original	550.7	117.6	0
HNC (RP)	498.2	1095	-52.5
Iterated model	482.0	105.9	-68.7

Results for Friedman's test and multiple comparison post hoc with Bonferroni correction are presented in table 21.

Table 21. *Friedman test's results for VMAT plans.*

Friedman test's p-value	Orig vs RP p-value	Orig vs iterated p-value	RP vs iterated p-value
0.004	0.074	0.003	0.922

In table 22 is presented the MU comparison results for the IMRT plans and the result for the paired sample t-test.

Table 22. *MU value comparison results for IMRT plans.*

Model	**Mean MU**	**SD**	**Δ*mean* from original**
Original	1197.9	194.7	0
HNC (RP)	1586.9	337.6	389.0
Iterated	1566.4	316.4	368.0

In table 23 are presented the Friedman and post hoc test results for IMRT plans.

Table 23. *Friedman test's results for IMRT plans.*

Friedman test's p-value	**Orig vs RP p-value**	**Orig vs iterated p-value**	**RP vs iterated p-value**
0.001	0.001	0.009	0.301

For IMRT plans 87 % of the cases, both RP created plans exceeded the original plan's MU values. VMAT plans the results were contrary to IMRT, and the original plan exceeded both RP model produced plans MU values in 75 % of the cases.

4.2.4 Statistical results for retrained training set plans

All the data was could have not been considered normally distributed due to Shapiro-Wilk test. The results for paired sample Wilcoxon signed rank test are presented in table 24.

Table 24. *The non-parametric paired sample test for all the HNC plans combined. Table presents the mean$\pm SD^{**}$ values for each model and structure, where the possible * or ** -superscript indicates the significance level. The statistical significance is *p<0.05 and highly significant deviation **p<0.01.*

Structure	Constraint [Gy]	Original plan ($\boldsymbol{mean \pm SD}$ [$\boldsymbol{Gy}$])	RP model ($\boldsymbol{mean \pm SD}$ [$\boldsymbol{Gy}$])	Df
PTV 70	$D_{95} \geq 66.5$	67.70±1.30	67.97±0.55	78
PTV 66	$D_{95} \geq 62.7$	62.37±8.11	62.47±8.17	55
PTV 63	$D_{95} \geq 59.9$	60.71±0.68	60.61±0.98*	55
PTV 60	$D_{95} \geq 57.0$	56.18±7.98	56.30±7.92	47
PTV 56	$D_{95} \geq 53.2$	53.94±0.59	53.60±1.63*	78
PTV 54	$D_{95} \geq 51.3$	50.32±7.64	50.41±7.44	57
Submandibular gland (R)	$D_{mean} < 45$	48.10±11.65	45.37±11.27**	55
Submandibular gland (L)	$D_{mean} < 45$	46.05±10.69	43.30±10.30**	61
Parotid (R)	$D_{mean} < 26$	28.40±9.41	30.52±9.06**	139
Parotid (L)	$D_{mean} < 26$	28.21±9.54	30.22±8.64**	135
Medulla	$D_{max} < 50$	42.88±5.45	43.57±7.95	139
Brainstem	-	47.00±5.45	47.46±4.36**	32
Oral Cavity	$D_{max} < 55$	34.04±10.35	33.32±10.91*	114
Mandible	$D_{mean} < 26$	41.21±9.77	40.55±9.88**	72
Larynx	$D_{max} < 70$	40.10±7.58	39.52±6.50	98
Inner ear (R)	$D_{mean} < 45$	15.31±9.33	11.50±7.81**	33
Inner ear (L)	$D_{mean} < 45$	14.72±7.99	9.57±6.26**	35

As seen in table 24. PTV 56 (PTV low) and PTV 63 (PTV intermediate) showed statistical significance between the plans. From OARs the statistical significance was high for 8 structures and only 2 structures had $p > 0.05$. Mean differences between OARs are $< 6.6\ \%$ from the higher dose for all except for inner years mean differences are relatively large ($> 30\ \%$). MU and CI values were not evaluated for the training set's plans.

5. DISCUSSION

In this thesis work the main objective was to find, test, improve and build KBTP models for prostate cancer and head and neck cancer with RP software integrated in Eclipse treatment planning software. Secondary objectives were to build a DVH analysis program for computer aided analysis and to build an iteratively retrained HNC model, and to study whether the iterative training increases OAR sparing or plan coherence. The model performance was studied using clinical treatment plans constructed as validation sets. In this chapter, the results presented in chapter 4 are further discussed with consideration of the objectives, previous research and Varian's proposed benefits for Rapidplan in clinical use.

5.1 Evaluation of the models

Prostate models

Model goodness and model fit statistics are comparable between the final prostate models only regarding rectum and bladder structures, since only the TAYS coh (coherent) model included femoral heads for line objective generation in RP. Model fit statistics presented in table 4 indicate that the coherent model explains the data's variance better (≈ 12 % higher). The same was true for the rectum structure where the rectum's R^2-value was almost twice as high (≈ 46 %) which is a substantial increase. The explanation for this is possibly related to the training set's higher coherence between plans, because the values are determined from the training set itself as explained in chapter 2.4.1. This also led to the conclusion that model fit statistic do not describe the true model performance for plans outside of the training set. These values should thus be used only as indicators, especially for models including small and coherent training sets. The robust model included also varying amount of bladder and rectum filling. This reflects to model fit statistics by increasing the variance and number of potential outliers, which were 10 for the TAYS tot and 0 for TAYS coh model for bladder. In general, one should consider the relevance of a given model fit statistic for structures, which's size is dependent mostly on physiology and may change between subsequent treatment fractions.

The model performance was measured with MSE, which had no major differences between the RP models; approximately 1 % for bladder and 8.6 % for rectum, i.e. TAYS tot's MSE -values were lower indicating higher predictive power. This was predictable, because training set's size also contributes to descriptive model statistics (see chapter 3.4), and TAYS tot model's training set was almost 4 times larger than in coherent model, including plans with higher variation in geometry.

When comparing the model fit statistics to previous research, the robust prostate model showed lower or similar fit for bladder and lower fit for rectum. The coherent model showed similar or higher results for prostate. The averages from previous research were $R^2 = 0.556$, $X^2 = 1.095$ for rectum and $R^2 = 0.832$ and $X^2 = 1.071$ for bladder. Femoral head model fit statistics presented in table 5 were similar between the model generated plans and previous research with less than 15 % difference. (Aviles 2018) (Botti 2015) The TAYS tot model was built by mostly random selection. Plan exclusion was done with loose boundaries compared to previous research or TAYS coh models, which could partly explain the differences in model fit. Though, high coherence in training set may lead to data overfitting in the regression model and affect negatively to the model's predictive power. Comparison of the model fit statistics and previous inspection of the regression data (not presented) showed that none of the models validated in this thesis had data overfitting.

As mentioned before, the model should not be modified just to get better descriptive model fit statistics. Also, rectum and bladder are hollow structures, which vary daily in size and shape. Aiming for high model fit statistics may thus not be meaningful especially for the robust model. The results show that higher model fit for TAYS coh model did not improve the model performance considering mean doses nor the dose variance compared to TAYS tot model.

Head and neck models

Model fit for HNC RP model can be considered optimal ($R^2 > 0.7$, $X^2 < 1.1$) or close to optimal to all OARs except for medulla and larynx. For the iteratively trained model only medulla ($R^2 = 0.479$) showed considerably lower model fit. Again, the model fit statistics should be and were used only as guiding indicators in model training, and in outlier detection and -validation. The iteratively trained model showed generally higher model fit, which is partly caused by further outlier exclusion. Matched OARs differed also slightly because of the additional neck structure in iteratively trained model. In comparison to previous research, R^2 was higher for the HNC RP model and considerably higher for HNC iter model ($\Delta R^2 \approx 0.2$).

MSE values were considerably higher for the HNC RP model for every OAR structure, excluding larynx, mandible and brainstem. The latter two yielded similar results between the two RP models. Only larynx's MSE was higher in HNC iter model. Lower MSE values again indicate higher predictive power. The mean difference between OAR MSE values was 116 % in favor of HNC iter model. One could expect that such high decrease in MSE would translate to smaller variation in the final DVHs, but the dose comparison results showed no sign of such benefit.

5.2 Prostate model dose comparison

IMRT plans

The TAYS tot model showed statistically significant deviations from the clinical IMRT plans for bladder's 38.5 Gy dose level with higher volume coverage. Also, rectum's high dose (60 Gy) showed statistically (highly) significant difference between RP and clinical plan with higher dose-volume.

TAYS coh model deviated also significantly from the clinical plans with all bladder dose levels and high rectum dose with higher mean values (table 6). For lower dose levels the TAYS coh model's volume coverage was lower with high statistical significance. This means that the RP models could not limit the dose spreading from the PTV to rectum as efficiently as the clinical optimization. The increased high dose coverage is then compensated in optimization by limiting the spread of low dose levels.

Femoral heads in both models showed no statistically significant deviations from the clinical plans. Unlike rectum and bladder, femoral heads are structures with constant geometry with similar features between patients. Also, the physical distance from PTV is relatively large. Thus, the structure is not as problematic for RP DVH optimization.

VMAT plans

VMAT plan comparison presented statistically significant deviations between TAYS tot and clinical plans for PTV structures and highest rectum dose. Again, the clinical plans are optimized with high priority for rectum's high dose sparing and RP models are unlikely to produce same level of sparing near PTVs.

For VMAT, only the TAYS tot (robust) model was compared to clinical plans and both measured volumes (D_2 and D_{98}) had slightly higher mean dose coverage for the RP generated plan. This also explains the increased (60 gy) rectum dose-volume. Otherwise the VMAT plan doses were statistically similar to the clinical plans. TAYS coh model was not validated with VMAT plans.

Further discussion

VMAT results showed overall surprisingly high similarities with the clinical plans. The prostate model was constructed using almost only IMRT plans, which would explain the differences in VMAT plan validation. Even though RP algorithm can construct VMAT plans from IMRT-trained model, the characteristic dose distributions are different between VMAT and IMRT plans, and the DVH prediction model may not apply sufficiently in these cases. However, this does not explain the large differences between clinical and RP based IMRT plans. The result could be explained by the high IMRT planning quality and routine in TAYS with high priority in OAR sparing. Furthermore, RP algorithms are

designed to generate coherent plans reflecting the hospitals conventions, rather than to produce highly individualized plans competing in OAR sparing with their clinical counterparts.

Generally, the clinical IMRT plans showed better OAR sparing compared to RP generated plans. As mentioned before, only for rectum's lower doses the volume exposure was smaller for RP plans when considering the mean values. The Bland-Altman plots constructed for additional mean-dose comparison for similar plans showed either no bias or unclear bias. Because of the small validation data set the plots did not bring additional insight to the statistical testing.

The results for prostate correspond to previous research in case of PTV coverage, which was generally found to be either the same (statistical sense) or higher in comparison with original plans. OAR sparing was also generally found higher in previous research, which was not the case in this thesis for prostate model. One reason might be the optimization objectives for PTVs, which were prioritized higher relative to OARs than generally in previous research. Also, only upper- and lower objectives were used as OAR optimization objectives instead e.g. combination of mean and gEUD objectives. All objectives were also chosen manually instead of letting RP to generate the objectives, which would have been justifiable considering the goals of this thesis work. Even though statistically the RP prostate model could not produce higher OAR sparing, the significance testing against the dose constraints set by the hospital showed that all prostate models are able to produce clinically acceptable plans.

5.3 HNC model dose comparison

PTV structures

The HNC models showed no statistically significant differences for IMRT nor VMAT plans for PTV structures in paired testing against the clinical plan (table 14 and table 15). The HNC comparison with combined VMAT and IMRT plans (table 16) showed statistically significant deviation for both RP plans with one dose level ($63\ Gy$), and for the original HNC RP plan with second dose level ($54\ Gy$). RP plans had higher mean dose coverage compared to clinical plans with similar deviations.

OAR structures

HNC model comparisons for OAR structures showed no statistical significance except for submandibular glands, larynx, and brain. From the basis of dose-means, OAR sparing was higher for both RP model generated plans for VMAT and IMRT. The HNC RP model had no statistical significant difference in OAR sparing (excluding brain) to clinical plans.

Paired tests and multiple comparisons results showed highly statistically significant differences for brain structure in all cases. However, the relevance of the results for brain should be reconsidered for 3 reasons:

1. The brain's dose measured only as mean value instead of maximum dose, which would be reasonable for serial organs.
2. The data set size for brain was small for statistical testing.
3. Brain structure geometry, optimization relevance and size vary much between different cancer cases.

Nevertheless, the brain structure with mean objectives was included in the models and had an effect to the plan optimization. For validation, the effect though remains relatively small, because the number of validation plans including brain structure was also relatively small (7 from 28 plans) and the upper dose objective was set as high ($55\ Gy$) with average priority. The mean dose for brain was considerably smaller than the placed objective, which means that the objective is easily achieved during the optimization without high expenses in PTV filling nor OARs sparing. Also, the RP generated line objective for brain and brainstem had the lowest priorities compared to other OAR line objectives (table 3).

Further discussion

The results for the HNC models were comparable to previous plans. Previous research mainly indicated either no statistically significant difference between RP and clinical plans or RP creating superior plans in PTV coverage and OAR sparing. The latter can be considered true only for few structures in this thesis work considering the statistical significance together with mean values and individual DVHs. The statistical testing indicated that both, HNC RP and HNC iter models can produce clinically acceptable plans with proper OAR sparing. Take note that in some individual RP HNC plan OARs with high dose variation (gland structures) usually exceeded the hospital's dose constraints. High variation between plans was caused by different cancer cases, where the OAR structures are physically close or overlapping the PTV and by different organ geometries. (Eclipse Algorithms 2015) Since additional structures were not included for OARs, e.g. due to overlapping, the total OAR received higher doses because PTV coverage had the highest priority. The same was true for clinical plans. Though, clinical plans had majorly higher level of OAR sparing for parotid glands compared to RP. Gland structures are spared during the manual optimization with high priority and the optimization is usually proceeded until the OAR sparing is maximized. The operation principle of RP does not always support same level of OAR optimization polishing which mostly explains the difference. RP plans spared the submandibular gland receiving lower total dose with statistical significance compared to clinical plans. Submandibular gland receiving higher dose was usually spared more by the clinical plan with high statistical significance (tables 14-16).

The multiple comparison results showed no statistical differences between any of the models for PTVs (figure 19). Though, it is important to notice that for the lower dose level PTVs the sample size was relatively small, and the data points classified as possible outliers in RP generated plans may thus have excessive leverage in statistical analysis.

One could consider dividing the validation set to more coherent cancer cases and then conclude the analysis case-wise. In this thesis work this was not possible for HNC model because of limited validation set size. The relevance of the constraint comparison results for parallel organs should be reconsidered for the reasons mentioned above, and because the constraints are not always considered as absolute, but rather as guidelines. Thus, the conclusion should be based more on the comparison between the accepted clinical plan and RP plans and with the serial organ constraints, which are considered more as absolute maximum values.

5.4 Dose conformity in target structures and MU values

Target conformity

The CI values in tables 18-19 were found statistically similar between all HNC model generated plans and original plans for the highest dose level (PTV high). For PTV intermediate and PTV low the CI values could be considered similar only for the TAYS iter model-generated VMAT plans. Otherwise the CI value significance level was $0.01 < p < 0{,}05$ for VMAT plans and $p < 0{,}01$ for all IMRT plans. The difference could be explained by the fact that the HNC models were trained using mostly VMAT plans. Creating an IMRT plan from VMAT based model might not be optimal considering the dose distributions and conformity, especially in lower PTV dose levels. When considering the RTOG protocol (see chapter 2.1.5) for CI, every plan created by each model stayed inside the proposed constraints for acceptance ($1 \leq CI \leq 2$). The results for conformity indices are in line with previous research, where the CI values were also in acceptable limits.

MU values

Finally, the MU multiple comparison results in tables 9-11 and 20-13 and in figure 18 and figure 22, showed that for prostate and HNC RP -model's plans had generally higher MU values compared to the original plans. Only exception was the TAYS tot model for prostate which had statistically similar MU values compared to original plans. Between the RP models, TAYS tot had significantly lower MU values compared to the coherent plan for prostate cancer. HNC models did not differ from each other. One possible reason for these results are the multiple optimization objectives including gEUD objectives and other high priority objectives. This may result to higher number of leaf positions and thus increasing delivery time which in turn increases the MU -values.

5.5 Effects of HNC model's iterative retraining

The HNC iter model was trained iteratively by re-planning the original model's training set as described in chapter 3.2. Varian's consultation suggested that this may increase the plan quality and coherence. As described in chapter 5.1, HNC iter had better model fit compared to the HNC RP model and generally lower MSE values, which indicated higher predictive power. The multiple comparison post-hoc test results (figures 19-21 & table 17) showed that HNC iter-generated plans had statistically same level of OAR sparing except for parotid glands which HNC RP spared slightly more (mean comparison). Otherwise the dose results were similar for both models. Similar to HNC RP model, HNC iter had higher (mean) OAR sparing for submandibular glands and larynx with high statistical significance compared to the clinical plan. This was not predicted because as mentioned before, submandibular gland's dose optimization has high priority in clinical planning.

CI values yielded similar results to clinical plans with no statistically significant difference for VMAT- and highest dose PTV for IMRT plans. IMRT plan PTVs with lower dose levels showed statistically significant difference between HNC iter and clinical plan. However, all CI values were inside the RTOGs limits for acceptance (chapter 2.1.5). In general, HNC iter produced more similar plans to clinical versions than the original HNC RP model. MU value comparisons did not show statistical significance between HNC RP and HNC iter models. The average (absolute) MUs for HNC iter- were slightly lower that for HNC RP model.

5.6 Limitations and recommendations

Specification of an optimal- or simply a better plan is not straight-forward, because of the multiple factors affecting to the plan evaluation. Individual cancer cases deviate largely and building a model, which covers even some of the deviation is a challenging process. The most challenging part is to define the plans which construct the model and to choose the features and cases, which are included and excluded from the model. In this thesis work the initial specification for each training set could have been defined more accurately for consistency. Also, in later testing of the model it was noticed that better results are achieved if additional ring structures are added around PTVs for normal tissue dose minimization. Additionally, HNC iter model's neck structure objectives would have been more sufficient to define with maximum dose constraint instead of mean dose.

As emphasized before, RP does not aim to compete with absolute OAR sparing with the clinical plans and in many cases, this is not possible with the program's architecture. If a given plan is manually optimized with special consideration to certain OARs, it is very probable that RP based plan yields lower OAR sparing for the same structures. When considering the coherence of RP model's, the standard deviation for OAR structures was smaller in majority of the cases for both, HNC and prostate models. The coherent RP

model for prostate was expected to show higher coherence, but in table 6 the SD values are comparable to the robust model. However, this kind of result was predicted because higher planning coherence demands also coherence in the validation set, and in this thesis the validation set was robustly chosen for prostate cancer. For HNC model the iteratively trained model showed no sign for better coherence in means of SD for OAR structures compared to the original RP model. Again, the validation set included several different types of cancer in head and neck region without further classification, which could be an explanatory factor together with small number of validation plans.

The final models have been used for some further testing by TAYS physicist and it is probable that the models will be used in future as support in clinical planning. Before this, the HNC model's user defined optimization objectives should be reconsidered especially for the brain structure. Also, additional testing and modifications are recommended.

Some of the previous research included validation groups, which include plans from the model's training set. This is not recommended in future model building. The model's training set and validation groups should be chosen so that they also represent the target group. This can be achieved either by constructing a robust general model or multiple coherent models. Plans generated using RP do not always correspond to manually optimized plans but can be used as a starting point to manual optimization.

6. CONCLUSIONS

The main objective of this master's thesis included prostate and HNC Rapidplan model implementation in Tampere University Hospital. Implementation was concluded by testing and modifying previously made models and by building new models based on TAYS constraints and planning conventions. Additional objectives were to create DVH data analysis program for data handling, presentation and analysis, and to build iteratively trained HNC model from the first model version by re-planning the model's training set.

Preliminary model testing yielded a suitable robust prostate model built by TAYS physicist. The model was then further trained and modified. Also, a prostate model with small and coherent training set was built as part of this thesis. Both models were able to generate DVH estimations and optimization objectives, which led to clinically acceptable VMAT plans based on statistical results. The RP generated IMRT plans met also all the DVH constraints, but the OAR sparing was found to be lower in RP generated plans.

HNC model building and training set's retraining resulted in 2 final models. The treatment plans generated using the HNC models were mostly comparable to previous research and clinical plans with proper OAR sparing and PTV filling. The iterative HNC model deviated only slightly from the original RP model. Only significant differences were found in PTV conformity indices and with dose differences mainly in parotid glands. Considering the required time for re-planning of the training set, these improvements are considered rather trivial. Data analysis was concluded with the miscellaneous DVH analysis program, which was built as part of this thesis and successfully used throughout the project.

As conclusion all the objectives set for this master's thesis were met. In future the models are planned to be tested further and included as support in clinical planning. The results show that the models can generate optimization objectives leading to clinically acceptable treatment plans after optimization. Depending on the case and the model training, the plan based on the RP model could be used clinically as such, or as starting point for standard AIO optimization.

REFERENCES

Apinorasethkul, O. 2017. Dosimetric Effects of Using Generalized Equivalent Uniform Dose (gEUD) in Plan Optimization. Penn Radiation Oncology. AAMD – Indianapolis. Presentation. [Accessed 5.4.2018]. Available at: http://pubs.medicaldosimetry.org/pub/8c9c9133-782b-cb6e-2763-bd441ca5d3ce.

Aviles, J. E., Marcos, M., Sasaki, D., Sutherland, K., Kane, B. & Kuusela, E. 2018. Creation of knowledge-based planning models intended for large scale distribution: Minimizing the effect of outlier plans. J Appl Clin Med Phys.

Baumann, M. & Petersen, C., 2005. TCP and NTCP: a basic introduction. Rays 30(2), pp. 99–104.

Bakiu, E., Telhaj, E., Kozma, E., Ruçi, F. & Malkaj, P. 2013. Comparison of 3D CRT and IMRT Tratment Plans. Acta Informatica Medica 21(3), pp. 211–212.

Bortfeld, T. 2006. IMRT: a review and preview. Phys Med Biol 51(13), pp. 363–379.

Bortfeld, T. 2010. The number of beams in IMRT—theoretical investigations and implications for single-arc IMRT. Physics in Medicine and Biology 55(1), pp. 83–97.

Botti, A., Cagni, E., Micera, R., Simoni, N., Orsingher, L., Orlandi, M., Iotti, C. & Iori, M. 2015. Rapidplan: 'knowledge-based' model with Tomotherapy plans. Arcispedale S. Maria Nuova, Medical Physics.

Bourland, D. 2016. Chapter 6 - Radiation Oncology Physics, In Clinical Radiation Oncology (Fourth Edition), edited by Leonard L. Gunderson and Joel E. Tepper, Elsevier, Philadelphia, pp. 93-147.e3, ISBN 9780323240987.

Boutilier, J., Craig, T., Sharpe, M. & Chan, T. 2016. Sample size requirements for knowledge-based treatment planning. Med. Phys 43(3) pp. 1212-1221.

British Institute of Radiology, Institute of Physics and Engineering in Medicine, National Patient Safety Agency, Society and College of Radiographers and the Royal College of Radiologists. 2015. Radiotherapy Board - Intensity Modulated Radiotherapy (IMRT) in the UK: Current access and predictions of future access rates. Report. [Accessed 15.3.2018]. Available: https://www.ipem.ac.uk/Portals/0/Documents/Partners/Radiotherapy%20Board/imrt_target_revisions_recommendations_for_colleges_final2.pdf.

Burnet, N., Thomas, S., Burton, K. & Jefferies, S. 2004. Defining the tumour and target volumes for radiotherapy. Cancer Imaging 4(2), pp. 153–161.

Cao, D. Volumetric Modulated Arc Therapy (VMAT): The future of IMRT? The American Association of Physicist in Medicine: Meetings. Presentation. [Accessed 27.4.2018]. Available: https://www.aapm.org/meetings/amos2/pdf/34-8076-8479-796.pdf.

Dahiru, T. 2008. P – Value, a True Test of Statistical Significance? A cautionary Note. Annals of Ibadan Postgraduate Medicine 6(1), pp. 21–26.

Eclipse Photon and Electron 13.6 Instructions of use. 2015. Varian Medical Systems. Varian Medical Systems, Inc. Palo Alto, CA. USA.

Eclipse Photon and Electron Algorithms Reference Guide. 2015. Varian Medical Systems. Varian Medical Systems, Inc. Palo Alto, CA. USA.

Feuvret, L., Noël, G., Mazeron, J. & Bey, P. 2006. Conformity index: A review. International Journal of Radiation Oncology*Biology*Physics 64(2), pp. 333-342, ISSN 0360-3016.

Fiorino, C., Rancati, T. & Valdagini, R. 2009. Predictive models of toxicity in external radiotherapy. Cancer 115, pp. 3135-3140.

Fogliata, A., Belosi, F., Clivio, A., Navarria, P., Nicolini, G., Scorsetti, M., Vanetti, E. & Cozzi, L. 2014. On the pre-clinical validation of a commercial model-based optimisation engine: Application to volumetric modulated arc therapy for patients with lung or prostate cancer. Radiotherapy and Oncology 113(3), pp. 385-391.

Fogliata, A., Nicolini, G., Bourgier, C., Clivio, A., De Rose, F., Fenoglietto, P., Lobefalo, F., Mancosu, P., Tomatis, S., Vanetti, E., Scorsetti, M. & Cozzi, L. 2015. Performance of a Knowledge-Based Model for Optimization of Volumetric Modulated Arc Therapy Plans for Single and Bilateral Breast Irradiation. PLoS ONE 10(12).

Fogliata, A., Reggiori, G., Stravato, A., Lobefalo, F., Franzese, C., Franceschini, D., Tomatis, S., Mancosu, P., Scorsetti, M. & Cozzi, L. 2017. RapidPlan head and neck model: the objectives and possible clinical benefit. Radiation Oncology 12(73).

Fogliata, A., Thompson, S., & Stravato, A., Tomatis, S., Scorsetti, M. & Cozzi, L. 2017. On the gEUD biological optimization objective for organs at risk in Photon Optimizer of Eclipse treatment planning system. Journal of Applied Clinical Medical Physics 19(1) pp. 106-114.

Galimberti, V., Veronesi, P., Arnone, P., De Cicco, C., Renne, G., Intra, M., Zurrida, S. Sacchini, V., Gennari, R. Vento, A., Luini, A. & Veronesi, U. 2002. Annals of Surgical Oncology 9(9), pp. 924-928.

Ghandour, S., Matzinger, O. & Pachoud, M. 2015. Volumetric-modulated arc therapy planning using multicriteria optimization for localized prostate cancer. Journal of Applied Clinical Medical Physics 16(3), pp. 258–269.

Ghandour, S., Oscar Matzinger, O. & Pachouda, M. 2015 Volumetric-modulated arc therapy planning using multicriteria optimization for localized prostate cancer. Journal of Applied Clinical Medical Physics 16(3).

Gibbons, J. & Roback, D. Monitor Unit Calculations for Photons and Electrons Report of TG-71. The American Association of Physicist in Medicine: Meetings. Presentation. [Accessed 27.4.2018]. Available at: https://www.aapm.org/meetings/2001AM/pdf/7214-33276.pdf.

Gossman, M. & Bank, M. 2010. Dose-volume histogram quality assurance for linac-based treatment planning systems. Journal of Medical Physics 35(4), pp. 197–201.

Grosu A., Sprague L. & Molls M. 2006. Definition of Target Volume and Organs at Risk. Biological Target Volume. New Technologies in Radiation Oncology. Springer, Berlin, Heidelberg.

Henseler, J., Ringle, C. & Sinkovics, R. 2009. The use of partial least squares path modeling in international marketing. Advances in International Marketing (AIM) 20, pp. 277-320

Herman, T., Hibbitts, K., Herman, T., & Ahmad, S. 2011. Evaluation of pencil beam convolution and anisotropic analytical algorithms in stereotactic lung irradiation. Journal of Medical Physics 36(4), pp. 234–238.

Humphrey, P., Liu, W. & Buote, A. 2009. X^2 And Poissonian Data: Biases Even in the High-Count Regime and How To Avoid Them. The Astrophysical Journal 693(1), p. 822.

IBM Corp. Released 2013. IBM SPSS Statistics for Windows, Version 23.0. Armonk, NY: IBM Corp.

Ivanescu, A., Li, P., George, B., Brown, A. W., Keith, S. W., Raju, D., & Allison, D. 2016. The Importance of Prediction Model Validation and Assessment in Obesity and Nutrition Research. International Journal of Obesity 40(6), pp. 887–894.

Lu, J., Zhang, J., Li, M., Cheung, M., Li, Y., Zheng, J., Huang, B. & Zhang, W. 2015. A simple optimization approach for improving target dose homogeneity in intensity-modulated radiotherapy for sinonasal cancer. Nature. Scientific Reports 5(15361).

Kim, T. 2015. T test as a parametric statistic. Korean Journal of Anesthesiology 68(6), pp. 540–546.

Krayenbuehl, J., Norton, I., Studer, G. & Guckenberger, M. 2015. Evaluation of an automated knowledge-based treatment planning system for head and neck. Radiation Oncology 10, p. 226.

Legates, D. & McCabe, G. 1999. Evaluating the use of "goodness-of-fit" measures in hydrologic and hydroclimatic model validation. Water resources research 35(1), pp. 233-241.

Liu, Q., Lee, J. & Jordan, M. 2016. "A Kernelized Stein Discrepancy for Good-ness-of-fit Tests". Proceedings of the 33rd International Conference on Machine Learning. The 33rd International Conference on Machine Learning. New York, USA: Proceedings of Ma-chine Learning Research. pp. 276–284.

Ma, L., Yu, C., Earl, M., Holmes, T., Sarfaraz, M., Li, X., Shepard, D., Amin, P., DiBiase, S., Suntharalingam, M. & Mansfield, C. 2001. Optimized intensity-modulated arc therapy for prostate cancer treatment. Int. J. Cancer, 96, pp. 379-384.

Maier, J., Knott, B., Maerz, M., Loeschel, R., Koelbl, O. & Dobler, B. 2016. Simultaneous integrated boost (SIB) radiation therapy of right sided breast cancer with and without flattening filter - A treatment planning study, Radiation Oncology 11(1), p. 111.

Mao, Y., Yin, W., Guo, R., Zhang, G., Fang, J., Chi, F., Qi, Z., Liu, M., Ma, J. & Sun, Y. 2015. Dosimetric benefit to organs at risk following margin reductions in nasopharyngeal carcinoma treated with intensity-modulated radiation therapy. Chinese Journal of Cancer 34(5), pp. 189-197.

Mayles, P., Nahum, A. & Rosenwald, J. 2007. Taylor & Francis Group. Handbook of Radiotherapy Physics: Theory and Practice. CRC Press, Boca Raton, FL, pp. 944 ISBN: 9780750308601.

Moonen, L. & Bartelink, H. 1994. Fractionation in radiotherapy. Cancer Treatment Reviews pp. 365-378.

Moore, D., Notz, W. & Flinger, M. 2013. The basic practice of statistics (6) p. 138.

MathWorks. 2018. Matlab Documentation - Multiple Comparisons: Boferroni Method. [Accessed 3.5.2018]. Available at: https://se.mathworks.com/help/stats/multiple-comparisons.html#bum7ugh.

Nithya, L., Raj, N., Rathinamuthu, S., Sharma, K. & Pandey, M. 2014. Influence of increment of gantry angle and number of arcs on esophageal volumetric modulated arc therapy planning in Monaco planning system: A planning study. Journal of Medical Physics 39(4), pp. 231–237.

Oh, C., Antes, K., Darby, M., Song, S. & Starkschall, G. 1999. Comparison of 2D conventional, 3D conformal, and intensity-modulated treatment planning techniques for patients with prostate cancer with regard to target-dose homogeneity and dose to critical, uninvolved structures. Medical Dosimetry 24(4), pp. 255-263.

Orlandi, E., Palazzi, M., Pignoli, E., Fallai, C., Giostra, A. & Olmi, P. 2010. Radiobiological basis and clinical results of the simultaneous integrated boost (SIB) in intensity modulated radiotherapy (IMRT) for head and neck cancer: A review. Critical Reviews in Oncology Hematology 73(2), pp. 111-125.

Palma, D., Vollans, E., James, K., Nakano, S., Moiseenko, V., Shaffer, R., McKenzie, M., Morris, J. & Otto, K. 2008. Volumetric Modulated Arc Therapy for Delivery of Prostate Radiotherapy: Comparison with Intensity-Modulated Radiotherapy and Three-Dimensional Conformal Radiotherapy. International Journal of Radiation Oncology*Biology*Physics 72(4), pp. 996-1001.

Petkovska, S., Tolevska, C., Kraleva, S. & Petreska, E. 2010 Conformity index for brain cancer patients. Conference on Med Phys and Biomed Engineering 43(1) pp. 56-58.

R Development Core Team. 2008. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. [Accessed 14.3.2018], Available: http://www.R-project.org

Sasane, J. & Iyer, P. 1981. Relevance of radiation penumbra in highenergy photon beam therapy. Strahlentherapie 157(10), pp. 658–661.

Schubert, C., Waletzko, O., Weiss, C., Voelzke, D. & Toperim, S. 2017. Intercenter validation of a knowledge-based model for automated planning of volumetric modulated arc therapy for prostate cancer. The experience of the German RapidPlan Consortium. PLOS ONE. 12(5).

Shepart, D. 2007. IMRT Optimization Algorithms. Swedish Cancer Institute. Seattle WA. The American Association of Physicist in Medicine: Meetings. Presentation. [Accessed: 1.5.2018] Available: https://www.aapm.org/meetings/amos2/pdf/49-14369-92189-877.pdf

Sievinen, J., Ulmer, W. & Kaissl, W. AAA Photon Dose Calculation Model in Eclipse™. Varian Medical Systems. Technical Documentation. [Accessed: 10.4.2018]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.468.1557

Snyder, K., Kim, J., Reding, A., Fraser, C., Gordon, J., Ajlouni, M., Movsas, B. & Chetty, I. 2016. Development and evaluation of a clinical model for lung cancer patients using stereotactic body radiotherapy (SBRT) within a knowledge-based algorithm for treatment. Journal of Applied Clinical Medical Physics. 17. pp. 263-275.

Stanley, J., Breitman, K., Dunscombe, P., Spencer, D. & Lau, H. 2011. Evaluation of stereotactic radiosurgery conformity indices for 170 target volumes in patients with brain metastases. Journal of applied clinical medical physics / American College of Medical Physics. 12(3449)

Boyd, S. & Vandenberghe, L. 2004 Convex Optimization. Cambridge University Press, New York, NY, USA.

Søvik, Å., Ovrum, J., Olsen, D. & Malinen, E. 2008. On the parameter describing the generalised equivalent uniform dose (gEUD) for tumours. Physica medica : PM : an international journal devoted to the applications of physics to medicine and biology : official journal of the Italian Association of Biomedical Physics (AIFB). 23, pp. 100-106.

Tehhunen, M. 2007. Sädehoidon fysiikka ja tekniikka. Helsinki.

Teoh, M., Clark, C., Wood, K., Whitaker, S. & Nisbet, A. 2011. Volumetric modulated arc therapy: a review of current literature and clinical use in practice. The British Journal of Radiology 84(1007), pp. 967–996.

Ting, J., Wu, X., Fiedler, J., Yang, C., Watzich, M. & Markoe, A. 1997. Dose-Volume Histograms for bladder and rectum. International Journal of Radiation Oncology • Biology • Physics 38(5), pp. 1105-1111.

Tol, J., Dahele, M., Peltola, J., Nord, J., Slotman, B. & Verbakel, W. 2015. Automatic interactive optimization for volumetric modulated arc therapy planning. Radiation Oncology 10(75).

Tol, J., Dahele, M., Slotman, B. & Verbakel W. 2015. Increasing the number of arcs improves head and neck volumetric modulated arc therapy plans. Acta Oncologica 54(2), pp. 283-287.

Wackerly, D., Mendenhall, W. & Scheaffer, R. 2008. Mathematical Statistics with Applications. Belmont, CA, USA: Thomson Higher Education. ISBN 0-495-38508-5. 7th edition.

Wang, J. Hu, W., Yang, Z., Chen, X., Wu, Z., Yu, X., Guo, X., Lu, S., Li K. & Gongyi Y. 2017. Is it possible for knowledge-based plan-ning to improve intensity modulated radiation therapy plan quality for planners with differ-ent planning experiences in left-sided breast cancer patients? Radiation Oncology 12.

Zinchenko, Y., Craig, T., Keller, H., Terlaky, T. & Sharpe, M. 2008. Controlling the dose distribution with gEUD-type constraints within the convex radiotherapy optimization framework, Physics in Medicine & Biology 53(12), p. 3231.

Yoon, M., Park, S., Shin, D., Lee, S., Pyo, H., Kim, D. & Cho, K. 2007. A new homogeneity index based on statistical analysis of the dose–volume histogram. Journal of Applied Clinical Medical Physics 8, pp. 9-17.

Zacarias, A. & Mills, M. 2009. Algorithm for correcting optimization convergence errors in Eclipse. Journal of Applied Clinical Medical Physics 10, pp. 281-289.