Automated planning for image-guided radiotherapy

AIRPLAN*: Automated Image-guided Radiotherapy PLANning*

by

Elisabetta Cagni

School of Engineering
CARDIFF UNIVERSITY

A thesis submitted to fulfil the requirements for the degree of DOCTOR OF PHILOSOPHY in the School of Engineering.

JANUARY 2022

Family means nobody gets left behind, or forgotten.
Past, present and future without any separation.

Dedicated to my father Cagni Pietro.
Dec. 1939 – Oct. 2020

Dedicated to my future daughter.
2022 – ...

THESIS ABSTRACT

Advanced radiotherapy delivery approaches have substantially increased opportunities for sparing organs at risk with proven clinical impact. Ideally, for each individual patient, the treatment plan maximally exploits the full potential of the applied delivery technique. Currently, most treatment plans are generated with interactive trial-and-error planning (‘manual planning’). It is well-known that plan quality in manual planning may be sub-optimal, e.g. depending on experience and ambition of the planner, and on allotted planning time. In recent years, several systems for automated plan generation have been developed, often resulting in enhanced plan quality compared to manual planning. Both in manual and automated planning, human evaluation and judgement of treatment plans is crucial. During plan generation, planners usually develop a range of plans, but generally only one or two competing plans are discussed with the radiation oncologist (RO). A necessary assumption for this process to work well, is that (unknown) disparity between planners and ROs on characteristics of good/optimal plans is absent or minor. Radiotherapy is gradually evolving towards real-time adaptive radiotherapy (ART). ART has the clinical rationale of reducing normal tissue toxicity and improving tumour control through plan adaptation. In this thesis the research in ART was focused on automated methods to standardize ART in predicting the eventual need for re-planning and to assess the goodness of the process. In this thesis the differences between users in perceived quality of plans has been quantified and analysed. Inter-observer differences in plan quality scores were substantial and may result in inconsistencies in generated treatment plans. A method for ART verification, with the ability to quantify registration spatial errors and assess their dose impact at the voxel level, is presented. A systematic workflow to identify effective OAR sparing in re-planning using knowledge-based methods, has been established as a step toward an on-line ART process.

ACKNOWLEDGEMENTS

Thanks to my family,
who are present, who were present and who will be present,
all belong to the same core.

Thanks to Cardiff University,
for this beautiful opportunity, to attend the PhD as in work and in distance student.
Thanks to all the staff who help me in these years..

Thanks to Emiliano Spezi, my supervisor,
for his support in making this initial dream, a real experience.

Thanks to Geraint Lewis,
for his always elegant support and help, for his patience with my English, for his effective advices.

Thanks to Ben Heijmen,
for all things I have learned from him, about good research and style.

Thanks to my colleagues at the AUSL-IRCCS hospital of Reggio Emilia,
Thanks to all the collegues of CIDA group, of the School of Engineering
for their support and collaboration.

Thanks to my friends,
Who've been there for a while, who're still here, who're gone.

Ode to All of You.

PUBLICATIONS AND OUTPUT

Key publications

- Cagni E, Botti A, Orlandi M, Galaverni M, Iotti C, Iori M, Lewis G, Spezi E. Evaluating the Quality of Patient-Specific Deformable Image Registration in Adaptive Radiotherapy Using a Digitally Enhanced Head and Neck Phantom. Applied Sciences. 2022; 12(19):9493.
  https://doi.org/10.3390/app12199493.

- Cagni E, Botti A, Rossi L, Iotti C, Iori M, Cozzi S, Galaverni M, Rosca A, Sghedoni R, Timon G, Spezi E, Heijmen B. Variations in Head and Neck Treatment Plan Quality Assessment Among Radiation Oncologists and Medical Physicists in a Single Radiotherapy Department. Front Oncol. 2021 Oct 12;11:706034.
  doi: 10.3389/ fonc.2021.706034.PMID: 34712606; PMCID: PMC8545894.

- Cagni E, Botti A, Chendi A, Iori M, Spezi E. Use of knowledge based DVH predictions to enhance automated re-planning strategies in head and neck adaptive radiotherapy. Phys Med Biol. 2021 Jun 23;66(13).
  doi: 10.1088/1361-6560/ac08b0.PMID: 34098549.

Additional co-authored

- Clemente S, Falco MD, Cagni E, Talamonti C, Boccia M, Gino E, Lorenzini E, Rosica F, Russo S, Alparone A, Zefiro D, Fiandra C. The influence of small field output factors simulated uncertainties on the calculated dose in VMAT plans for brain metastases: a multicentre study. Br J Radiol. 2021 Mar 1;94(1119):20201354.
  doi: 10.1259/bjr.20201354.PMID: 33481637.

- Chendi A, Botti A, Orlandi M, Sghedoni R, Iori M, Cagni E. EPID-based 3D dosimetry for pre-treatment FFF VMAT stereotactic body radiotherapy plan verification using dosimetry CheckTM. Phys Med. 2021 Jan;81:227-236. doi: 10.1016/j.ejmp.2020.12.014. Epub 2021 Jan 20. PMID: 33485140.

- Loi G, Fusella M, Vecchi C, Menna S, Rosica F, Gino E, Maffei N, Menghi E, Savini A, Roggio A, Radici L, Cagni E, Lucio F, Strigari L, Strolin S, Garibaldi C, Romanò C, Piovesan M, Franco P, Fiandra C. Computed Tomography to Cone Beam Computed Tomography Deformable Image Registration for Contour Propagation Using Head and Neck, Patient-Based Computational Phantoms: A Multicenter Study. Pract Radiat Oncol. 2020 Mar-Apr;10(2):125-132.
  doi: 10.1016/j.prro.2019.11.011. PMID: 31786233.

- Talamonti C, Russo S, Pimpinella M, Falco MD, Cagni E, Pallotta S, Stasi M, Mancosu P. Community approach for reducing small field measurement errors: Experience over 24 centres. Radiother Oncol. 2019 Mar;132:218-222.
  doi: 10.1016/j.radonc.2018.10.012. PMID: 30385173.

- Isolan L, De Pietri M, Iori M, Botti A, Cagni E, Sumini M. Analysis of the bias induced by voxel and unstructured mesh Monte Carlo models for the MCNP6 code in orthovoltage applications. Radiation Effects and Defects in Solids. 2019 Mar 29.

Conference presentations/posters

- ESTRO (European SocieTy of Radiotherapy Oncology) newsletter physics corner autumn 2021.
  www.estro.org/About/Newsroom/Newsletter/Physics/Use-of-knowledge based-DVH-predictions-to-enhance

- ESTRO oral contribution OC-0469 Tudda, A., Castriconi, R., Cagni, E., Benecchi, G., Dusi, F., Esposito, P., Guernieri, M., Ianiro, A., Landoni, V., Mazzilli, A., Moretti, E. Placidi, L. , Trojani, V., Scaggion, A. and Fiorino, C. (2021) Inter-institute variability of kb-models for whole breast irradiation with tangential field Radiotherapy and Oncology, Volume 161, S355 - S356.

- ESTRO oral contribution OC-0105 Cagni, E., Rossi, L., Botti, A., Iori, M., Sghedoni, R., Iotti, C. , Rosca, A., Timon, G., Cozzi, S., Galaverni, M., Orlandi, M. ,Spezi, E. and Heijmen, B. J. M. (2020) Inter-observer variability in quality scores of Pareto optimal plans Radiotherapy and Oncology, Volume 152, S51.

- ESTRO electronic poster contribution EP-2010 Cagni, E., Botti, A., Orlandi, M., Galaverni, M., Sghedoni, R., Iotti, C. , Spezi, E. and Iori, M. (2019) A QA method for evaluation of deformable image registration in head and neck adaptive radiotherapy Radiotherapy and Oncology, Volume 133, S1101.

- ESTRO physical poster contribution PO-0996 Cagni, E., Botti, A., Orlandi, M., Sghedoni, R., Spezi, E. and Iori, M. (2019) A knowledge-based tool to estimate the gain of re-planning strategy for Head and Neck (HN) ART Radiotherapy and Oncology, Volume 133, S548–S549.

- ESTRO electronic poster contribution EP-2114 Botti, A., Cagni, E., Orlandi, M., Sghedoni, R., Lambertini, D., Barani, A., Bertolini, V. and Iori, M. (2019) Predicting inaccuracy of overmodulated RapidArc plans using Machine Learning model Radiotherapy and Oncology, Volume 133, S1170—S1171.

CONTRIBUTIONS

This thesis is in my own words. The research has been developed in Azienda USL-IRCCS of Reggio Emilia (Reggio Emilia, Italy), the hospital where I am working as medical physicist and in collaboration with Cardiff University. My PhD has been attended as undertaken as an in-work, distance learning full PhD course.
Chapters within this thesis include published material (listed in the previous section) in which I was a lead or co-author. In particular, key aspects of this work fed into the international collaborative effort called AIRPLAN (Automation In radiotherapy PLANning and evaluation) project, which is mainly discussed in chapter 6 and chapter 7, applied to head and neck cancer radiotherapy treatments. I am the principal investigator of this project, that involved Erasmus Cancer Institute (Rotterdam, The Netherlands) and Azienda Unità Sanitaria Locale - Istituto di Ricerca a Carattere Scientifico (IRCCS) (AUSL-IRCCS) of Reggio Emilia (Reggio Emilia, Italy). The main results of this study were published in the Frontiers in Oncology Journal in 2021 (see previous section). Although the entire project is based on team work (radiation oncologists and medical physicists working at AUSL-IRCCS of Reggio Emilia), I mainly contributed to the design of the study in collaboration with Erasmus Institute, I produced all the automated plans used in both chapters 6 and 7 and I performed the main part of the data analysis. I also supervised the development of the study in its clinical aspects (plans evaluation procedure).
A second multi-centre project, using the methods developed in the AIRPLAN project and applied to breast cancer (AIRPLAN Breast (AIRPLAN B)), is on-going. The AIRPLAN B project is described in chapter 8 and chapter 11. I am the principal investigator. This project involves Erasmus Cancer Institute, Azienda USL-IRCCS of Reggio Emilia and another 2 Italian hospitals: Careggi Hospital and USL of Piacenza. AIRPLAN B was approved by local ethical committee in August 2021 (see chapter 11 for more details). I mainly contributed to the design the study, I produced all the automated plans presented in chapter 8 and I performed the main part data analysis that it is an on-going work.

The other part of the thesis is related to adaptive radiotherapy planning (Part IV). Two studies are presented in chapter 9 and chapter 10, respectively. For both I performed the design of the study and the data analysis. I performed all the registrations using reference and third party software and the data analysis presented in chapter 9. The manuscript related to chapter 9 is ready to be submitted in a peer review journal. I am contributing as first and corresponding author. I mainly performed and evaluated knowledge-based models using in chapter 10, with the help of the AUSL-IRCCS of Reggio Emilia staff, used for plans generation and for DVH predictions. The work presented in chapter 10 was published in 2021, and I contributed as the first and corresponding author.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

GLOSSARY AND LIST OF ABBREVIATIONS

> “When I use a word,” Humpty Dumpty said in rather a scornful tone, “it means just what I choose it to mean — neither more nor less.”
> “The question is,” said Alice, “whether you can make words mean so many different things.”
> “The question is,” said Humpty Dumpty, “which is to be master – that’s all.”
> *Lewis Carroll-Alice in Wonderland*

ART	Adaptive Radiotherapy.
AUC	Area Under The Curve ROC.
ASR	Absolute Sum of Residuals between DVHs. Used to estimate the gain of replanning, see chapter 10.
CBCT	Cone Beam Computed Tomography.
CI	Conformity Index.
CLIN	Manual clinical plans used for patient treatment, see chapters 3, 6.
CT	Computer Tomography.
DIR	Deformable Image Registration.
DVF	Deformable Vector Field.
DVH	Dose Volume Histogram.
DVH$_{d}$	Delivered DVH, see chapter 10.
DVH$_{fKBP}$	Final KBP DVH after optimisation, see chapter 10.
DVH$_{pKBP}$	Predicted KBP DVH by KBP tool, see chapter 10.

EBRT	External Beam Radiotherapy.
FDVH	Feasibility DVH of PlanIQ tool, see chapter 5, 7 and 8.
FN	False Negative, see ROC analysis in chapter 10.
FP	False Positive, see ROC analysis in chapter 10.
GED	Geometry based Expected Dose, see RapidPlan algorithm in chapter 4.
GEDVH	GED cumuluative Volume Histogram, see RapidPlan algorithm in chapter 4.
gUIDE	Generalized Uniform Ideal Dose using Exponential function, see chapter 5.
HN	Head and neck cancer.
KBP	Knowledge Based Planning.
IGRT	Image Guided Radiation Therapy.
IMRT	Intensity-Modulated RadioTherapy.
LINAC	LINear ACcelerator.
MA	Model Analytic platform, see RapidPlan model validation.
MCO	MultiCriteria Optimisation.
MCOa	Automated plans (based on MCO) using WLa. See chapters 3, 6.
MCOx	Automated plans (based on MCO) using WLx. See chapters 3, 6.
ML	Machine Learning.
MLC	Multi-Leaf Collimator.
MP	Medical Physicist.
MRI	Magnetic Resonance Imaging.
NTCP	Normal Tissue Complication Probability.

OAR	Organ At Risk.
OOB	Out-Of-Bag, see ML.
OP	Predicted Operating Point, see ROC curve.
PCA	Principal Component Analysis.
PDD	Percentage Depth Dose curve.
PET	Positron Emission Tomography.
PTV	Planning Target Volume.
PQM	Plan Quality Metric, see PlanIQ tool in chapter 8.
$\text{refDVF}_{\text{pt}}$	Reference DVF obtained using the third-party algorithm from patient image, see chapter 9.
RO	Radiation Oncologist.
ROC	Receiver Operating Characteristics curve or discriminant analysis, see chapter 10.
RP	RapidPlan tool, see chapter 4.
SD	Standard Deviation.
SIB	Simultaneous Integrated Boost.
SR	Sum of Residual between DVHs. To estimate DVHs difference, see chapter 10.
TCP	Tumor Control Probability.
$\text{testDVF}_{\text{ph}}$	Test DVF obtained using the clinical DIR algorithm to be evaluated, see chapter 9.
TF	Tangential Fields.
TN	True negative, see ROC analysis.
TP	True positive, see ROC analysis.

TPS	Treatment Planning System.
VMAT	Volumetric Modulated Arc radioTherapy.
WBI	Whole Breast Irradiation.
WL	Wish-list for MCO optimisation, see chapter 3.
WLa	Wish-list optimal based on clinical goals.
WLx	Wish-list suboptimal derived from WLa changing the endpoints priority order randomly, see chapter 6.

Part I

OVERVIEW

1

OUTLINE

> "Begin at the beginning,". the King said gravely, "and go on till you come to the end: then stop."
> *Lewis Carroll-Alice in Wonderland*

1.1 AIM OF THE WORK

The main object of this thesis is the automation of the radiation oncology treatment planning process. In more details, this thesis investigates the implementation and use of automated tools, such as automation in plan generation, plan evaluation and plan adaptation into radiation therapy clinical practice and their impact on treatment quality. Two aspects are focused on, the first being modulated radiation therapy treatment plans (a term including both Intensity Modulated Radiation Therapy, IMRT, and Volumetric Modulated Arc Therapy, VMAT) and the second being to adaptive radiotherapy, including image registration and plan modification. The study was centred mainly on head and neck cancer treatment, however ongoing research using the methods developed during the thesis work, applied to breast cancer treatment is presented (chapter 8 and chapter 11).

1.2 THESIS STRUCTURE

Figure 1.1 gives a schematic thesis map to show the flow and relationships between chapters. Chapter 2 provides background to the work presented. An introduction to key concepts of radiotherapy and a brief general literature overview are provided, with more critical discussion of the specific literature in subsequent chapters. Part II, composed of chapter 3, 4 and 5 describes the tools used in the research projects of this thesis. Chapter 3 describes multicriteria optimisation

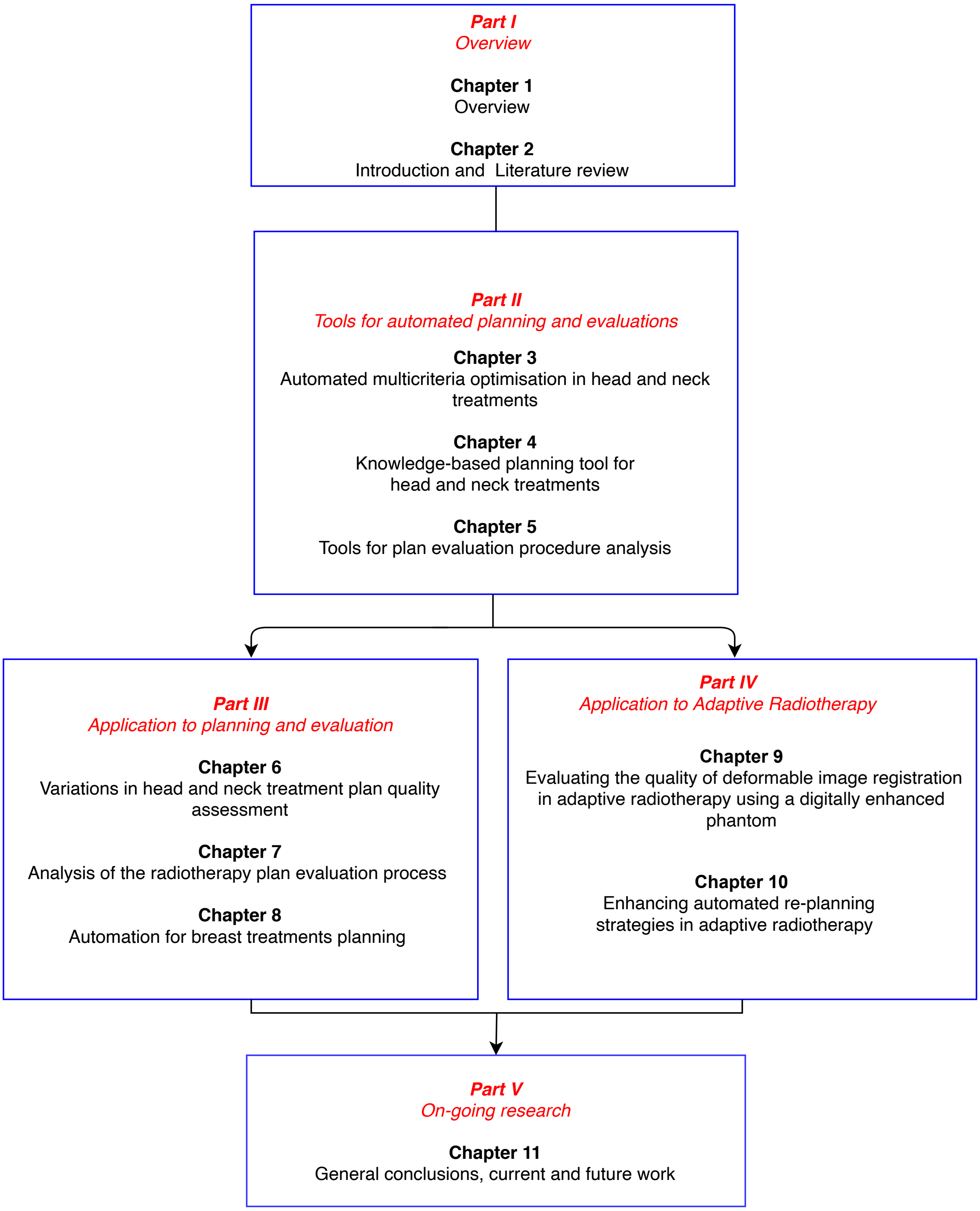

Part I
Overview
Chapter 1
Overview
Chapter 2
Introduction and Literature review
Part II
Tools for automated planning and evaluations
Chapter 3
Automated multicriteria optimisation in head and neck treatments
Chapter 4
Knowledge-based planning tool for head and neck treatments
Chapter 5
Tools for plan evaluation procedure analysis
Part III
Application to planning and evaluation
Chapter 6
Variations in head and neck treatment plan quality assessment
Chapter 7
Analysis of the radiotherapy plan evaluation process
Chapter 8
Automation for breast treatments planning
Part IV
Application to Adaptive Radiotherapy
Chapter 9
Evaluating the quality of deformable image registration in adaptive radiotherapy using a digitally enhanced phantom
Chapter 10
Enhancing automated re-planning strategies in adaptive radiotherapy
Part V
On-going research
Chapter 11
General conclusions, current and future work

Figure 1.1: Thesis map

(MCO) tool used for automated planning used for the research presented in Part III -chapter 6, chapter 7 and chapter 8. Chapter 4 reports a knowledge based radiotherapy planning tool, Rapidplan tool, used in a research study described in Part IV- chapter 10. Chapter 5 describes the tools used in the plan evaluation assessment analysis presented Part III -chapter 7. These included a commercial tool for plan evaluation (planIQ) and an in-house developed ideal dose concept, used as baseline for real plan evaluation (gUIDE).

The thesis is then split into two parallel parts: Part III – Application to planning and evaluation (chapters 6, 7 and 8) and Part IV – Application to adaptive radiotherapy (chapters 9 and 10).

Chapter 6 reports on a multicentre study performed at AUSL-IRCCS of Reggio Emilia. The aim of this chapter is to gain an understanding of the level of agreement between medical physicists, who perform treatment plans, and radiation oncologists on plan quality criteria. Good agreement is essential for consistent planning. In this chapter the use and application of automated treatment planning, and its impact on plan quality is introduced. Differences between radiation oncologists (ROs) and planning medical physicists (MPs) in perceived quality of head and neck plans were assessed using both automated and manual planning. The results were used to consider beam modelling investigations in the following chapter. This chapter was published in the journal Frontiers in Oncology in September 2021.

Chapter 7 analyses the data of the previous chapter and investigates the reasons for large variation in scoring of different evaluators using both statistical and machine learning tools. The specific aim of this chapter is to determine which parameters in the plans have the greatest impact on the quality judgement and to explore differences to improve consistency between users in evaluation.

Chapter 8 This chapter uses the methods developed in chapter 6 for head and neck automated treatment planning applied to breast cancer. This study is part of a multi-institute plan quality assessment evaluation for the breast site (described in the final chapter 11).

Chapter 9 presents a registration-based method for deformable image registration quality assurance for head and neck patients using digitally post-processed anthropomorphic phantom image sets. The findings of this work underline that spatial and dose errors are a function of the magnitude of the deformation and of the gradient of the dose distribution. This emphasizes the importance of performing patient specific image registration verification. This work contributes to the standardization and automation of verification methods for deformable image registration accuracy. The aim of this work is to automate the step of verification of image registration in adaptive head and neck planning. This chapter is intended to be submitted in a peer-reviewed journal.

Chapter 10 presents a novel application of a commercial knowledge-based planning (KBP) tool for radiotherapy planning, using potential organ at risk sparing estimation in the replanning strategy for head and neck adaptive radiotherapy. The work demonstrates the utility of a KBP model trained with Pareto optimised plans to estimate the potential organ sparing gain in a replanning strategy. A systematic workflow for identifying effective organ sparing in replanning strategies based on KBP prediction is presented. This method could provide an important KBP application for adaptive radiotherapy and estimation of organ sparing. This chapter has been published in the Physics in Medicine and Biology in June 2021. Finally, Part V, contains the general conclusions of the thesis, along with current and future work.

Chapter 11 discusses the body of work as a whole, highlighting on-going and potential future research resulting from this work.

2 INTRODUCTION AND LITERATURE REVIEW

> When radium was discovered, no one knew that it would prove useful in hospitals. The work was one of pure science. And this is a proof that scientific work must not be considered from the point of view of the direct usefulness of it.
> *Marie Curie*

Globally, cancer is diagnosed in around 15 million patients each year, and radiotherapy is used in $\approx 50\%$ of cases, sometimes in combination with chemotherapy or surgery ([1, 2]). Radiation therapy (radiotherapy) is one of the main treatment modalities, together with surgery and chemotherapy. It can be used for curative tumour eradication, tumour size reduction, tumour bed cleansing or palliative purposes. Radiotherapy uses ionizing radiation to inflict damage on tumour cells. The main delivery technique involves the use of external X-ray photon beams that interact with tissue, resulting in the deposition of radiation dose within the patient. This is called external beam radiation therapy (EBRT), which was the technique used in all the studies described in this thesis.

2.1 EXTERNAL BEAM RADIATION THERAPY WORKFLOW

The delivery of a radiotherapy treatment is a complex process which involves several parties, including physicians, physicists and technicians. Figure 2.1 shows a scheme of radiotherapy workflow.

When a patient is diagnosed with cancer and radiation therapy is a selected treatment, a computerised tomography (CT) scan is acquired with a specific immobilisation technique to obtain a precise localization of the anatomical region to

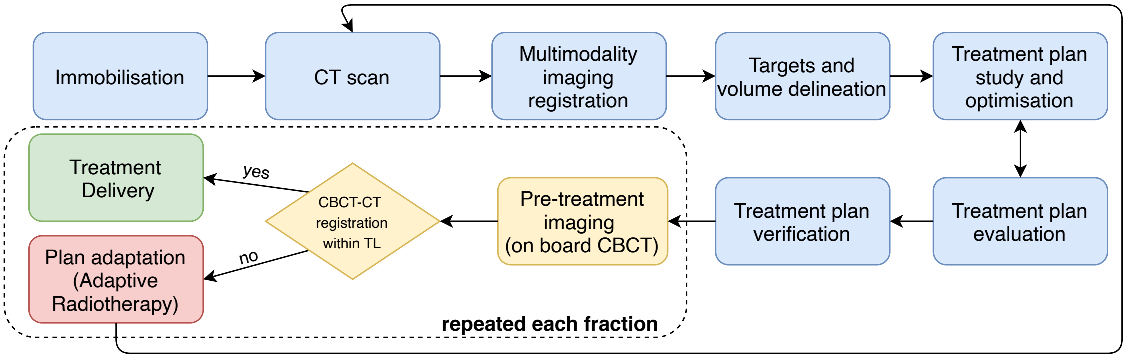

Immobilisation
CT scan
Multimodality imaging registration
Targets and volume delineation
Treatment plan study and optimisation
Treatment Delivery
yes
CBCT-CT registration within TL
Pre-treatment imaging (on board CBCT)
Treatment plan verification
Treatment plan evaluation
no
Plan adaptation (Adaptive Radiotherapy)
repeated each fraction

Figure 2.1: Radiation treatment planning workflow.

treat. On this acquisition, all regions of interest, i.e. the organs at risk (OARs) and the tumour (also referred to as the target) are delineated by the radiation oncologist. This is one of the most crucial phases of treatment planning, as it actually determines which area is going to be irradiated. During target delineation, there are precise protocols to be followed. If necessary, the CT scan is registered with other types of imaging such as magnetic resonance imaging (MRI) or positron emission tomography (PET). ICRU (International Commission on Radiation Units and Measurements) guidelines are widely accepted to prescribe the dose delivered during the treatment [3]. Specific margins are to be added to the contoured region to treat in order to account for uncertainties that could arise during the next phases of the treatment. After this, the medical physicist is tasked to design a treatment plan with the aim to deliver enough dose to the target, while keeping the dose to the healthy tissues below acceptable levels. To achieve this, modifications to a treatment plan depend on the radiation beam quality and its energy, collimation, intensity and interactions within the patient's local environment (such as position of targets, density along beam path, scattering conditions,...). After the medical physicist has devised an acceptable plan, the dose distribution associated with the plan is analysed with a radiotherapist. This important phase of the process determines whether a plan is suitable to be delivered; the steps of treatment plan optimisation and plan approval can be done iteratively until a final solution is found. Then, after the necessary dosimetric verifications following specific protocols in each institution, the plan is ready to be delivered. Generally, setup verification before delivery using, on-board, imaging cone beam CT (CBCT) is performed and verified. The total plan dose is delivered by a series of (usually) daily ’fractions’. During the course of the treatment, the patient can undergo internal and external anatomical modifications. In this case, the plan adaptation task is required.

2.2 LINEAR ACCELERATOR

The ionising radiation beams used in EBRT are generated by a linear accelerator (LINAC) and directed toward the patient. LINACs are machines that consist of a number of discrete components Figure 2.2 functioning together to accelerate electrons to a high energy using radiofrequency (RF) waves before the electrons hit a target to produce X-rays. After this the X-ray profile is flattened, shaped (collimated) and measured before clinical use. LINACs are now also capable of producing X-ray beams of different energy (multi-energy units) and/or producing both X-rays and electrons (multi-modal units) [4]. The LINAC has undergone many innovations in technology during the years since its initial medical application [4]. Among these, the most important ones are:

- the standard application of multi-leaf collimators (MLC) to define highly conformal fields;
- intensity modulated radiation therapy (IMRT) allowing MLC to generate complex shapes of dose distributions through not uniform beam fluences, inversely optimized for each patient, introduction of kilovoltage imaging systems built in standard clinical accelerators, enabling high quality image guided radiation therapy (IGRT). They are usually coupled with image registration with CBCT imaging acquired at the time of treatment for confident and accurate patient positioning;
- volumetric modulated arc therapy (VMAT) which extended the degrees of modulation in IMRT to include the gantry angle entry, dose rate and gantry speed between several 3D points in space (control points).

The described developments permitted the implementation of precise and personalized radiotherapy which improved outcomes and quality of life after treatment by conforming the dose distributions to the desired target and sparing the

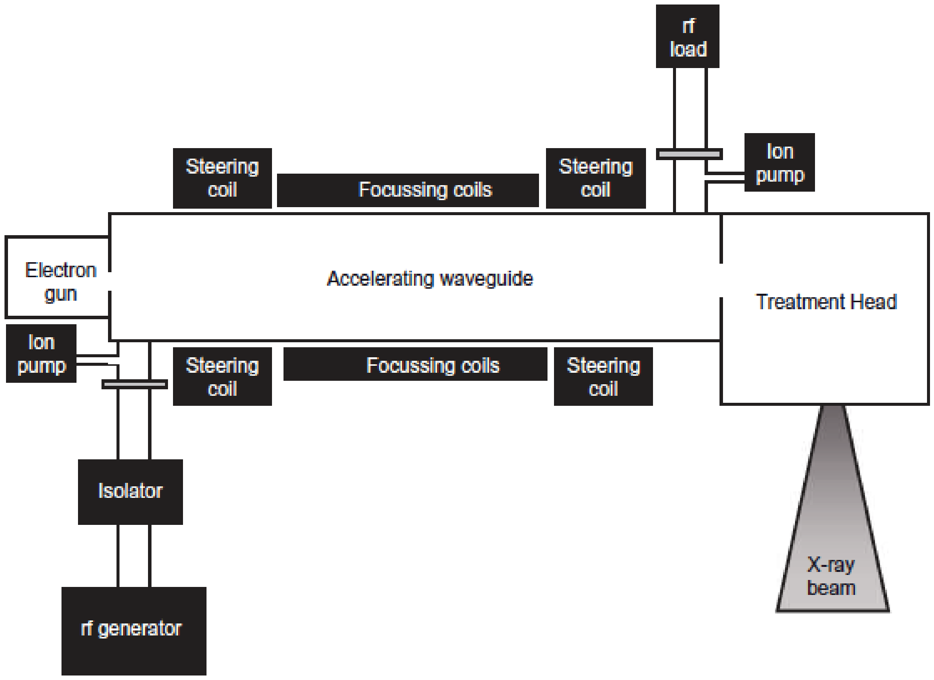


Figure 2.2: Major components of LINAC.

healthy organs around it. As such, modern RT makes use of many complex machines and techniques and, over the past decades, many planning techniques have been developed and implemented for routine clinical use. Given the potential risks associated with any RT method, it is essential that treatments are planned and delivered safely and correctly.

While the beam passes through the patient, it interacts and delivers dose to all tissues, not only the malignant ones. Healthy cells can therefore also be affected by the treatment. It is physically impossible to fully spare them while also delivering a dose to eradicate the tumour. Thus, an important goal of a treatment is to minimize the possible negative impact of the irradiation on the patient's quality of life by limiting dose delivery to healthy tissues.

In order to minimize dose to healthy tissues, multiple beams are targeted at the tumour, essentially creating a 'cross-fire' arrangement. As a result, the surrounding dose is relatively low, which contributes to reducing the damage to healthy tissues. An important challenge lis 1) to deliver the desired minimum dose to the tumour, and 2) maximally reduce the dose to the surrounding healthy tissues. Since some healthy tissues are more radiation sensitive than others, different trade-offs are generally required, e.g. it may be desirable to reduce the radiation on specific tissues. This type of knowledge on the healthy tissues surrounding the tumour has to be taken into consideration when beam geometry and beam contributions (intensity profiles) are defined.

2.3 RADIOTHERPY TREATMENT PLAN EVALUATION

It is important to evaluate the plan as a whole before its approval for clinical use. There are three principal tools to aid the evaluation of the plan quality and decision-making [5]:

- Numerical data
- Dose-volume histograms (DVH)
- Visual inspection of the dose distribution on patient's CT slices.

The numerical data either result directly from the criteria used for plan optimization or are based on evaluating the (final) dose distribution. Sometimes advanced models are available that link delivered dose to a probability of developing a certain complication. To evaluate the therapeutic ratio, defined as the relationship between the probability of tumour control and the likelihood of normal tissue damage, generally radio-biologically indices, such as the tumour control probability (TCP) and normal tissue complication probability (NTCP) are used [6]. TCP and NTCP are mathematical models, describing the probability of complete tumour eradication and any complication resulting from the radiotherapy treatment, respectively; for more details see [6]. In specific cases this metric could also be added to the list above as a tool for plan evaluation.

Some aspects of the 3D dose distribution can be summarized in a 2D plot by using Dose-Volume Histograms (DVH) (see Figure 2.3). Each DVH curve represents a structure, depicting the portion of the volume of that structure which receives that amount of dose or higher. The DVH is equal to the complement of the observed cumulative distribution function (CDF) if the dose distribution is viewed as an empirical probability function. For OARs, the curves are ideally close to the origin, where the part of the volume that receives a high dose is minimal. For PTVs,

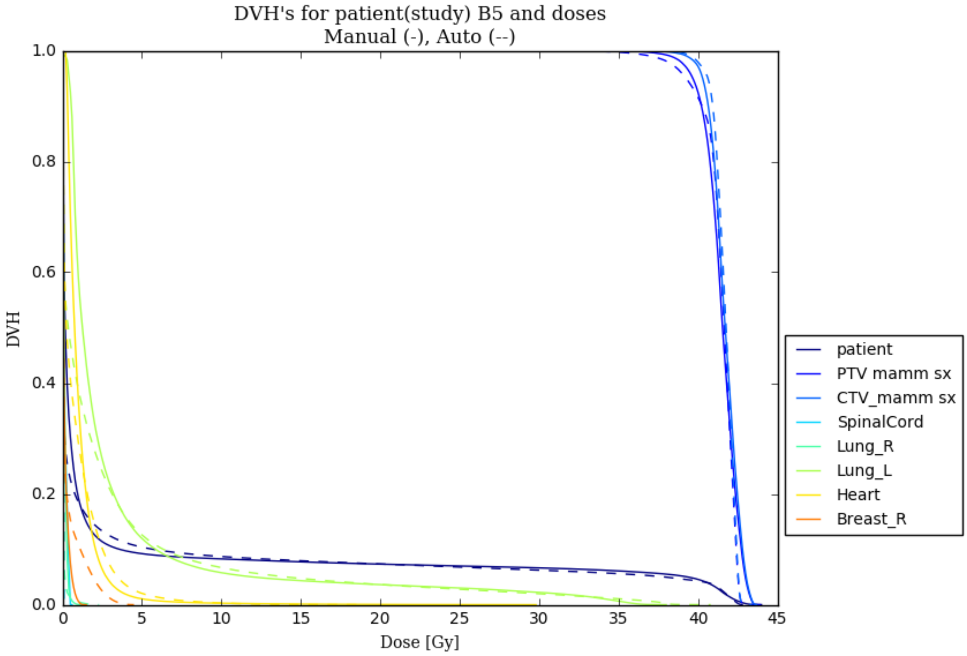


Figure 2.3: An example of DVH comparisons for a left breast cancer patient belong to patient set of PartIII study (see chapter 8). 2 plans are compared in term of DVH: a manual plan (solid line) and a Pareto optimal plan generated using automated tool (MCO).

on which often a minimum and a maximum dose is imposed, the curve is expected to be as far as possible towards at the right, with a steep slope downward at the end (see Figure 2.3).

A DVH is a convenient tool to compare two or more treatment plans for the same patient. In general, the closer the correspondence of the DVHs for two plans, the more likely it is that both plans are similar. The representation is not unique because spatial information is lost in the conversion. Two completely different dose distributions can in theory have identical DVHs, so a visual inspection of the spatial 3D dose distribution is always necessary to fully assess acceptability or favourably of a plan. Nevertheless, the DVH it generally gives a concise overview of a 3D dose. When visually assessing dose distributions there are several undesirable aspects to be considered, e.g. conformality and how the high dose “leaks” outside the PTV into the normal tissue, hot spots (isolated high dose points distant from the PTV), high dose streaks, and high doses in regions where it is not expected. Sufficient coverage of the PTV should also always be visually inspected.

2.4 RADIOTHERAPY TREATMENT PLAN OPTIMISATION

The goal of treatment plan optimisation, or treatment planning, is to find settings of the applied treatment unit that result in an optimal therapeutic balance for the patient. Treatment plans are generated with the aid of a dedicated software application, called a Treatment Planning System (TPS). Traditionally, most conventional planning is done in an interactive trial-and-error procedure (manual planning). Based on the initially selected beam geometry, the planner defines a mathematical optimization problem (i.e. cost functions, objectives, weights and/or additional parameters) (see Figure 2.4 and Figure 2.5) that is subsequently used by the computer to generate beam intensity profiles. If the result is a not high quality plan, the planner could modify the optimization problem or change beam geometry for another run of optimization. This interactive and iterative process stops if the plan is considered adequate, or if there are no more ideas or time,or if significant improvements with further optimization are considered unlikely.

2.5 AUTOMATION IN TREATMENT PLANNING

Efforts to streamline and standardise the treatment planning process are ongoing. In the last few years, there has been significant progress into research and development of automated inverse treatment planning approaches, with most commercial manufacturers now offering some form of solution. There is a rapidly growing body of research published in the literature. These algorithms could significantly improve the efficiency, consistency, and quality of treatment planning, leading potentially to improved patient access and improved patient outcome through maintaining and improving high-quality radiotherapy. In 2014, the National Health Service in England and Cancer Research UK published a 10 year 'Vision for Radiotherapy' in the UK to allow patients to receive advanced and innovative radiotherapy that is cost-effective, and one suggestion to facilitate this is through the implementation of software that automate parts of the planning process [7].

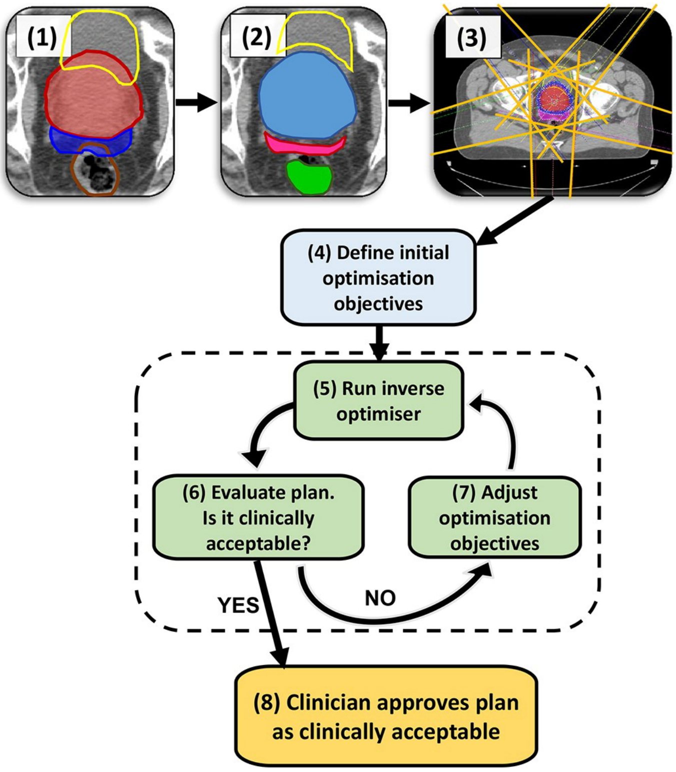
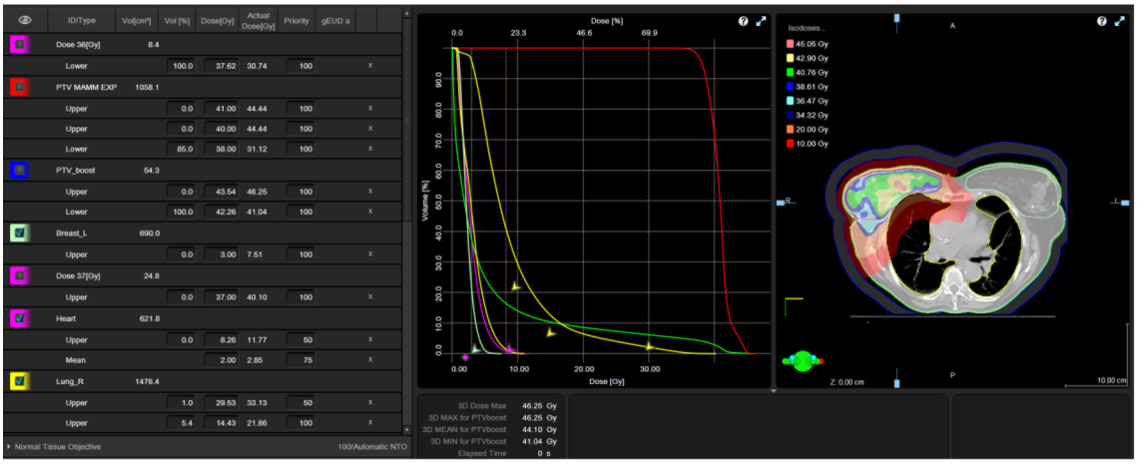

(1)
(2)
(3)
(4) Define initial optimisation objectives
(5) Run inverse optimiser
(6) Evaluate plan. Is it clinically acceptable?
(7) Adjust optimisation objectives
YES
NO
(8) Clinician approves plan as clinically acceptable

Figure 2.4: A typical manual IMRT treatment planning pathway. The example shown is for a prostate + seminal vesicle case. The steps are as follows: (1) CT scan with PTVs and OARs delineated; (2) create a range of "helper" (ROI) to aid the optimiser. Step (3) set-up beam geometry. (4) Define the initial optimisation objectives either from scratch or from a class solution. (5) Run the inverse optimiser until it converges to a solution, calculate dose distribution. (6) Evaluate the resulting plan, if it is clinically acceptable proceed to Step 8, otherwise go to Step 7 to adjust the optimisation objectives. The part shaded in green (steps 5, 6, 7) is the iterative process of optimisation required by the planner to arrive at a clinically acceptable treatment plan to be approved by the clinician in Step 8 [7].

2.5.1 *Multi-criteria treatment planning*

Radiotherapy planning is a multi-criteria problem, balancing the dose between the tumour and different healthy organs with the aim of achieving the highest possible quality of life for the patient. This involves balancing up to 30 highly correlated criteria, and because each patient is anatomically unique, requires an individualized solution for each patient.

Therefore, in multi-criterial optimization, multiple objectives are in competition with each other, so that reaching one objective can lead to not fulfilling the others. A solution is called Pareto-optimal, described in Figure 2.6, if none of the objective functions can be improved without detriment to other objective values. Without



	ID/Type	Vol[cm³]	Vol [%]	Dose[Gy]	Actual Dose[Gy]	Priority	gEUD a	
☐	Dose 36[Gy]	8.4						
	Lower		100.0	37.62	30.74	100		x
☐	PTV MAMM EXP	1058.1						
	Upper		0.0	41.00	44.44	100		x
	Upper		0.0	40.00	44.44	100		x
	Lower		85.0	38.00	31.12	100		x
☐	PTV_boost	54.3						
	Upper		0.0	43.54	46.25	100		x
	Lower		100.0	42.26	41.04	100		x
☑	Breast_L	690.0						
	Upper		0.0	3.00	7.51	100		x
☐	Dose 37[Gy]	24.8						
	Upper		0.0	37.00	40.10	100		x
☑	Heart	621.8						
	Upper		0.0	8.26	11.77	50		x
	Mean			2.00	2.85	75		x
☑	Lung_R	1476.4						
	Upper		1.0	29.53	33.13	50		x
	Upper		5.4	14.43	21.86	100		x

Normal Tissue Objective 100/Automatic NTO

3D Dose Max 46.25 Gy
3D MAX for PTVboost 46.25 Gy
3D MEAN for PTVboost 44.10 Gy
3D MIN for PTVboost 41.04 Gy
Elapsed Time 0 s

Figure 2.5: Example of TPS module (Eclipse) for inverse planning optimization. The DVH endpoints with their relative importance in the whole treatment planning are showed on the left and on the DVH which is produced during the iterative process.

additional preference information, all Pareto optimal solutions can be considered mathematically equally good. Thus, it is necessary to introduce user-defined preferences to differentiate between various solutions.

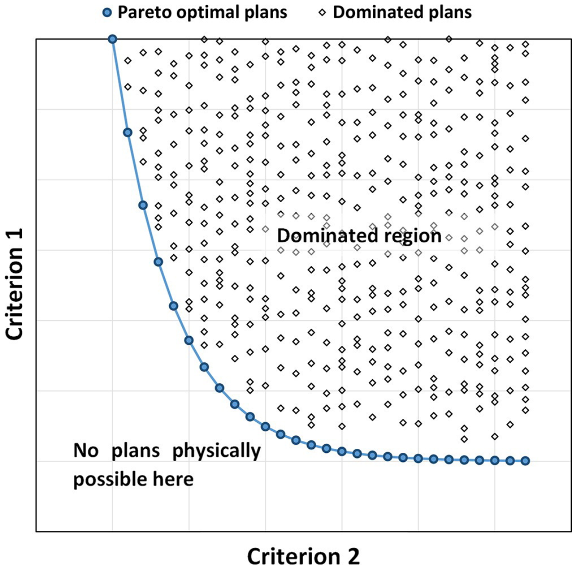


Figure 2.6: Schematic diagram of two competing criteria. The graph shows a large number of different feasible planning solutions, representing a variety of different permutations for criteria 1 and 2. The solid line represents the pareto front where improving one criterion inevitably leads to the worsening of the other and vice versa. Plans that lie on this front are the "pareto optimal solutions", shown as blue circles in the schematic. The plans shown as diamonds are referred to as "dominated" because there is always a solution on the pareto front where at least one criterion can be improved. Reference [7].

Usually, in a manual planning scenario, it is not even assured that one is going to reach the Pareto surface. In fact, this surface is to be considered as a lower bound in terms of possible achievable solutions (in our case, plan dose distributions). It is true that, with a sufficiently high number of iterations, it is possible to get sufficiently close to the surface; however, the navigation on its boundary is made manually and as in the previous section, it is subject to variability according to the different planners involved. In contrast, in an automatic planning optimization scenario, an algorithm to reach the Pareto surface is employed [8].

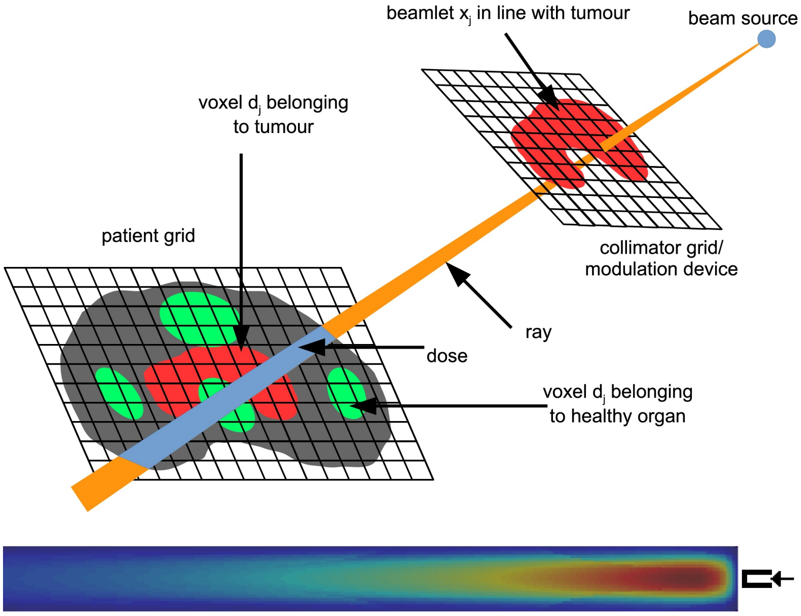

beamlet x_j in line with tumour
beam source
voxel d_j belonging to tumour
patient grid
collimator grid/ modulation device
ray
dose
voxel d_j belonging to healthy organ

Figure 2.7: Radiotherapy problem decomposition. Top panel: Ionizing radiation originates from the beam source point and falls onto a collimator which allows shaping of the beam and its discretisation into beamlets. The longer a beamlet is "open", the higher the intensity through that beamlet, and the higher the resulting dose in the patient. Dose, measured in Gray (Gy), is equivalent to the absorbed radiation, and a higher dose results in more cell damage. The patient is discretized in voxels. Delivering a series of different shapes allows intensity modulation. Bottom panel: physics of a photon-based pencil beam, where red indicates highest dose. The irradiation is from the right for a single beamlet opening with the width indicated by the size of the black square. Due to the particle scattering effect, the pencil beam dose is wider than the beamlet opening. As a consequence, tissue at the lateral sides of the pencil beam will also be damaged. Reference [5].

The numerical decomposition of the radiotherapy problem is based on 'beamlets' (see Figure 2.7). The radiation beam is discretized into beamlets, which are the fundamental decision-variables for the numerical optimization problem. The numerical value of the decision variable represented by a single beamlet is equal to the fluence, defined as the time-integrated flux of radiation, that passes through the grid element of that beamlet; fluence results in dose in the patient. The patient is discretized into voxels. The relation between the beamlets x (from the entire set of fluence maps) and the voxel doses to the patient is a linear relation:

$$d(x) = \mathbf{A}x \tag{2.1}$$

where **A** is called the pencil-beam matrix or dose influence matrix.

The MCO algorithm considers the treatment planning problem as a multi-objective optimization exercise. In such a problem, a vector of objective functions is optimized instead of a single objective function (manual optimisation), as described in the equation 2.2. Contrary to optimization of a single objective function, no single best objective function value exists, but a set of best-compromise points which constitute the Pareto surface of the problem.

$$\min_{x \in X} \overrightarrow{f(x)}, \overrightarrow{g(x)} \leqslant 0 \tag{2.2}$$

Where, x is the optimization parameters (fluence), X is the set of available optimization parameters, $\overrightarrow{f(x)}$ represents the vector of n objective functions f1...fn and $\overrightarrow{g(x)}$ a vector of r constraint functions g1...gr.

2.5.2 *Knowledge based treatment planning*

An approach to improving the speed, efficiency and reducing variability in treatment planning is using a 'knowledge-based planning' (KBP) optimization approach. KBP is defined as any approach which directly utilises prior knowledge and experience to either predict an achievable dose in a new patient of a similar population or to derive a better starting point for further trial-and-error optimisation by a planner. There are two distinct approaches to this: the atlas-based approach and the model-based approach. In the atlas-based method, the knowledge base is used to select the closest matching patient(s) to give a better starting point for the inverse optimisation than would be provided by conventional template-based approaches [7].

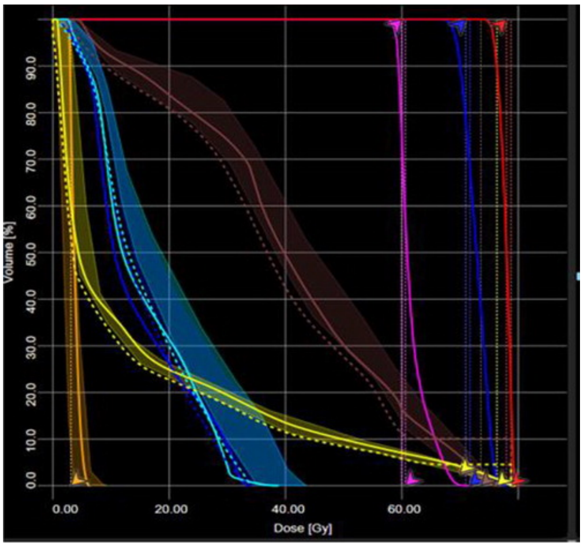


Figure 2.8: An example of DVH prediction KBP in a 3-dose level localised prostate cancer case. The shaded lines are the predicted range of achievable DVHs for the different OARs. The solid lines are the actual achieved DVH in the plan. This example is from Varian RapidPlan and the dashed lines and arrows are the optimisation objectives that have been generated by RapidPlan. Courtesy: Royal Surrey County Hospital NHS Foundation Trust, Guildford, UK. DVH, dose-volume histogram; KBP, knowledge-based planning; OAR, organ at risk. Reference [7].

An example of DVH prediction KBP in a 3-dose level localised prostate cancer case is shown in Figure 2.8. The shaded lines are the predicted range of achievable DVHs for the different OARs. The solid lines are the actual achieved DVH in the plan. This example is from Varian RapidPlan and the dashed lines and arrows are the optimisation objectives that have been generated by RapidPlan.

Dose-volume histogram (DVH)-guidance is one of the approaches of model-based KBP [3, 7, 9–15], [16–24]. In this approach, a large number of clinically accepted treatment plans and contours are used to characterise the relationships between anatomical and geometric features for a given anatomical site to build a predictive DVH model for that site. For any new patient treated in the same anatomical site, this knowledge can be used to predict the achievable DVH based on the features of similar contours and quality of treatment plan; see an example in Figure 2.8. A range of different implementations of DVH-guided KBP has been proposed and developed. Commercially, the DVH-guidance KBP approach is utilised by the Varian Eclipse Treatment Planning System as RapidPlan (Varian Medical Systems, Palo Alto, Cl). This module is described in chapter 4.

2.6 ADAPTIVE RADIATION THERAPY

Adaptive radiotherapy (ART) enables the treatment to be changed, or adapted, to respond to a signal that additional information is known about the patient or that the patient has changed from the original state at the time of planning [25]. In an adaptive workflow, the main steps are: creating a treatment plan, performing periodic imaging (e.g. cone beam CT (CBCT)), and deciding to create a new or modified treatment plan when deemed necessary by the clinical team. This may be performed without any of the sophisticated tools now available (such as deformable image registration (DIR), automated planning, dose accumulation or decision-making) with the new treatment plan generated using the same clinical criteria as the original plan. However, this process is typically *ad hoc* and does not allow the clinical team to gain knowledge about the delivered dose, the toxicity rates, and the benefit of adaptation. In current practice, RT treatment plans are designed individually for each patient as was described in Figure 2.1). However, several factors can lead to anatomic changes of both the target volumes and OARs leading to deviations between the actual anatomy of the patient and the one represented on the planning CT. These factors include daily setup variations, primary tumour or nodal volume regression or progression, alteration in muscle mass and/or fat distribution, fluid shift within the body and weight loss [26, 27]. If unnoticed or unattended, these changes might lead to discrepancies in dose delivery, with loss of tumour control, and/or to overdosage of the normal structures, potentially producing unexpected side effects [28, 29]. A possible solution to this problem is ART, which aims at correcting anatomical modifications by adapting the initial dose plan to the current patient status [30]. This process requires repeated imaging with sufficient quality for treatment planning, re-contouring and re-planning.

As radiotherapy and its associated information technology have developed, the sophistication of adaptive therapy has increased accordingly. Volumetric CBCT imaging and DIR process allow decisions on adaptation to be made based on dosimetric information rather than geometric information alone. The procedure consists of deformably registering the planning CT on the daily CBCT and applying the deformation matrix (deformation vector field (DVF)) obtained from DIR to planning CT contours to transpose them into daily CBCT. DIR allows also for the accumulation of dose over the course of treatment, adaptation based on the accumulated dose rather than independent snapshots, by means of applying the DVF to daily dose to warp it into planning CT, and the recording of the final estimated delivered dose on planning CT, including adaptations.

2.6.1 *Image Registration*

Image registration is the process of determining the geometric transformation that relates identical (anatomic) points in two image series: a ‘moving’ dataset and a ‘stationary’ source dataset [25].

The general image registration process can be illustrated as shown in Figure 2.9. In this process, the new image will be transformed to the reference image space iteratively by an optimization process. The success of the registration will be measured by a similarity metric. According to the nature of the geometric transformation, the image registration methods are normally categorized into rigid image registration methods and deformable image registration methods.

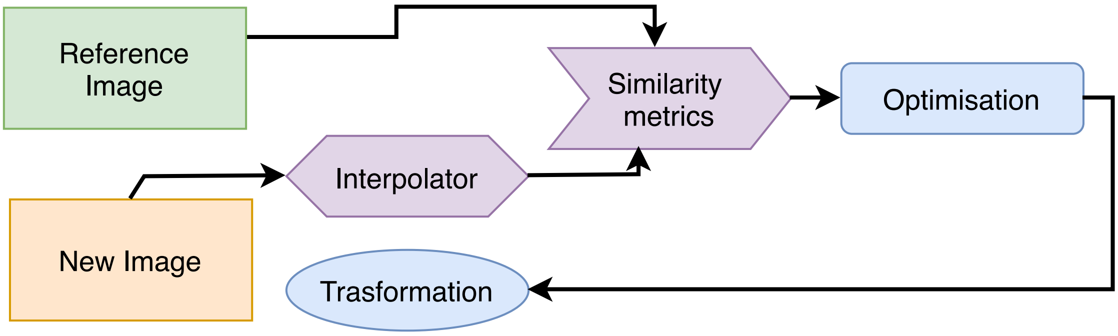


Figure 2.9: The generalized image registration process consists of the following components: (1) a pair of images to be registered, (2) a similarity metric to measure the success of registration, (3) an optimization algorithm to drive the direction and the magnitude of transformation, and (4) the transformation and interpolation modules to change one image to match with the other image. The transformation can be either rigid or deformable.

Rigid registration between two images only allows rotations and translations. Rigid transformation is a special case of a more general transformation, that is, a global or affine transformation. An affine transformation may be composed of rotations, translations, scaling, and shearing.

DIR is now playing a central role in modern ART, as it was described in section 2.6 [31], [32–34], [35]. Deformable image registration is essential to link the anatomy at one time point to another, while maintaining the desirable one-to-one geographic mapping. In addition, DIR can be used to map secondary images or treatment parameters, such as structure set and dose distribution.

2.6.2 *Re-planning*

Generating new plans, once deemed necessary, is time consuming – especially for complex IMRT and VMAT plans which require running optimisation algorithms. Clearly, methods of automatic planning would help improve the situation and vendors are starting to release these systems [36–39]. More details on automatic solutions are presented in chapter 3 and chapter 4. Moreover chapter 10, reports a study performed in this thesis work to automate the re-planning phase.

2.7 CLINICAL RADIATION ONCOLOGY

Approximately 50–60% of all cases of cancer require radiotherapy at some stage during their treatment [40]. The radiation oncologist decides whether radiation therapy is indicated. However, it is best to take a multidisciplinary approach (surgery, medical oncology, nuclear, medicine, radiology) when deciding on the final treatment in clinical practice.

We can define 3 types of radiotherapy according to aim [40]:

- Curative radiotherapy. This is the application of radiotherapy to cure. Used in cases of early-stage Hodgkin's lymphoma, head and neck cancer, prostate cancer, breast cancer and some skin cancers.
- Palliative radiotherapy. This is the alleviation of cancer symptoms by applying palliative doses of radiation. Used in cases of brain and bone metastases for example.
- Prophylactic (preventative) radiotherapy. This is the prevention of possible metastases or recurrences through the application of radiotherapy. An example is whole-brain radiotherapy for small cell lung cancer.

According to timing, we can define 3 type of radiotherapy [40]:

- Adjuvant radiotherapy. Radiotherapy given after any kind of treatment modality (i.e. If given after surgery, postoperative radiotherapy).
- Neoadjuvant radiotherapy. Radiotherapy given before any kind of treatment modality (i.e. if given before surgery, preoperative radiotherapy).
- Radiochemotherapy (chemoradiotherapy). Radiotherapy given concurrently with chemotherapy.

And finally, the factors that should be taken into account for a radiotherapy treatment [40]:

- Aim: palliation or cure;
- tumour: stage, histology, location, radio-sensitivity, previous treatments;
- patient: age, performance, morbidity, cosmesis, patient preference.

In the next subsection a brief overview of the cancers that are investigated in this thesis work, head and neck and breast, is provided.

2.7.1 *Head and Neck cancer radiotherapy*

Head and neck (HN) cancer is one of the most common types of cancer worldwide, with approximately 1.5 million new cases and 0.9 million deaths in 2018 alone [41]. In current clinical practice, a vast majority of patients with locally advanced HN cancer require radiotherapy, with or without concomitant chemotherapy. In the past decades, the treatment paradigm for radiotherapy, has evolved to IMRT, which is the current gold standard [42]. The highly conformal dose distributions produced by IMRT lead to steep dose gradients surrounding the target volumes, which are extremely sensitive to positional errors and anatomic changes. This is particularly critical in HN cases, since there are several structures at risk very close to, and sometimes overlapping with, the target volumes [43]. The proximity of up to 25 organs at risk, with recommendations of critical organs at risk to contour, multiple target dose levels, complex anatomy, and multiple tissue/air/bone interfaces, make the head and neck one of the most challenging sites in treatment planning [40].

To achieve adequate target coverage while protecting numerous OARs, IMRT plans for HN cancer require highly conformal dose distributions and a steep dose fall-off between the boundary of tumour volumes and sensitive structures. With limited clinical resources (time and manpower), a major challenge in HN IMRT planning is large variations in plan quality between different treatment planners, in part due to varied planning skills and limited planning time [42], [43], [44]. For any radiation oncology clinic, particularly in a low-resource setting, these factors may hinder the creation of high-quality HN radiation treatment plans and delay the ability to start treatment while waiting for a treatment plan. Automatic planning could help to overcome these issues.

2.7.2 *Breast cancer radiotherapy*

Breast cancer is the most commonly occurring cancer in women and the second most common cancer overall, making it one of the main causes of mortality and morbidity in females worldwide [45]. Breast-conserving therapy with limited surgery followed by homogenous irradiation of the whole breast (WBI) is often the procedure of choice for management of early-stage breast cancer [46].

The conventional radiotherapy technique for WBI consists of two opposing tangential fields (TF) with wedges. Wedge filters are commonly used to improve dose uniformity within the target volume and are of two types: physical and nonphysical. A physical wedge is usually constructed from a high-density material, such as lead or steel, which attenuates the beam progressively across the entire field. Nonphysical wedges produce modulated dose distributions that are similar to those of physical wedges, for instance by dynamic movements of a pair of independent collimating jaws during beam delivery.

IMRT with static beams, and more recently also VMAT, have been proposed to improve breast dose homogeneity and possibly reduce dose to OARs. IMRT is mostly delivered with two opposing tangential fields with patient-specific intensity modulated profiles, sometimes combined with two open tangential fields (hybrid approach [47, 48]). For VMAT, the two static tangential IMRT fields are often replaced by two small tangential arcs or by a single, larger partial arc [49], [50]. Improving dose homogeneity in the target and reducing OARs doses in WBI can be clinically advantageous [51–58]. So far, clinical trials comparing IMRT and VMAT for WBI have not been performed. There are few published treatment planning studies for left-sided WBI (left side breast tumour where the sparing of OARs is more critical than right side breast due to the presence of heart close to the target) that compare tangential IMRT with tangential VMAT, all with low numbers of patients [59–63]. Overall, the literature is inconclusive regarding the choice of IMRT or VMAT for WBI. Apart from the low patient numbers, this may also be related to the applied conventional trial-and-error treatment planning with well-known challenges for consistent high-quality plan generation [64].

Part II

TOOLS FOR AUTOMATION IN PLANNING AND EVALUATION

3

AUTOMATED MULTI-CRITERIA OPTIMISATION WITH ERASMUS-ICYCLE

> In mathematics, computer science, physics and economics, an optimization problem is the problem of finding the best solution from all feasible solutions.
> *Wikipedia*

3.1 PREVIEW

Chapter 3 is preparatory to the next chapters of Part III of the thesis (chapter 6, chapter 7 and chapter 8) and presents the methods used in the studies described in that part.

3.2 ERASMUS-ICYCLE: SYSTEM FOR AUTOMATED PLANNING

Erasmus-iCycle is an algorithm for multi-criteria optimization of beam intensity profiles (see also chapter 2) and beam angles. Erasmus-iCycle has been developed at Erasmus Medical Center Cancer Institute since 2012 [65, 66] and represents a powerful tool in an automated planning strategy. Erasmus-iCycle is currently the most well known system with these features [5]. The plan generation is multi-criterial and the generated solutions are Pareto-optimal. Of course, the generation of high quality plans is dependent on the setting of the constraints and the objectives with their priorities. Several studies have demonstrated the consistently high quality of Erasmus-iCycle plans, superior to conventional 'trial and error' planning [67–69].

3.2.1 *Wish-list*

In Erasmus-iCycle the optimization operation is based on a user defined wish-list which contains hard constraints and objectives with given priorities. Every element of the wish-list consists of a specific cost function, a priority and a goal:

- Constraints must be strictly met, otherwise the plan is considered invalid. It is proper to consider that too strict constraints may limit possibilities to generate acceptable plans.
- Objectives are cost functions whose goals have to be met as much as possible (without violating the imposed constraints).

In Erasmus-iCycle the objectives are optimized one by one in order of priority. Priorities play a crucial role in determining the outcomes. In Erasmus-iCycle, a later objective is optimized as far as it does not affect the result of the previous ones: even inverting priorities of two objectives, especially for complex anatomies, may strongly affect the entire outcome.

3.2.2 *Brief introduction to 2pec optimization method*

Erasmus-iCycle uses the *2pec* algorithm for prioritized optimization. This algorithm, proposed by Breedveld et al. [66], is an extension of the *e* constraint method [70], in which one objective at a time is optimized while keeping the others (higher prioritized) constrained. The method is extended to a 2-phase constraint optimization (hence *2pec*), where a goal can be assigned to each objective in the first phase of the optimization, while in the second phase a full optimization of the objectives is applied. Hard constraints and prioritized objectives are given in a wish-list which completely regulates the whole optimization process. The idea at the base of the approach is that when it is possible to minimize the dose below a certain threshold (i.e., its goal) for one objective, it is often more desirable to minimize the dose for the other (lower prioritized) objectives first than to directly minimize the dose for the higher objectives to its fullest extent. The wish-list defines hard constraints and a list of n objectives $f_i(x)$ characterized by a priority i and a goal b_i. During the first iteration of the first phase, the objective having highest priority is optimized:

$$\begin{aligned} &\text{subject to } \overrightarrow{g(x)} \leqslant 0 \\ &\text{minimize } f_1(x) \end{aligned} \tag{3.1}$$

$f_1(x)$ is the objective with priority 1 and $\overrightarrow{g(x)}$ is a vector which represents the list of the constraints which are to be met at all times. Based on the solution x^* of this optimisation, a new bound for the optimised objective is defined and it is set as constraint during the optimisation of the following objective. The new bound is chosen according to the following rule:

$$\epsilon_1 = \begin{cases} b_1, & \text{if } f_1(x^*) < b_1 \\ f_1(x^*)\delta, & \text{if } f_1(x^*)\delta \geqslant b_1 \end{cases} \tag{3.2}$$

where b_i and ϵ_1 are respectively the goal and the new bound of the objective with priority 1, $f_1(x^*)$ is the obtained value for the objective , $f_1(x)$ and δ is a slight relaxation to create some space for the subsequent optimisation (usually set to 1.03, i.e. 3 %). In practice this relaxation is mandatory to avoid the optimisation algorithm from stalling due to a numerical problem. The next step is the optimisation of the second objective, $f_2(x)$, while the obtained result for f_1 is added to the constraints list:

$$\begin{aligned} &\text{minimize } f_2(x) \\ &\text{subject to } \overrightarrow{g(x)} \leqslant 0,\ f_1(x) \leqslant \epsilon_1 \end{aligned} \tag{3.3}$$

This is repeated for all n objectives. In the second phase of the multi-criterial optimisation, all the objectives which met their goals are minimized to their fullest, while keeping all others constrained: so, for each f_i which met its goal b_i, the following problem is solved:

$$\begin{aligned} &\text{minimize } f_i(x) \\ &\text{subject to } \overrightarrow{g(x)} \leqslant 0,\ f_k(x) \leqslant \epsilon_k,\ k \in 1, ..., n \end{aligned} \tag{3.4}$$

3.2.3 *Wish-list generation optimisation functions*

Many cost functions are available, allowing a large degree of freedom in the generation of the wish-list.

- Linear: when used in conjunction to minimize maximum, this cost function regulates the maximum allowed dose. It is mainly employed to settle undesired overdoses and to strictly govern the maximum allowed dose to organs at risk characterized by serial complication mechanism (i.e., organs in which the cost of increasing the dose to an already dosimetrically ‘hot’ sub-volume rises tremendously, following a strongly non-linear behaviour, such as the spinal cord [71].
- Mean dose cost function.
- Logarithmic tumor control probability (LTCP):

$$\mathrm{LTCP} = 1/m \sum_{j=1}^{m} \exp(\alpha(d_j - D^P)) \tag{3.5}$$

  LTCP is defined as above, where m is the number of voxels in the target structure, D^P is the prescribed dose, d_j is the dose in voxel j and α is cell sensitivity parameter. LTCP is able to guarantee a proper target coverage.

- Quadratic overdose (QUOP) that allows a maximum dose to the target to not exceed a defined root mean square value of tolerance over all target voxels.
- Equivalent uniform dose (EUD):

$$\mathrm{EUD} = \sqrt[k]{1/m \sum_{j=1}^{m} d_j^k} \tag{3.6}$$

This function is regulated by the parameter k in the formula where m is the number of voxels in the target structure, d_j is the dose in voxel j. For k=1 the meaning of EUD is the same as the arithmetic mean; for k>1 the high dose regions gain more importance in the weighted sum and, ideally, for k=∞, EUD is equal to the maximum dose in the structure.

- Dose-volume reference points: these cost functions try to force the DVH related to a certain structure to achieve the requested value. Due to non-convexity, the use is discouraged because it may lead to sub-optimal solutions of the problem. A convex optimization problem is a problem where all of the constraints are convex functions, and the objective is a convex function if minimizing, or a concave function if maximizing. With a convex objective and a convex feasible region, there can be only one optimal solution, which is globally optimal. A non-convex optimization problem is any problem where the objective or any of the constraints are non-convex. Such a problem may have multiple feasible regions and multiple locally optimal points within each region.

  Because radiotherapy treatment planning is in general a large-scale nonconvex optimization problem, it is often split into several (often convex) subproblems. Typical cost-functions used in radiotherapy are linear mini- mum/maximum dose functions, (generalized) mean dose (which is termed generalized equivalent uniform dose (gEUD) in the radiotherapy field), single or double-sided quadratic penalty functions. There are also functions particularly used for radiotherapy, such as a nonconvex DVH cost-function (the fraction of an organ that receives more than a certain pre-set dose level, as mentioned above), and biological functions that relate physical dose to probability of developing a complication [67], [68, 69].

Wish-lists are stored in .xml format by the Erasmus Institue's in-house developed software, called Lucy, part of the Erasmus-iCycle module.

3.3 ERASMUS-ICYCLE HN PLAN GENERATION

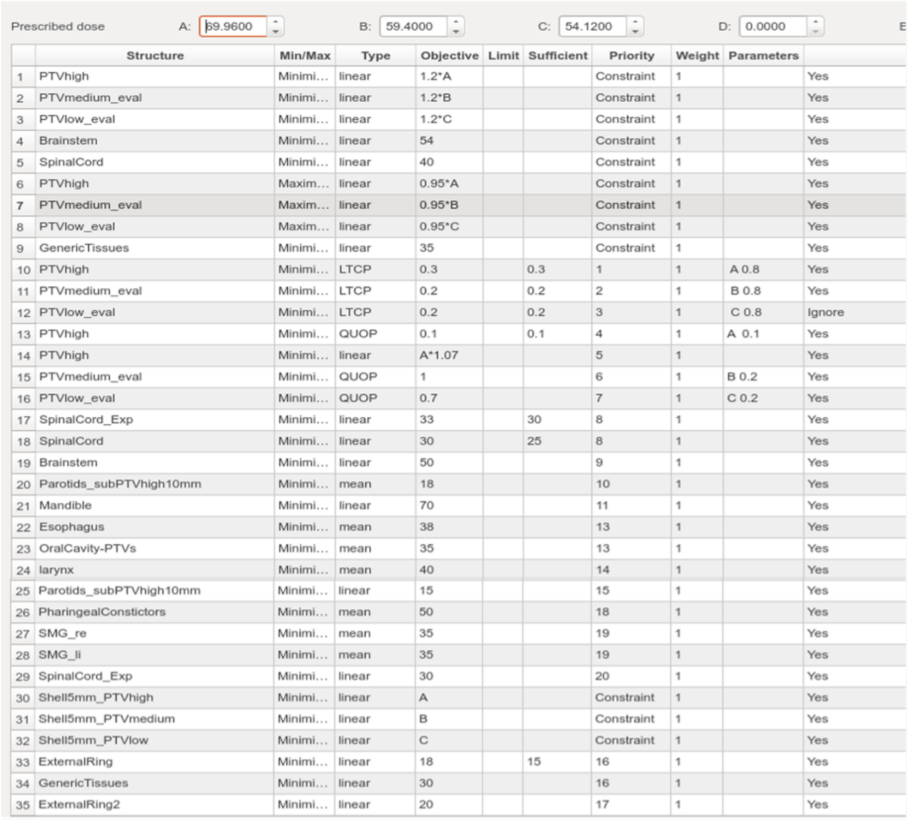
Erasmus-iCycle is specifically suitable to IMRT plans because it does not need to define *a priori* the number of orientations in a plan. Nevertheless, it is possible to simulate VMAT delivery mode with adequate accuracy: in this study, the pre-selected configuration consisted of 23 equi-angular beams. It has previously been proven that further increasing the number of beams increased calculation time but did not lead to an improvement in plan quality [72].

3.3.1 *Patients selection and target delineation*

Between January 2015 and December 2018, 15 patients with histologically confirmed cancer of the oropharynx underwent radiotherapy at Azienda USL-IRCCS Hosptial (AUSL) of Reggio Emilia (Italy). Patients were immobilized in the supine position with a 5-point thermoplastic head-neck-shoulder mask to ensure daily reproducibility of treatments. For each patient a CT image set was acquired both with and without contrast medium. FDG-PET scans were obtained using the same CT acquisition patient set-up. All patients were treated using a Simultaneous Integrated Boost (SIB) technique: the prescribed dose was 70 Gy in 2.12 Gy daily fractions over 33 days to the PTVhigh; 59.4 Gy in 1.80 Gy daily fractions to the PTVmedium; and 54 Gy in 1.64 Gy daily fractions to the PTVlow [73–75]. Surrounding critical normal structures, including the spinal cord, mandible, parotid glands, oral cavity, larynx, oesophagus, pharyngeal constricted muscles and submandibular gland were considered [73] and outlined for oropharyngeal cancer. In nasopharyngeal patients right and left cochlea, brainstem, eyes, optic nerves and chiasm were also outlined in addition to the aforementioned critical organs. In comparison to international protocols [73, 74],[76] and JAVELIN protocols [75], branchial plexus was not considered because it is difficult to localize the brachial plexus on CT. Pharyngeal constrictor muscles were added to the list of protocol's suggested OARs to reduce the risk of dysphagia, which remains a side effect influencing the quality of life of HN patient after radiotherapy [75].

3.3.2 *Construction of the head and neck WLa best wish-list*

An initial wish-list was composed based on previous clinical experience, the planning protocol, and intent of treating physicians on how to improve clinically applied plans. It was used to automatically generate a plan for the first 10 patients included in this study. These plans were then evaluated together with physicians, and the wish-list was modified according to their input. Several optimisation functions, described in subsection 3.2.3, were used in a trial-and-error process, until the expected solution was obtained. This iterative procedure continued until no further improvements in plan quality were achieved for the 10 training patients. The final wish-list, referred to in this study as the 'best' wish-list, is shown in Figure 3.1. This wish-list referred in its major clinical constraints to RTOG protocols [74] and the JAVELIN protocol [75].



Prescribed dose A: 69.9600 B: 59.4000 C: 54.1200 D: 0.0000 E

	Structure	Min/Max	Type	Objective	Limit	Sufficient	Priority	Weight	Parameters	
1	PTVhigh	Minimi...	linear	1.2*A			Constraint	1		Yes
2	PTVmedium_eval	Minimi...	linear	1.2*B			Constraint	1		Yes
3	PTVlow_eval	Minimi...	linear	1.2*C			Constraint	1		Yes
4	Brainstem	Minimi...	linear	54			Constraint	1		Yes
5	SpinalCord	Minimi...	linear	40			Constraint	1		Yes
6	PTVhigh	Maxim...	linear	0.95*A			Constraint	1		Yes
7	PTVmedium_eval	Maxim...	linear	0.95*B			Constraint	1		Yes
8	PTVlow_eval	Maxim...	linear	0.95*C			Constraint	1		Yes
9	GenericTissues	Minimi...	linear	35			Constraint	1		Yes
10	PTVhigh	Minimi...	LTCP	0.3		0.3	1	1	A 0.8	Yes
11	PTVmedium_eval	Minimi...	LTCP	0.2		0.2	2	1	B 0.8	Yes
12	PTVlow_eval	Minimi...	LTCP	0.2		0.2	3	1	C 0.8	Ignore
13	PTVhigh	Minimi...	QUOP	0.1		0.1	4	1	A 0.1	Yes
14	PTVhigh	Minimi...	linear	A*1.07			5	1		Yes
15	PTVmedium_eval	Minimi...	QUOP	1			6	1	B 0.2	Yes
16	PTVlow_eval	Minimi...	QUOP	0.7			7	1	C 0.2	Yes
17	SpinalCord_Exp	Minimi...	linear	33		30	8	1		Yes
18	SpinalCord	Minimi...	linear	30		25	8	1		Yes
19	Brainstem	Minimi...	linear	50			9	1		Yes
20	Parotids_subPTVhigh10mm	Minimi...	mean	18			10	1		Yes
21	Mandible	Minimi...	linear	70			11	1		Yes
22	Esophagus	Minimi...	mean	38			13	1		Yes
23	OralCavity-PTVs	Minimi...	mean	35			13	1		Yes
24	larynx	Minimi...	mean	40			14	1		Yes
25	Parotids_subPTVhigh10mm	Minimi...	linear	15			15	1		Yes
26	PharingealConstictors	Minimi...	mean	50			18	1		Yes
27	SMG_re	Minimi...	mean	35			19	1		Yes
28	SMG_li	Minimi...	mean	35			19	1		Yes
29	SpinalCord_Exp	Minimi...	linear	30			20	1		Yes
30	Shell5mm_PTVhigh	Minimi...	linear	A			Constraint	1		Yes
31	Shell5mm_PTVmedium	Minimi...	linear	B			Constraint	1		Yes
32	Shell5mm_PTVlow	Minimi...	linear	C			Constraint	1		Yes
33	ExternalRing	Minimi...	linear	18		15	16	1		Yes
34	GenericTissues	Minimi...	linear	30			16	1		Yes
35	ExternalRing2	Minimi...	linear	20			17	1		Yes

Figure 3.1: Erasmus-iCycle wish-list for HN cancer treatment in Lucy module, that is part of iCycle system.

For each patient in the database one plan was generated with the same wish-list (with no adjustment by planners). This optimal wish-list is referred to "WLa".

Figure 3.1 shows the WLa wish-list built into the Lucy module. The wish-list consists of 12 hard constraints that are met per definition and 23 objectives that are optimized in order of priority. Once the goal of an objective is achieved or when further optimization is no longer possible, the optimizer fixes the achieved value of the objective as a constraint and continues with the next objective. 5 constraints were used to minimize the maximum dose to PTVs, brainstem and spinal cord. The other 3 constraints maximize the minimum dose to PTVs (95 % of the prescription dose). Then, constraints to a maximum dose to the generic tissue outside the PTVs (ring of 1.5 cm thick up to 4cm from PTVs) was considered. Finally, 3 more constraints require a steep dose fall-off outside PTV: shell5mmPTVhigh, shell5mmPTVmedium and shell5mmPTVlow. The first priority was PTV coverage, ensured by the use of the LTCP and QUOP cost functions. When the spinal cord was close to the target or overlapped it, the objective was applied to the target volume from which the spinal cord or brainstem was subtracted. There is an objective only for PTVhigh which aims to reduce the maximum dose within 107% of the prescribed dose. The next objectives try to reduce the dose to several OARs considered in the optimisation. Parotids were optimized together using the parts of the parotids external to PTVhigh expanded by 10 mm. Two priorities were assigned to this structure, 20 and 25, for each priority different mean dose reduction was asked. Similarly, the oral cavity was optimised in Erasmus-iCycle using the external structure to PTVs, Oralcavity-PTVs. At the end of the lists 3 more objectives (ExternalRing, GeneralTissues, ExternalRing2) were used to reduce the isodoses 30 and 45 Gy in the general tissues outside PTVs.

3.3.3 *Generation of Erasmus-iCycle Pareto MCOa plans*

Pareto-optimal 23-beam IMRT plans with 6 MV beams were generated with Erasmus-iCycle, using WLa. The process of optimisation with Erasmus-iCycle has been described in detail previously [66], [65, 77–79] and is briefly summarized here. For each patient, the template ensures generation of a clinically deliverable VMAT plan that mimics the 23-beam Erasmus-iCycle plan. Consequently, the dose distributions used in this study are highly similar to VMAT dose distributions. Plan generation with Erasmus-iCycle was performed using WLa, described in the previous section. Plans generated in this way were labelled as MCOa plans.

3.3.4 *Construction of suboptimal wish-lists and generation of Erasmus-iCycle MCOx plans*

Several alternative wish-lists WLs, “WLx” (x = b, c, d, . . .),, referred as suboptimal wish-lists, were also generated with different levels of agreement starting from best wish-list (WLa). The WLx were derived from WLa by randomly varying the priorities of PTVmedium and PTVlow objectives and of the OARs (object priorities from 4 to 20 (see Figure 3.1)). As for WLa, the 20 WLx enforced adherence to the hard planning constraints for brainstem, optic chiasm, and spinal cord, as in clinical planning. Moreover, the highest priorities (1-3) referred to PTVhigh coverage and homogeneity which were left unchanged in WLx to generate clinically acceptable plans. This process was done to generate plans with different levels of quality in an automatic way. These MCOx solutions are applied in the studies described in Part III of this thesis, see chapter 6.

4

KNOWLEDGE BASE PLANNING TOOL FOR HEAD AND NECK TREATMENTS

Experience is not what happens to you; it's what you do with what happens to you.
Aldous Huxley

The only source of knowledge is experience.
Albert Einstein

4.1 PREVIEW

Chapter 4 is preparatory to the chapters of Part IV of the thesis and presents the planning automation methods used in the studies described therein.

4.2 KBP RAPIDPLAN TOOL

RapidPlan (RP) is a commercially available KBP tool (Varian Medical System (USA)), implemented into the Eclipse TPS as an optional module [80]. Given a certain delivery/planning technique and treatment site, existing clinical treatment plans may be modelled in the form of DVH-estimation model to individually estimate the most likely dosimetric features expected in new patients (Figure 4.1); RP is actually configured to model plans delivered with IMRT or VMAT. The data from the existing treatment plan are extracted and used first to train the DVH-estimation model. Then, for a new patient the DVH-estimation model generates an estimated DVH-range that shows where the DVH curve of a structure will most likely land; the plan may be automatically optimized using automatic line constrains or building a template based on the KBP individually optimized constraints.

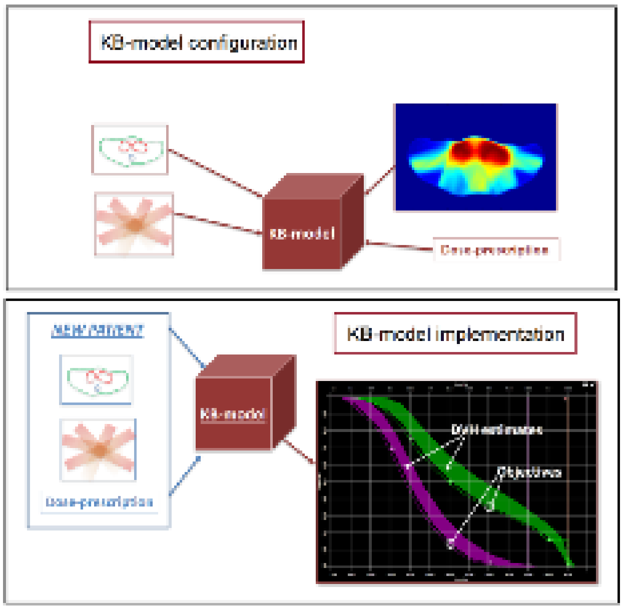


Figure 4.1: Typical KBP workflow: existing clinical treatment plans are used to train the DVH-estimation model (configuration phase). From this modelled patient data, the KB-model generates an estimated DHV range suggesting where the DVH of a structure will most likely be (implementation phase). Reference [80].

4.3 KBP-RP DVH-ESTIMATION ALGORITHM

The DVH-estimation model is site-specific and contains several separate single OAR models. The estimation models of different OARs are entirely separated; correlations with different OARs are not considered. For this reason, a single structure can be removed from a model without affecting the estimates. During the extraction phase, the OAR structures and the target structures are treated differently. For each training plan and for each structure matched to any model OAR structure, the data extraction phase first divides the volume of the structure into functionally different regions (Figure 4.2 a-b)):

- Out-of-field region, is the part of the structure not visible from the jaw aperture of any of the fields (or in case of an arc-field, from any of the control points).

- Leaf-transmission region, is the part of the structure visible at least from one jaw aperture, but has no overlap with the target projection from any of the fields.

- In-field region, is the part of the structure that has overlap with target projection at least from one field.

- Overlap region, is the part of the volume that is also inside any structure matched to one of the model target structures.

Together the four regions (overlap, in-field, leaf transmission and out-of-field) define all structure volumes inside the body. During the training phase, the DVH estimates are generated separately for each sub-volumes and the total DVH estimates are built by combining the sub-volume estimates, weighting different sub-volumes with their relative volumes (Figure 4.2 c). All OARs are evaluated on the basis of their relative position to the targets. A maximum of three target structures may be handled by the RP-tool. DVH-estimation model generates an estimated DVH-range (upper and lower bounds as in Figure 4.1) to show the one confidence interval around the most likely estimate: with both upper and lower bounds one standard deviation away from the most probable estimates. The confidence intervals are determined separately for each sub-volume.

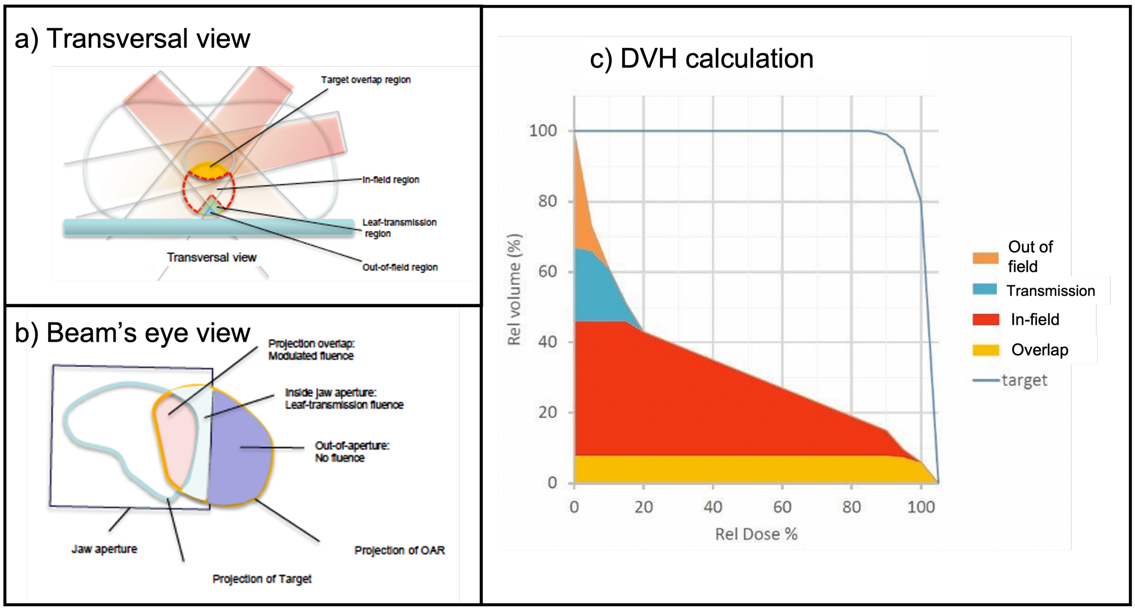


Figure 4.2: a) and b) DVH-estimation algorithm for the in-field partition: during the extraction phase GEDVH and DVH for a given OAR are retrieved; during the training phase a PCA is performed both for GEDVH and DVH; after parametrization of geometry and dose features for the whole training set, the regression model is trained. c) DVH computation for each region of OAR showed in a). Reference [80].

For out-of-field, leaf-transmission and target-overlap sub-volumes a simplified model is used to evaluate their contribution to the estimate of the DVH-range. The model is only based on just the observed variation of DVH of that particular sub-volume in the training set: mean and standard deviations curves are calculated for each sub-volume. At least two instances in the training set needs to have non-zero sub-volume of each kind in order to configure the models: otherwise, default models are used. In contrast with previous cases, the in-field sub-volume is subject to the greatest modulation during the optimization, thus it requires the most complex estimation process. The DVH-estimation algorithm for the in-field sub-volume implements a regression model that permits estimation of the parameters describing the DVH-shapes once certain geometric parameters are known. The evaluation of the geometry for the in-field partition is based on the so-called 'geometry based expected dose' (GED) which is actually an effective distance metric, where the distance to target is converted into dose units in order to handle situations where multiple targets with different dose levels are present. Once the GED field is constructed, the geometrical position of individual OARs is evaluated by calculating a GED cumulative volume histogram (GEDVH) inside the OAR volume (Figure 4.3). Once both GEDVH and DVH for a given OAR in-field partition are retrieved, the algorithm tries to find correlation between them. During the training phase, the parametrization is done by performing PC analysis for both GEDVH and DVH. This process is described in Figure 4.3.

PC analysis involves a set of curves that represent the variation of DVHs observed in the training set: together with the mean DVH curve (see Figure 4.3 a), the PCs can be used to decompose any DVH curve to a couple of parameters representing the multipliers of each DVH-PC. The number of PCs needed to decompose a DVH curve may vary depending on how large a variation occurs in the DVH set: in practice, two or three PCs are generally sufficient. Once the DVH parameters have been estimated, the DVH curves are determined by multiplying each DVH-PC by the estimated score and summing them together with the mean DVH curve. GEDVH curve parameterization is done in the same way. After both geometry and dose features have been parameterized for the whole training set, the regression model is trained. The regression model is determined in a step-wise process and the coefficients, needed to calculate best fitting DVH parame-

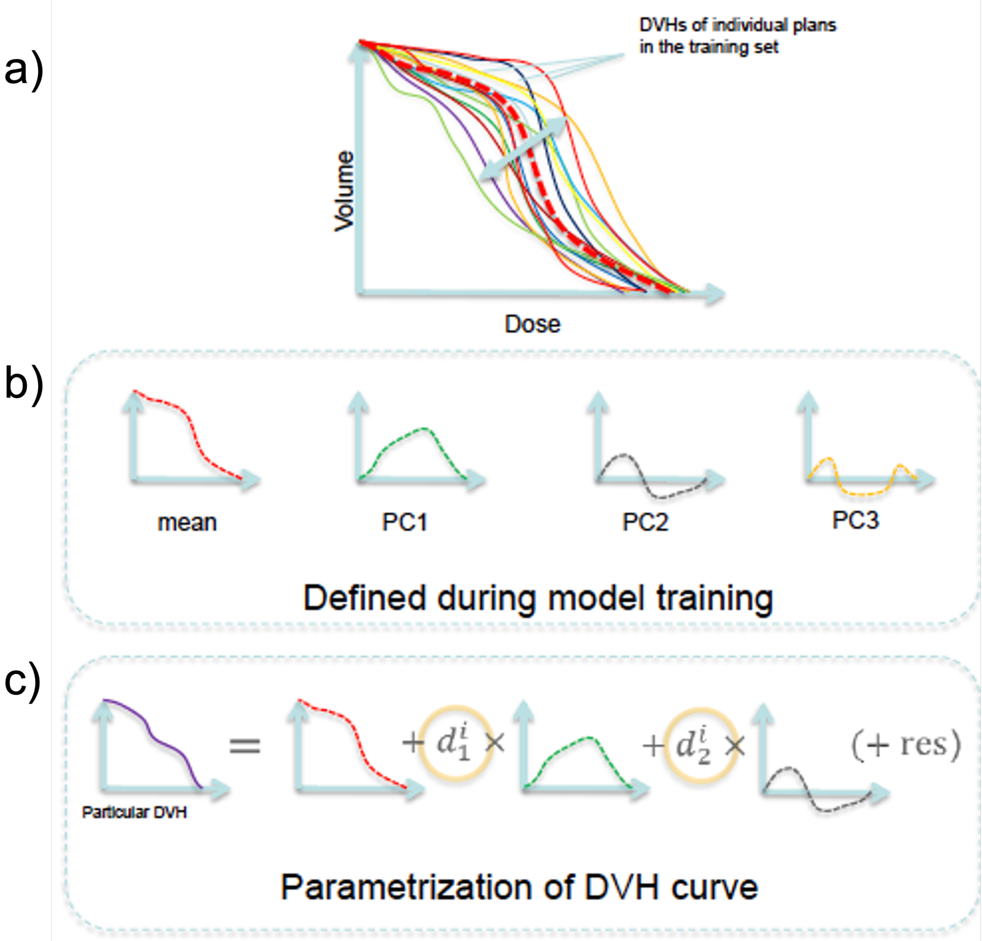

a)
DVHs of individual plans in the training set
Volume
Dose
b)
mean
PC1
PC2
PC3
Defined during model training
c)
Particular DVH
=
$+ d_1^i \times$
$+ d_2^i \times$
(+ res)
Parametrization of DVH curve

Figure 4.3: DVH-estimation algorithm for the in-field partition (a): during the extraction phase GEDVH and DVH for a given OAR are retrieved (b); during the training phase the PC analysis is performed both for GEDVH and DVH (c); after parametrization of geometry and dose features for the whole training set, the regression model is trained (c). Reference [80].

ters estimates based on the geometric parameters, are retrieved (Figure 4.3). The standard error of the regression is used to calculate the upper and lower bounds of the DVH-estimate.

4.4 KBP-RP DVH ESTIMATES AND OPTIMIZATION OBJECTIVES

Once the regression model is trained, for a new patient the model should be able to predict the corresponding achievable dose starting from the geometric features. Schematic process of DVH estimation for a new patient is shown in Figure 4.4.

For instance, considering a new patient and the contours of PTVs and OARs, the geometry information is calculated: partitions, volumes and GED. Then, the GED PCs from the training set are calculated and the parameterization of the GEDVH for the in-field partitions of OARs of the current plan are optimized. Once the coefficients of the GED-PC for the current plan are known, the regression model is used to determine the DVH principal component coefficients from the training set. Then, the PCs of the DVHs are used to obtain the most probable in-field partitions of the OAR DVH. For the other sub-volume partitions, the mean value of the training set is used. Therefore, the most probable DVH is determined: the estimation range (upper and lower bounds) is created by considering the variation in the training data set for each OAR volume partition separately. The standard deviation of the training set is added and subtracted from the most probable DVH for each partition regions. However, for the in-field partition where the regression model is used to create the most probable DVH, the standard error of the regression is considered in creating the bands. Furthermore, the estimated DVHs may be used to generate automatic objectives for the planning optimization. All objectives are positioned below the estimated range. The most general objective is the automatically generated line, otherwise point objectives can be added. Priorities can be user defined or generated by the DVH-estimation algorithm. The system considers the entered prescribed dose as the 100% dose and all DVHs are rescaled using this estimate. Target upper and lower bounds depend only on the prescribed dose; no machine-learning is used to predict them as is the case for the OARs. At least 20 plans are required in order to generate the

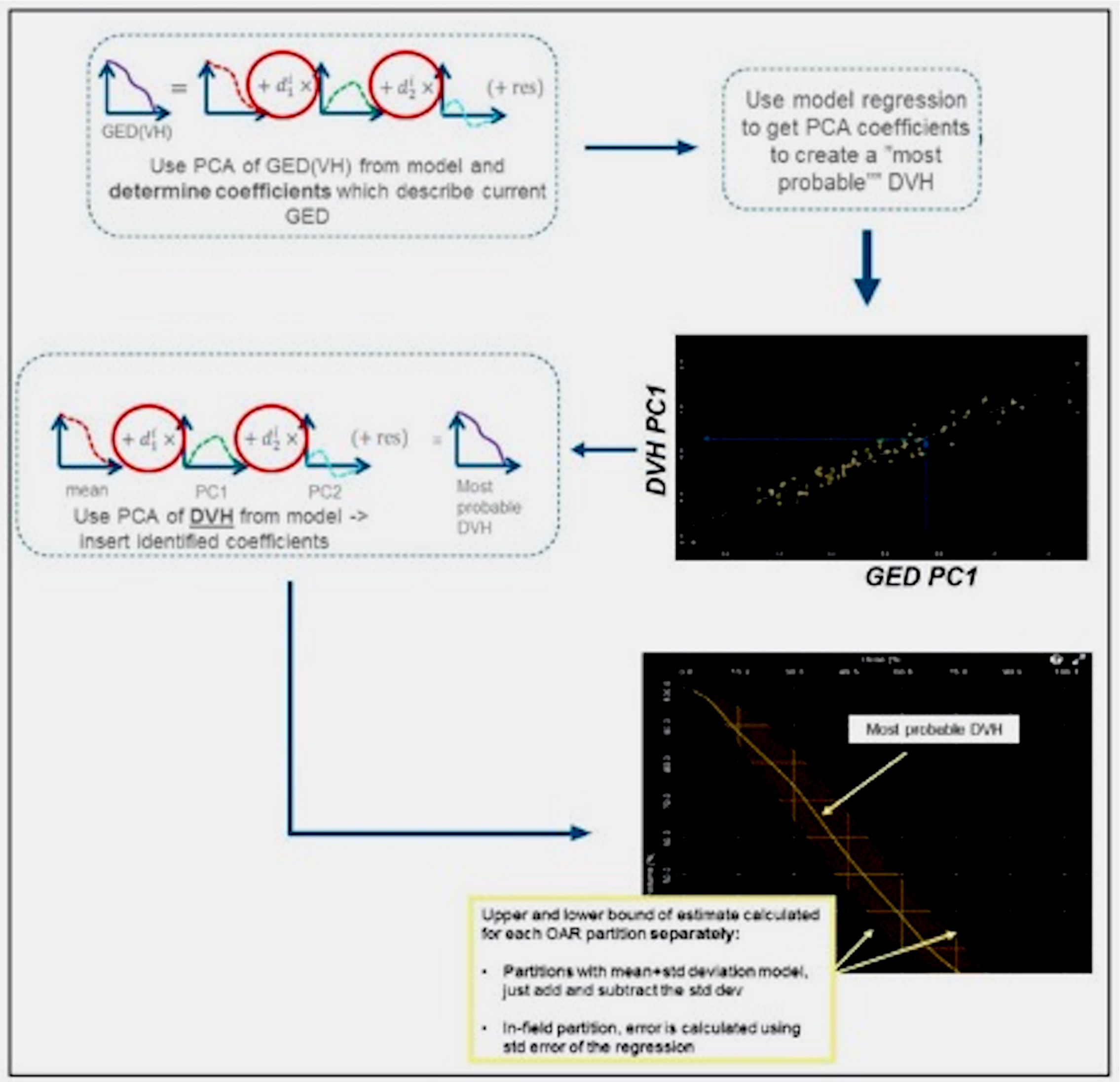

GED(VH)
=
$+d_1^i \times$
$+d_2^i \times$
(+ res)
Use PCA of GED(VH) from model and **determine coefficients** which describe current GED
Use model regression to get PCA coefficients to create a "most probable"" DVH
DVH PC1
GED PC1
$+d_1^i \times$
$+d_2^i \times$
(+ res)
=
mean
PC1
PC2
Most probable DVH
Use PCA of DVH from model -> insert identified coefficients
Most probable DVH
Upper and lower bound of estimate calculated for each OAR partition **separately**:
• Partitions with mean+std deviation model, just add and subtract the std dev
• In-field partition, error is calculated using std error of the regression

Figure 4.4: Schematic process of DVH estimation for a new patient: the GED-PCs are calculated and the GEDVH for the in-field partitions of OAR is retrieved from the model; the regression model is used to determine the DVH-PC and using to obtain the most probable DVH. The standard deviation of the training set is added and subtracted from the most probable DVH for each partition region. Reference [80].

regression model. However, as discussed in the literature, training sets with less than 30 plans need to be avoided [81, 82].

4.5 VERIFYING AND VALIDATING THE KBP-MODEL

Once a KBP-model has been configured, one should verify the performance of the model. During model configuration, the model needs to be interactively fine-tuned, aiming at maximizing its robustness. Then, an internal validation of the model should be performed in order to verify how well the model is able to estimate DVHs (and consequently to optimize plans) that were included in the training set. Finally, an external validation should be carried out to verify how well the model is able to estimate new patient cases. The tuning of the model can be performed by using statistical tools available in the RP system [80] and the so called *Model Analytic platform* (MA) [80]. Both systems permit the operator to evaluate and possibly exclude potential outliers. These "outlier" plans are expected to deviate from the general trend of the model because they do not meet the clinical goal of dosimetric parameters or show a geometry that significantly differs from the rest of the training set [82, 83]. Generally, in the first case the outliers are sub-optimal plans and they should be removed from the training set. In the second situation, plans with geometrical outliers may provide useful information for the model to estimate DVHs in future patients with similar properties: they should not be removed when building the model, apart from selected *atypical* (far from standard) anatomies, such as for instance a full rectum, or rectum contining a large air cavity or an empty bladder. After this fine tuning process, the resulting KB-model may be used to predict DVHs and for each a boundary of expected values is individually generated. DVH estimates may be used to produce automatic objectives for the planning optimization or to generate an individually optimized template for the optimization. Also, this phase requires an appropriate fine tuning to find the right compromise between the priority of PTV coverage and OAR sparing [18, 24, 83–85]. Generally, the resulting RP-based optimization has been shown to generate acceptable quality plans, at least comparable to previously optimized clinical plans [3, 16, 84, 86–89]. An interaction with the planner may further improve planning performance [84, 90]. Furthermore, intra-operator and inter-operator variability has been shown to be significantly reduced by the assistance of RP during the optimization [90], as well as providing a reduction of the planning time [20, 22, 91, 92] and the avoidance of sub-optimal plans [93].

4.6 KBP-RP MODEL CONFIGURATION USING MANUAL PLANS FOR HN

In line with examples used in previous work [24, 81, 82], a dataset of 80 Head and Neck (HN) VMAT patients previously treated at AUSL-IRCCS of Reggio Emilia was used as training set. Treatment plans following two different fractionation schemes were included in this work, 69.96 Gy/59.4 Gy/54.12 Gy in 33 fractions (46 patients) and 66 Gy/60 Gy/54 Gy in 30 fractions (34 patients), both schemes using a simultaneous integrated boost (SIB) technique. For all plans, the goal was to deliver 100% of the prescribed dose to 95% of every PTV. All plans were generated with the Eclipse Treatment Planning System (TPS) (Varian Medical Systems, Palo Alto, CA) using 3 fully coplanar arcs with collimator rotated to 30°, 315° and 90°, with 6 MV energy. The dataset was used to train a KBP RapidPlan DVH prediction model using Eclipse v.15.6. Using the training set, KBP DVH prediction models were created for the following OARs: brainstem, spinal cord, left parotid gland, right parotid gland, mandible, oral cavity, oesophagus, and larynx. This model trained with manual plans was identified as the KBP model. Once the model was trained, selected DVH constraints may be extracted from the KBP prediction model to generate an individually optimized template for plan optimization. The template created for this study is shown in Table 4.1.

4.7 TRADE-OFF MULTICRITERIA OPTIMISATION

A MCO approach based on trade-off exploration modules is implemented in the Eclipse TPS (MCO Trade-Off). The inclusion of MCO in radiotherapy planning aims to allow the exploration of the trade-offs of the treatment objectives in an efficient way to then select a plan that best fulfils the prescribed clinical goals. In MCO, a range of different Pareto solution plans is generated, based on a selection of optimisation objectives. The priority of each objective may vary from plan to plan but all plans belong to a ‘Pareto surface’. The user can explore the trade-offs

Structure	Objective	Volume(%)	Dose	Priority
PTVhigh	upper	0	103%	185
	lower	100	99%	180
PTVmedium	upper	0	101%	100
	lower	100	99%	180
PTVlow	upper	0	101%	100
	lower	100	99%	180
brainstem	upper	0	generated	generated
esophagous	mean		generated	70
larynx	mean		generated	70
	upper	10	generated	70
	upper	70	generated	70
	upper	80	generated	80
mandible	upper	0	generated	generated
	upper	1	generated	generated
	upper	5	generated	generated
oral cavity	mean		generated	70
parotids	upper	generated	30 Gy	generated
	mean		generated	100
	line	generated	generated	generated
parotids	upper	generated	30 Gy	generated
	mean		generated	100
	line	generated	generated	generated
spinal cord	upper	3	generated	generated
	upper	0	generated	130
RINGa	upper	0	64 Gy	150
	upper	10	generated	generated
RINGb	upper	0	55 Gy	120
	upper	10	generated	generated
RINGc	upper	0	50 Gy	120
	upper	10	generated	generated
Generic Tissue	upper	0	generated	generated
	upper	10	generated	generated
	mean		generated	generated

generated = determined by model and estimated DVH

Table 4.1: The KBP-based template for automatic planning optimization in the RapidPlan models. Lower, upper and mean objectives and priorities were selected for each structure according to the table.

along the Pareto surface and select the plan that best fulfils the treatment goals. With the use of graphical ‘slider bars’, dynamic DVHs and dynamic 3D dose distributions, the TPS allows users to visually review and evaluate plans along the Pareto surface in ‘real time’ [18, 19]. An example of the use of the Trade-Off MCO module is illustrated in Figure 4.5). To commence, it is required to have a starting plan (that will be at the centre of the approximation of the Pareto surface). Figure 4.5 a displays the slider bars for several organs and corresponding DVHs of the approximated Pareto plan corresponding to the manual starting input solution plan. The user can move the slider corresponding to one or several endpoints navigating along the Pareto surface. This action will affect the corresponding DVHs of all structures involved, in this case right and left parotids, spinal cord and all PTVs (see Figure 4.5 b).

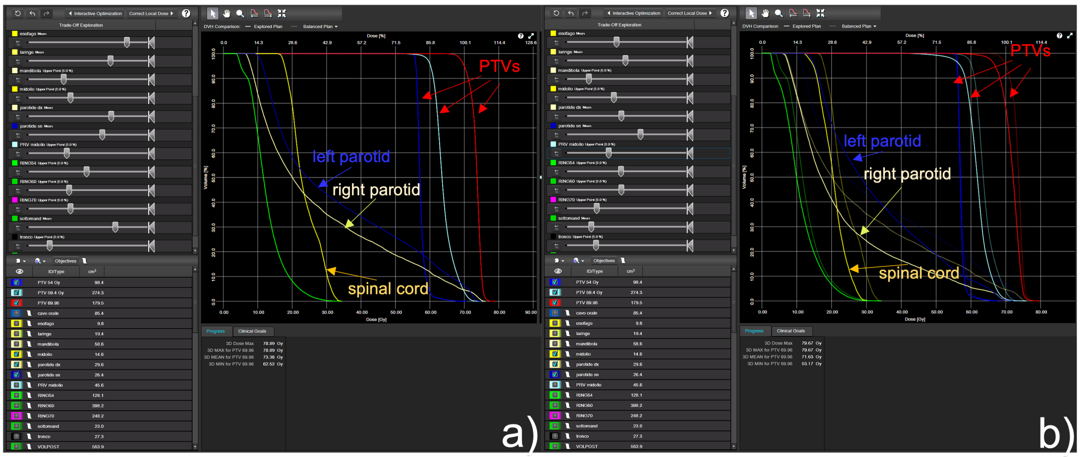


Figure 4.5: Eclipse real time plan navigation screen view during Trade-Off optimisation for one of the head and neck plans. a): Trade-Offs exploration with slider bars for each selected objective for the initial plan and corresponded DVHs for the structures considered. b): Graphical feedback of the modifications instantly monitored in the DVHs view displayed for spinal cord, parotids and PTVs.

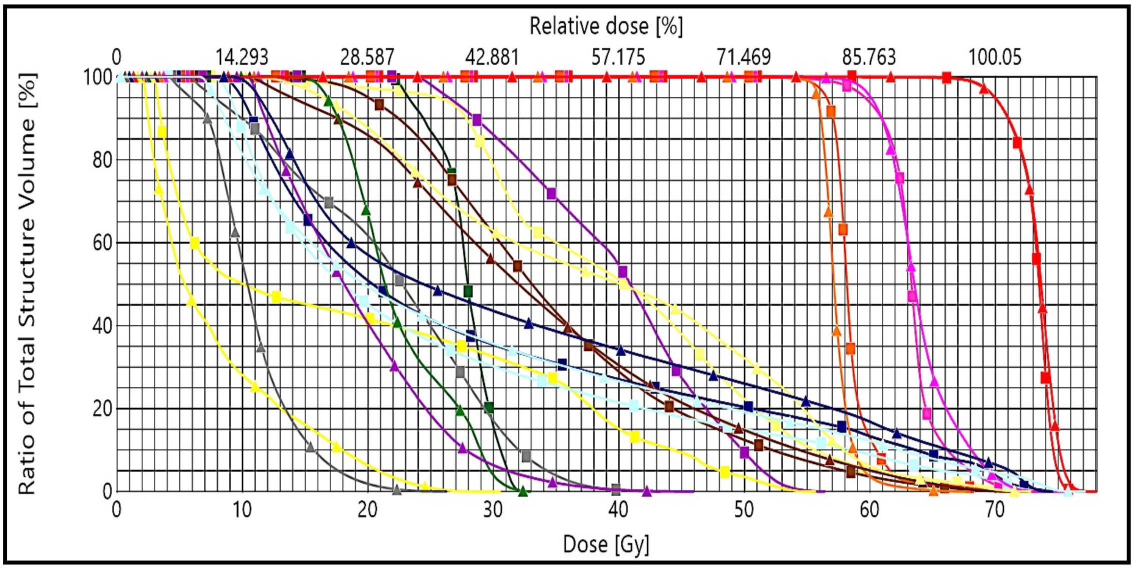
4.8 KBP-RP MODEL CONFIGURATION USING AUTOMATED PLANS FOR HN

All 80 patients used in the training set were automatically re-optimised using the KBP model. The solution found after optimisation without manual intervention was used as a starting plan in the MCO Trade-Off module. A wish list of objectives to fulfil was used to consistently select a solution on the Pareto surface in the MCO module. This wish list was created together with radiation oncologists at AUSL-IRCCS, based on previous clinical experience and planning protocols [74, 75]. In the MCO module, once the goal of an objective in the wish list was obtained, the optimiser fixed the achieved value (slider restrictor) as a constraint and continued with the next objective. Once the MCO module had completed the optimisation for every objective on the wish list, the DVHs and the spatial distribution of the dose that resulted from the wish list trade-off process were selected from the Pareto surface, and the final dose was calculated. This wish-list is reported in Table 4.2.

Priority	Structures	Objective	
1	PTVhigh	Dmin	V66.46Gy>99.6%
2	PTVhigh	Prescription	V69,96Gy=95%
3	PTVhigh	Dmax	V107%<1%
4	Brainstem	Dmax	V54Gy<0,03cc
5	spinal cord	Dmax	V45Gy<0,03cc
6	PTVmedium	Prescription	V59,4Gy>99%
7	PTVlow	Prescription	V54,12>99%
8	PTVmedium	Dmax	V69,96Gy<3%
9	PTVlow	Dmax	V59,4Gy<3%
10	parotid gland controlateral	Dmean	20Gy
10	parotid glands both	Dmean	26Gy
11	spinal cord	Dmax	ALARA
12	spinal cord exp	Dmax	42
13	Mandible/TM joint	Dmax	V70Gy≤1cc
14	Esophagus	Dmean	50Gy
15	Oral cavity ext. (excluding PTVs)	Dmean	40Gy
16	Larynx	Dmean	40Gy
17	RINGS	Dmax	A,B,C
18	VOL ANT/VOLPOST	Dmax	45Gy
19	Pharingeal Constictors	Dmean	50Gy
19	Submandibular Glands	Dmean	Dmean <35 Gy

Table 4.2: Applied wish-list to MCO Trade-Off module. A, B, C referred to the prescription dose for PTV high, PTV medium and PTV low in the 2-fractionation regimen used in this study: 69.96Gy/59.4Gy/54.12 Gy in 33 fractionsand 66Gy/60Gy/54Gy in 30 fractions.

Due to the use of the same optimisation scheme for all patients, plan generation was highly consistent across the entire cohort, with no plans adjusted by the planning team. An example of DVH difference between manual and MCO Trade-Off plans for PTVs and several OARs is reported in Figure 4.6. All plans resulting from both methods were deemed clinically acceptable according to the criteria of PTV coverage and OAR doses. While maintaining comparable target coverage,



DVH	Structure	Min Dose	Max Dose	Mean Dose
■	Mandibula	9.7 Gv	74.0 Gv	40.5 Gv
■	Oesophagus	2.9 Gy	55.7 Gy	19.2 Gy
■	PTV 69.96	55.3 Gy	77.6 Gy	73.1 Gy
■	Left parotid	8.2 Gy	75.9 Gy	29.2 Gy
■	Brainstem	5.7 Gy	43.3 Gy	21.8 Gy
■	Spina cord	21.6 Gy	32.5 Gy	27.8 Gy
■	PTV 59.4 Gy	45.9 Gy	74.3 Gy	63.4 Gy
■	PTV 54 Gy	47.9 Gy	73.9 Gy	58.5 Gy
■	Right parotid	7.4 Gy	76.4 Gy	25.9 Gy
■	Larynx	23.5 Gy	56.4 Gy	39.8 Gy
■	Oral cavity	11.1 Gy	73.0 Gy	35.2 Gy
▲	PTV 69.96	61.3 Gy	78.1 Gy	73.3 Gy
▲	Mandibula	10.5 Gy	72.9 Gy	39.0 Gy
▲	Oesophagus	2.1 Gy	30.6 Gy	7.9 Gy
▲	Left parotid	9.3 Gy	74.6 Gy	32.9 Gy
▲	Brainstem	3.9 Gy	26.3 Gy	10.9 Gy
▲	PTV 59.4 Gy	51.1 Gy	75.5 Gy	64.0 Gy
▲	PTV 54 Gy	51.5 Gy	68.0 Gy	57.3 Gy
▲	Larynx	10.1 Gy	46.0 Gy	19.2 Gy
▲	Oral cavity	10.5 Gy	71.8 Gy	34.1 Gy
▲	Right parotid	6.6 Gy	77.2 Gy	28.0 Gy
▲	SpinalCord	14.6 Gy	33.0 Gy	22.5 Gy

■ Manual plan
▲ MCO Trade-off plan

Figure 4.6: DVH comparison between manual and MCO plan, the latter produced using Trade-Off module starting from manual plan solution, for one representative case belong to the training set used for KBP model configuration.

the sparing of the OAR resulted in differences between manual and MCO plans for several structures. However, the major improvement of MCO plans resulted in a higher consistency of plan quality among all training sets considered, as reported above.

4.9 KBP AND KBP-MCO MODEL RESULTS

The verification of the DVH estimation for each model was performed using both the RapidPlan Model Configuration and Model Analytics tool [80], as described in section 4.5.

Figure 4.7 and Figure 4.8 show the resulting regression models for each OAR for the two models, KBP and KBP-MCO: each graph shows the correlation between the dosimetric and the geometric components, as parameterized during the training phase. The trend line (dash line) with the correlation R^2 value and the two standard deviation of the regression (straight lines) are also shown.

The model quality was evaluated by checking the model goodness of fit statistics for each structure, with the coefficient of determination R^2 (between 0 and 1: the larger, the better) and the average Pearson's chi square χ^2 (the closer to 1, the better). Those parameters, together with the number of potential outliers or influential points are reported in Table 4.3 for all models.

The potential outliers identified in the MA tool were evaluated case by case. They were judged as not real being outliers, in the majority of the cases related to some anatomical differences with respect to the rest of the population in the model, all plausible and not anomalous anatomies. These parameters, together with the number of potential outliers (also known as influential points), are reported in Table 4.3. No particular trends were observed for χ^2 and R^2. A mean χ^2 of 1.08±0.04 and 1.11±0.05 and a mean R^2 of 0.54±0.14 and 0.83±0.10 were found for KBP and KBP-MCO modules respectively, showing an improvement of the KBP-MCO module, especially for R^2. This improvement in the regression modesl of KBP-MCO is also evident by comparing Figure 4.7 and Figure 4.8 . By comparing Figure 4.7 and Figure 4.8, it is possible to observe that for two models (spinal cord and oesophagus) RP tool chosen a different combination of gemetrical features (x-axis) for the regression model. The process of choosing the bet-

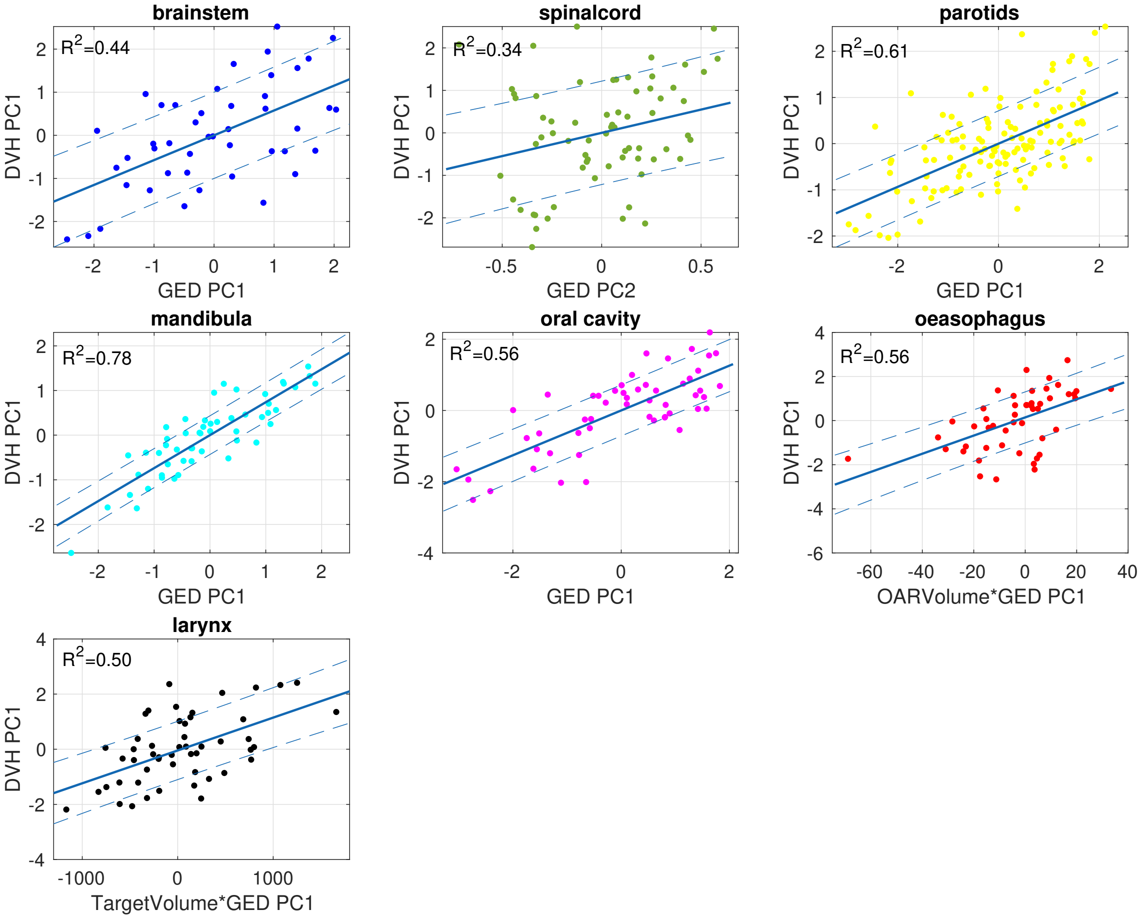


Figure 4.7: Regression model for each OAR trained in the manual model: each graph shows the correlation between the dosimetric and the geometric components, as parameterized during the training phase; the trend line (solid line) with the correlation R^2 value and the two standard deviation of the regression (dash lines) are also shown.

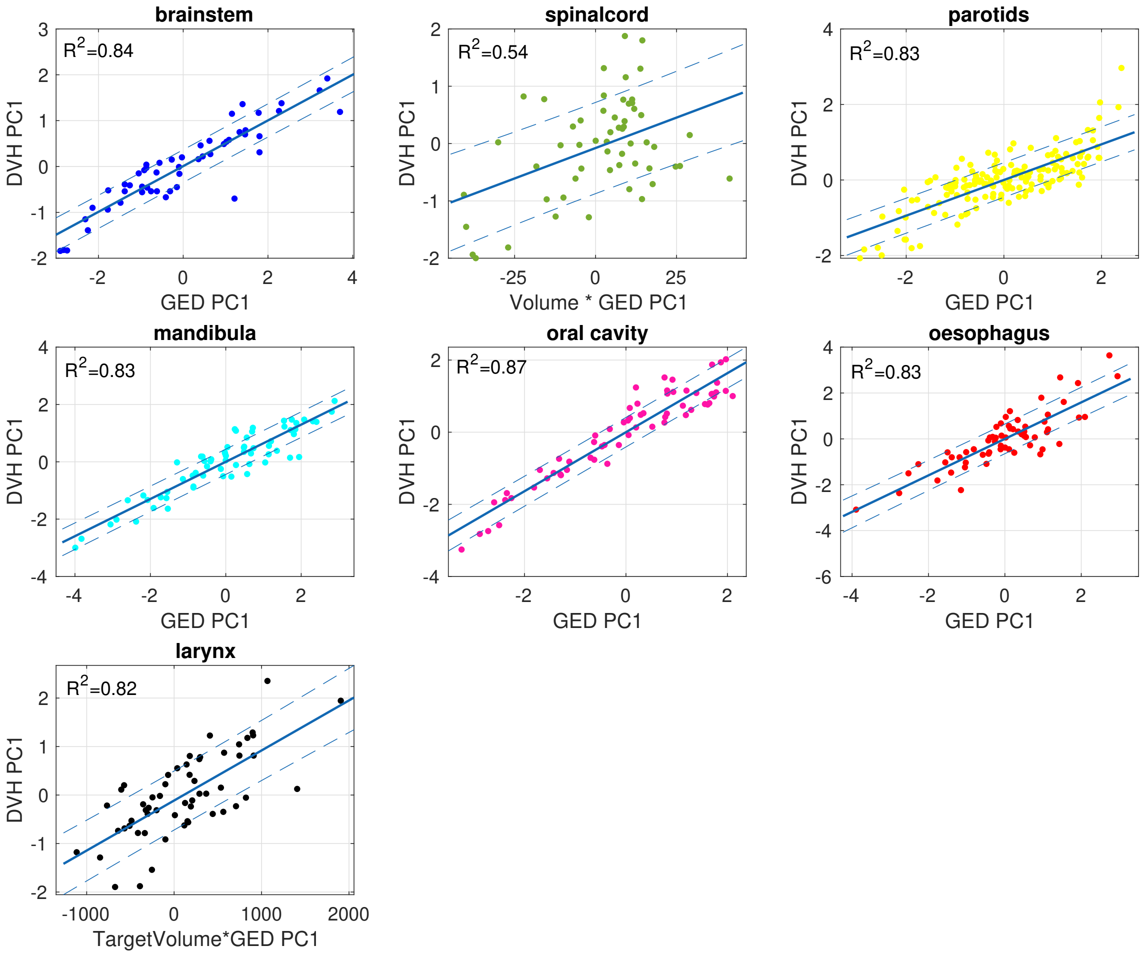
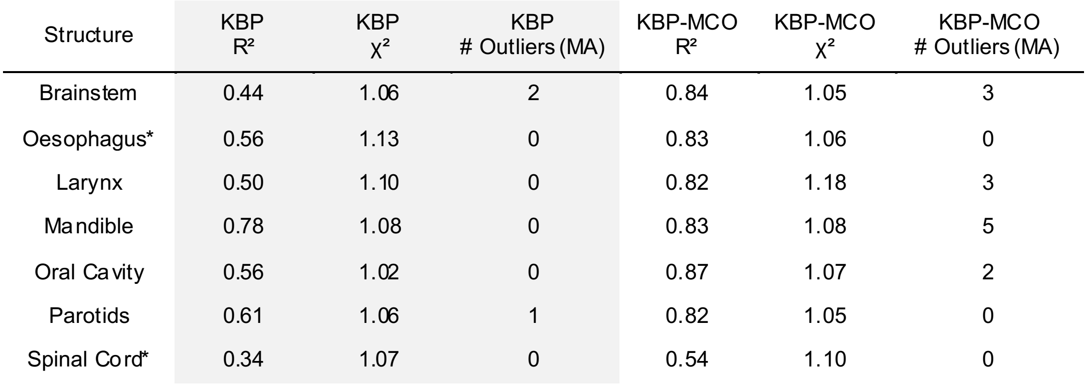


Figure 4.8: Regression model for each OAR trained in the MCO model: each graph shows the correlation between the dosimetric and the geometric components, as parameterized during the training phase; the trend line (solid line) with the correlation R^2 value and the two standard deviation of the regression (dash lines) are also shown.



Structure	KBP R²	KBP χ²	KBP # Outliers (MA)	KBP-MCO R²	KBP-MCO χ²	KBP-MCO # Outliers (MA)
Brainstem	0.44	1.06	2	0.84	1.05	3
Oesophagus*	0.56	1.13	0	0.83	1.06	0
Larynx	0.50	1.10	0	0.82	1.18	3
Mandible	0.78	1.08	0	0.83	1.08	5
Oral Cavity	0.56	1.02	0	0.87	1.07	2
Parotids	0.61	1.06	1	0.82	1.05	0
Spinal Cord*	0.34	1.07	0	0.54	1.10	0

Table 4.3: Goodness of the prediction models in terms of coefficient of determination, R^2,, average Pearson's chi square, χ^2, and number of potential outliers (model analytics, MA, suggested plans to be removed and plans to be checked). KBP refers to model trained using manual plans and KBP-MCO refers to model trained using consistently generated Pareto solutions using MCO Trade-Off tool. The asterisk [*] indicates that the combination of geometrical features in the regression model is different in the two models.

ter geometrical features is completely automated and it is independent from the setting that the user can do. RP tool automatically selects the best combination of geometrical features to maximize the goodness of the model. Thus, a simple comparison of R^2 between KBP and KBP-MCO for oesophagus and spinal cord models is not possible. However, it is possible to assume that since the R^2 and χ^2 improved in MCO-KBP for these two organs, as indicated in Table 4.3, these are improved models but using a different combination of geometrical features. A (*) symbol was used to indicate that the combination of geometrical features was different for the regression models in Table 4.3,

5

TOOLS FOR PLAN EVALUATION PROCEDURE ANALYSIS

ideal:
— perfect, or the best possible;
— a principle or a way of behaving that is of a very high standard;
— a perfect thing or situation.
Cambridge English Dictionary

5.1 PREVIEW

Chapter 5 is preparatory to Part III of the thesis. Planning evaluation methods are presented in this chapter. In particular a tool called gUIDE (generalized Uniform Ideal Dose using Exponential function) developed in this thesis to produce an 'ideal' dose distribution that could be useful as a baseline for a clinical plan. By comparing the obtained dose distribution with gUIDE dose a measure of plan quality can be derived. The gUIDE dose is compared for validation with a commercial tool, planIQ (see following sections for details), for plan evaluation based on a feasibility DVH. This feasibility DVH is considered the benchmark.

5.2 PLANIQ MODULE

The commercial software package called planIQ (Sun Nuclear Corp., Melbourne, FL), helps users to evaluate a clinical plan [44, 94]. Before treatment planning begins, PlanIQ analyzes the patient-specific feasibility of institute specific clinical goals, with insights on areas for improvement. Target, OAR and overall plan quality are summarized from "acceptable" to "ideal" for easy identification of weaknesses in the plan. Based on the clinical goals, every target and OAR receives a quality score. The treatment plan receives a Plan Quality Metric score and an Adjusted PQM score customized to the patient-specific feasibility analysis. In PlanIQ a tool called a feasibility DVH (FDVH), has the aim of helping the operator during plan generation in the achievement of challenging patient-specific dose objectives. In more detail, the FDVH tool, introduced by Ahmed textitet al. [13], uses the CT images and DICOM RT structure set of the patient to generate a synthetic dose distribution based on first principle assumptions and a series of energy-specific dose-spread calculations [13, 44, 94]. This 3D dose distribution is 'ideal' and is intentionally unachievable, such that each PTV is evenly populated has the prescription dose (the DVH of each PTV will therefore be a simple rectangle). A high dose gradient and moderate dose periphery is then added to the PTV dose cloud. Once the dose cloud is generated, for each individually considered OAR the lower possible boundary of its DVH is predicted [13, 44, 94]. A detailed explanation of the algorithm can be found in [44]. The PlanIQ software v2.1 implements the FDVH estimation algorithm. However, the user can only visualise and export the ideal FDVH and not spatial dose distribution. The FDVH calculation process is broken down into three steps:

i. specify target volumes, prescription doses, and calculation parameters,

ii. build the benchmark dose grid,

iii. generate the FDVH curves.

The following equation (eq. 5.1) describes the baseline dose concept implemented in PlanIQ module:

$$\mathrm{BD}_{\mathrm{final}}(x, y, z) = \max(\mathrm{D}_{\mathrm{GHDS}}[x, y, z], \mathrm{D}_{\mathrm{LDS,mid}}[y], [x, z], \mathrm{D}_{\mathrm{LDS,far}}[y], [x, z]) \quad (5.1)$$

In eq. 5.1 $\mathrm{D}_{\mathrm{GHDS}}[x, y, z]$, describes general penumbral effects at a beam edge, tangential to the target surface. $\mathrm{D}_{\mathrm{LDS,mid}}[y][x, z]$ and $\mathrm{D}_{\mathrm{LDS,far}}[y][x, z]$ contains both the percentage depth dose curve (PDD) effect and the low dose outside the steep portion of the penumbra due to scatter out-of-field. They involve 2D convolutions in axial planes of a signal function with energy-dependent 2D kernels. This benchmark dose is used to produce the *best possible sparing* FDVH for an OAR, and based on that, progressively more easily achievable FDVH curves can be estimated.

5.3 GENERALIZED UNIFORM IDEAL DOSE USING EXPONENTIAL FUNCTION (gUIDE) COMPUTATION

5.3.1 *Rationale behind gUIDE implementation*

As explained in the previous chapter, there are significant variations in plan quality evaluation, even among radiation oncologists and medical physicists belonging to the same department. One of the limitations of using only the DVHs of available plans is that information about the spatial position of the dose is lost to a large extent. As the process of evaluation also necessarily involves the visual assessment of a 3D dose distribution, it was thought that the computation of a 'baseline dose', which is not attainable but represents the closest option to the most ideal (but physically impossible) situation, could help improve the modelling of the evaluation process. Said situation is the scenario where the entire prescription dose is delivered to the target and there is null dose to voxels outside the target. This dose was not known beforehand by the automatic planning system or the evaluators, but takes into account the unique patient anatomy and how that plays a significant role in the best achievable doses to specific anatomic regions. The comparison of these theoretical and synthetic (but patient-specific limits) could give more insight into the evaluation process and could help in highlighting the different personal preferences that observers could employ when evaluating a plan.

5.3.2 *gUIDE implementation*

To fulfil the aims set out above, an initial version of a gUIDE tool was developed during this project. The primary goal of this tool was to estimate the best case of dose distribution, given the specific anatomy of any patient. The algorithm is a generalized version of the one proposed in Ahmed *et al.*'s work [13] implemented into a commercial tool called PlanIQ, described in previous section, section 5.2. From PlanIQ, the user can only visualise and export the ideal FDVH and not a 3D spatial dose distribution. This is the reason why the gUIDE function was developed in the current work in order to have an ideal spatial dose distribution as reference. As described in the previous section, the aim and the initial application of the gUIDE tool is somewhat different to the methodology of PlanIQ FDVH. As a first step a simpler description of the best achievable dose was implemented but, at the same time, it was also intended to maintain the spatial information which is lost in the DVH output of Ahmed *et al.*'s work. In this section, the algorithm employed to develop the tool is described, together with the optimization of the algorithmic parameters. Then, an application based on the previously described set of HN patients is presented. This tool was implemented using Matlab version R2020b (Mathworks, Natick,USA). The process of the gUIDE computation is composed of three steps and it is described in Figure 5.1:

i. Specification of target volume(s) and their prescription(s) together with calculation parameters;

ii. implementation of the initial ideal dose;

iii. generation of the gUIDE from the initial ideal dose.

The algorithm does not require any specification of the treatment machine or beam energy. The inputs needed for the gUIDE tool to generate the dose distribution are the CT simulation scan volume, with the 'masks' of the PTV(s) that need to be covered and their respective dose prescription(s). A mask of the external body of the patient also needs to be provided. The other calculation parameter that needs to be given to the tool is the dose grid spatial resolution. In this work the best dose grid resolution among the available plans was employed (0.97 mm in the in the anterior posterior (A-P) and left-right (L-R) direction and 3 mm in the superior-inferior (S-I) direction).

- *Initial gUIDE*

  The initial version of the ideal dose is a basic 3D dose grid (with the user-specified resolution parameters) which provides 100% coverage of each of the target volumes with its associated prescribed dose. As the PTV mask is initially specified in the coordinates and the resolution space of the CT simulation scan, an interpolation of the mask is carried out to map the mask in the 3D dose grid space. Then the dose grid points [x,y,z] corresponding to the voxels of the dose matrix, are assigned a dose value. In this first step,

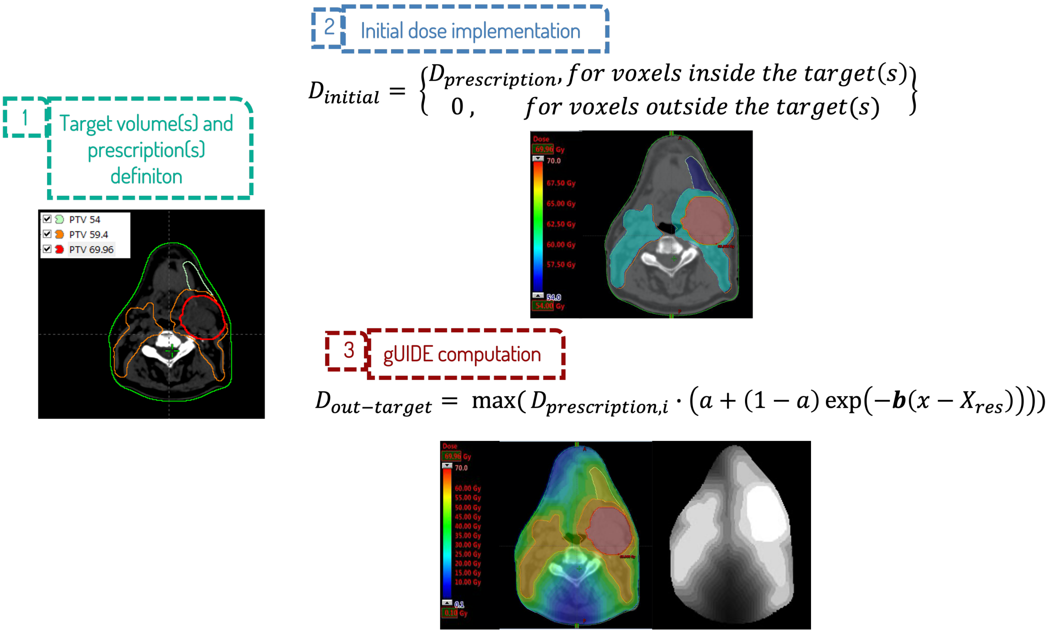

1
Target volume(s) and prescription(s) definiton
PTV 54
PTV 59.4
PTV 69.96
2
Initial dose implementation
$D_{initial} = \begin{cases} D_{prescription}, & for\ voxels\ inside\ the\ target(s) \\ 0, & for\ voxels\ outside\ the\ target(s) \end{cases}$
Dose
3
gUIDE computation
$D_{out-target} = \max(D_{prescription,i} \cdot (a + (1-a)\exp(-\boldsymbol{b}(x - X_{res}))))$
Dose

Figure 5.1: gUIDE computation process based on 3 steps. 1. specification of target volume(s) and their prescription(s) together with calculation parameters; 2. implementation of the initial ideal dose as uniform dose on each target equal to prescription dose and zero outside targets (equation 7.1); 3. generation of the gUIDE from the initial ideal dose outside targets using eq. 5.3.

the dose values are assigned following a simple binary target coverage grid (see Figure 5.1-panel 2).

$$D_{initial} = \begin{cases} D_{prescription}, & \text{for voxels inside the targets} \\ 0, & \text{for voxels outside the targets} \end{cases} \tag{5.2}$$

If there are several dose prescription levels, there are also several different 3D sub-doses that are generated (described in the next section steps). The final dose is composed of the maximum dose values from the generated sub-doses so that only the PTV that provides the highest contribution counts towards the gUIDE's generation.

- *gUIDE generation outside targets*

After assignment of dose to the PTVs, the algorithm assigns the dose to the non-target voxels. This is achieved by creating successive expansions of the target in an iterative process: the dimension of the expansion margins used in every iteration is equal to the highest dose grid resolution. Then, the voxels inside the expansion are given the prescribed target dose, multiplied by a negative exponential 'fall-off' factor depending on the distance of the specific expansion with respect to the target, with the following relationship (see Figure 5.1-panel 3).

$$D_{(\text{out–target})} = \max(D_{\text{prescription},i} \cdot (a + (1-a)\exp(-b(x - X_{\text{res}})) \qquad (5.3)$$

Where:

- $D_{(\text{out–target})}$ is the dose assigned to every voxel inside the n^{th} expansion;
- $D_{\text{prescription}}$ is the target prescription for that sub-dose;
- a is the plateau parameter describing the minimum percentage of dose showing in the 3D dose map; as our ideal dose needed to be as low as possible, this parameter was set to 0.01;
- b is the fall-off parameter, determining the steepness of the dose descent;
- X_{res} is the dose grid resolution, set equal to the maximum resolution of the map (in this work, 3 mm);
- x is the distance from the target for that specific expansion which is computed by $x{=}i \cdot X_{\text{res}}$, where i is the number of the iteration. Thus $(x - X_{\text{res}})$ is always a positive quantity.

As previously mentioned, if there are multiple dose prescriptions after all the associated sub-gUIDE doses are generated, the final gUIDE is composed of the maximum values among all gUIDE sub-doses. The resulting dose derived from the above formula is thus composed of dose steps as it is a collection of isodoses decreasing exponentially with distance from the target. After every iteration, the computed dose is multiplied by the binary mask function defined by the external body contour in order to speed up computational time, as voxels outside the body are set to receive null dose. Generally, as shown also in Ahmed's work [13], there are two factors in the dose distribution:

i. the high gradient component of the dose spread effect that is the predominant effect in the vicinity of the target;

ii. low-dose effect component is due to a combination of physical factors. It is a function beam energy, that influences dose attenuation along beam axes.

Low-dose effect was not accounted for in our 3D dose computation. For these reasons, gUIDE doses are sufficiently accurate only for regions of patient close to the target. In the current HN plan evaluation study, lower doses (under 20% of the prescription dose) were not included . For this reason, it was decided to exclude the modelling of low dose effects. The main scope of this gUIDE computation is intended only a benchmark for comparison with the actual dose distribution obtained in plans described in chapter 7.

5.3.3 *gUIDE tuning setup*

As the gUIDE dose distribution is parametrized with the fall-off variable detailed in eq. 5.3 this needs to be 'tuned'. A tuning and validation setup was devised, using a simplified geometry: a cylindrical phantom virtual scan with homogeneous density equal to water, with a diameter of 32 cm and a length of of 34 cm (see Figure 5.2 for transversal view). Following Ahmed *et al.*'s example [13], two model geometries were employed. In both configurations, the centres of the targets were placed at the centre of the phantom, with the OAR placed next to the target. The two studied configurations are shown in Figure 5.2. In the first , (configuration A) the PTV (i.e. the target) consisted of a cylinder with a diameter of 8 cm, with a cylindrical OAR next to it with a diameter of 4 cm. The OAR was placed tangential to the target's surface to explore how steep the dose descent would be if the system's priority forced it to a single direction. The second setting (configuration B) had a similar geometry, with a cylindrical PTV with a diameter of 5 cm, with a cylindrical OAR next to it with a diameter of 3 cm. These two configurations were used to model the dimensions of PTVs and OARs typical of a HN site, where the OARs are very close to the targets. For other sites, different configuration setups to tune the dose fall-off parameter might need to be considered (i.e. a different target site with different OARs for sites like breast).

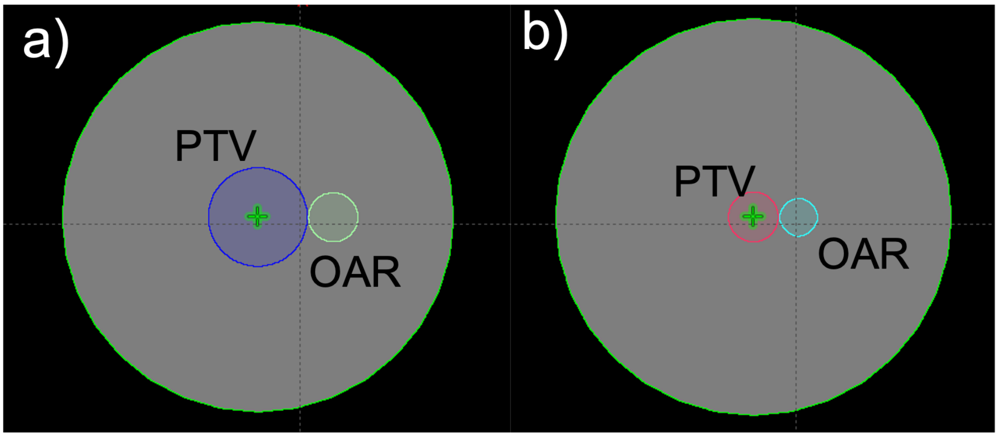


Figure 5.2: Configuration A (a) and configuration B (b) of simplified geometry in a homogeneous cylindrical phantom, used in the model tuning and validation.

The VMAT plans were generated using 6MV beams from a TrueBeam linear accelerator, using Eclipse TPS v17.1. Two full arcs were used for both configurations with collimator rotations of 330° and 30°. A 2 mm voxel grid was used for the plan dose calculation, employing the Acuros External Beam v17.1 dose calculation algorithm. The dose prescription in these model geometries was 2 Gy to be delivered to both configuration A and configuration B PTVs. The optimization objectives were the same for both configurations. The aim was to ensure near-perfect conformity of the prescription dose with its border facing the OAR, while maintaining acceptable prescription dose elsewhere. This point was achieved by requiring dose homogeneity in the PTV voxels to be maintained within ±10%. This value is higher than that normally sought in clinical practice, but it was the best compromise between the conflicting objectives to minimize the OAR dose and PTV dose homogeneity with a large range of trade-offs. For the OAR, the goal was to drive the mean dose as low as possible. The parameters reported in Table 5.1 were used in the optimization module of the TPS. As shown in the table, priority was given to the PTV. The use of the Eclipse TPS Normal Tissue Objective (NTO) function was also employed.

After the manual optimization phase, the MCO trade-off module (see chapter 3 for details) was used to extract a Pareto solution between the two trade off parameters (target homogeneity and OAR sparing). After attaining homogeneity,

Structure	Volume (%)	Dose (Gy)	Priority
PTV	0	2.03	150
PTV	100	2.97	145
OAR (mean dose)		0.4	100
BODY	0	2.03	90
NTO	--	--	130

Table 5.1: Optimization parameters used in both configurations. A and B during the gUIDE tuning.

the OAR sparing objective function element was given the most importance in the Pareto surface navigation (see chapter 3). The plan generation for each configuration of Figure 5.3 was repeated 3 times (thus in total, six plans) to assess the consistency of the solutions. After the final dose calculation was completed, the dose profiles taken in the perpendicular directions passing through the middle of the OAR (starting from the target) were extracted. All the obtained plan doses were normalized so that 95% of the total PTV volume would received the total prescribed dose (in this case, 2 Gy). Then, the gUIDE profiles were fitted using an exponential function having the same form and parameters as the one described in eq. 5.3. The steepest dose profiles were then recorded and fitted using the gUIDE equation, parametrized with the a and b parameters of eq. 5.3. Parameter a was set to a constant value of 0.01 as it takes into consideration the low-gradient effects which were not of interest in modelling as a first step, so was not used in the fit. Only the fall-off parameter of eq. 5.3, used in the dose descent in gUIDE, was employed in our tuning strategy. Said parameter was set as the mean among the values found from the 2 configurations repeated 3 times (N=6).

5.3.4 *gUIDE computation for clinical case*

The gUIDE for all of the 15 patients enrolled in the study described in chapter 3 was computed. To benchmark the gUIDE results for clinical cases, gUIDE DVHs were compared with the FDVH tool discussed previously in section 5.2. For all 15 patients, CT images, structure set and gUIDE doses were imported into PlanIQ in DICOM format and DVH comparisons were performed with feasibility dose, to check the gUIDE for clinical cases with a commercial benchmark. The DVH comparison was performed for spinal cord, brainstem, left and right parotids, oral cavity, madible, oesophagus and larynx.

5.4 RESULTS

5.4.1 *gUIDE tuning and validation for a simple case*

In Figure 5.3, the obtained dose distributions based on the two configurations described in subsection 5.3.3 are presented. The arrow shows the direction where the dose profile was recorded for the fitting.

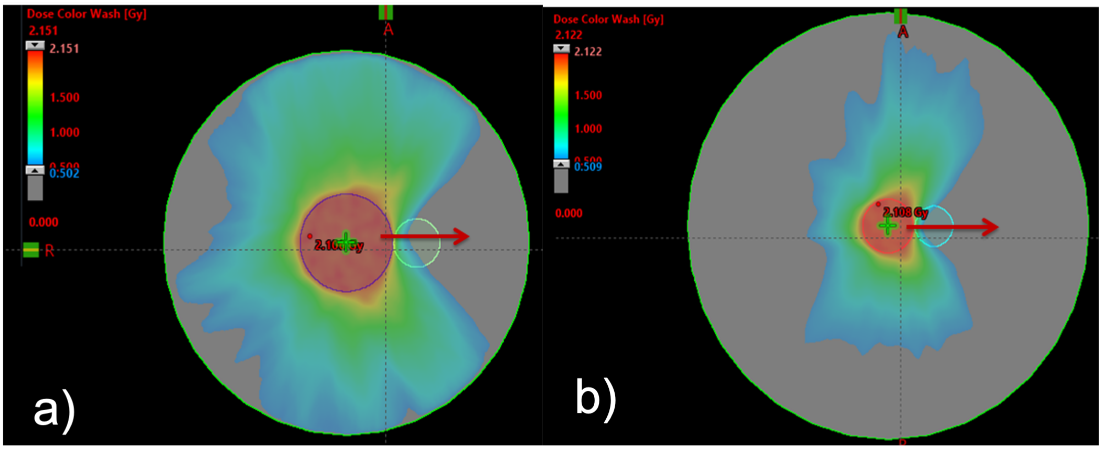


Figure 5.3: Dose distributions from two examples of the dose calculation in configuration A (a, PTV diameter= 8 cm) and B (b, PTV diameter= 4 cm). The red arrow shows the direction the dose profile was taken for the tuning of the gUIDE fall-off parameter.

In Figure 5.4 the obtained dose profiles associated with the 2 configurations (repeated 3 times, thus in total 6 lines) are reported. The graph can be divided in two parts. The first one (above the black line, shaded in orange) concerns the doses over 0.4 Gy, i.e. 20% of the maximum value (2 Gy in this case). This is the part of the graph which was used in the fit and in the validation, as it describes the fall-off parameter. The second region shows the low-gradient part which was not used in the gUIDE tuning because of the missing low gradient effect as described

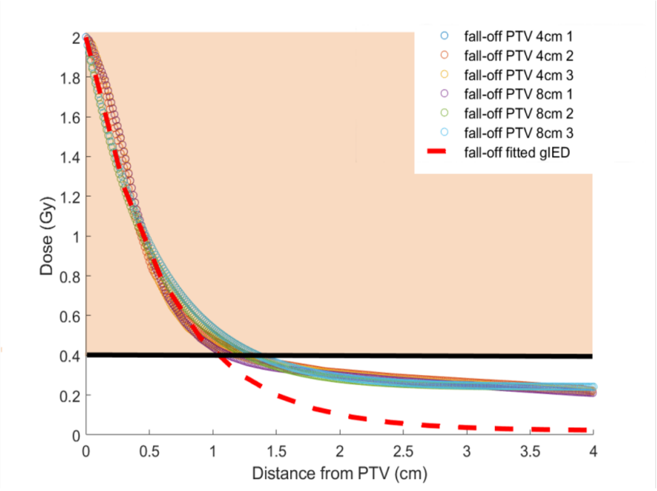


Figure 5.4: Results from the fall-off parameter tuning. The curves agree up until 0.6 – 0.4 Gy, which, in relative dose, means 30% - 20% of the maximum dose to the target (2 Gy in our setup). The shaded orange zone is the data used in the actual fit, as it described the fall-off parameter in modeling for the gUIDE tuning. The other part of the graph simply shows that, for lower doses, the gUIDE differs from the experimental data, but it is expected as in the gUIDE modeling, the low-gradient effects were not taken into consideration.

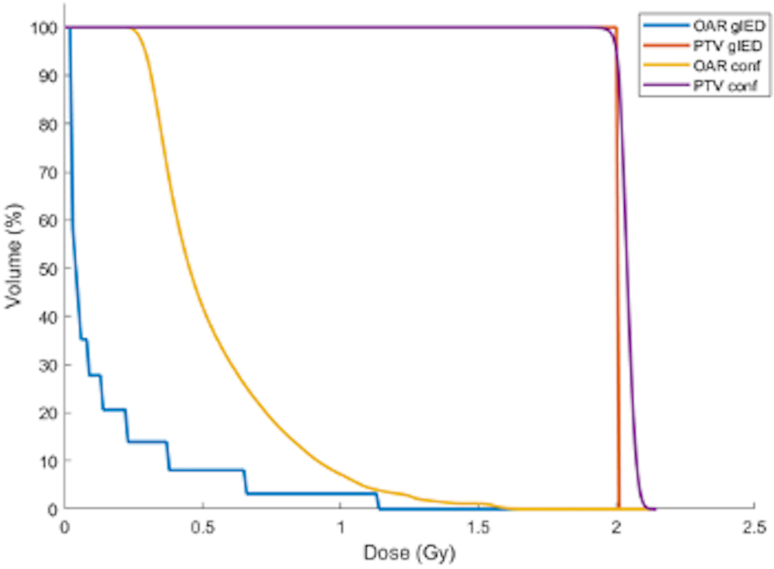


Figure 5.5: gUIDE DVHs for PTV and OAR structures (blue and red line), and configuration A DVHs for PTV and OAR (yellow and purple line). The gUIDE OAR's DVH is comprised of steps as a result of the implementation of the gradient descent using isodoses.

above. All six fits exhibited a mean $R^2 > 0.98 \pm 0.01$ and the final fall-off parameter was thus set as the mean of all the 6 fall-off parameters of the curves, i.e. b=1.9.

Figure 5.5 shown for configuration A (PTV=8 cm), the obtained DVHs related to the two involved structures. As expected, the PTV DVH for the gUIDE is a step function where all of the prescription dose is delivered to the PTV, while the OAR DVH is composed of steps as a result of the isodoses with descending values implementation. The PTV DVH of configuration A cannot reach the step function given to the gUIDE, by definition (Figure 5.1). Regarding the OAR DVHs, it is expected that the gUIDE would be lower as the tuning was performed using the steepest one dimensional dose descent in the OAR while in the real dose distribution the OARs receive the sum of various profile contributions.

5.4.2 *gUIDE validation for clinical cases*

An example of gUIDE distribution for one of the 15 patients (see details in chapter 3) is showed in Figure 5.6. Profile A and B indicated the gUIDE fall-off over targets (profile-A) and from target to normal tissue (profile-B). Profile-A showed a step function, one level for each target dose (66 Gy, 60 Gy and 54 Gy) due to the definition of gUIDE inside targets (eq.5.2). Profile-B, reflects the fall-off behaviour described in eq. 5.3 using the fall-off parameter tuned in previous subsection.

The gUIDE and planIQ doses were compared with each other in term of DVHs for the 15 patients described in subsection 3.3.1). The median DVHs for both cases, together with their 10-90 percentiles are showed for the principal OARs in Figure 5.7a. Overall, the results are quite similar, even if for some OARs the differences between the two methods are more evident, such as larynx. However, the paired two-sided Wilcoxon signed rank test on the mean doses of all 8 OARs considered in the comparison (area under the DVH curve) showed different median values for the two DVH sets, with a p-value $<<$ 0.05. Figure 5.7b shows this comparison in terms of boxplots commercial PlanIQ and gUIDE mean dose values. Although the statistical test showed significant difference between the two groups, overall values are quite similar for all OARs in the two approaches.

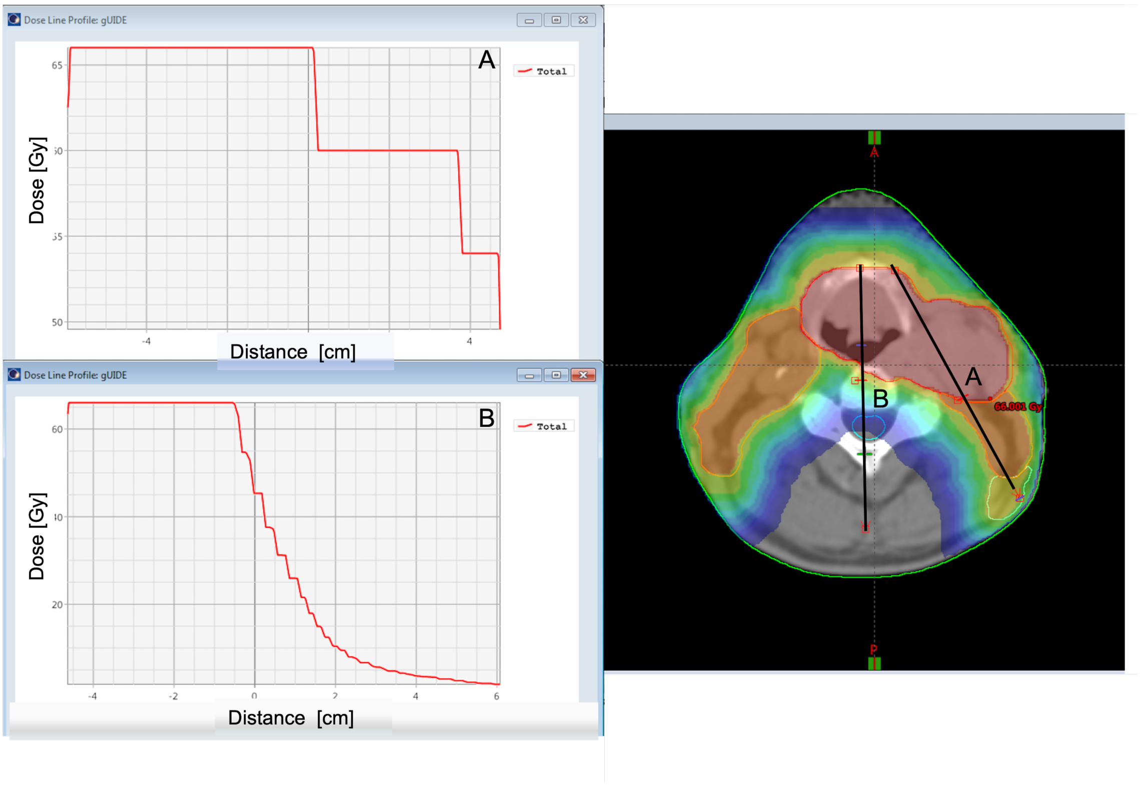


Figure 5.6: An example of gUIDE distribution for a HN patient. Profile A and B showed the fall off of the gUIDE: profile-A over the targets (3 different dose levels, 66Gy, 60Gy and 54 Gy with step profile) and profile-B from target 66Gy to normal tissue.

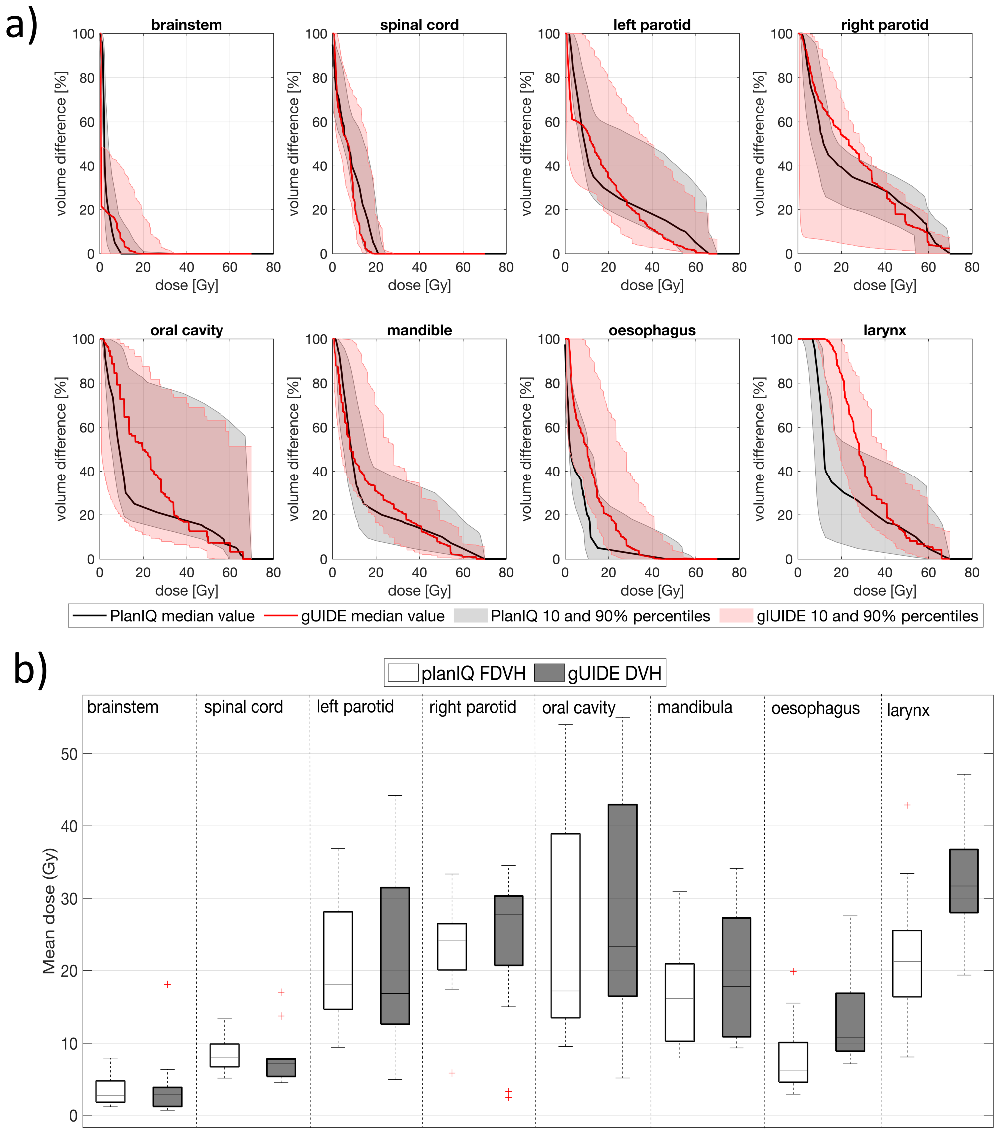


Figure 5.7: a) PlanIQ and gUIDE DVHs difference for principal OARs structure (black and red line) in term median DVHs for the 15 patients with 10-90% percentiles. b) Boxplots of planIQ Feasibility DVH area (white boxes) and gUIDE DVH area (gray boxes) distribution sorted by OAR. For each box, the central mark represents the median value, while the bottom and top edges of the box are the 25th and 75th percentiles over 15 patients, respectively. The whiskers represent the range of values. Observations beyond the whisker length are marked as outliers (+). By definition, an outlier is a value that is more than 1.5 times the interquartile range away from the bottom or top of the box. An outlier appears as a red + sign.

Part III

APPLICATION TO PLANNING AND EVALUATION

6

VARIATIONS IN HEAD AND NECK TREATMENT PLAN QUALITY ASSESSMENT

Knowledge is a deadly friend
When no one sets the rules.
King Crimson - In the court of Crimson King

6.1 PREVIEW

Work from this chapter was published in Cagni *et al.* (Front. Oncol. 2021) [95]. The figures and tables that are shown in this chapter are drawn from that published work. In this chapter a study evaluating the variability between users from a single department in plan quality assessment, is presented. There are several steps in the radiotherapy process, where human actions can bring variability in quality. One key area is contouring variations (OARs and PTV) between radiation oncologists (ROs). Concerning medical physicists (MPs), the major variation step is in the manual planning quality. Plan quality differences are usually attributed to differences between planners in planning skills, dedication, and ambition, and in time spent on planning. Another factor of variation is how perceptions of plan quality and the choice of the best plan to go to treatment could have impact in the radiotherapy process. To the best of current knowledge, this is the first study that quantitatively evaluates variations in subjective assessments of the same treatment plans by various observers (ROs and MPs) in the same department. This is also a first study showing reduced inter-observer variation in subjective plan scores for automatically generated plans compared to corresponding manual plans.

6.2 INTRODUCTION

Advanced radiotherapy delivery approaches such as Intensity Modulated Radiation Therapy (IMRT) and Volumetric Modulated Arc Therapy (VMAT) have substantially increased opportunities for sparing organs at risk (OARs) with proven clinical impact [54, 96–99]. Ideally, for each individual patient, the applied treatment plan maximally exploits the full potential of the applied delivery technique. Currently, most treatment plans are generated with interactive trial-and-error planning ('manual planning'). It is well-known that plan quality in manual planning may be sub-optimal, e,g. depending on experience and ambition of the planner, and on allotted planning time [44, 100]. In recent years, several systems for automated plan generation have been developed, often resulting in enhanced plan quality compared to manual planning [7, 82, 101–104]. Both in manual- and automated planning, human evaluation and judgement of treatment plans is crucial. Normally, plans are produced by MPs or dosimetrists and presented to treating ROs for approval. During manual plan generation, planners usually develop a range of (intermediate) plans, but generally only a single plan or sometimes two competing plans are discussed with the RO. Prior to approval, the RO may request for adaptation of presented plans. A necessary assumption for this workflow to work well, is that (unknown) disparity between planners and ROs on characteristics of good/optimal plans is absent or minor. In case of large disparity, a plan with high quality from the planner's point of view may be presented to the RO, while a different plan with lower quality according to the planner, but clearly more attractive to the RO if she/he would have been aware of it, is intentionally not generated or presented. In such cases, there is no guarantee that plan modifications are requested, and if requested, to what extent the adapted plans would satisfy the needs of the RO. This study has systematically investigated differences between five ROs and four planning MPs, all working in a single radiotherapy department, in perceived quality of head and neck (HN) cancer plans. Using automated planning, multiple plans were generated per patient. Plan quality was scored using visual analogue scales.

6.3 MATERIALS AND METHODS

6.3.1 *Patients and clinical (CLIN) treatment plans*

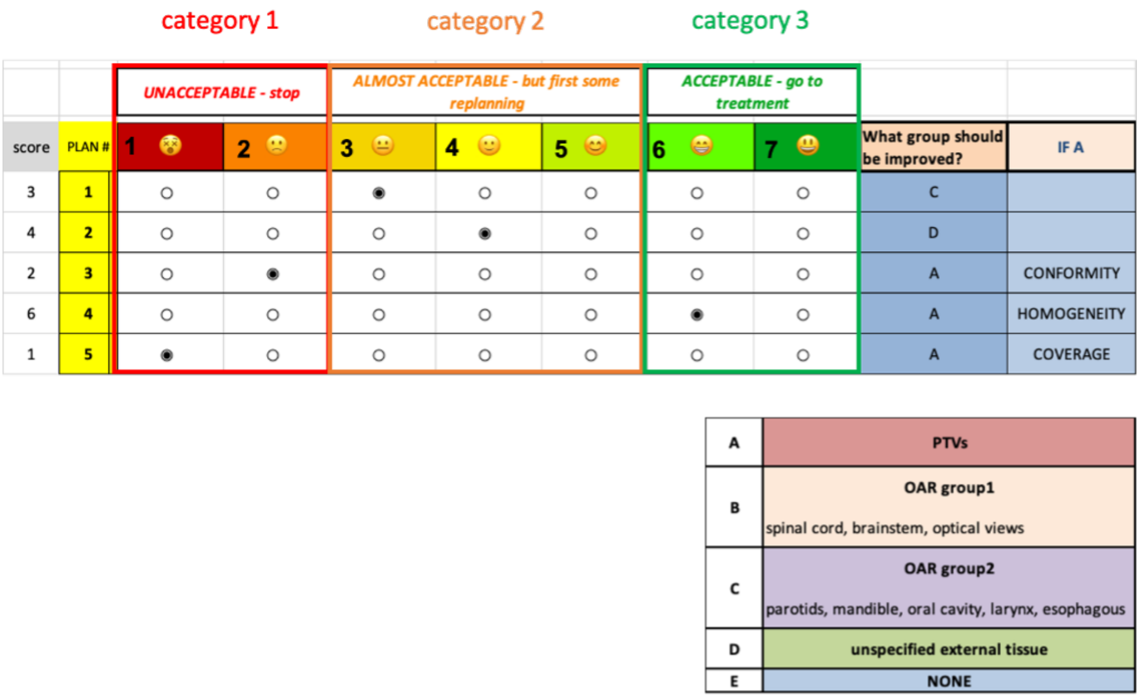
Planning CT data, contoured structures and the CLIN plan of 15 arbitrarily selected oropharyngeal HN cancer patients, recently treated with radiotherapy at Azienda USL-IRCCS Hospital (AUSL) in Reggio Emilia (Italy), were included in this study. Following AJCC TNM staging (7th edition)[105], 6 patients were classified as T2N2, 3 as T1N2, 3 as T2N1 and 3 as T4N2. Bilateral neck was irradiated in all patients. A Simultaneous Integrated Boost (SIB) technique was used for all patients, delivering the prescribed doses in 33 daily fractions. Total doses for PTVhigh, PTVmedium and PTVlow were 69.96 Gy, 59.4 Gy and 54 Gy, respectively [73–75]. For each PTV, the goal was to deliver 100% of the prescribed dose to 95% of the volume. All plans were normalized so that exactly 95% of PTVhigh received the prescription dose. Sizes of the involved planning target volumes (PTVs) were (mean±SD [min, max]): 178.5±97.3cm^3 [63.3,409.6], 208.4±105.7 cm^3 [39.8,431.7] and 184.8±51.0 cm^3 [95.2, 248.7] for PTVhigh, PTVmedium and PTVlow, respectively. OARs considered in planning were spinal cord, brainstem, left and right parotid, oesophagus, oral cavity, larynx, mandible, pharyngeal constrictor muscles, and submandibular glands [73]. Plans were generated using the following priorities for achieving planning objectives: 1) sparing of brainstem, optic chiasm, and spinal cord (so higher priority than PTV coverage), 2) achievement of PTV dose objectives in the order PTVhigh, PTVmedium, PTVlow, 3) parotid glands sparing, 4) sparing of other OARs and healthy tissues. The clinical planning protocol was largely in line with international protocols, such as RTOG [73, 74, 76] and JAVELIN protocols [75]. Patients were treated with 3-arc 6MV VMAT delivered with a Truebeam STx linac (Varian Medical Systems, Palo Alto, USA) (10 patients), or using Tomotherapy (Accuray Inc, Sunnyvale, USA) (5 patients). Clinical planning was performed with the Eclipse treatment planning system (TPS) vs.13 (Varian Medical Systems, Palo Alto, USA) or Tomoplan v. 3-4 (Accuray Inc, Sunnyvale, USA).

6.3.2 *Global study design*

Apart from the CLIN plan, 2 (for 5 patients) or 4 (for 10 patients) additional VMAT plans were evaluated in this study, resulting in a total of 65 evaluable plans. The extra plans had variable plan quality and were generated with automated planning (details in subsection 6.3.5). Each of the 65 available plans was evaluated by 5 departmental ROs (3 with more than 5 years of experience in HN radiotherapy and 2 with less than one year of experience) and 4 MPs (all with more than 5 years of experience), resulting in a total of 585 subjective plan evaluations. For each patient, every observer independently gave a score to each of the 3 or 5 available plans in a single session (details in subsection 6.3.3). Scoring was blinded, i.e. observers did not know how the plans were generated. Apart from giving a quality score to each plan, observers were also asked what change they considered most desirable for improvement of the plan (without knowing whether this would be feasible or not), see also subsection 6.3.5. To assess intra-observer variability in quality scoring, 1RO and 1MP performed the entire scoring process for 65 plans a second time, with a delay of at least a month. Previous results were blinded.

6.3.3 *Plan Scoring Procedure*

For each patient, all available dose distributions were simultaneously imported into the Eclipse TPS and linked to a virtual plan without any mentioning of the original delivery approach (VMAT or Tomotherapy), plan geometry, machine parameters, etc. With all plans simultaneously open, the observer gave a separate 1-7 score to each plan, following the routine procedure for plan evaluation (inspection of 3D dose distribution, DVH data, etc.), with higher scores pointing at perceived higher quality: 1-2: unacceptable (plan category 1), 3-5: acceptable if further planning would not have resulted in a better plan (this planning was not performed in this study) (plan category 2), 6-7: acceptable, no further planning needed (plan category 3). A 7-point scale was chosen because of good performance in psychometric literature [106–108]. In the remainder of this paper, the 1-7 scores are denoted ‘raw’ scores, while plan categories 1-3 define the more intuitive ‘category’ scores. The applied division of the raw scores in categories



		category 1		category 2			category 3			
		UNACCEPTABLE - stop		ALMOST ACCEPTABLE - but first some replanning			ACCEPTABLE - go to treatment			
score	PLAN #	1 😲	2 ☹	3 😐	4 🙂	5 😊	6 😁	7 😃	What group should be improved?	IF A
3	1	○	○	●	○	○	○	○	C	
4	2	○	○	○	●	○	○	○	D	
2	3	○	●	○	○	○	○	○	A	CONFORMITY
6	4	○	○	○	○	○	●	○	A	HOMOGENEITY
1	5	●	○	○	○	○	○	○	A	COVERAGE

A	PTVs
B	OAR group1 spinal cord, brainstem, optical views
C	OAR group2 parotids, mandible, oral cavity, larynx, esophagous
D	unspecified external tissue
E	NONE

Figure 6.1: Sheet used for plan scoring and for indicating the most important suggestion for plan improvement. a) Left part: filled-in scoring sheet of one of the observers for one of the study patients with 5 available plans. The same (empty) sheet was used by all observers for all patients. For each patient plan (first, yellow column), an observer had to choose a score between 1 and 7, with 1 and 7 lowest and highest quality, respectively. The scores were divided in three categories with explanations in the first row of the sheet. A score in category 2 meant that the plan would be acceptable if further planning would not result in desired improvements. Right part: the observer also had to express the most desired plan improvement (without knowing whether it would be feasible or not). b) Possible choices for plan improvements. PTVs (coverage, conformity, or homogeneity), OAR group1 (spinal cord, brainstem, optical system), OAR group 2 (parotids, mandible, oral cavity, larynx, oesophagus), unspecified external tissue or NONE.

was made before the start of subjective plan scoring. As visible in Figure 6.1, this division was also explicitly shown to the observers while giving scores to plans. For the analyses, another scoring system was introduced as well, the so-called ‘binary’ scoring system: raw scores 1 and 2 were grouped as binary score 0 (plan is unacceptable) and raw scores 3-7 were given binary score 1 (plan is in principle acceptable). To express the most urgent need for plan improvement, the observers could choose from: (A) PTVs (coverage, conformity, homogeneity), (B) OAR group1 (spinal cord, brainstem, optical system), (C) OAR group 2 (parotids, mandible, oral cavity, larynx, oesophagus), (D) unspecified normal tissue, or (E) none. See Figure 6.1.

6.3.4 *Evaluation of inter-observer differences in plan scoring*

With 9 observers, there were in total 36 unique combinations of two observers, here designated ‘pairs’. To analyze inter-observer differences in perceived plan quality, for all these observer pairs, percentages of agreement and disagreement in the scores given to the 65 evaluated plans were established. Analyses were partially based on raw scores, category scores and on binary scores. Observed percentages of agreement in RO-RO pairs and MP-MP pairs were compared to percentages of agreement in RO-MP pairs. Suggested most desired plan improvements were used to generate for each observer separately a frequency analysis of provided suggestions for the 65 evaluated plans.

6.3.5 *Automatically generated MCOa and MCOx plans*

Autoplans were generated with the Erasmus-iCycle system for fully-automated multi-criterial optimization (MCO) [65, 103]. Plan optimization in Erasmus-iCycle is based on so-called wish-lists (WL), containing hard planning constraints, and planning objectives with goal values and assigned priorities. A dedicated wish-list is needed for every treatment site. In essence, the wish-list defines an optimization protocol for automated multi-criterial generation of a single Pareto-optimal treatment plan for each patient. The aim in wish-list creation is to maximally ensure highest clinical quality of the generated Pareto-optimal plans, in line with the clinical planning protocol and tradition [Appendix of [103]]. Also, in this study such a wish-list was created with input of all ROs and MPs involved in the study (WLa). In the remainder of the chapter, plans generated with WLa are denoted ‘MCOa’. These MCOa plans consisted of 23 equi-angular IMRT beams, with high similarity to VMAT and avoiding time for segmentation [77–79]. With WLa as a starting point, twenty alternative wish-lists, ‘WLx’ (x=b,c,d, . . .), were created for generation of ‘MCOx’ plans. The WLx were derived from WLa by randomly varying the priorities of PTVmedium and PTVlow objectives and of the OARs. For generation of an MCOx plan for a patient, one of the 20 WLx was randomly selected, and in addition the number of beams was randomly varied between 10 and 23. As for WLa, the 20 WLx enforced adherence to the hard planning constraints for brainstem, optic chiasm, and spinal cord, as in clinical planning (above). At the same time, the WLx allowed generation of MCOx plans with a spread in dosimetric differences compared to the corresponding MCOa plans. For patients 1-10, the CLIN plan was supplemented with the MCOa plan and 3 MCOx plans (in total 5 evaluable plans). For patients 11-15, apart from the CLIN and MCOa plan, there was 1 additional MCOx plan used in this study (3 evaluable plans in total). The switch from 5 to 3 plans is considered in the Discussion section. For putting the subjective scoring of plan quality by observers in context, dosimetrical characteristics of CLIN, MCOa and MCOx plans were analysed by mutual comparisons of dosimetric plan parameters and DVHs.

6.3.6 *Statistical Analysis*

The Shapiro test and the Student's T-test were used to assess the normality of distributions and statistical significance of dosimetric differences between plans generated with different planning approaches, i.e. CLIN, MCOa and MCOx. Wilcoxon two-sided signed-rank tests were used to assess statistical significance of mean score differences between CLIN, MCOa and MCOx. Differences were considered significant if $p<0.05$. To assess statistical significance (0.05 level) of observed percentages of agreement for the 65 plan scores of the two observers in an observer pair, binomial distributions were used to calculate probabilities of percentage agreements in case of complete uncorrelated (random) choices of the two observers in a pair. To this end, success probabilities p of 1/7, 1/3 and 1/2 were used for raw, category and binary scores, respectively. The percentages of agreement in plan scores between the two observers in observer pairs were also analysed with the Cohen coefficient (K) [109]. The relative strength of agreement between the two observers in a pair is dependent on the calculated K value. Landis and Koch [110] have proposed the following classification: $K < 0$, agreement 'poor', $0 \leqslant K \leqslant 0.2$ agreement 'slight', $0.2 < K \leqslant 0.4$ agreement 'fair', $0.4 < K \leqslant 0.6$ agreement 'moderate', $0.6 < K \leqslant 0.8$ agreement 'substantial and $0.8 < K \leqslant 1$ agreement 'almost perfect'. For binary scoring the number of samples for unapproved status was not enough to achieve significant confidence limits in Cohen coefficients for many evaluators [111]. Therefore, Cohen analyses were only performed for raw and category scores. One-way Anova tests were performed to assess statistical significance of differences in percentages of agreement between subgroups of observers: 1) only RO-RO, 2) only MP-MP and 3) only RO-MP pairs, after having assessed the normality of the distribution with the Kolmogorov-Smirnov test. The Bartlet test was used to test the homogeneity of variance. When ANOVA assumptions were not met, the Kruskal-Wallis rank sum test was used as non-parametric alternative to one-way ANOVA. The Wilcoxon signed-rank test was used to test significance of differences in agreement between CLIN and MCOa plans.

6.4 RESULTS

6.4.1 *Dosimetric differences between CLIN, MCOa and MCOx plans*

In panels a) and c) of Figure 6.2, median DVHs for the CLIN, MCOa and MCOx plans are presented, showing for each dose, the corresponding median volume in the considered plans.

For individual patients, the DVH differences between the CLIN, MCOa and MCOx plans were pairwise quantified by generating differential DVHs: volume differences as a function of dose. Median volume differences and 10% and 90% percentiles are presented in panels b) and d) of Figure 6.2. The 10% and 90% percentile curves point at large inter-patient variations in DVH differences between CLIN, MCOa and MCOx plans. Table 6.1a shows how the DVH differences translate into differences in dosimetric plan parameters. Only a few of the differences between CLIN, MCOa and MCOx plan parameters were statistically different, while ranges were very broad. This is in line with the observations in Figure 6.2. Figure 6.7 P1-P15 in the (section 6.7) of this chapter I presents for each of the 15 study patients an overview of the dosimetric differences between the included 3-5 treatment plans.

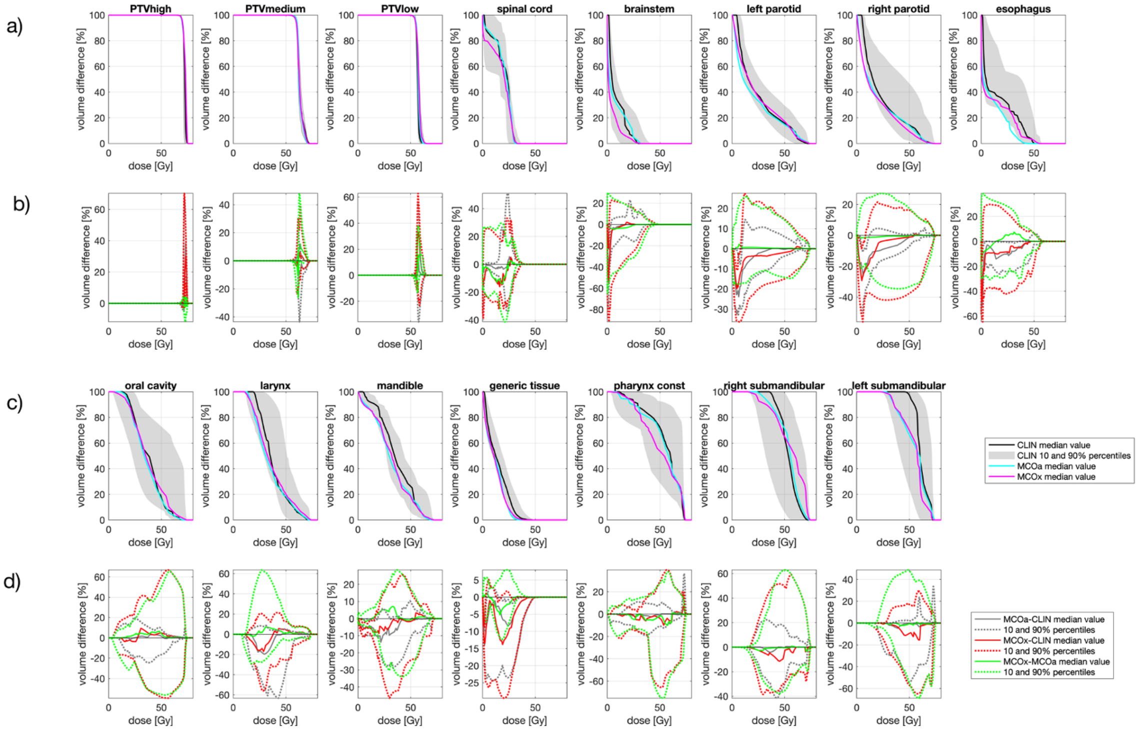
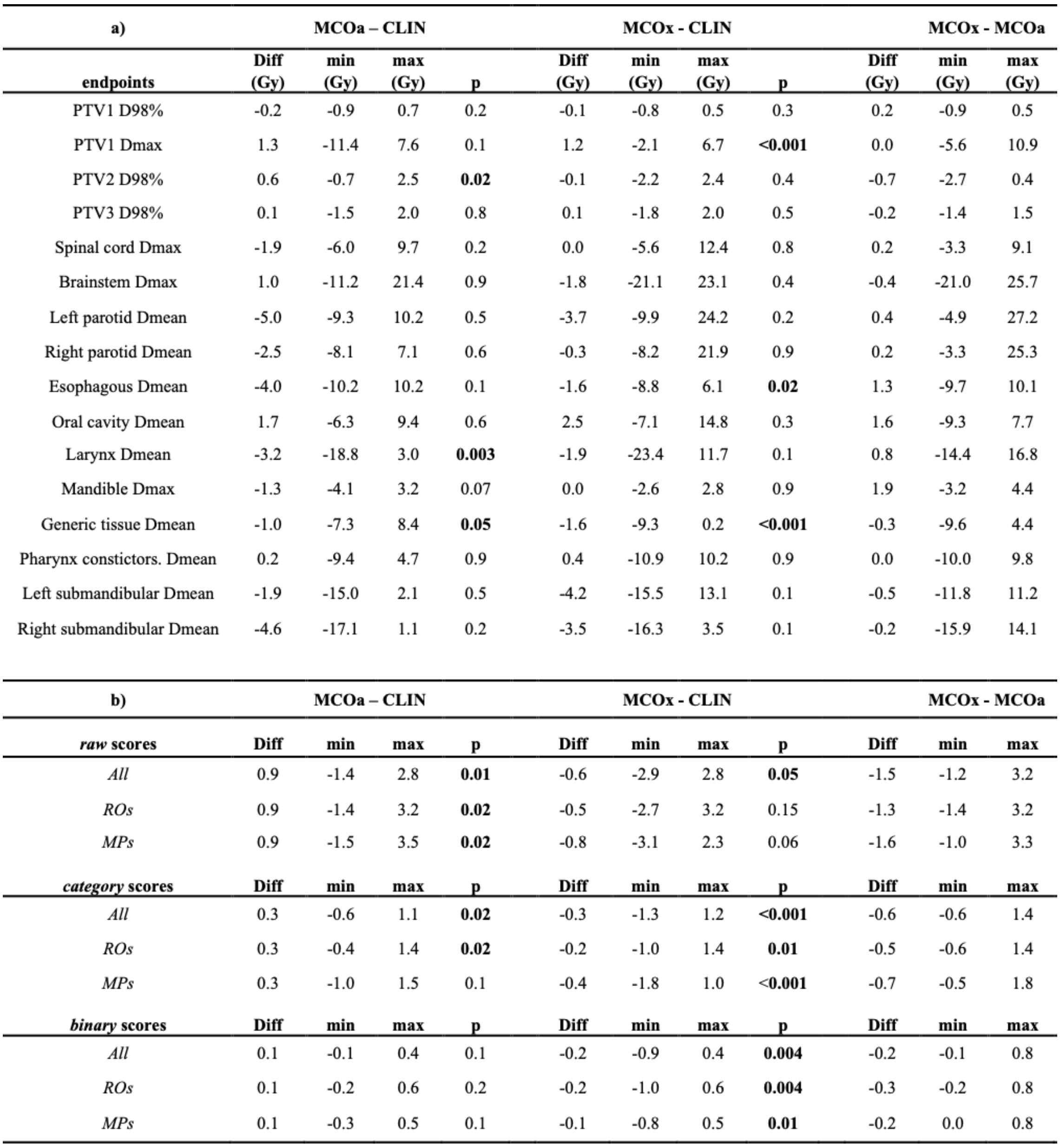


Figure 6.2: a) and c): median DVHs for the 15 CLIN with 10-90% percentiles, 15 MCOa and 35 MCOx plans. b) and d): for each dose level (x-axis), median differences in DVH volumes with 10 and 90% percentiles.

6.4.2 *Scoring for an example patient*

To introduce the type of scoring data obtained for each patient, Figure 6.3 shows the raw scores of the 9 observers for the CLIN, MCOa and MCOx plans of study patient 13, a patient showing large scoring variations. The majority of observers (6/9) selected MCOa as the best plan, while MCOx was selected most as the worst plan (5/9). This ranking of MCOa and MCOx is in line with the applied wish-lists for generation of these plans (subsection 6.3.5). However, for all three plans, there were large inter-observer differences in raw scores, (2-5 for MCOx and 2-6 for CLIN and MCOa). RO4 scored the clinically delivered CLIN plan as unacceptable, while for MP1 this plan was acceptable without further planning attempts. For RO3, MCOa was unacceptable, while for MP2 it could be delivered straightaway. Figure 6.3 also shows large inter-observer differences in score ranges. As demonstrated in the group analyses below, large scoring variations were observed for all patients and the vast majority of plans.



a)	MCOa – CLIN				MCOx - CLIN				MCOx - MCOa		
endpoints	Diff (Gy)	min (Gy)	max (Gy)	p	Diff (Gy)	min (Gy)	max (Gy)	p	Diff (Gy)	min (Gy)	max (Gy)
PTV1 D98%	-0.2	-0.9	0.7	0.2	-0.1	-0.8	0.5	0.3	0.2	-0.9	0.5
PTV1 Dmax	1.3	-11.4	7.6	0.1	1.2	-2.1	6.7	**<0.001**	0.0	-5.6	10.9
PTV2 D98%	0.6	-0.7	2.5	**0.02**	-0.1	-2.2	2.4	0.4	-0.7	-2.7	0.4
PTV3 D98%	0.1	-1.5	2.0	0.8	0.1	-1.8	2.0	0.5	-0.2	-1.4	1.5
Spinal cord Dmax	-1.9	-6.0	9.7	0.2	0.0	-5.6	12.4	0.8	0.2	-3.3	9.1
Brainstem Dmax	1.0	-11.2	21.4	0.9	-1.8	-21.1	23.1	0.4	-0.4	-21.0	25.7
Left parotid Dmean	-5.0	-9.3	10.2	0.5	-3.7	-9.9	24.2	0.2	0.4	-4.9	27.2
Right parotid Dmean	-2.5	-8.1	7.1	0.6	-0.3	-8.2	21.9	0.9	0.2	-3.3	25.3
Esophagous Dmean	-4.0	-10.2	10.2	0.1	-1.6	-8.8	6.1	**0.02**	1.3	-9.7	10.1
Oral cavity Dmean	1.7	-6.3	9.4	0.6	2.5	-7.1	14.8	0.3	1.6	-9.3	7.7
Larynx Dmean	-3.2	-18.8	3.0	**0.003**	-1.9	-23.4	11.7	0.1	0.8	-14.4	16.8
Mandible Dmax	-1.3	-4.1	3.2	0.07	0.0	-2.6	2.8	0.9	1.9	-3.2	4.4
Generic tissue Dmean	-1.0	-7.3	8.4	**0.05**	-1.6	-9.3	0.2	**<0.001**	-0.3	-9.6	4.4
Pharynx constictors. Dmean	0.2	-9.4	4.7	0.9	0.4	-10.9	10.2	0.9	0.0	-10.0	9.8
Left submandibular Dmean	-1.9	-15.0	2.1	0.5	-4.2	-15.5	13.1	0.1	-0.5	-11.8	11.2
Right submandibular Dmean	-4.6	-17.1	1.1	0.2	-3.5	-16.3	3.5	0.1	-0.2	-15.9	14.1

b)	MCOa – CLIN				MCOx - CLIN				MCOx - MCOa		
***raw* scores**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**
All	0.9	-1.4	2.8	**0.01**	-0.6	-2.9	2.8	**0.05**	-1.5	-1.2	3.2
ROs	0.9	-1.4	3.2	**0.02**	-0.5	-2.7	3.2	0.15	-1.3	-1.4	3.2
MPs	0.9	-1.5	3.5	**0.02**	-0.8	-3.1	2.3	0.06	-1.6	-1.0	3.3
***category* scores**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**
All	0.3	-0.6	1.1	**0.02**	-0.3	-1.3	1.2	**<0.001**	-0.6	-0.6	1.4
ROs	0.3	-0.4	1.4	**0.02**	-0.2	-1.0	1.4	**0.01**	-0.5	-0.6	1.4
MPs	0.3	-1.0	1.5	0.1	-0.4	-1.8	1.0	**<0.001**	-0.7	-0.5	1.8
***binary* scores**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**	**p**	**Diff**	**min**	**max**
All	0.1	-0.1	0.4	0.1	-0.2	-0.9	0.4	**0.004**	-0.2	-0.1	0.8
ROs	0.1	-0.2	0.6	0.2	-0.2	-1.0	0.6	**0.004**	-0.3	-0.2	0.8
MPs	0.1	-0.3	0.5	0.1	-0.1	-0.8	0.5	**0.01**	-0.2	0.0	0.8

Table 6.1: a) Median differences between the 65 included CLIN, MCOa and MCOx plans in dosimetric plan parameters with ranges and p-values (b) corresponding differences in raw, category and binary scores assigned by the 5 ROs, 4 MPs and all 9 observers combined (All). Significant p-values are reported in bold.

6.4.3 *RO experience in HN radiotherapy and scoring*

As mentioned in subsection 6.3.2, three participating ROs had more than five-year experience in HN radiotherapy, while the other two had less than 1 year experience for this tumor site. The choice of include also ROs with limited experience (1

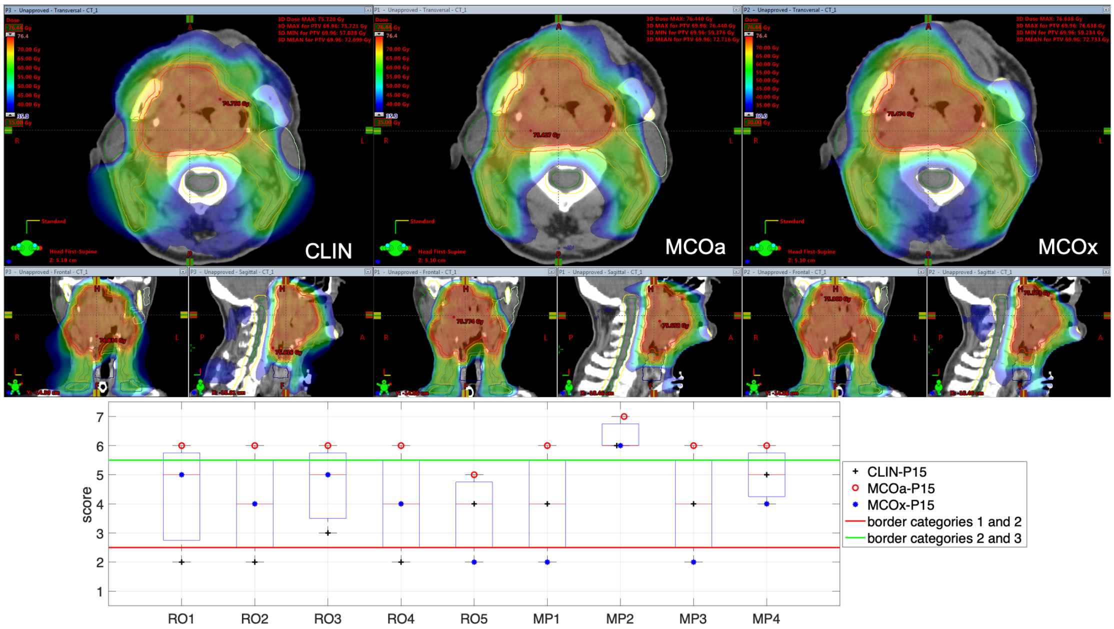


Figure 6.3: Differences in subjective plan scores among the 9 observers in the study, illustrated with an example patient (#15). Upper panels: dose distributions of the evaluated CLIN, MCOa and MCOx plans in axial, sagittal and frontal planes. Lower panel: subjective plan quality scores for the CLIN, MCOa and MCOx plans for each of the 9 observers, 5 radiation oncologists (RO1-RO5) and 4 medical physicists (MP1-MP4). Plans below the horizontal red line are considered unacceptable (Category 1). Above the green line are plans that can straightaway be delivered without any attempt to further improve the plan (Category 3). In the middle are the plans that are acceptable if further planning would not result in significant improvements (Category 2).

year) in the plan quality assessment was based on the need to consider a large number of evaluators in the study to improve the robustness of the results. The 5 ROs represent the doctors who worked at that time in HN field clinical practice at AUSL-IRCCS of Reggio Emilia at the study time. When considering the raw, category and binary scores of all 65 plans, median values for all 5 ROs/only 3 expert ROs were 28.5%/36.9% (p=0.5), 56.2% /61.6% (p=1.0) and 75.4%/75.4% (p=0.7), respectively. Based on these observations, it was decided that in further group analyses, the five ROs in this study were considered as a single group.

6.4.4 *Differences between CLIN, MCOa and MCOx plans in observer scores*

Table 6.1b reports differences between CLIN, MCOa, and MCOx in subjective scores, complementary to the dosimetrical differences in Table 6.1b. The automat-

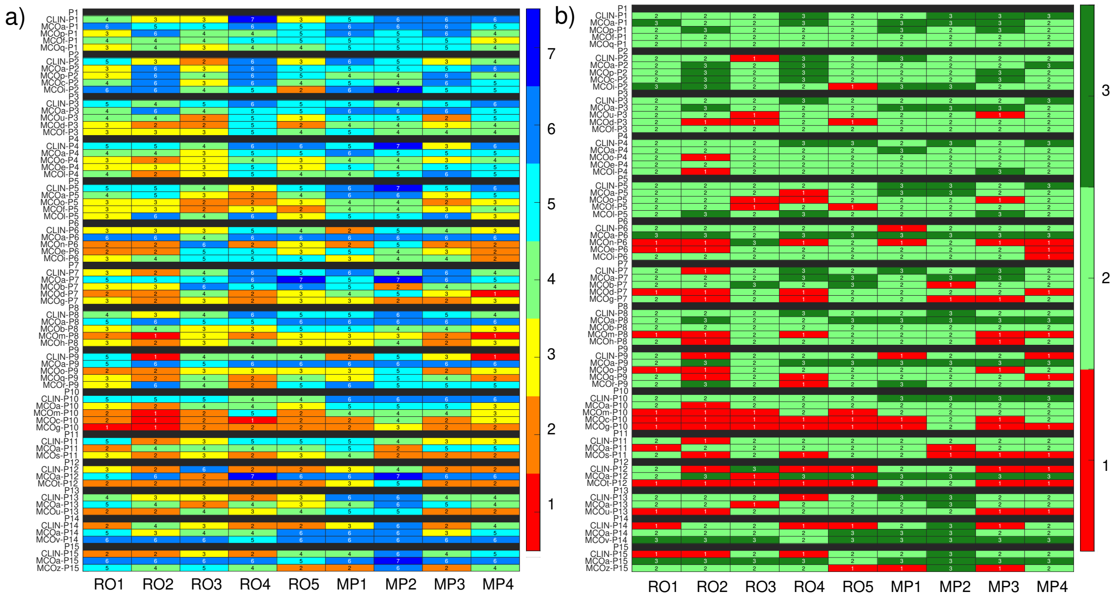


Figure 6.4: a) Raw plan quality scores (1-7, 7 indicating highest quality) of the 9 observers (x-axis) for all 65 included plans (y-axis). b) Raw plan quality scores derived from the raw scores. In b) the colour red indicates that the plan is considered unacceptable (binary score 0), while light and dark green (categories 2 and 3, respectively) indicate that the plan is in principle acceptable (binary score 1).

ically generated MCOa plans outperformed the clinically delivered CLIN plans, but for the binary scores this was not statistically significant. Score differences were overall largest between MCOa and MCOx and with smallest p-values, with the former showing highest scores, as to be expected from the respective wish-lists used for automated plan generation (subsection 6.3.5).

6.4.5 *Inter-observer variability in plan quality scores*

In line with the observations for patient #15 (above), for the majority of plans, inter-observer variations in assigned scores were large (Figure 6.4).

For the 65 evaluated plans, the average standard deviation (SD) for the nine raw observer scores was 1.06 [0.33,1.56]. For 29 of the 65 plans, all category scores (1,2, and 3) were present in the 9 scores (Figure 6.4b). For 15/65 plans, there was at least one observer that scored category 3 (acceptable without further planning attempts) while at the same time there were also observers that considered the plan unacceptable (category 1). Considering all 65 plans, the median percentage of plans declared unacceptable by an observer was 18.8±8.6% [6.2,35.4%]. For

CLIN, MCOa and MCOx plans separately, these percentages were 14.8±9.9% [0.0,33.3], 4.4±4.7% [0.0,13.3] and 26.7±12.3% [8.6,48.6], respectively. Kruskal-Wallis rank tests resulted in a statistically significant difference, with p=0.005. The Wilcoxon signed rank test showed a statistically significant difference between MCOa and MCOx (p=0.005), while for CLIN vs. MCOa p=0.1, and for CLIN vs. MCOx, p=0.2. Figure 6.5a-c show for unique pairs of two observers, the percentages of plans for which they agreed in plan score. Considering all 36 unique observer pairs in this study, the median percentage of agreement in raw plan scores was 27.7% [6.2,40.0] ('all' boxplot in Figure 6.5a).

In case of complete randomness in the scoring of two observers in a pair, an agreement percentage of 14.3% would be expected (horizontal solid line in grey zone). For category (Figure 6.5b) and binary scores (Figure 6.5c), these median percentages were 58.5% [35.4,73.8] (33.3% expected in case of randomness) and 78.5% [63.1,86.2] (50% in case of randomness), respectively. The vast majority of percentages of agreement in Figure 6.5a-c are outside the grey zones, meaning that they are statistically significantly different from the corresponding expected values for random scoring, indicated by the horizontal solid lines. With one-way Anova p-values of 0.3, 0.6 and 0.4, there were no differences between the observer pair subgroups RO-RO, MP-MP and RO-MP in the agreement distributions in Figure 6.5a-c, respectively. Cohen‘s coefficient analyses for raw scores resulted in median K-values [range] of 0.46 [0.12,0.68] when considering all observer pairs, 0.47 [0.17,0.56] for ROs, 0.51 [0.33,0.64] for MPs and 0.46 [0.12,0.68] for RO-MP. Following the labelling by Landis and Koch, the overall agreement is 'moderate'. In more details, considering all 36 observer pairs, 11% (N=4) resulted in slight agreement, 25% (N=9) in fair agreement, 47% (N=17) in moderate agreement and 17% (N=6) in substantial agreement. For category score analyses, Cohen's median K-values [range] were 0.40 [0.03,0.66] for All, 0.35 [0.04,0.53] for ROs, 0.44 [0.37,0.54] for MPs and 0.39 [0.03-0.66] for RO-MP pairs. The overall agreement, in Landis and Koch scale, resulted in 'fair'; 19% (N=7) resulted in slight agreement, 31% (N=11) in fair agreement, 47% (N=17) in moderate agreement and 3% (N=1) in substantial agreement. Figure 6.5f-h present scoring agreements for CLIN and MCOa plans separately, showing substantially better agreements for the automatically generated MCOa: when considering all

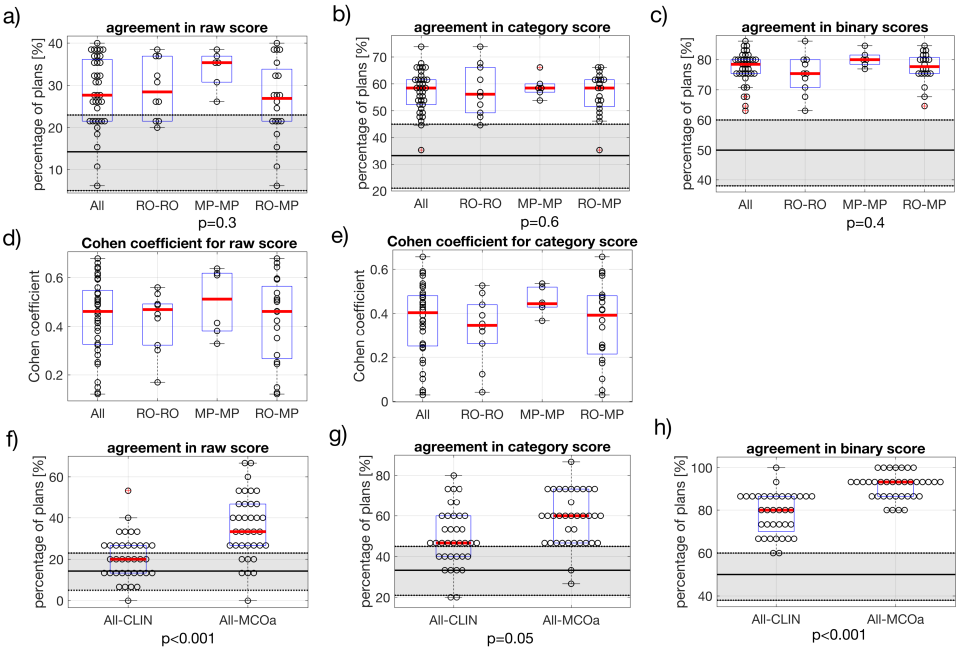


Figure 6.5: panels a), b), c): Each marker shows for 1 of the unique 36 observer pairs in this study the percentage of 65 evaluated plans for which they agree in a) raw score, b) category score and c) binary score. In each panel, the first boxplot includes the data for all 36 observer pairs (All). For the other three boxplots the data is split ac-cording to subgroups of observer pairs; RO-RO: pairs consist of 2 radiation oncologists, MP-MP: pairs consist of 2 medical physicists, RO-MP: pairs consist of 1 radiation oncologist and 1 medical physicist. Dash black line inside the dotted lines represented the Binomial distribution (expected value and 95% confidence limits), thus the random probabilities. P-values of Anova test between groups (RO-RO. . .) were reported for each score agreement type. Panels d) and e): Corresponding Cohen coefficients for raw and category scores. Panel f), g), h) agreement comparison for All pairs between CLIN and MCOa plans (automated plans with consistent wish-list) for agreement in plan raw score (f)), plane category score (g)) and plan binary score (h)). The p-values were established with 2 tailed Wilcoxon's signed rank tests. In each panel, horizontal red lines in the boxplots show median values, while the edges of the boxes are the 25th and 75th percen-tiles, respectively. The whiskers extend to the most extreme data points not considered outliers, and the outliers are plotted indi-vidually using the '+' symbol.

36 observer pairs, agreement percentages for CLIN/MCOa were 20.0%/33.3% ($p<0.001$), 46.7%/60.0% ($p=0.005$) and 80.0%/93.3% ($p<0.001$) for raw, category and binary scores, respectively.

6.4.6 *Intra-observer variability in plan quality scores*

For the RO and MP involved in the intra-observer analyses, agreement percentages for the 65 initial raw plan scores and the 65 repeat raw scores were 40.0%/52.3% for RO/MP (N=65). This is substantially higher than the expected percentage for random scoring (14.3%) and the median percentage of inter-observer score agreement of 27.7%, see Figure 6.5a. The repeat category agreements for the RO/MP were 70.8%/89.2% (N=65) with corresponding expected random agreements and median inter-observer agreements of 33% and 58.5% Figure 6.5b, respectively. For binary scoring, the RO/MP agreements were 86.2% /96.2%, with expected random and median inter-observer agreements of 50% and 78.5% Figure 6.5c, respectively.

6.4.7 *Suggested plan improvements*

Large variability between observers was also observed in the suggestions for plan improvement. Figure 6.6 shows the variability between observers for each of the possible options for improvement. Overall, the most chosen options were PTV conformity and dose reduction in parotids with median percentages of 24.6% [0.0,38.5] and 21.5% [13.8,47.7], respectively. In the intra-observer evaluations, the participating RO and MP showed agreement percentages in the request for plan improvement of 28% and 46%, respectively.

6.5 DISCUSSION

In most centres, treatment plans are prepared by dosimetrists or MPs, and evaluated for final approval by the treating RO. The process, often denoted as manual planning or trial-and-error planning, may have several iterations in which the planner adjusts intermediate plans, based on feedback by the RO. Limited common understanding or agreement between planners and ROs on how good plans should look can result in suboptimal dose distributions, even with iteration loops. This study has systematically investigated differences between five ROs and 4 planning MPs from a single radiotherapy department in perceived quality of

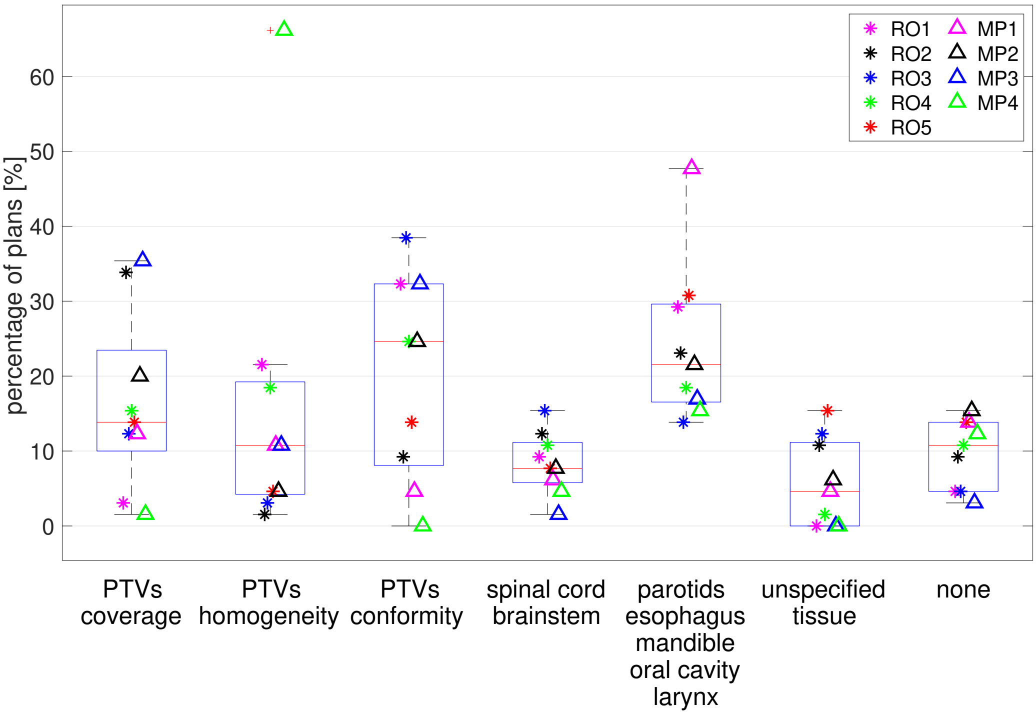


Figure 6.6: Percentages of plans (y-axis) for which plan approvement options along the x-axis were requested. Each marker indicates for a selected observer the percentage of plans for which the corresponding option for plan improvement was selected.

oropharyngeal cancer plans. To the best of current knowledge, this is the first study that systematically investigates variations in subjective plan quality assessment among ROs and MPs working in a single department. Even in this relatively small centre with ROs and MPs working closely together based on the centre's planning protocol (which is in line with international protocols, see section 6.3), large variations in subjective plan scores were observed. Considering all 36 unique observer pairs, the median percentage of plans for which they disagreed on clinical acceptability was 21.5% (Figure 6.5c), with minimum/maximum disagreements between pairs of 13.8%/36.9%. Based on Landis and Koch's labelling of Cohen's Kappa-values, the overall agreements in raw and category scores were 'moderate' and 'fair', respectively, but large variations between observer pairs were observed, going from 'slight agreement' to 'substantial agreement'.

As shown in Figure 6.5b-d and Table 6.1a, dosimetric differences between the CLIN, MCOa and MCOx plans could be substantial. As demonstrated in Figure 6.4a, for many observer-patient combinations these dosimetric variations resulted in large variations in the 3 or 5 plan scores. On the other hand, different observers did often substantially disagree on the score of a patient plan (see rows in Figure 6.4a). As can be observed in Figure 6.2, Figure 6.3 (section 6.7 of this chapter), and Table 6.1, dosimetric differences between patient plans, both positive and negative, were generally not restricted to one parameter or one structure. Probably, different observers often appreciated dosimetric pluses and minuses rather differently, contributing to the large disagreements between observers in assigned scores. This would be in line with the large inter-observer variations in suggested plan improvements (subsection 6.4.7). Figure 6.5a-c shows that agreement percentages for RO-RO, MP-MP and RO-MP pairs were similar (no statistically significant differences). This implies that despite large differences in training and clinical roles of ROs and MPs, there were no enhanced rates of score mismatches in RO-MP pairs compared to RO-RO pairs. Possibly, renewed, broad departmental discussions on plan requirements, aiming at a widely shared, and precisely defined view on plan quality, could improve the current large inter-observer variation in plan quality assessments. Probably also automated planning could result in improvements: as visible in Figure 6.5f-h, scoring agreements were better for the MCOa plans than for the CLIN plans, possibly related to more consistent automated generation of the MCOa plans. Apart from the better agreement between observers, MCOa scores were overall also higher than CLIN scores (Table 6.1b), and MCOa plans were less frequently considered unacceptable than CLIN plans (4.4% vs. 14.8%, p=0.1 (subsection 6.3.5). Enhanced plan quality with automated planning compared to manual planning has been observed previously (see e.g.[7, 101–104]), but to our knowledge this is the first study showing also reduced inter-observer variations in subjective plan scores for the autoplans compared to corresponding manual plans. Other studies have pointed at the use of numerical plan quality assessment tools to enhance treatment plan quality [112]. The 70.8% and 89.2% agreements in repeated category scoring and 86.2% and 96.2% in repeated binary scoring (subsection 6.3.6), point at an option for high-accuracy score prediction for single observers with machine learning. This is a topic of on-going research. Although the observers were asked to give an absolute score (1-7) to each plan, the scoring of all 3 or 5 plans of a patient in a single session could have influenced the scores for the individual plans. For example, a plan could be perceived as unacceptable in the presence of a very good alternative plan, while when scored separately, the former plan could possibly have been acceptable for the observer. Such a mechanism could in part explain the observation that 14.8% (median percentage for the 9 observers, subsection 6.3.5) of the clinically delivered plans (CLIN) were scored as unacceptable, while all CLIN plans fulfilled the clinical hard constraints on PTV coverage, spinal cord Dmax, etc. It could also explain the large difference between MCOa and MCOx in unacceptability rate (4.4% vs. 27.7% p=0.005, subsection 6.3.5), while also the intentionally suboptimal MCOx plans were generated while obeying all hard constraints (PTV, spinal cord, etc). These observations point at a weakness of current manual planning: evaluating a plan is extremely difficult if there are no alternative plans.

In this study, oropharynx cases were considered with 3 dose levels and many OARs. The complexity of these cases could have contributed to the observed large and frequent disparities in observer scores. Possibly, for less complex tumor sites, agreement in plan scores could be better, which is a topic for further research.

This is the first study that has quantitatively evaluated variations in subjective assessments of the same treatment plans by various observers (ROs and MPs) in the same department. This study is very different from, but complementary to, other studies that demonstrate that different planners can generate very different plans for the same patient, even with very detailed, quantitative instructions on how the plan should look [44]. In the latter studies, plan quality differences are usually attributed to differences between planners in planning skills, dedication, and ambition, and in time spent on planning. On the contrary, in this study all observers evaluated the same plans, and the study tested how well these plans fit the observer-specific ideas on how good plans should look. The results of the current study could stimulate similar studies in other departments as they seem to point at an important weak link in radiotherapy planning. It is commonly recognized that variations between ROs in delineating targets is a major concern in clinical radiotherapy. This study suggests that large inter-observer variations in plan quality assessments (even in a single department), could be another ‘Achilles heel’ for compromising optimal treatment.

6.6 CONCLUSIONS

Inter-observer differences in treatment plan quality assessments in radiotherapy can be substantial and could hamper consistent preparation of high-quality plans, even in a single radiotherapy department. Agreements between ROs and MPs in plan assessments were similar to agreements among ROs only, despite large differences between ROs and MPs in training and clinical role. Automatically generated plans (MCOa) showed highest median scores and best inter-observer score agreements, indicating the potential for automated planning to improve clinical practice.

6.7 CHAPTER 6 SUPPLEMENTARY MATERIAL

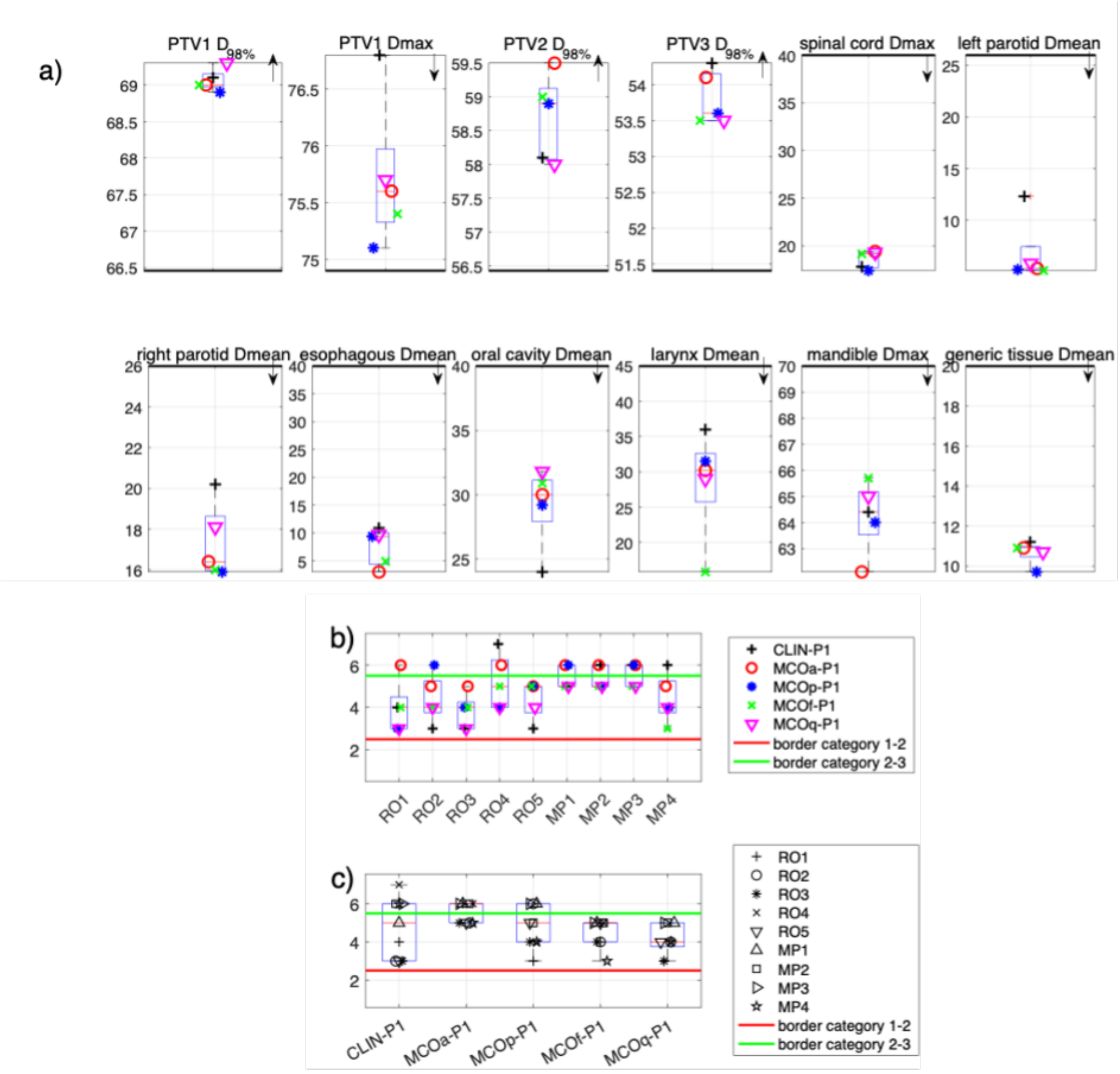


Figure 6.7.P1: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

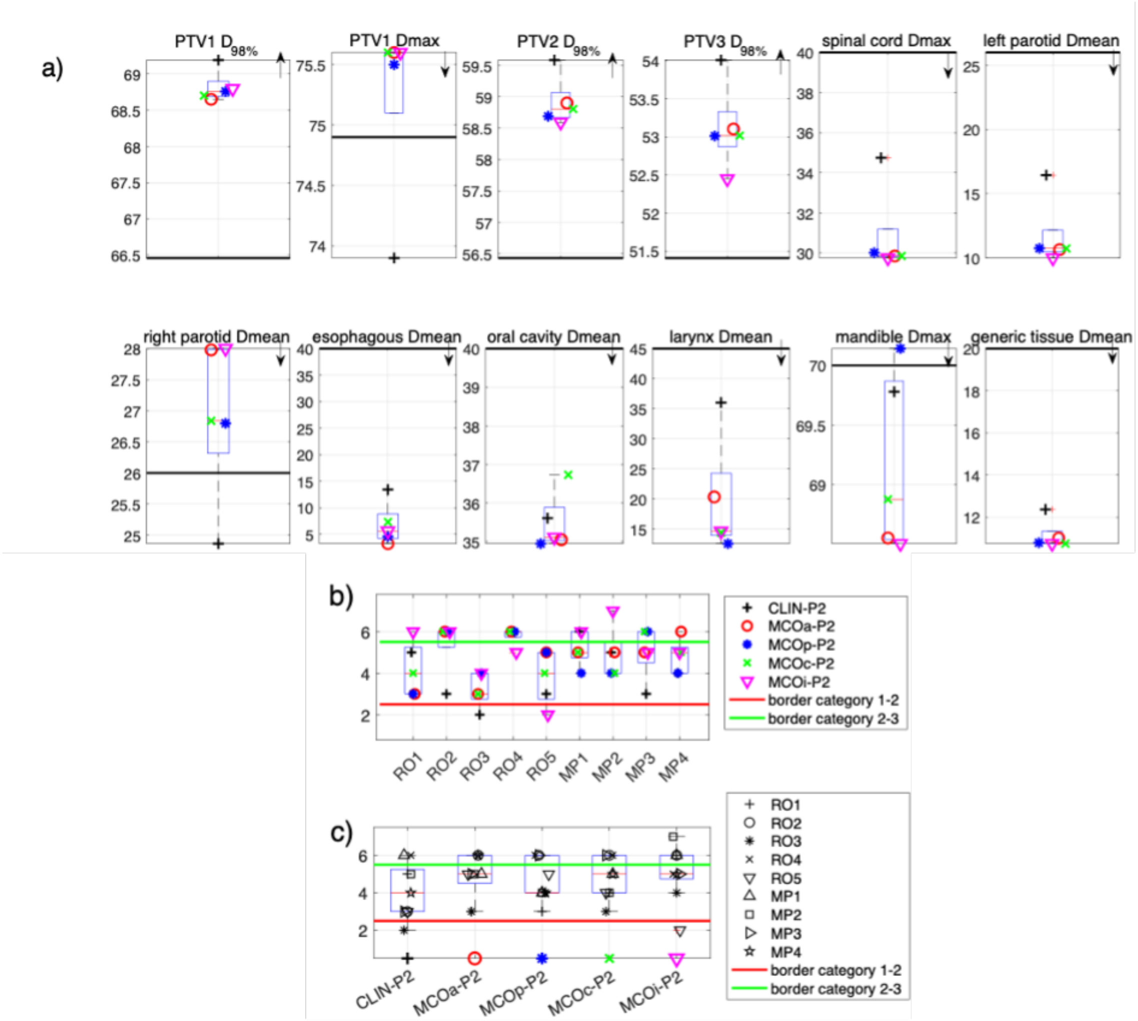


Figure 6.7.P2: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

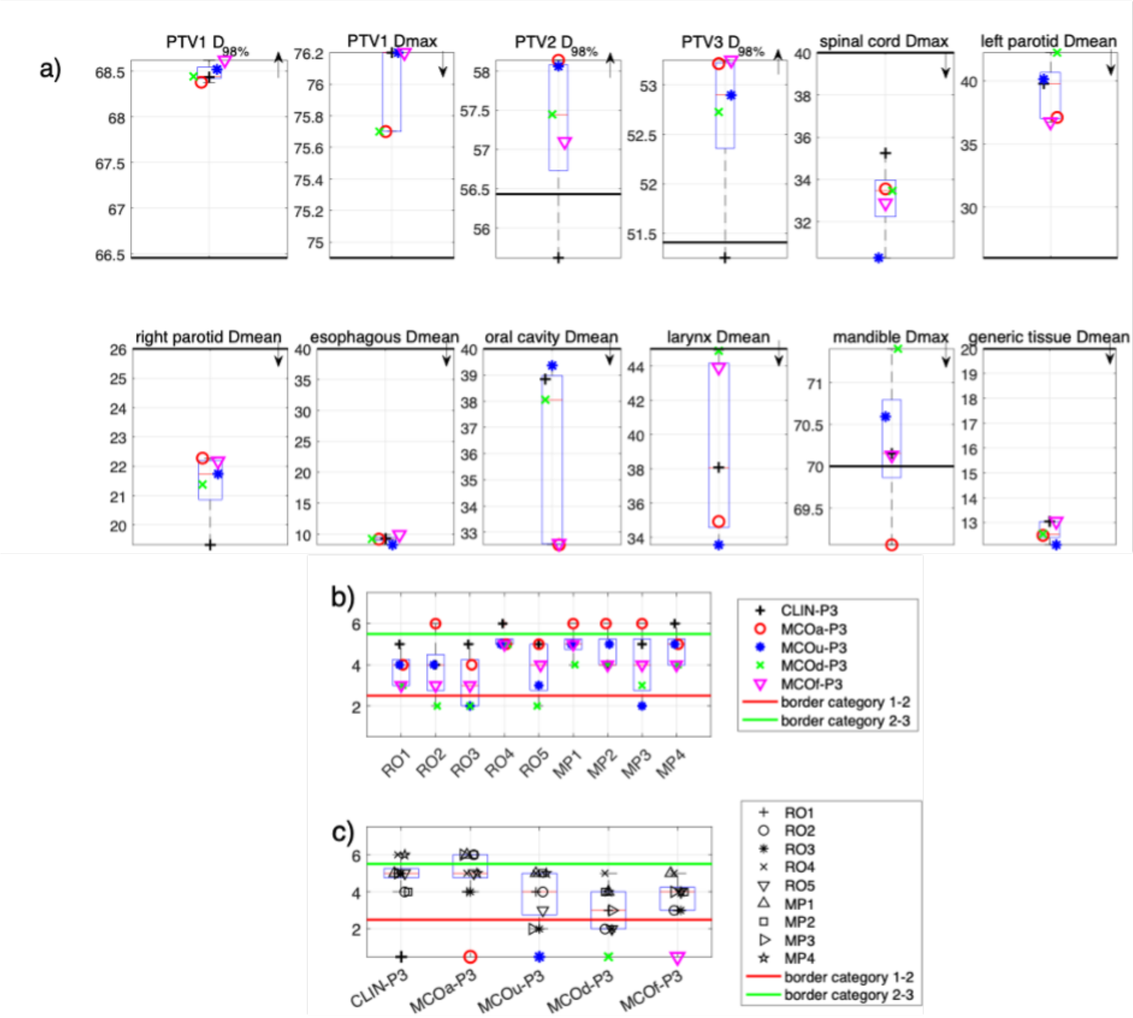


Figure 6.7.P3: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

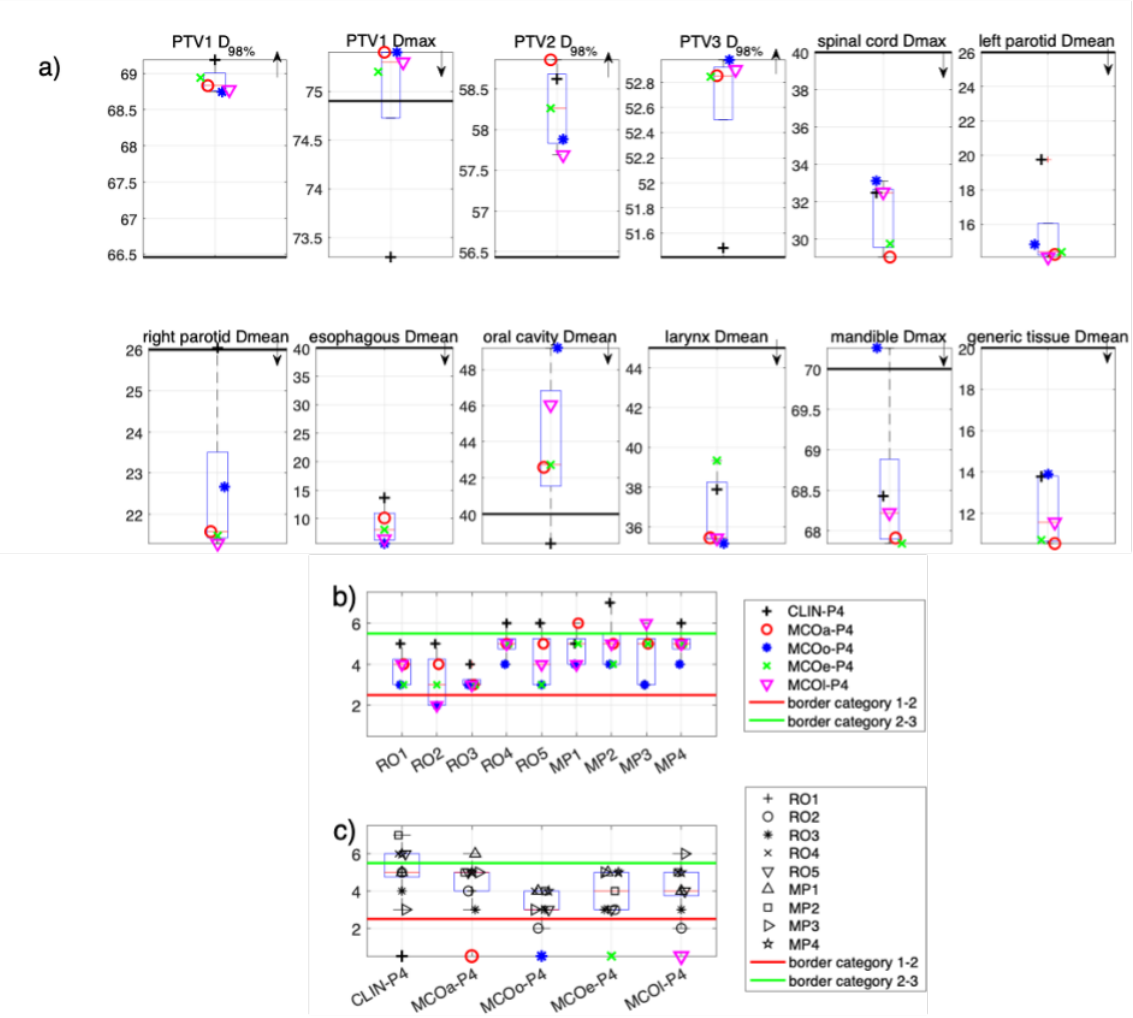


Figure 6.7.P4: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

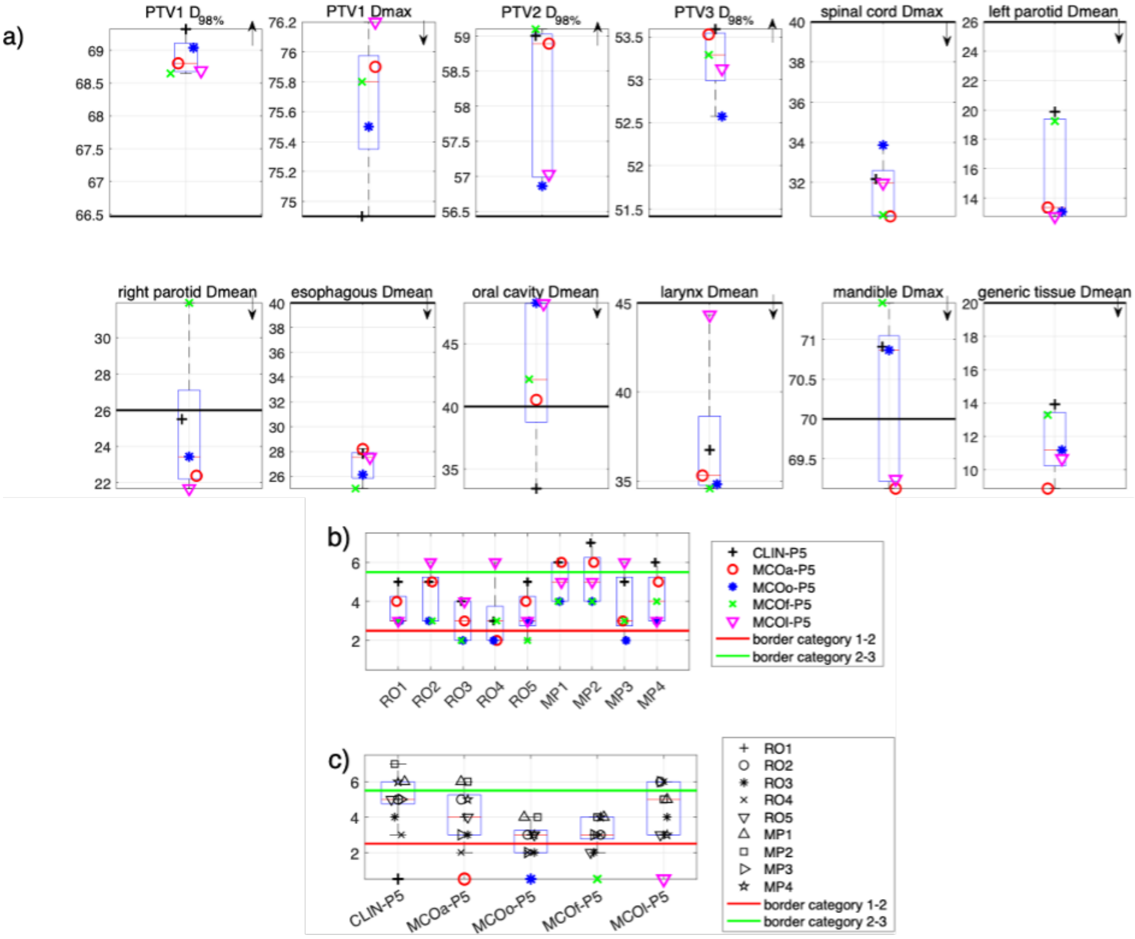


Figure 6.7.P5: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

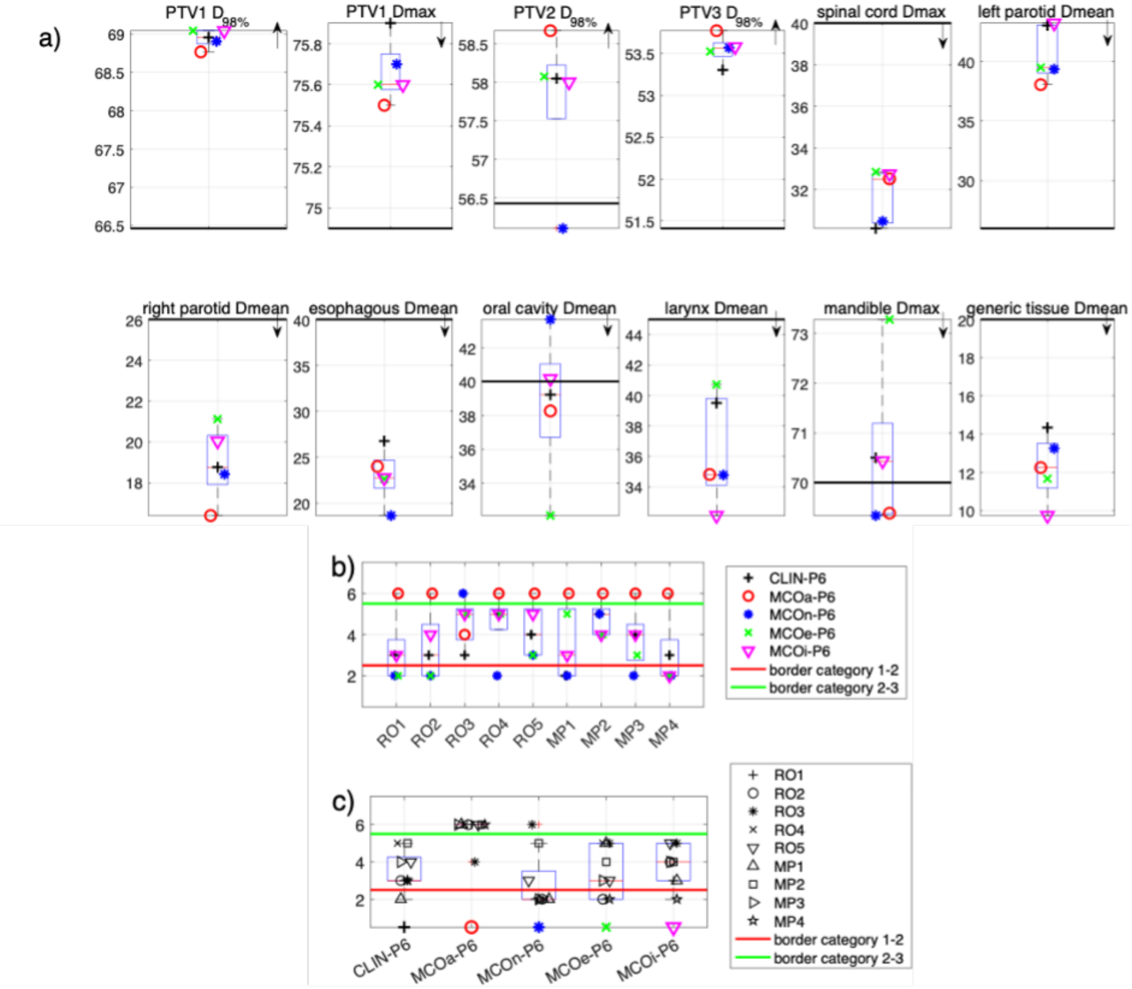


Figure 6.7.P6: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

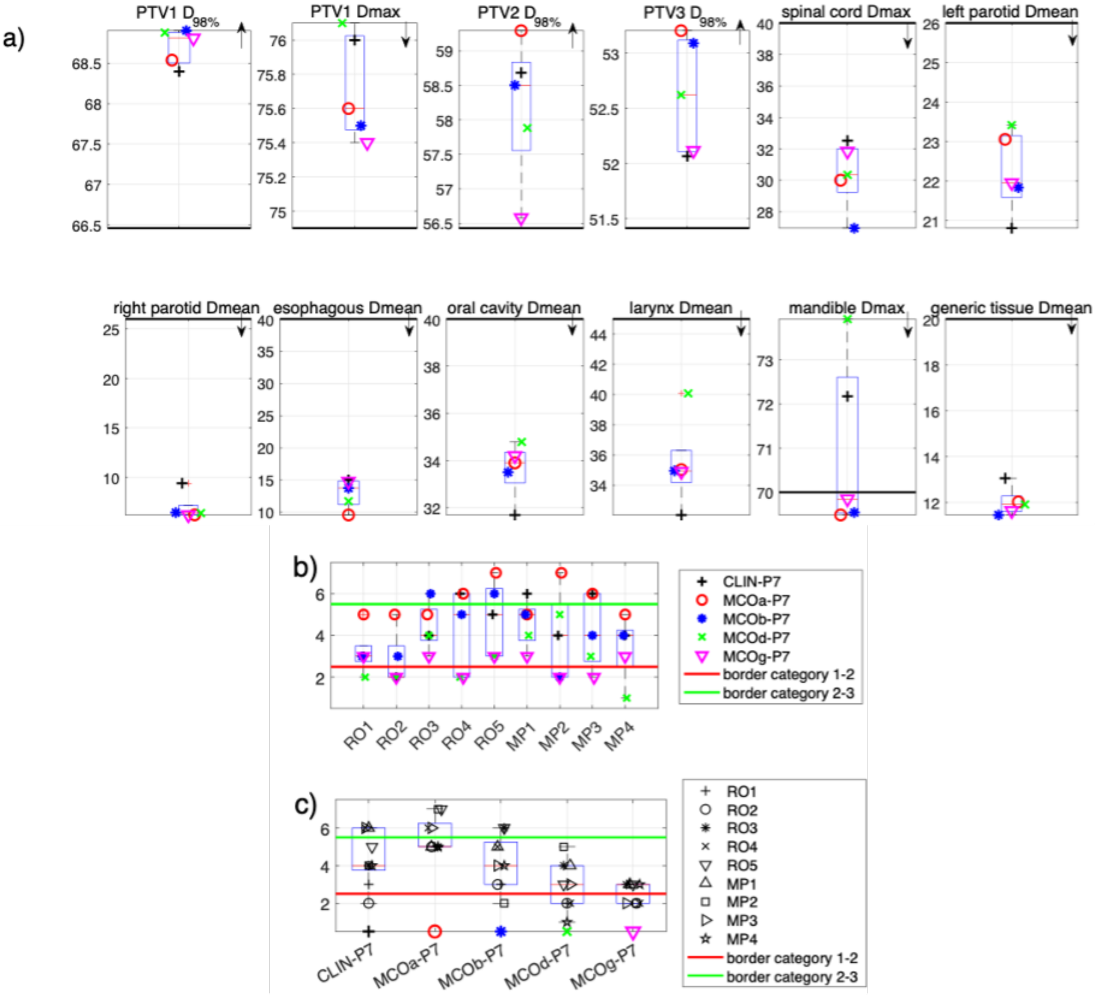


Figure 6.7.P7: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

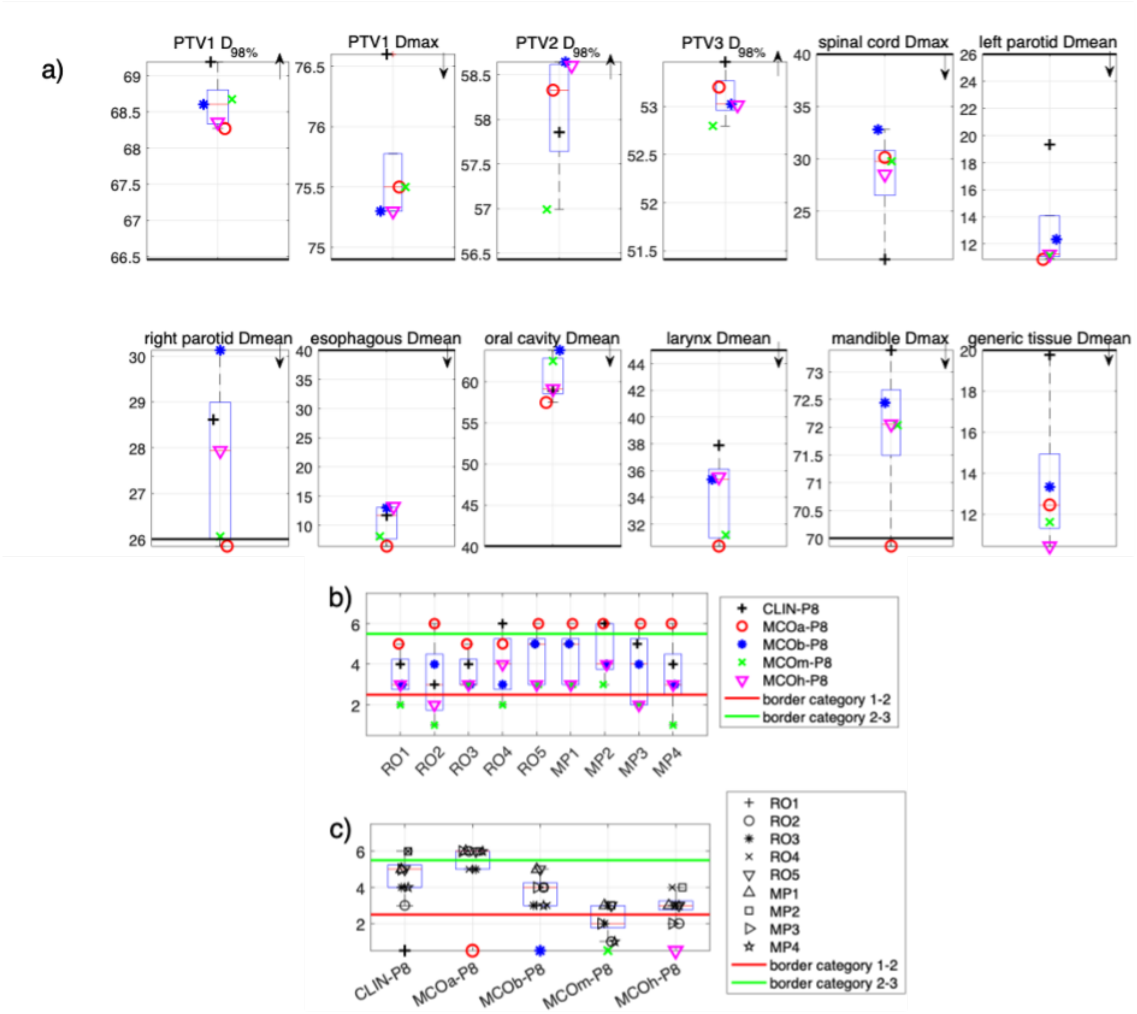


Figure 6.7.P8: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

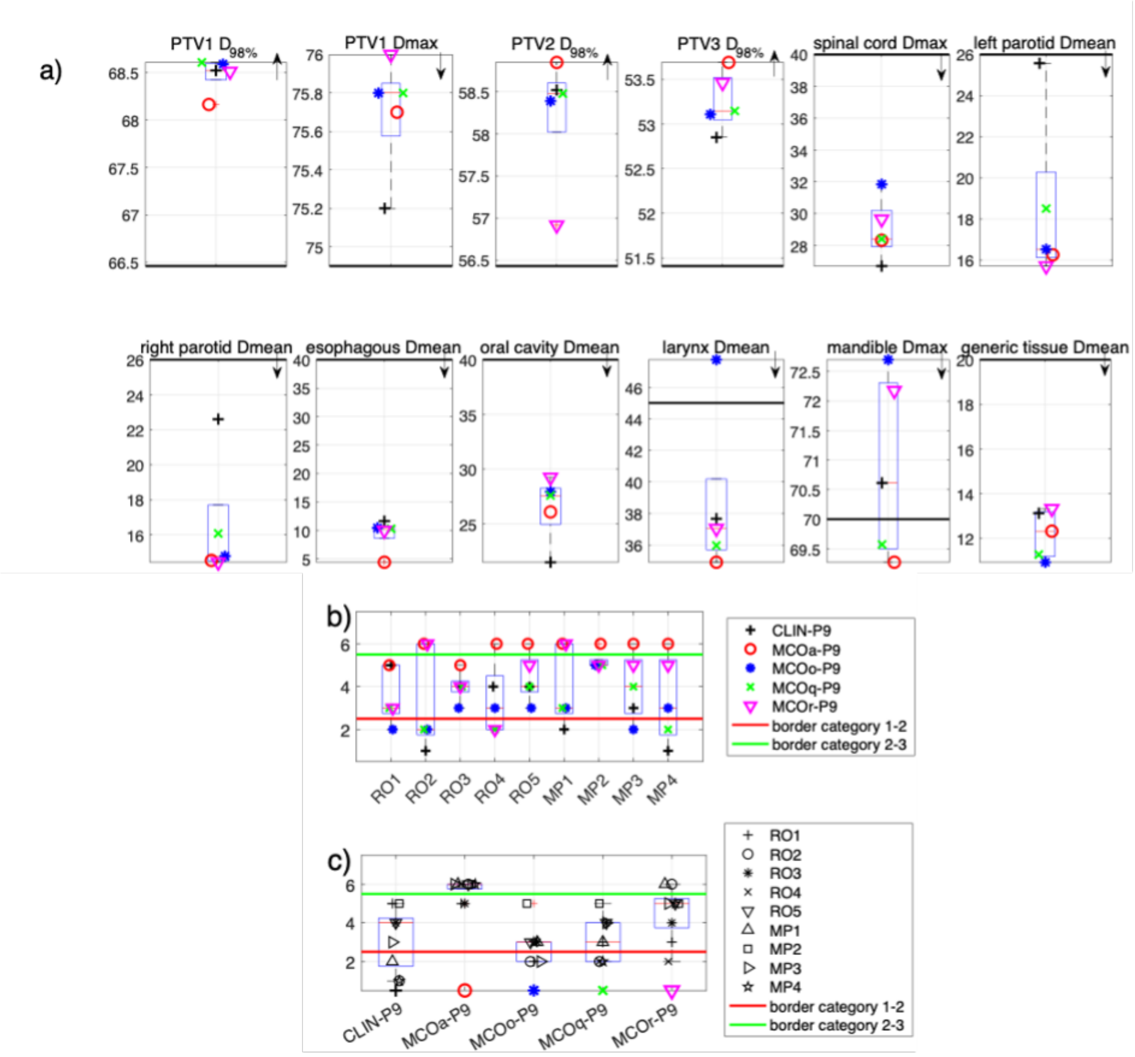


Figure 6.7.P9: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

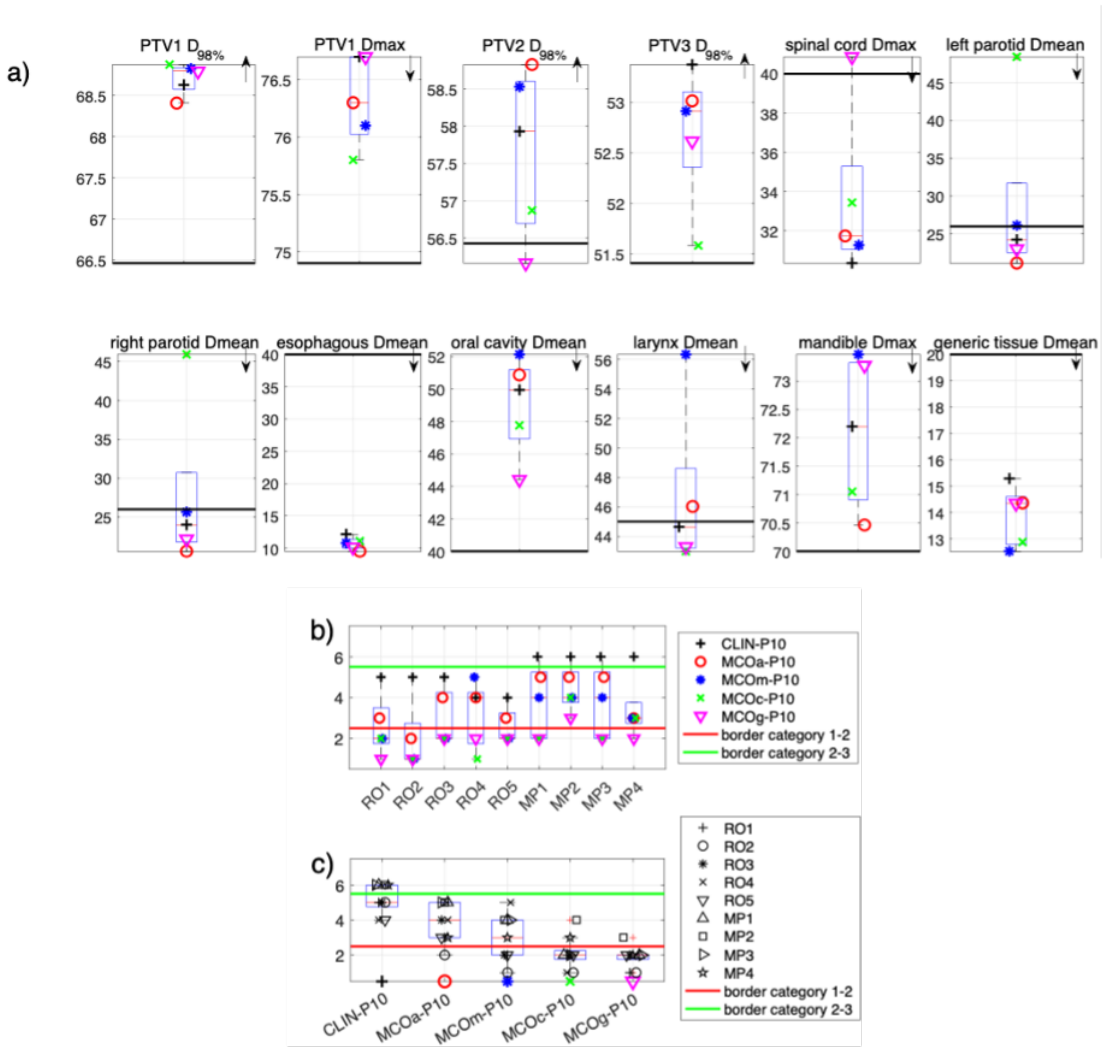


Figure 6.7.P10: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

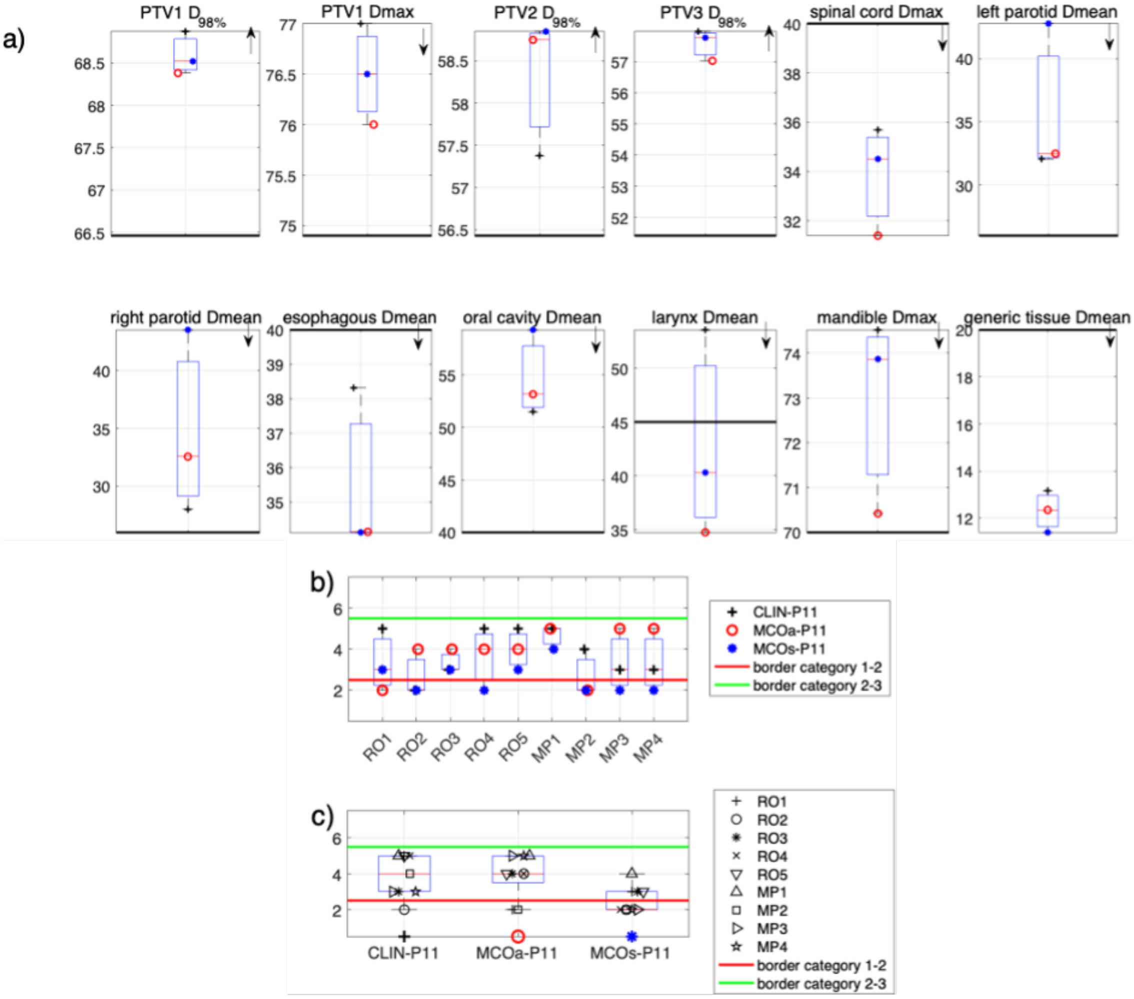


Figure 6.7.P11: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

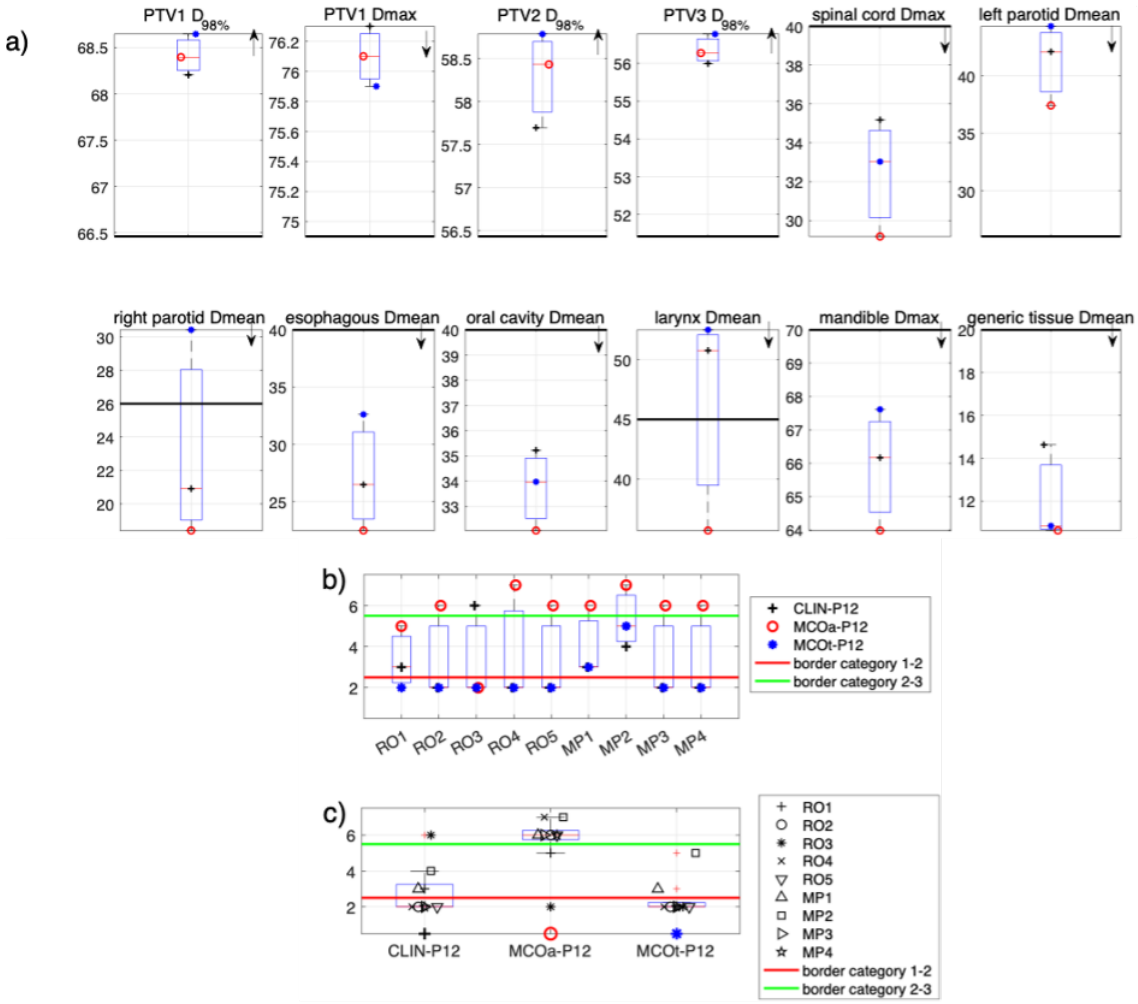


Figure 6.7.P12: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

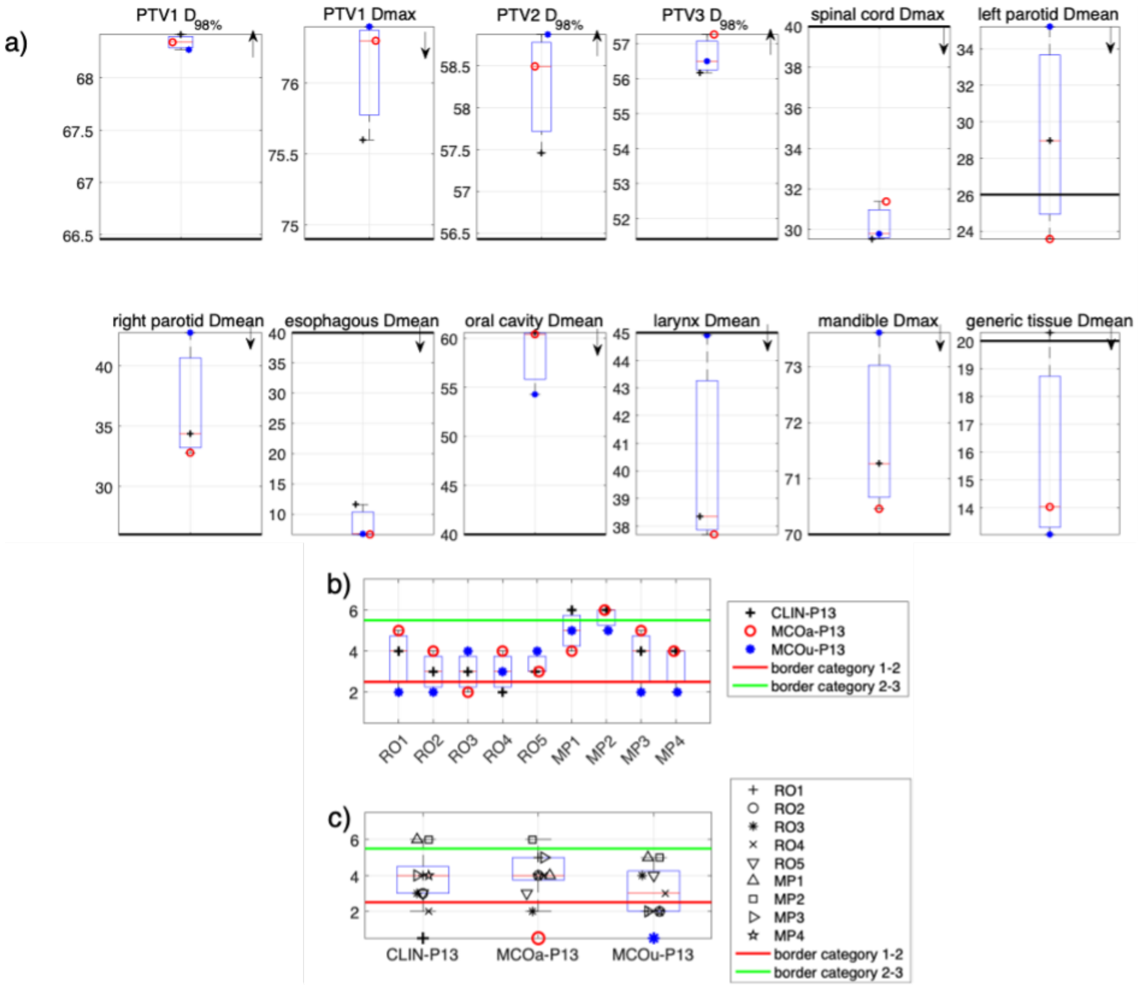


Figure 6.7.P13: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

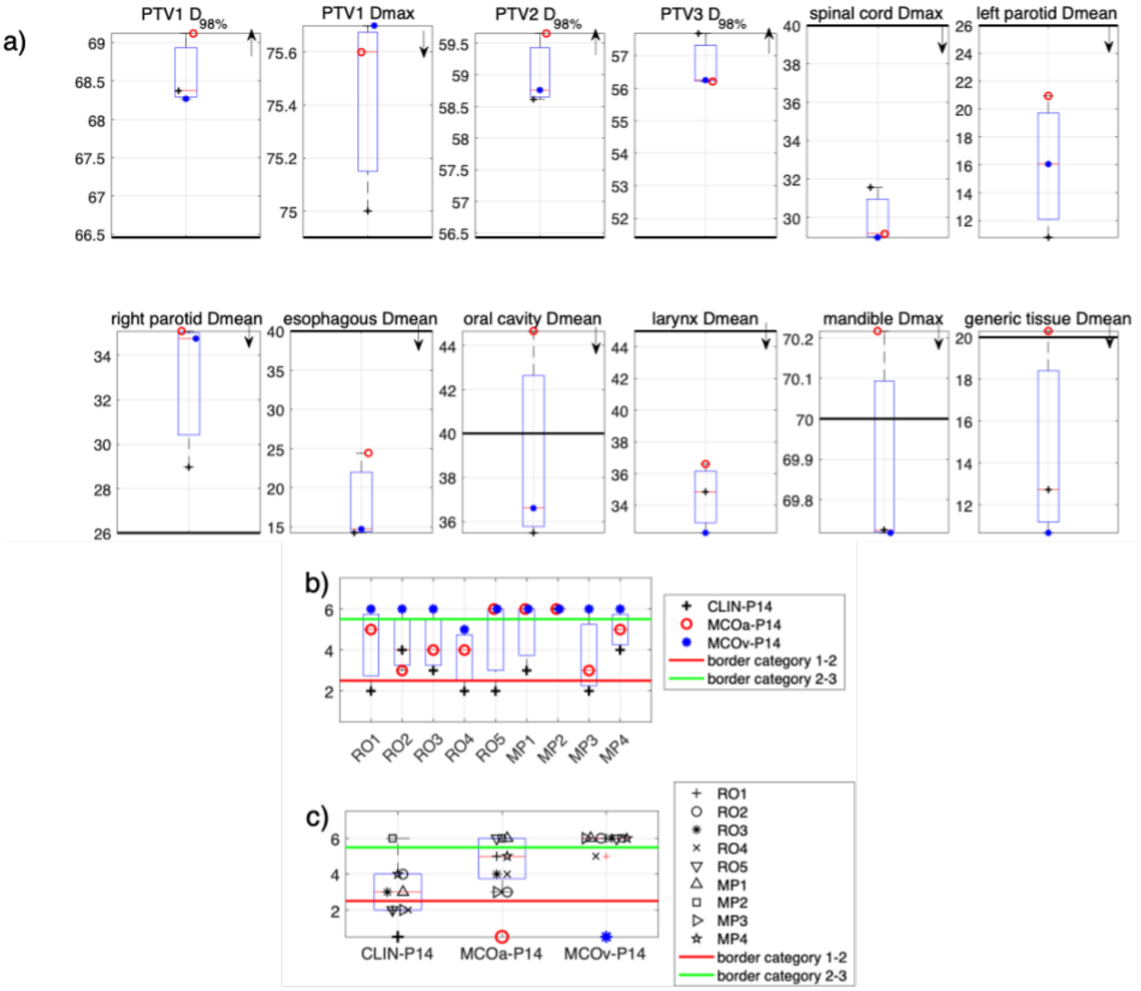


Figure 6.7.P14: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

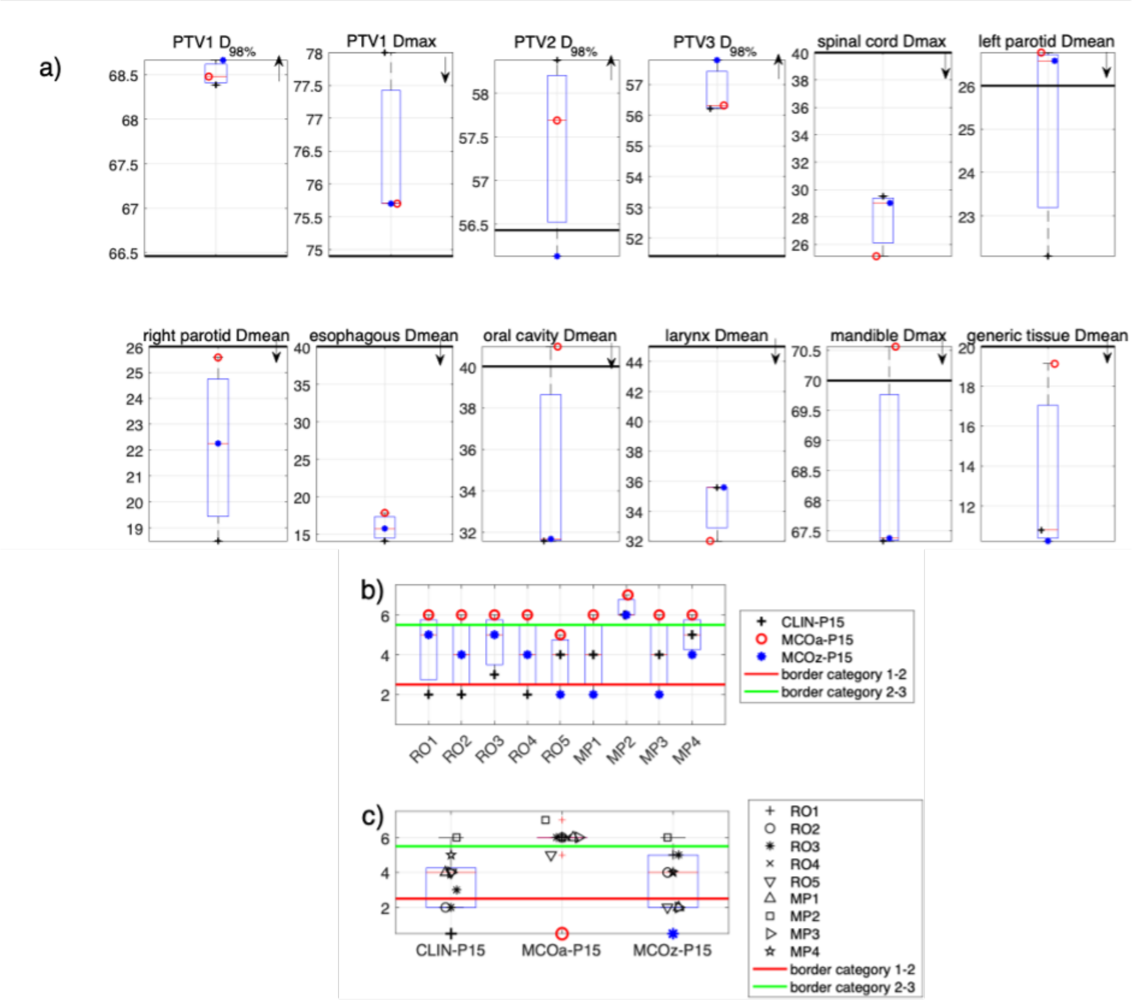


Figure 6.7.P15: For each of the 15 study patients, dosimetric parameters (a)) and subjective scores (b) and c)) of the 5 or 3 available treatment plans. In a), arrows indicate where plans should be relative to the black horizontal constraint lines. b) shows for each observer the scores for all available plans, while c) shows for each available plan the scores by all observers. Patients number (from P1 to P15) is indicate both in figure label and in figure legend.

7

ANALYSIS OF VARIATIONS IN THE RADIOTHERAPY PLAN EVALUATION PROCESS

> Often in judging a thing we get carried away more by the opinion than by the true substance of the thing itself.
> *Lucio Anneo Seneca*

7.1 PREVIEW

The chapter contains follow-up to the study described in chapter 6, which was focused on investigating the differences between a group of evaluators' scores assigned to different treatment plans. The available plans included in that study consisted of the clinical plan delivered to the patient and a set of automatically generated plans where different priorities were assigned to the relevant structures (see chapter 4). The aim of the work described in this chapter is to understand whether there are some dosimetric features associated with specific structures that bear more importance to some evaluators than others, during the scoring process described in the study of chapter 6. This analysis should help to understand better the large inter-user variability found. The analysis was performed by means of development and implementation of a tool estimating a 'generalized uniform Ideal Dose based on Exponential function' (gUIDE), described in chapter 5. By using this ideal dose, based on the anatomy of a single patient, together with the dosimetric features belonging to the different plans (which are strongly influenced by each patient's unique anatomy), the behavioral patterns of the evaluators during the scoring process can be investigated.

7.2 INTRODUCTION

It could be argued that any automation strategy might benefit if the system had *a priori* expectations of the ideal achievable results based on each patient's unique anatomy. Such knowledge could be used in two main ways:

1. during optimization, to provide superior inputs (as compared to standard tolerances used across all patients) that might allow the automated planning to exceed standard goals while avoiding pursuing impossible ones;

2. after optimization, to help gauge plan quality by comparing achieved results to theoretical but patient specific limits.

This information could be used to help gauge plan quality by comparing achieved results to theoretical but patient-specific limits [13] . These theoretical patient-specific plan limits are strictly related to patient geometry, i.e. target size and distance of the OARs from targets volumes. Thus, to mathematically understand plan quality, it is not enough to know the absolute values of obtained endpoints for several organs and targets, as information is also required on how close the considered dose distribution is compared to the theoretically achievable one. The focus of this work is not, however, to provide a tool that evaluates plans instead of the qualified people, but, with a sufficient accuracy, it is to objectively assess which features are relevant in the decision-making process of the evaluators and whether they differ between personnel working in the same department. This analysis was performed with and without the use of gUIDE, to quantify the improvement of this ideal dose in plan evaluation analysis. By analyzing the behaviour of the two groups of evaluators, i.e. the medical physicists (MPs) and radiation oncologists (ROs) (see chapter 6 for details), differences and similarities were identified in order to devise strategies to align priorities and improve the clinical quality of the approved and delivered plans. To best of current knowledge, this is the first study investigating the factors that could influence plan quality judgement in a group of several evaluators. For this study machine learning combined with the gUIDE tool are used as methods for the analysis. This information could be used in the future for specific training courses with the aim of reducing the user variability in plan quality assessment.

7.3 MATERIAL AND METHODS

The materials and methods used in this chapter, such as gUIDE and PlanIQ tools are described in chapter 5.

7.3.1 *Introduction to machine learning tools*

As stated in the Preview and Introduction sections, one aim of this chapter was to identify observer-related patterns behind plan evaluations. In order to investigate how the various endpoints related to the included OARs influenced the evaluator's decision, a machine learning (ML) method was employed based on tree classification. A ML algorithm is a computational process that uses input data to achieve a desired task without being literally programmed (i.e., "hard coded") to produce a particular outcome [113, 114]). These algorithms are in a sense "soft coded" in that they automatically alter or adapt their architecture through repetition (i.e., experince) so that they become better and better at achieving the desired task. The process of adaptation is called training, in which samples of input data are provided along with desired outcomes. The algorithm then optimally configures itself so that it can not only produce the desired outcome when presented with the training inputs, but also can generalize to produce the desired outcome from new, previously unseen data. This training is the "learning" part of machine learning. The training does not have to be limited to an initial adaptation during a finite interval. As with humans, a good algorithm can practice "lifelong" learning as it processes new data and learns from its mistakes [113]. There are many ways that a computational algorithm can adapt itself in response to training. The input data can be selected and weighted to provide the most decisive outcomes. The algorithm can have variable numerical parameters that are adjusted through iterative optimization. It can have a network of possible computational pathways that it arranges for optimal results. It can determine probability distributions from the input data and use them to predict outcomes. The ideal of machine learning is to emulate the way that human beings (and other sentient creatures) learn to process sensory (input) signals in order to accomplish a goal. This goal could be a task in pattern recognition, in which the learner wants to distinguish apples from oranges.

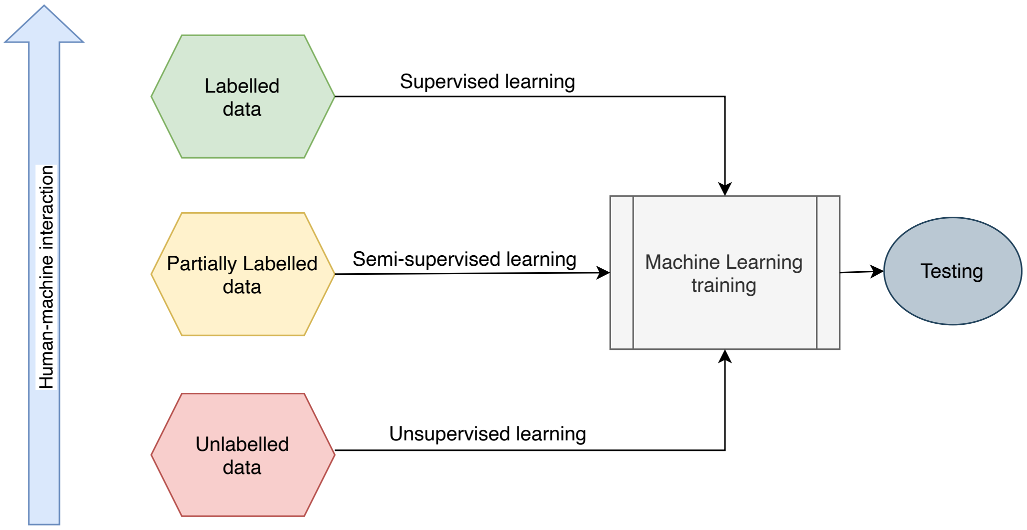


Figure 7.1: Categories of machine learning algorithms according to training data nature.

Every apple and orange is unique, but we are still able (usually) to tell one from the other. Rather than hard code a machine with many, many exact representations of apples and oranges, it can be programmed to learn to distinguish them through repeated experience with actual apples and oranges. This is a good example of supervised learning, in which each training example of input data (color, shape, odour, etc.) is paired with its known classification label (apple or orange). It allows the learner to deal with similarities and differences when the objects to be classified have many variable properties within their own classes but still have fundamental qualities that identify them. Most importantly, the successful learner should be able to recognize an apple or an orange that it has never seen before. A second type of machine learning is the so-called unsupervised algorithm. This might have the objective of trying to throw a dart at a bull's-eye. The device (or human) has a variety of degrees of freedom in the mechanism that controls the path of the dart. Rather than try to exactly program the kinematics *a priori*, the learner practices throwing the dart. For each trial, the kinematic degrees of freedom are adjusted so that the dart gets closer and closer to the bull's-eye. This is unsupervised in the sense that the training doesn't associate a particular kinematic input configuration with a particular outcome. The algorithm finds its own way from the training input data. Ideally, the trained dart thrower will be able to adjust the learned kinematics to accommodate, for instance, a change in the position of the target. A third type of machine learning is semi-supervised learning, where part of the data is labelled and other parts are unlabelled. In such a scenario, the part can be used to aid the learning of the unlabeled part. This kind of scenario lends itself to most processes in nature and more closely emulates how humans develop their skills. Figure 7.1 reports a scheme of main types of ML algorithms. There are two particularly important advantages to a successful algorithm. First, it can substitute for laborious and repetitive human effort. Second, and more significantly, it can potentially learn more complicated and subtle patterns in the input data than the average human observer is able to do. Both of these advantages are important to radiation therapy. For example, the daily contouring of tumours and organs at risk during treatment planning is a time-consuming process of pattern recognition that is based on the observer's familiarity and experience with the appearance of anatomy in diagnostic images. That familiarity, though, has its limits, and consequently, there is uncertainty and inter-observer variability in the resulting contours. It is possible that an algorithm for contouring can pick up subtleties of texture or shape in one image or simultaneously incorporate data from multiple sources or blend the experience of numerous observers and thus reduce the uncertainty in the contour. More than half of the patients with cancer receive ionizing radiation (radiotherapy) as part of their treatment, and it is the main treatment modality at advanced stages of disease. Radiotherapy involves a large set of processes that not only span the period from consultation to treatment but also extend beyond, to ensure that the patients have received the prescribed radiation dose and are responding well. The complexity of these processes can vary and may involve several stages of sophisticated human-machine interactions and decision making, which would naturally invite the use of machine learning algorithms to optimize and automate these processes, including but not limited to radiation physics quality assurance, contouring and treatment planning, image-guided radiotherapy, respiratory motion management, treatment response modelling, and outcome prediction.

7.3.2 *Bagged trees classification ensembles to estimate feature importance*

Classification is a process that is broadly divided into two steps. The first one is the learning step, the second is the step where prediction is done. In the stage

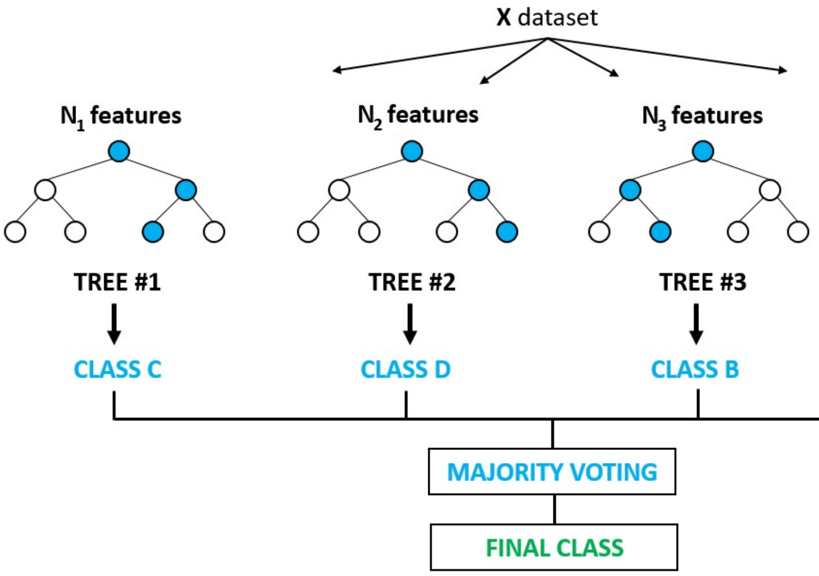


Figure 7.2: Bagged on process of a classification tree ensemble for a multi class problem.

of learning the model gets developed on the training data that is given. In the prediction stage, the model is used in order to predict the response to the data given. A very popular and easy example of the classification trees algorithm is the decision tree [115]. A classification tree is a model with a tree-like structure [116, 117]. It contains nodes and edges. There are two types of nodes:

- Intermediate nodes - An intermediate node is labelled by a single attribute, and the edges extending from the intermediate node are predicates on that attribute
- Leaf nodes - A leaf node is labelled by the class label which contains the values for the prediction.

The attributes that appear near the top of the tree typically indicate they are more important in making the classifications. See Figure 7.2 for an example of baggage tree classification.

In this study, model trees are coupled with bagged process for solving classification problems. Model trees are binary decision trees with linear regression functions at the leaf nodes: thus they can represent any piecewise linear approximation to an unknown function. The process starts from the root nodes and gets repeated until it reaches the leaf node. When this occurs then depending on the classification problem the predicted category will be the mode of the categories on the leaf node. A test sample that will reach this node has the maximum probability of belonging to the class with the training samples.

For the regression tree, the prediction is made at the end and this is the mean of the values for the variable that thus targets the leaf node.

In the first step, the training set gets created where the classification label is known for every record. The algorithm will systematically assign each record to one of the two subsets that are available on the basis of some factor. This helps to get a set of homogeneous labels in every partition. The splitting is applied to every new partition and the process will continue till there are no more splits found.

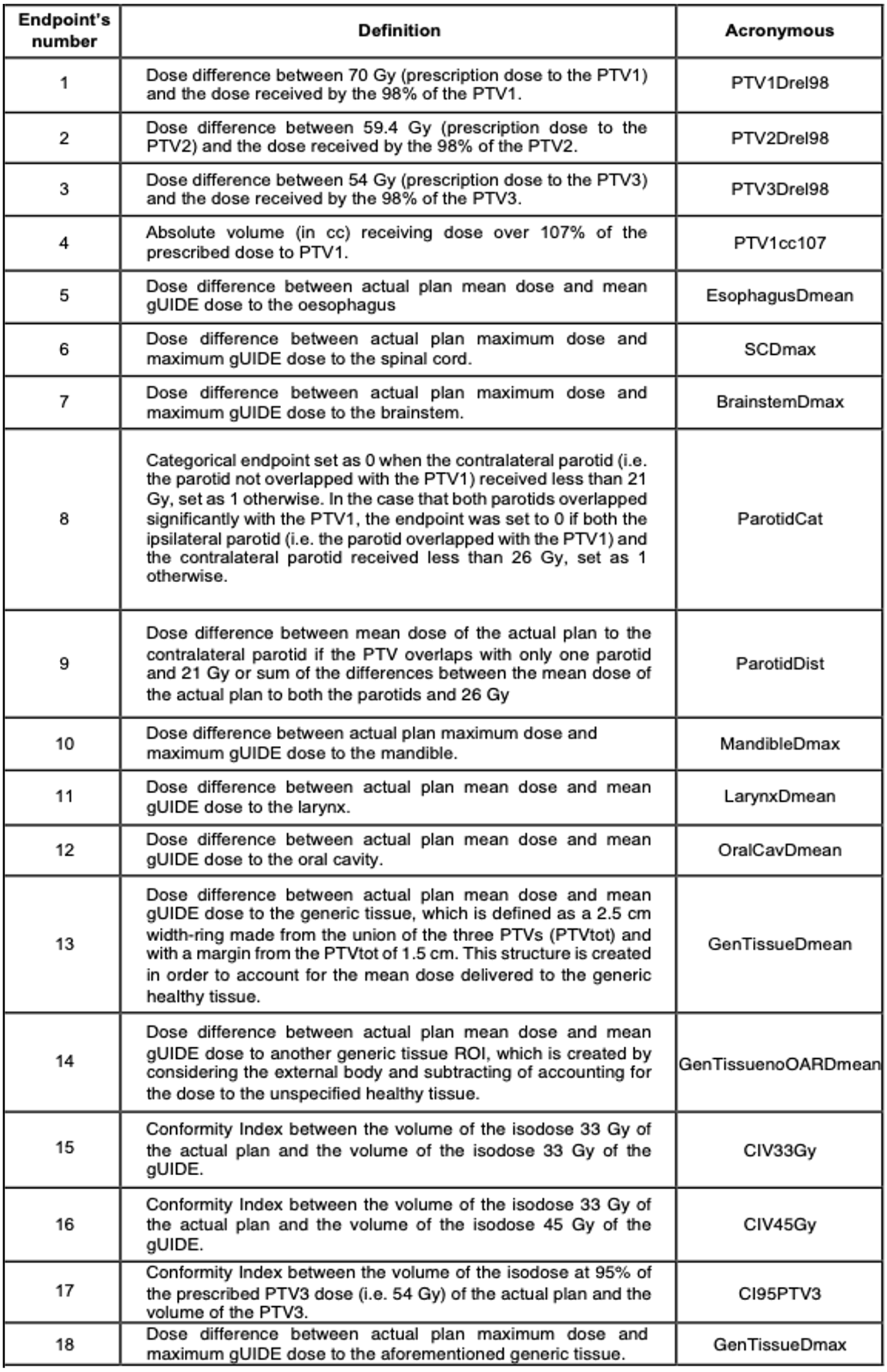
Due to their structure, ML trees are very suitable for an analysis that is more oriented to seek out what are the influential factors, rather than wanting to attain very high accuracies. As is well known, trees suffer greatly from the influence of the training set, due to being 'greedy' ML algorithms. To overcome this issue, bagged ensembles, have been used in this work. Through bootstrapping (i.e. creating several training subsets with $m<65$ observations) and aggregation (i.e. using the whole ensemble comprised of the single models trained on the bootstrapped sets), the final classification (computed through the voting of the single models) should be more stable than a single tree classification. The bagged process is briefly explained in Figure 7.2. The maximum number of splits allowed in the models was optimized and the model ensemble with the lowest validation loss was selected as the model with the optimal maximum number of splits parameter. This selected model was used in the actual training and prediction processes. This ML method aims to investigate if a specific ensemble model variable was influential in predicting the outcome; the higher the measure of associated importance is, the more influential the variable. The basis behind this algorithm resides on the assumption that if a predictor is influential in prediction, then permuting its values should affect the model error; otherwise, the effect of the permutation should not significantly affect the model error. For every tree in the ensemble, the Out-Of-Bag (OOB) observations are used as a test set to compare the error of the model with the variables in the correct order and the model with permuted variables. The importance is a mean value among all the learners and it is proportional to the difference between the errors associated with the two aforementioned models.

7.3.3 *Modelling the dosimetric endpoints used as input*

- *Input parameters: gUIDE-based endpoints*
  Different dosimetric endpoints were computed from the gUIDE and the available plan's dose (manual (CLIN), automated with clinical wish-list (MCOa) and with suboptimal wish-list (MCOx), see chapter 6) for each patient. The endpoints considered in this analysis, used as inputs to the selected ML mode, represent the most used during the plan approval process. They were reported in Table 7.1.

  The choice of the maximum number of endpoints/features (N=18) was dependent on the size of the training set (N=65 plans, see chapter 6). The isodoses 33 Gy and 45 Gy were chosen as they were commonly analyzed visually during the evaluation by the observers. The gUIDE is not involved in some of the endpoints, i.e. the PTV-related ones and two associated endpoints regarding the parotids (endpoints #7,8,9).

- *Input parameters: DVH-based endpoints*
  In order to effectively assess whether the gUIDE actually provides added value to the observer score modelling, the performance of another set of 9 machine learning models provided with endpoints extracted directly from the available plan DVHs (thus, not employing the gUIDE) was considered. Thus, the input parameters of this second training process were the 18 endpoints reported in Table 7.1 but directly computed from the DVH without the subtraction of the gUIDE contributions. Reference to, gUIDE based, indicates the gUIDE tool is considered as baseline and, raw models, relates to when gUIDE is not considered in the simulations.



Endpoint's number	Definition	Acronymous
1	Dose difference between 70 Gy (prescription dose to the PTV1) and the dose received by the 98% of the PTV1.	PTV1Drel98
2	Dose difference between 59.4 Gy (prescription dose to the PTV2) and the dose received by the 98% of the PTV2.	PTV2Drel98
3	Dose difference between 54 Gy (prescription dose to the PTV3) and the dose received by the 98% of the PTV3.	PTV3Drel98
4	Absolute volume (in cc) receiving dose over 107% of the prescribed dose to PTV1.	PTV1cc107
5	Dose difference between actual plan mean dose and mean gUIDE dose to the oesophagus	EsophagusDmean
6	Dose difference between actual plan maximum dose and maximum gUIDE dose to the spinal cord.	SCDmax
7	Dose difference between actual plan maximum dose and maximum gUIDE dose to the brainstem.	BrainstemDmax
8	Categorical endpoint set as 0 when the contralateral parotid (i.e. the parotid not overlapped with the PTV1) received less than 21 Gy, set as 1 otherwise. In the case that both parotids overlapped significantly with the PTV1, the endpoint was set to 0 if both the ipsilateral parotid (i.e. the parotid overlapped with the PTV1) and the contralateral parotid received less than 26 Gy, set as 1 otherwise.	ParotidCat
9	Dose difference between mean dose of the actual plan to the contralateral parotid if the PTV overlaps with only one parotid and 21 Gy or sum of the differences between the mean dose of the actual plan to both the parotids and 26 Gy	ParotidDist
10	Dose difference between actual plan maximum dose and maximum gUIDE dose to the mandible.	MandibleDmax
11	Dose difference between actual plan mean dose and mean gUIDE dose to the larynx.	LarynxDmean
12	Dose difference between actual plan mean dose and mean gUIDE dose to the oral cavity.	OralCavDmean
13	Dose difference between actual plan mean dose and mean gUIDE dose to the generic tissue, which is defined as a 2.5 cm width-ring made from the union of the three PTVs (PTVtot) and with a margin from the PTVtot of 1.5 cm. This structure is created in order to account for the mean dose delivered to the generic healthy tissue.	GenTissueDmean
14	Dose difference between actual plan mean dose and mean gUIDE dose to another generic tissue ROI, which is created by considering the external body and subtracting of accounting for the dose to the unspecified healthy tissue.	GenTissuenoOARDmean
15	Conformity Index between the volume of the isodose 33 Gy of the actual plan and the volume of the isodose 33 Gy of the gUIDE.	CIV33Gy
16	Conformity Index between the volume of the isodose 33 Gy of the actual plan and the volume of the isodose 45 Gy of the gUIDE.	CIV45Gy
17	Conformity Index between the volume of the isodose at 95% of the prescribed PTV3 dose (i.e. 54 Gy) of the actual plan and the volume of the PTV3.	CI95PTV3
18	Dose difference between actual plan maximum dose and maximum gUIDE dose to the aforementioned generic tissue.	GenTissueDmax

Table 7.1: Dosimetric endpoints computed from plans as input for ML simulation. The choice of the endpoints is based on the most generally used parameters during plan approval procedure.

7.3.4 *Modelling the evaluators' scores used as output*

As shown in chapter 6, there is great variability between the observers' scores. Due to limited number of plans (N=65), to simplify the problem the 7 point scale used in the study was converted into binary output according to the following:

$$\mathrm{new-score}_{(\mathrm{i,j})} = \begin{cases} 0, & \text{if raw score} < \text{median score for obs}_\mathrm{j} \\ 1, & \text{if raw score} \geqslant \text{median score for obs}_\mathrm{j} \end{cases} \tag{7.1}$$

7.3.5 *ML models evaluation*

In order to investigate how the various endpoints related to the included OARs influenced the evaluator's decision, a ML method was employed based on tree classification. As there were 9 observers, 9 different bagged tree ensemble models were trained using every evaluator's binary scores as label. A representative observer (referred to as 'oss evaluator'), composed of the modal results across all of the evaluators was also trained and evaluated, for a total of 10 ensemble models. The area under the receiver operating characteristic (ROC) curve (AUC)[118], was used to test their performance; the AUC values for the 10 models were obtained after a repeated 5-fold cross validation (20 times), in order to reduce the estimated errors in the models' performance. After the optimization and the repeated 5-fold cross validation (20 times), the mean AUC value and its associated standard deviation on the repeated cross validations were computed for both sets of models (raw and gUIDE-based), in order to evaluate their performance. Feature importance was then evaluated for the set of models with the overall highest AUC, with the aim of investigating what feature(s) were the most influential in the models and thus what endpoints were the most important for each evaluator when making their choices. The feature importance was computed through OOB predictor importance estimates by permutation, implemented in Matlab (MathWorks, USA). The extracted feature importance is then used as a basis to understand how the decision process of the nine observers is correlated with the chosen endpoints. The feature importance related to the different observers was used to rank them; however, the importance values are greatly affected by noise (as a result of many uncorrelated features in the training set). In order to understand how the found endpoint importance values varied among the observers, it was decided as first analysis to choose the three endpoints (from the 9 observers' models) that had the highest mean importance values and it was noted (in percentage terms) how many times a certain endpoint was featured among the three best for every observer. This analysis had the aim of understanding the difference between the evaluators and whether the ROs and MPs (as groups) followed specific patterns during the evaluation process.

7.3.6 *Statistical analysis*

The statistical dosimetric differences between the groups of binary scores for all the observers were also analysed. The continuous variables were analysed in their distributions between the two groups (0 and 1, see section above) by comparing their median values using the Kruskal-Wallis test. P-values less than 0.05 meant that the median values between the two groups were deemed significantly different. For the ParotidCat variable, which is categorical, an exact Fisher test was employed to test the differences between the two groups, as the sample size is small. All the statistical analysis was univariate, as the relationship between the single endpoint was investigated with respect to the two groups of scorers' labels. Then, the significantly different variables found with the statistical methods were compared to the three most important features from the feature importance analysis to investigate the consistency between the ML model and the statistical model.

7.4 RESULTS

7.4.1 *gUIDE ideal dose prediction for plans of Chapter 6*

Despite the statistical difference between the planIQ and gUIDE FDVHs found in chapter 5, the gUIDE prediction was found to be useful as a baseline dose to

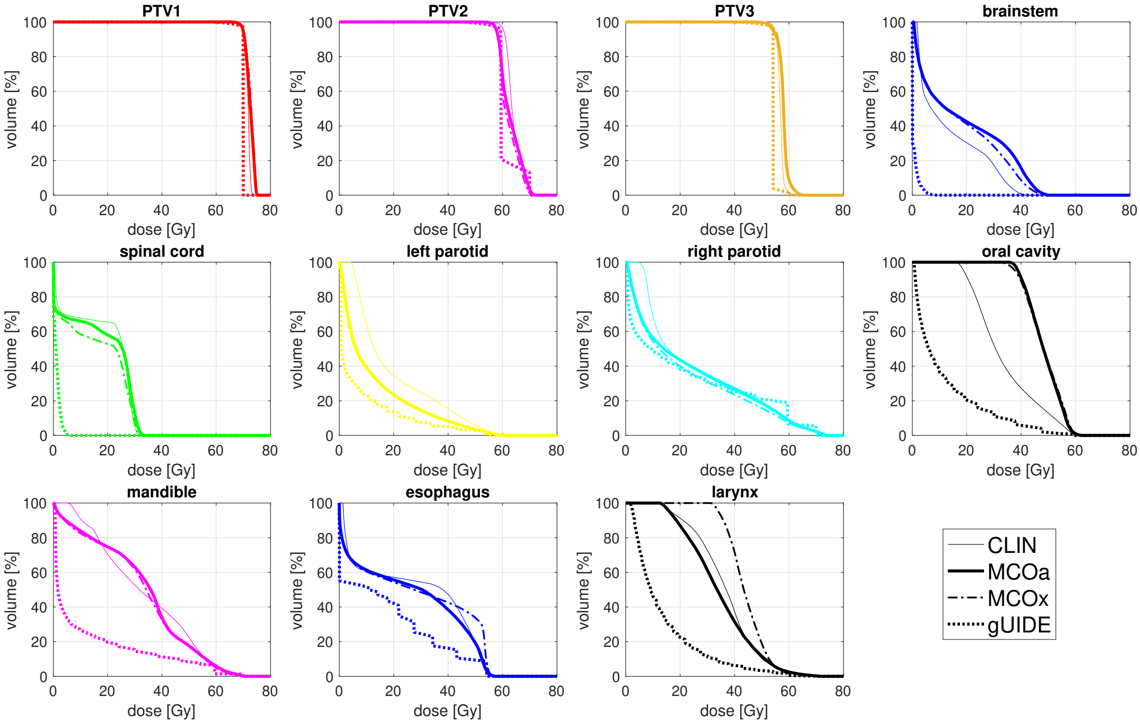


Figure 7.3: DVH comparisons for one patient of the set for CLIN, MCOa, MCOx and gUIDE.

estimate the difference with CLIN, MCOa and MCOx plans (chapter 5) as reported in Figure 7.3. In this figure it is possible to observe the plan comparison DVHs for a sample of patients: gUIDE DVHs for all main OARs were always lower than real plans (CLIN, MCOa and MCOx) as a baseline.

7.4.2 *Bagged trees ensemble results: comparison of raw and gUIDE based model simulations*

Table 7.2 reports the AUC values and standard deviation (SD) for the model results, both raw and gUIDE based. Using only the endpoints coming from the DVH (raw), overall AUC values are lower than 0.66 and SDs among the repeated cross validations reach values of 0.05. One of the observers, RO3, is not modelled at all, as their associated AUC is 0.44. The two best performing observers are MP2 and RO2 with AUCs of 0.67. The other observers showed poor AUCs, lower than 0.65.

For the model gUIDE-based endpoints, overall AUC values were larger than 0.55 with a large variability over all the evaluators but with a small standard devia-

Evaluators	AUC_gUIDE	std	max - min	AUC_raw	std	max - min
MP1	0.67	0.02	0.71 - 0.62	0.62	0.03	0.66 - 0.56
MP2	0.71	0.03	0.77 - 0.67	0.66	0.05	0.77 - 0.58
MP3	0.66	0.03	0.71 - 0.58	0.64	0.05	0.70 - 0.55
MP4	0.66	0.03	0.74 - 0.63	0.62	0.03	0.69 - 0.56
RO1	0.79	0.03	0.82 - 0.71	0.64	0.04	0.71 - 0.58
RO2	0.66	0.03	0.72 - 0.61	0.66	0.03	0.72 - 0.61
RO3	0.55	0.03	0.59 - 0.51	0.44	0.05	0.54 - 0.32
RO4	0.70	0.03	0.77 - 0.66	0.63	0.03	0.70 - 0.58
RO5	0.57	0.04	0.64 - 0.50	0.58	0.04	0.64 - 0.52
O_mode	0.73	0.02	0.76 - 0.68	0.64	0.03	0.69 - 0.58

Table 7.2: AUC values for the ML models for the available observers. AUC_{raw} refers to the AUC values for the DVH-based endpoints used as inputs, while $\text{AUC}_{\text{gUIDE}}$ regards the models where gUIDE-based endpoints were used as input.

tion for the single observer model. For three evaluators, MP2, RO1 and RO4 the model showed an AUC greater than or equal to 0.70. On the other side, RO3 and RO5 showed a poor AUC, less than 0.60. RO5 also showed the highest standard deviation among the repeated AUCs. The performances of the gUIDE-based models were higher than the raw models, for MP1, MP2, MP4, RO1, RO3 and RO4. The gUIDE models also exhibited a reduced standard deviation among the repeated cross validations for MP2, MP3, RO1, RO3 and RO5. The RO3 scores showed low AUC values both with raw and gUIDE-based endpoints. Regarding MP3, RO2 and RO5, their scores did not increased with the use of gUIDE-based endpoints. Paired two-sided Wilcoxon signed rank test on the mean doses of AUC over all 9 evaluators showed statistically different median values for the two sets (raw and with gUIDE baseline) with a p-value = 0.0078. As the models employing the gUIDE-based endpoints had better (or at least equal) AUCs compared to the raw endpoints set, this approach was considered for features importance.

7.4.3 *Feature importance*

From the obtained models, the feature importance computation procedure was repeated 20 times and the mean value over the 20 iterations was plotted in the following bar graphs. These parameters (20 times and 20 iterations) come from empirical considerations derived from various simulations. Ten different feature importance graphs (9 observers, plus the oss-mode (composed of the modal results across all of the evaluators' scores, see subsection 7.3.5)) were plotted and the results are reported in Figure 7.4. The error bar refers to 1.96 times the standard deviation from the mean importance value, which is represented by the height of the bars. It is evident that the pattern of feature importance varied among evaluators without any correspondence between MPs and ROs.

Regarding the negative values displayed in the graphs, they should be considered as noise. They come from the mathematical definition of the importance (see subsection 7.3.6 from this chapter) and it is clearly referable to as a situation where the involved feature was not relevant in the model's computation of the outcome. From the feature importance of Figure 7.4, it is possible to observe that the different evaluators had different features influencing their score. The number of times a feature was chosen among the best (first, second and third) was reported in Figure 7.5, expressed in percentage divided by MP, RO and all groups.

From Figure 7.5, the MPs focus their attention mostly on the coverage of PTV2, while the most important features (as a group) for the ROs is ParotidCat, which describes if the dose to a parotid has exceeded the clinical endpoint figure in the dedicated protocols.

7.4.4 *Comparison with the statistical analysis results*

In Table 7.3, the statistical analysis results are summarized for the available evaluators. Values in bold correspond to p-values$<$ 0.05.

There were several significantly different endpoints for the binary score (0 and 1) between the two methods of analysis (ML and statistic). The only evaluator showing just one significantly different endpoint is RO3, whio was also one of the observers with the lowest AUC in the ML analysis. Table 7.4 shows the compar-

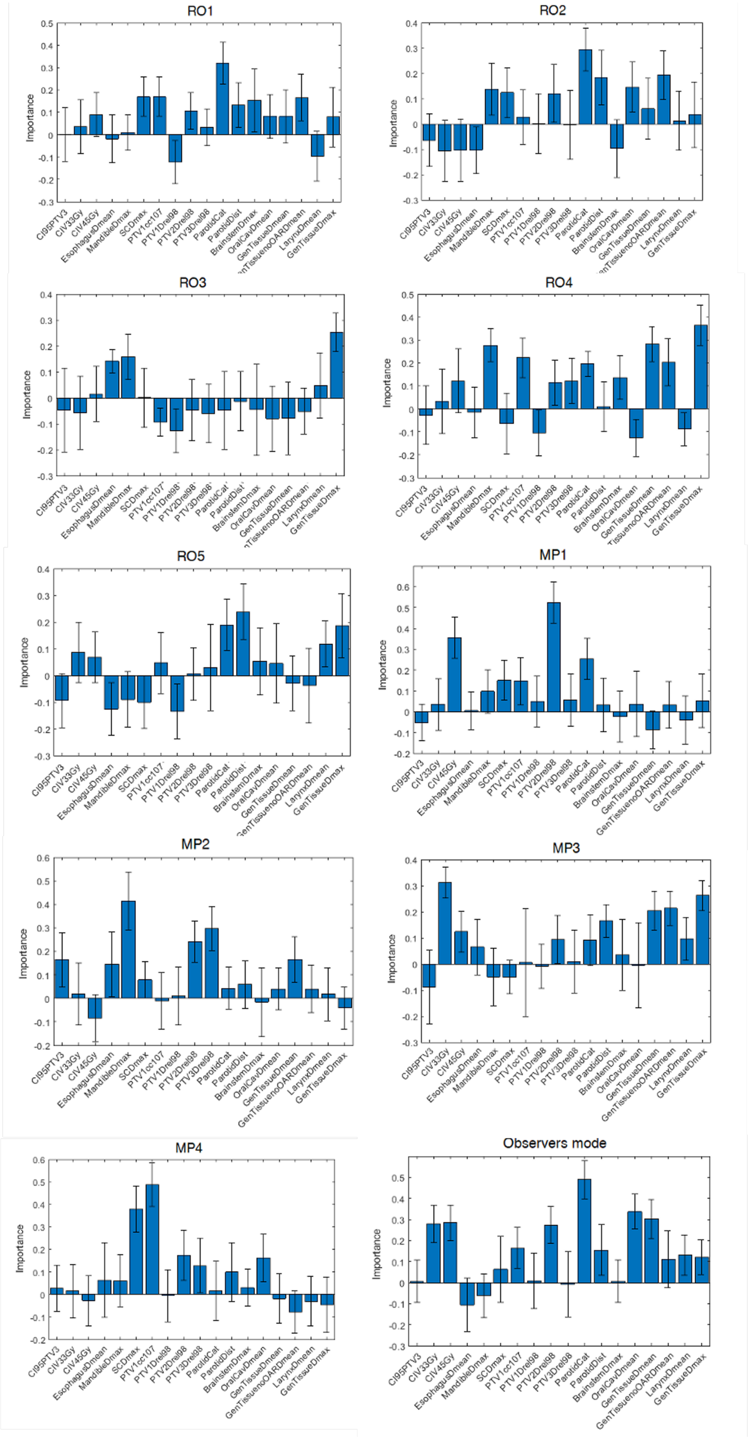


Figure 7.4: Feature importance computation for all the 9 observers. MP and RO indicate medical physicist and radiation oncology, respectively (see chapter 6). The displayed value is the mean among all the iterations (20) and the error bar refers to 1.96 times the standard deviation among the iterations. The negative importance bars displayed in the graphs should be regarded as noise.

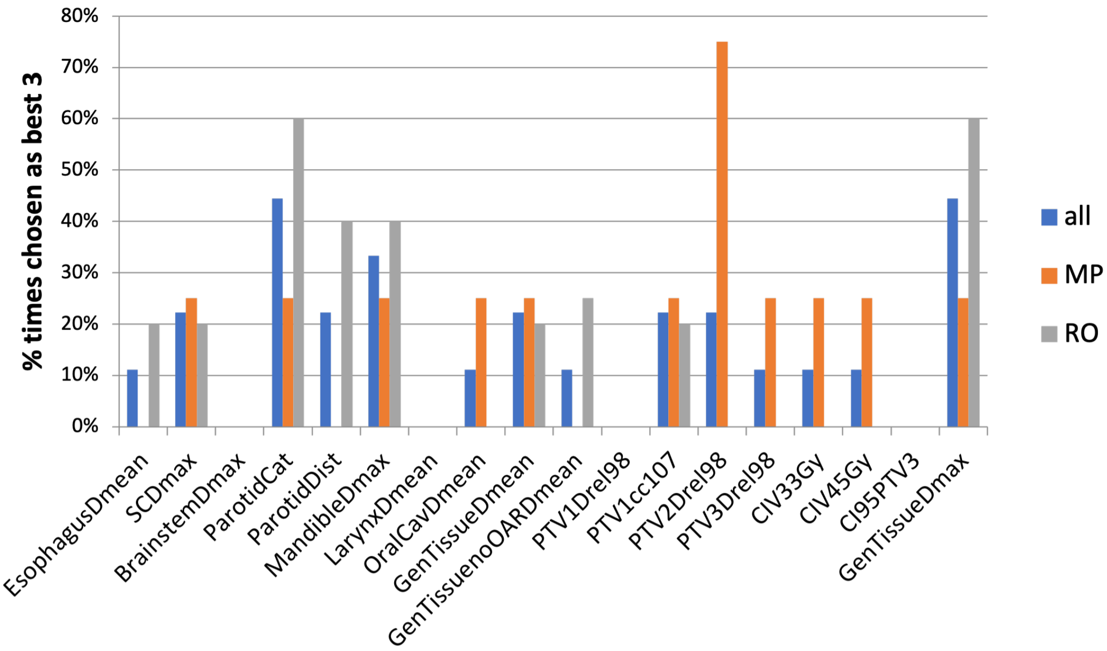


Figure 7.5: Number of times (%) a certain feature was chosen as one of the most important three for an observer. The results are expressed for the whole observer set, and divided by the MP and RO groups.

	RO1	RO2	RO3	RO4	RO5	MP1	MP2	MP3	MP4	O_mode
Oesophagus Dmean	0.173	0.358	0.181	0.803	0.761	0.565	0.097	**0.017**	0.131	0.088
SCDmax	0.566	0.150	0.118	0.470	0.812	0.087	0.978	0.335	0.174	0.637
BrainstemDmax	0.244	0.501	0.713	**0.018**	0.791	0.315	0.248	**0.036**	0.438	0.228
ParotidCat	**<0.001**	**<0.001**	0.16	**<0.001**	**0.001**	**<0.001**	0.38	**<0.001**	**0.04**	**<0.001**
ParotidDist	**0.016**	0.094	0.081	0.222	**0.035**	0.130	0.471	**0.048**	0.505	**0.028**
MandibleDmax	0.743	0.092	0.115	**0.027**	0.662	0.957	**0.002**	0.789	0.892	0.667
LarynxDmean	0.731	0.711	0.213	0.783	0.072	0.486	0.744	**0.023**	0.541	0.065
OralCavDmean	0.302	0.192	0.338	0.591	0.662	0.659	0.683	0.768	0.050	0.901
GenTissueDmean	**0.033**	**0.006**	0.793	**0.043**	0.781	0.218	0.714	**0.015**	0.328	**0.003**
GenTissuenoOARDmean	**0.036**	**0.006**	0.803	**0.020**	0.491	0.199	0.755	**0.005**	0.232	**0.001**
PTV1Drel98	0.629	0.696	0.655	0.230	0.952	0.723	0.714	0.606	0.362	0.961
PTV1cc107	**0.023**	0.222	0.665	**0.049**	0.064	0.134	0.197	0.189	**0.011**	0.149
PTV2Drel98	0.600	**0.018**	0.072	0.177	**0.024**	**<0.001**	**0.003**	0.054	0.069	**0.005**
PTV3Drel98	0.577	0.291	0.679	0.145	0.596	0.429	**0.013**	0.718	0.903	0.989
CIV33Gy	1.000	0.197	0.345	**0.010**	0.076	0.077	0.903	**0.002**	0.277	**0.020**
CIV45Gy	0.441	0.945	0.470	0.072	0.397	0.111	0.166	0.052	0.095	0.256
CI95PTV3	0.544	0.752	0.990	0.306	0.771	0.602	0.055	0.883	0.644	0.598
GenTissueDmax	0.491	0.064	**0.023**	**0.007**	**0.042**	0.057	0.978	**0.008**	0.860	**0.011**

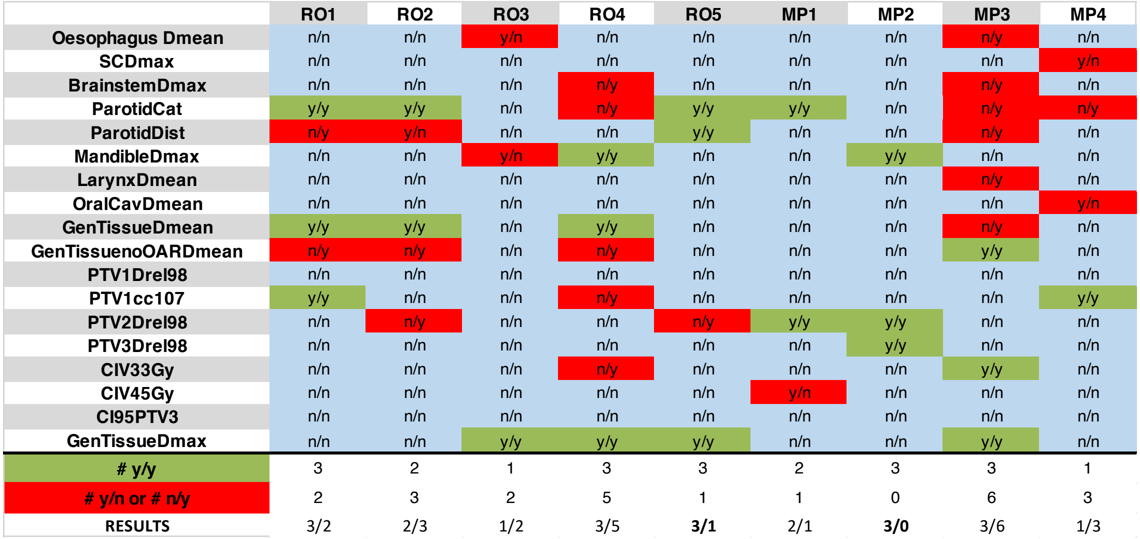
Table 7.3: P-values associated to the endpoints based on statistical analysis.

ison between the statistically different endpoints and the three endpoints having the highest importance for all the observers using the ML approach. There are four different values shown in this table:

- n/n (highlighted in cyan) means that the considered endpoint was not among the three best features with the highest importance in the ML model (key features) for its respective observer and it was also not found as significantly different between the two group's scores (0 and 1) in the statistical analysis.
- n/y or y/n (highlighted in red) means a disagreement: the endpoint was not among the three key features for that observer in their ML model but the statistical analysis found it significantly different.
- y/y (highlighted in green) means that the endpoint was both among the three key features for that observer's ML model and a statistical difference was also found between the binary score.

At the end of the Table 7.3, the total number of y/y, y/n and n/y instances is counted.

The highest accordance between the two methods (3/3 endpoints) was found for 5 out of 9 observers, while for MP4 and RO3 only one of the three best features found in the ML model was found to have a statistical difference between the binary scores groups. However, for a more complete analysis of agreement between the two methods, the disagreement in endpoints between the two methods should be considered (y/n or n/y). Since for the statistical analysis every endpoint showing a p-value$<$0.05 was included there are also several instances of an endpoint found to be statistically different but not bearing one of the three highest importance values in the ML model. There are also (fewer) instances of a feature having a high importance value in the ML model but not having a significant statistical difference between the scores. In the last row of Table 7.4 the ratio between agreement and disagreement occurrences (agreement/disagreement) was reported. For only two evaluators a good results were observed, RO5 (3/1) and MP2 (3/0).



	RO1	RO2	RO3	RO4	RO5	MP1	MP2	MP3	MP4
Oesophagus Dmean	n/n	n/n	y/n	n/n	n/n	n/n	n/n	n/y	n/n
SCDmax	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n	y/n
BrainstemDmax	n/n	n/n	n/n	n/y	n/n	n/n	n/n	n/y	n/n
ParotidCat	y/y	y/y	n/n	n/y	y/y	y/y	n/n	n/y	n/y
ParotidDist	n/y	y/n	n/n	n/n	y/y	n/n	n/n	n/y	n/n
MandibleDmax	n/n	n/n	y/n	y/y	n/n	n/n	y/y	n/n	n/n
LarynxDmean	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/y	n/n
OralCavDmean	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n	y/n
GenTissueDmean	y/y	y/y	n/n	y/y	n/n	n/n	n/n	n/y	n/n
GenTissuenoOARDmean	n/y	n/y	n/n	n/y	n/n	n/n	n/n	y/y	n/n
PTV1Drel98	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n
PTV1cc107	y/y	n/n	n/n	n/y	n/n	n/n	n/n	n/n	y/y
PTV2Drel98	n/n	n/y	n/n	n/n	n/y	y/y	y/y	n/n	n/n
PTV3Drel98	n/n	n/n	n/n	n/n	n/n	n/n	y/y	n/n	n/n
CIV33Gy	n/n	n/n	n/n	n/y	n/n	n/n	n/n	y/y	n/n
CIV45Gy	n/n	n/n	n/n	n/n	n/n	y/n	n/n	n/n	n/n
CI95PTV3	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n	n/n
GenTissueDmax	n/n	n/n	y/y	y/y	y/y	n/n	n/n	y/y	n/n
# y/y	3	2	1	3	3	2	3	3	1
# y/n or # n/y	2	3	2	5	1	1	0	6	3
RESULTS	3/2	2/3	1/2	3/5	**3/1**	2/1	**3/0**	3/6	1/3

Table 7.4: Agreement between ML and statistical analysis results. Consistency table comparing the results from the models feature importance and the results from the statistical analysis. The table reports the times for which an agreement (y/y) or a disagreement (y/n or n/y) is observed between the two methods.

7.5 DISCUSSION

The aim of this work was to gain more knowledge about the treatment plan evaluation process in radiotherapy using ML models and an ideal dose as baseline of an independent analysis. The inputs given to the ML model were a set of selected dosimetric endpoints (regarding both the tumour target and the important surrounding organs) in combination with a purpose-built developed tool (called gUIDE) in their calculation. The gUIDE could provide partial but fundamental information about the quality of obtained dose distributions in different patient anatomies and geometries. Moreover, in the definition and formulation proposed in this work, the gUIDE was shown to be accurate for doses over 20% of the prescription value, which was within the scope of its use in this study. Due to the limited number of plans to model (65 plans), the raw scores were transformed into binary scores (0 and 1) in order to to simplify the problem in employing a ML classification. To test the importance of gUIDE as baseline dose to compare different dose distributions, two different ML tools were used for modeling the evaluators' scores: one having as input a set of endpoints extracted directly from the available plans, DVHs (raw) and the other having the same set of endpoints but computed also using the contribution from the gUIDE as baseline (gUIDE-based). The gUIDE-based ML model showed better AUCs and lower variability (SD) for the majority of the observers and it was thus used as the basis for the final ML analysis. Statistical tests confirmed the difference between AUC gUIDE-based and AUC raw-based. Two observations can be derived from this result. Firstly, the information provided by the DVHs is not enough to understand the decision-making process of the plan evaluators, since there are obviously other factors involved in that operation, such as the dose distribution and the patient geometry complexity. Secondly, gUIDE-based ML AUC values were different for the available observers; this means that the chosen input parameters are not significant for all evaluators, especially for low AUC model evaluators, or it might be that some other factor not accounted for should be taken into consideration. ML model results were also compared to a statistical analysis of the gUIDE-based endpoints and it was found that, while for some observers the two methods aligned, there were some differences. It can be concluded that the use of ML methods to investigate the plan evaluation could give a more complete insight on the evaluators' different scoring procedure. Generally, the major difference between ML and statistics is their purpose. ML models are designed to make the most accurate predictions possible. Statistical models are designed for inference about the relationships between variables. In this study, ML is used not for a predictions but for a deeper analysis of the results found in chapter 6. ML methods are particularly helpful when one is dealing with 'wide data', where the number of input variables exceeds the number of subjects, as it was in this case. ML makes minimal assumptions about the data-generating systems; they can be effective even when the data are gathered without a carefully controlled experimental design and in the presence of complicated nonlinear interactions. While statistical analysis gave only a p-value of significance, ML is able to report a level of accuracy (AUC and SD). Moreover the statistical analysis methods used in the study has the limitation that it was univariate, while the problem is clearly multivariate. However, the significant p-values in bold for each evaluator reported in Table 7.3 were few especially for evaluator RO2 (N=3), RO3 (N=1), MP1 (N=2), MP2 (N=3), MP4 (N=2).

This work has some limitations. As mentioned previously, in chapter 5, the gUIDE tool implementation does not take into account the low-gradient effect, which affects the lower doses; while this situation is not relevant for the HN site, it might be important and useful to include such an effect for other anatomical sites, such as the breast, as lower doses become clinically important for those patients. This was also evident in Figure 5.6 where planIQ and gUIDE DVHs and mean dose were compared for several organs The implementation of a low-gradient effect in the gUIDE for another anatomical site requires also the design of a new validation strategy and geometry capable of giving information on the new PTV/OAR dimensions, which are significantly different from the HN case. Nevertheless, for the purpose of using gUIDE as baseline dose to improve the ML model accuracy quantified by the AUCs, is accurate enough. But for further application such as feasibility dose its implementation should be improved. The association to the ML models showed, the use of bagged tree classification is also unable to properly model the decision-making process of some evaluators. The ML model performance was enhanced by employing the gUIDE-based dosimetric endpoints but there is still room for improvement. The use of neural networks could help improve the model's accuracy, but the amount of data (i.e. the number of plans to be used as a training set) needs to be much higher. Moreover, the endpoint features choice should be improved, considering endpoints more related to the spatial dose distribution instead of DVH parameters, such as Dmean and Dmax.

7.6 CONCLUSIONS

In this study a baseline dose gUIDE was implemented and evaluated. This tool has been shown to improve accuracy when using ML to model plan quality evaluation for several users. The dataset analysed was described in chapter 6. It was demonstrated that the ML approach with gUIDE gives more complete information compared to the use of the ML tool without any anatomical and dose distribution information. Large variability was found for the features of importance considered by each evaluator, however further analysis will be performed to improve the accuracy of the models.

8

AUTOMATION IN BREAST TREATMENT PLANNING

The Three Laws of Robotics:
— *First*: A robot may not injure a human being or, through inaction, allow a human being to come to harm;
— *Second*: A robot must obey the orders given it by human beings except where such orders would conflict with the First Law;
— *Third*: A robot must protect its own existence as long as such protection does not conflict with the First or Second Law;
— *Zeroth*: A robot may not harm humanity, or, by inaction, allow humanity to come to harm.
Isaac Asimov — I, Robot

8.1 PREVIEW

In this chapter the methods of automated planning using Erasmus-iCycle introduced in chapter 3 and applied in chapter 6 to head and neck treatment were applied to left breast cancer radiotherapy. This chapter is part of an ongoing international multicentre study on breast radiotherapy, described in chapter 11.

8.2 INTRODUCTION

As discussed previously in this thesis, in the Introduction section of chapter 6, in both planning and evaluation phases, human action in treatment plan quality is crucial. Breast cancer is by far the most common cancer in the female population [119] and whole breast irradiation following surgery has proved its benefit in terms of outcome for a significant fraction of patients [55] and is nowadays a well estab-

lished therapy. Indeed, breast radiotherapy makes up around 30% of most departments' workloads. Despite different techniques being available for the irradiation of the whole breast, many institutes still use the conventional tangential field (TF) arrangement to obtain an adequate dose delivery, either using 3-dimensional conformal radiation therapy (3DCRT) or intensity modulated radiotherapy (IMRT) [51, 55, 120–122]. Compared to rotational techniques [123, 124], the TF approach conserves the advantage of being "simple" and, above all, efficient in limiting the dose received by areas out of the breast region, avoiding the typical "low-dose" spread of rotational techniques [124]. Due to the still unresolved issues related to the potential clinical impact of the low-dose spread to heart, lungs and contralateral breast, especially in long-surviving patients [125, 126], TF are expected to remain among the most used techniques to treat breast cancer in the next decade. On the other hand, forward planned (and also inverse planned) optimization is time consuming and dependant on the planner's skill [127]. A few automatic solutions have been reported [127–130]: a relatively weak point of auto-planning for whole breast is the intrinsic difficulty of taking into account the inter-patient variations in assessing the best position of the fields to limit the dose to the adjacent organs, concomitantly assuring PTV coverage and highly homogenous dose distribution within the PTV. Left breast is particulary critical due to the close position of the heart to the PTV. This distance is strongly dependent on individual patient anatomy. Thus, this site is particulary difficult to standardize the plan with a totally automatic procedure. The main OARs considered for the left breast are heart, lung, contralateral breast and left anterior descending artery (LAD).
The goal of the study presented in this chapter was to configure, test, and implement Erasmus-iCycle for automated, multicriterial IMRT treatment planning for left breast patients.

8.3 MATERIAL AND METHODS

8.3.1 *Patients and clinical (CLIN) treatment plans*

Planning CT data, contoured structures and the clinical 3DCRT plan (CLIN) with TF of 18 arbitrarily selected oropharyngeal HN breast cancer patients, recently treated with radiotherapy at Azienda USL-IRCCS Hospital (AUSL) of Reggio Emilia (Italy), were included in this study. Nine patients were less than 50 years old and the other 9 older. The patients were divided in two prescription schemes based on age (see international protocols [131–133]): for patients with age $>$ 50 years, 39 Gy were delivered in 13 fractions, while for patients with age $\leqslant$ 50 years, 40.5 Gy were delivered in 15 fractions. The goal was to deliver 100% of the prescribed dose to 85% of the PTV . Left breast was irradiated in all patients. Sizes of the involved PTVs (mean±sd) were: 825.6±230.9 cm^3 [363.3,1224.31]. OARs considered in planning were spinal cord, heart, left and right lung, and contralateral breast [132, 133]. Plans were generated using the following priorities for achieving planning objectives: 1) sparing of brainstem, optic chiasm, and spinal cord (so higher priority than PTV coverage); 2) achievement of PTV dose objectives in the order PTVhigh, PTVmedium, PTVlow; 3) parotid gland sparing; 4) sparing of other OARs and healthy tissues. The clinical planning protocol was largely in line with international protocols, such as RTOG 1005 [131], StartB protocol [133] and Lee *et al.*'s published study [132]. Patients were treated with 2 to 4 TFs in a 3DCRT technique using 6MV and 10MV delivered with a Truebeam linac (Varian Medical Systems, Palo Alto, USA). Enhanced dynamic wedge with angles of 20 to 45 degrees were used. Clinical planning was performed with the Eclipse treatment planning system (TPS) v.13 (Varian Medical Systems, Palo Alto, USA).

8.3.2 *Breast institute wish-list definition*

As was described in chapter 3, in the Erasmus-iCycle module the optimization operation is based on a user defined wish-list which contains hard constraints and objectives with given priorities. To guarantee consistency of automated planning the optimal wish-list is unique for all patients. In the case of left breast cancer radiotherapy, the anatomy among patients varies considerably and some patients exhibits unfavourable anatomy, such as implants, large breasts, a heart very close to target, and limited upper-extremity range of motion. For these reasons priority order should be tuned carefully, considering different anatomies during the wish-list tuning. More details concerning the Erasmus-iCycle module are reported in chapter 3. An initial wish-list was composed based on previous clinical experience, the planning protocol, and intent of treating physicians on how to improve clinically applied plans. It was used to automatically generate a plan for the first 8 patients included in this study. These plans were then evaluated together with physicians, and the wish-list was modified according to their input. Several optimisation functions, described in chapter 3 were used in a trial-and-error process, until the expected solution was obtained. This iterative procedure continued until no further improvements in plan quality were achieved for the first 8 training patients.

8.3.3 *Generation of Pareto optimal plan with Erasmus-iCycle*

For this study, the defined wish-list based on institutional clinical goals was modified to minimise low dose (i.e. 5-10 Gy) outside PTV while not compromising PTV and OAR dose. To reduce the 'dose bath', the IMRT technique was chosen using tangential fields and mimicing TF 3DCRT dose distribution but improving dose conformity and creating convex isodose shapes. After several simulations on the first 8 patients, a configuration of 8 IMRT tangential fields was considered. The setting of the fields was based on the 2 clinical TF angles. From this, 2 fields were added with increased angles (medial and lateral) by 5 degrees (external fields) and 4 fields were added internally with steps of 3 degrees (see Figure 8.1). The 8 TF IMRT fields (4 for each side) covered an angle of 11 degrees for both medial and lateral directions. 6MV was chosen as unique energy for all fields.

This configuration was then applied to all 16 patients. By including more cases, it was found that 4 patients didn't fit well into this configuration. They all belonged to the cases in which the patient was overweight, the target volume was higher than the mean values and a considerable part of normal tissue, such as axilla, was inside the lateral TF. For this patient specific anatomical group a second configuration was considered, adding to the standard configuration a 90 degree field (9th field) and using 10MV energy for all 4 lateral TF. The isocentre was set in all plans as the clinical plan. Several 'dummy structures' were created to reduce the dose outside the target. Figure 8.1 illustrates the shape of these structures. These include a shell-PTV-contract, a ring structure internal to the body (3 cm within the body structure) of 3 cm thickness used to reduce the low dose inside

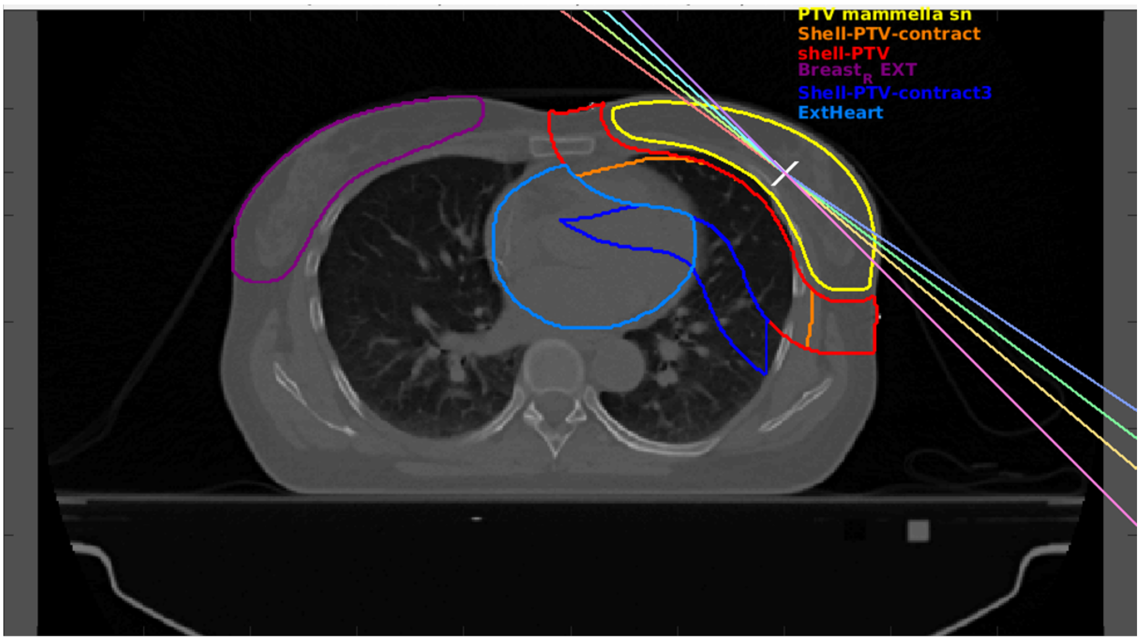


Figure 8.1: Transverse slice of a sample patient considered in this study. The angular direction of the 8 tangential IMRT fields used is displayed. The isocenter is indicated as a white cross inside PTV. PTV left breast (yellow) is reported as well as several dummy structures used to contain the dose.

lung and heart, shell-PTV, a ring structure of 3 cm thickness used to contain the hot spot in the axilla, or in the entrance of tangential fields. Other dummy structures considered were Shell-PTV-contract3 a second ring more distant than the others to contain the very low dose, Breast$_R$ EXT and ExtHeart, an external right breast structure 4cm from the PTV and an external heart structure 3 cm from PTV respectively, used to reduce the low dose in the distal part of right breast and heart.

8.3.4 *Comparison of auto and manual plans*

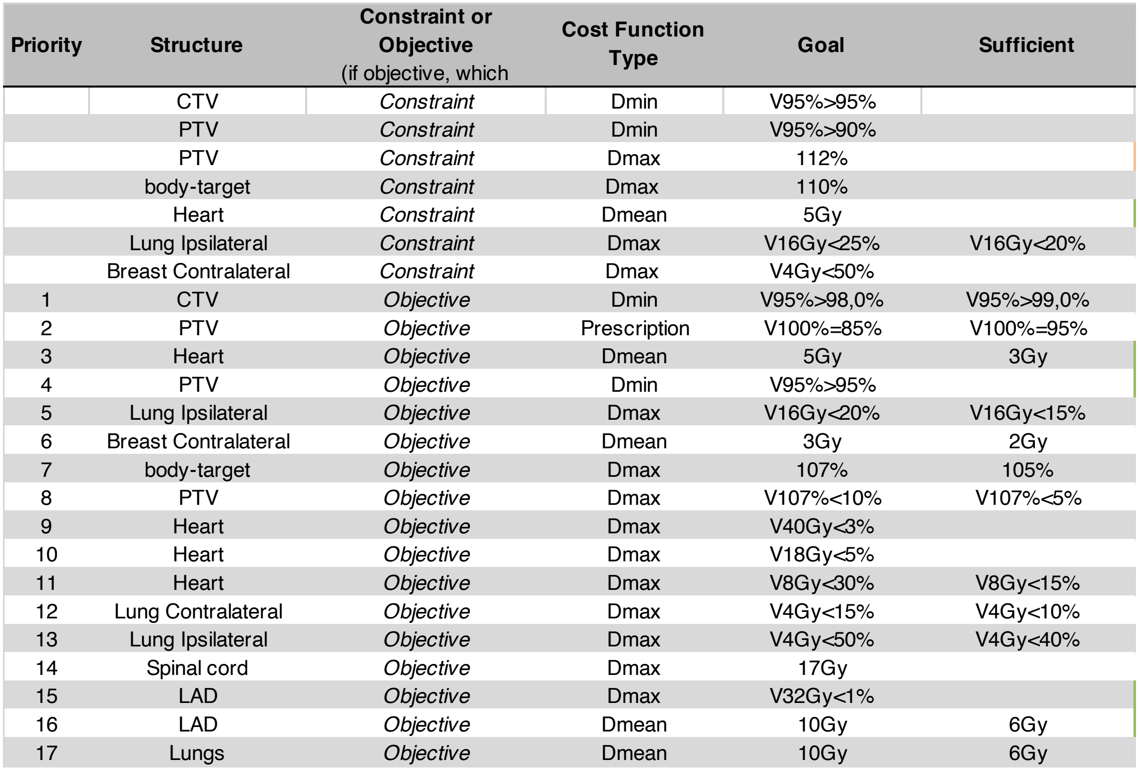
Pareto optimal plans, generated with the Erasmus-iCycle module, were compared with clinical plans manually generated using the AUSL-IRCCS clinical treatment planning system (TPS), Eclipse (Varian Medical Systems). Comparison between Pareto optimal Erasmus-iCycle plans and manual plans were made in terms of dosimetric endpoints and DVH metrics. Two-sided Wilcoxon signed-rank tests were used to analyse plan differences, using $p<0.05$ for statistical significance. The plan was scored with a plan quality metric (PQM) to evaluate the comprehensive quality of auto plans with respect to manual ones. The PQM is a user defined metric intended to quantify and compare the plan quality by mimicking the judgement of a physician, which consisted of a set of clear and specific plan objectives of the treatment. To each objective, the user associated a numerical scoring function to model as accurately as possible the judgement criteria of the clinicians. The PQM was the sum of the scores obtained by each objective and measured how much the plan adhered to the list of identified goals. The percentage PQM (PQM%) thus represented a relative measure of plan quality. In this work, the PQM% was calculated using PlanIQ software, described in chapter 5. A specific list of objectives was configured in the planIQ tool and a score was associated to each object. The list is based on clinical objectives reported in Table 8.1 (see subsection 8.4.1). The FDVH tool, implemented in PlanIQ, that was able to create a feasibility DVH based on an ideal dose fall-off from the prescription dose at the target boundary (see chapter 5), was used to evaluate the overall quality of manual plans with respect to auto plans. The inter-plan quality variations were also calculated and compared between the auto and manual plans using Student's t-test.

8.4 RESULTS

8.4.1 *Wish-list definition for AUSL-IRCCS Reggio Emilia hospital*

The clinical wish-list, is shown in Table 8.1. This referred in its major clinical constraints to RTOG and StartB protocols [131–133].

The converted wish-list in the Erasmus-iCycle module is reported in Table 8.2. The conversion from clinical to Erasmus-iCycle wish-list is not obvious, and the majority of objectives and constraints need to be refined to obtain the desired solution. The functions reported below, such as LTCP, EUD and QUOP are described in chapter 3.

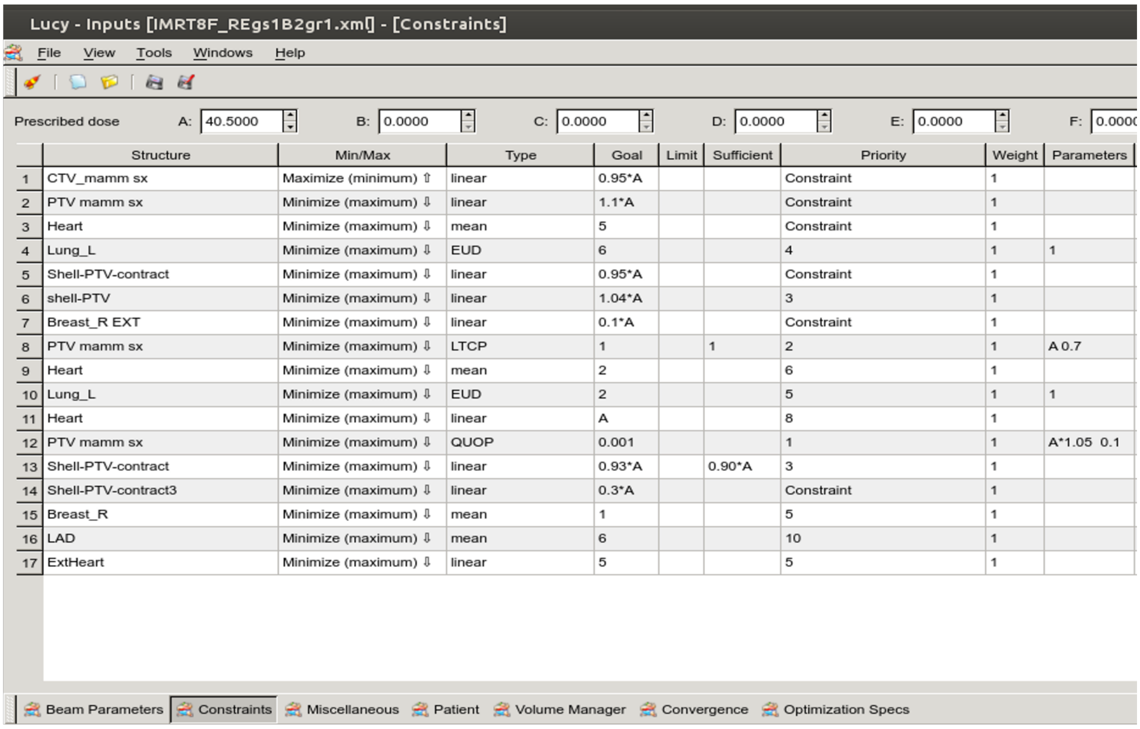


Priority	Structure	Constraint or Objective (if objective, which	Cost Function Type	Goal	Sufficient
	CTV	*Constraint*	Dmin	V95%>95%	
	PTV	*Constraint*	Dmin	V95%>90%	
	PTV	*Constraint*	Dmax	112%	
	body-target	*Constraint*	Dmax	110%	
	Heart	*Constraint*	Dmean	5Gy	
	Lung Ipsilateral	*Constraint*	Dmax	V16Gy<25%	V16Gy<20%
	Breast Contralateral	*Constraint*	Dmax	V4Gy<50%	
1	CTV	*Objective*	Dmin	V95%>98,0%	V95%>99,0%
2	PTV	*Objective*	Prescription	V100%=85%	V100%=95%
3	Heart	*Objective*	Dmean	5Gy	3Gy
4	PTV	*Objective*	Dmin	V95%>95%	
5	Lung Ipsilateral	*Objective*	Dmax	V16Gy<20%	V16Gy<15%
6	Breast Contralateral	*Objective*	Dmean	3Gy	2Gy
7	body-target	*Objective*	Dmax	107%	105%
8	PTV	*Objective*	Dmax	V107%<10%	V107%<5%
9	Heart	*Objective*	Dmax	V40Gy<3%	
10	Heart	*Objective*	Dmax	V18Gy<5%	
11	Heart	*Objective*	Dmax	V8Gy<30%	V8Gy<15%
12	Lung Contralateral	*Objective*	Dmax	V4Gy<15%	V4Gy<10%
13	Lung Ipsilateral	*Objective*	Dmax	V4Gy<50%	V4Gy<40%
14	Spinal cord	*Objective*	Dmax	17Gy	
15	LAD	*Objective*	Dmax	V32Gy<1%	
16	LAD	*Objective*	Dmean	10Gy	6Gy
17	Lungs	*Objective*	Dmean	10Gy	6Gy

Table 8.1: Clinical wish-list defined with the radiotherapy department staff of AUSL-IRCCS of Reggio Emilia (Italy) for left breast treatment, based on major clinical international protocols.

8.4.2 *Generation of Pareto optimal plan with Erasmus-iCycle*

The automatic plan (Auto) was generated in Erasmus-iCycle for all 16 patients considered using the wish-list reported in Table 8.2 and the 2 schemes of field setup described in the Materials and Methods section. The Auto plan dose was then exported in DICOM format from the Erasmus-iCycle tool and imported into the Eclipse TPS as a virtual plan, to be compared with the clinical manual plan (Manual) dose.



Prescribed dose A: 40.5000 B: 0.0000 C: 0.0000 D: 0.0000 E: 0.0000 F: 0.0000

	Structure	Min/Max	Type	Goal	Limit	Sufficient	Priority	Weight	Parameters
1	CTV_mamm sx	Maximize (minimum) ⇧	linear	0.95*A			Constraint	1	
2	PTV mamm sx	Minimize (maximum) ⇩	linear	1.1*A			Constraint	1	
3	Heart	Minimize (maximum) ⇩	mean	5			Constraint	1	
4	Lung_L	Minimize (maximum) ⇩	EUD	6			4	1	1
5	Shell-PTV-contract	Minimize (maximum) ⇩	linear	0.95*A			Constraint	1	
6	shell-PTV	Minimize (maximum) ⇩	linear	1.04*A			3	1	
7	Breast_R EXT	Minimize (maximum) ⇩	linear	0.1*A			Constraint	1	
8	PTV mamm sx	Minimize (maximum) ⇩	LTCP	1		1	2	1	A 0.7
9	Heart	Minimize (maximum) ⇩	mean	2			6	1	
10	Lung_L	Minimize (maximum) ⇩	EUD	2			5	1	1
11	Heart	Minimize (maximum) ⇩	linear	A			8	1	
12	PTV mamm sx	Minimize (maximum) ⇩	QUOP	0.001			1	1	A*1.05 0.1
13	Shell-PTV-contract	Minimize (maximum) ⇩	linear	0.93*A		0.90*A	3	1	
14	Shell-PTV-contract3	Minimize (maximum) ⇩	linear	0.3*A			Constraint	1	
15	Breast_R	Minimize (maximum) ⇩	mean	1			5	1	
16	LAD	Minimize (maximum) ⇩	mean	6			10	1	
17	ExtHeart	Minimize (maximum) ⇩	linear	5			5	1	

Table 8.2: Erasmus-iCycle wish-list for left breast treatment.

8.4.3 *Dosimetric comparison of manual and auto plans*

The Manual and Auto plans resulted in very similar DVHs. Figure 8.2 reports the DVH comparison for each patient separately. Average DVH comparison was reported in Figure 8.3. Right lung and spinal cord are the OARs for which the mean DVHs were visibly different (with Auto better than Manual). For the remaining OARs the average DVH results were quite similar.

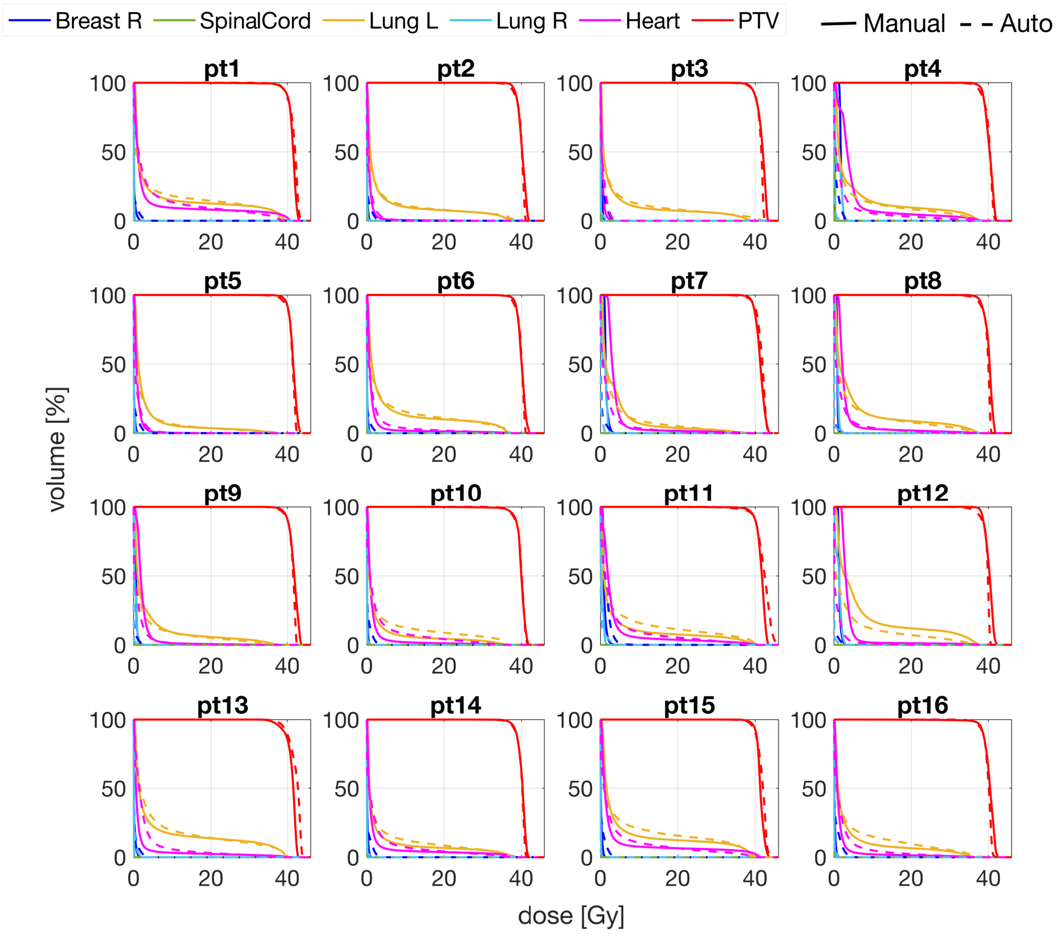


Figure 8.2: DVH comparison between manual clinical plan (Manual) and automated plan performed with Erasmus-iCycle module (Auto) for all 16 patients considered in the study (labelled from 1 and 16).

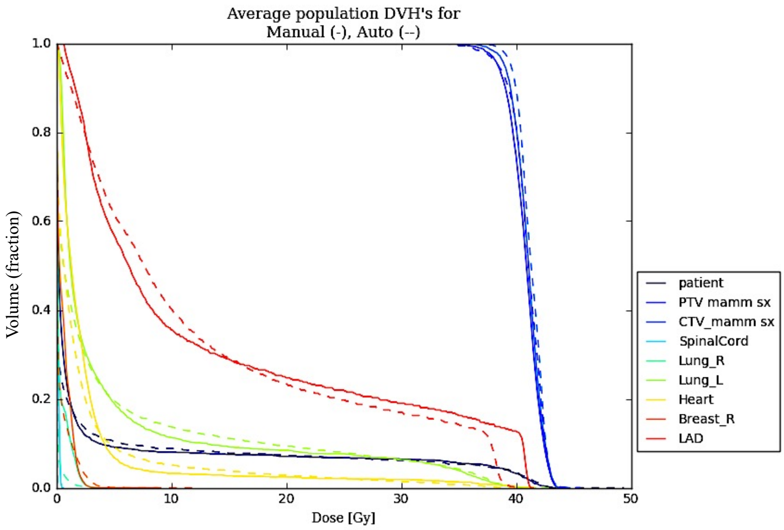


Figure 8.3: Average DVHs, over the 16 patients, comparison between manual clinical plans (continuous lines) and automated plans performed with Erasmus-iCycle (dashed lines).

All Manual and Auto plans were clinically acceptable with tumour coverage (99% of volume of CTV) covered by 95% of prescription dose while fulfilling all OAR constraints (See Table 8.3). Statistical tests showed significant differences in terms of PTV maximum dose and prescription volume, even if the values are comparable and within clinical goals.

Conformity Indexes (CIs) for Manual plans were slightly better than for Auto plans (0.96 vs 0.94), but the difference was insignificant (p=0.25). All OAR endpoints were within the clinical goals for both plan types. Manual and Auto plans were comparable in terms of OAR sparing: statistically significant differences occurred only for left Lung V(16 Gy) (p=0.01) for which the Manual plan is slightly but significantly better than the Auto plan. The dose in the axilla, quantified by shell-PTV is comparable and insignificant differences were observed.

Main endpoints are also reported in Figure 8.3 and Figure 8.4 for each single patient in terms of a spider plot. This is a two-dimensional chart type containing a series of values over multiple quantitative variables (main endpoints). Each endpoint has its own axis, all axes are joined in the centre of the figure. For endpoints such as right breast and heart, mean dose (Dmean), left lung V(16Gy) and V(5Gy) (i.e. volume receiving 16 Gy or 5 Gy, respectively) and PTV maximum dose (Dmax) the better plan is one with lower values, thus more close to the center of the graph. For PTV D(85%) (prescription endpoint) and CTV D(99%)

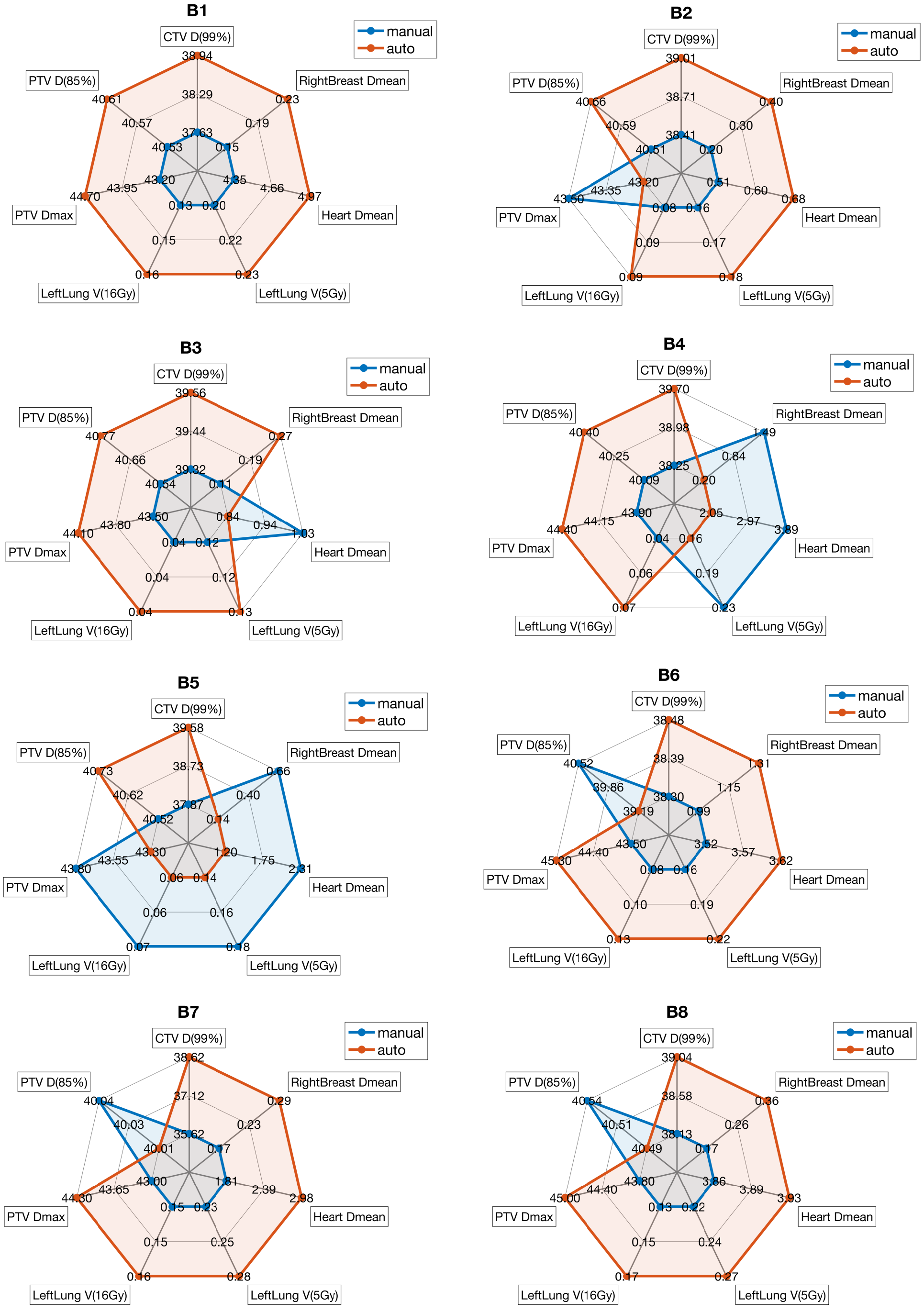


Figure 8.4: Main endpoints differences between Manual and Auto plans divided by patient (from patient#1 (B1) to patient#8 (B8))in term of spider plot.

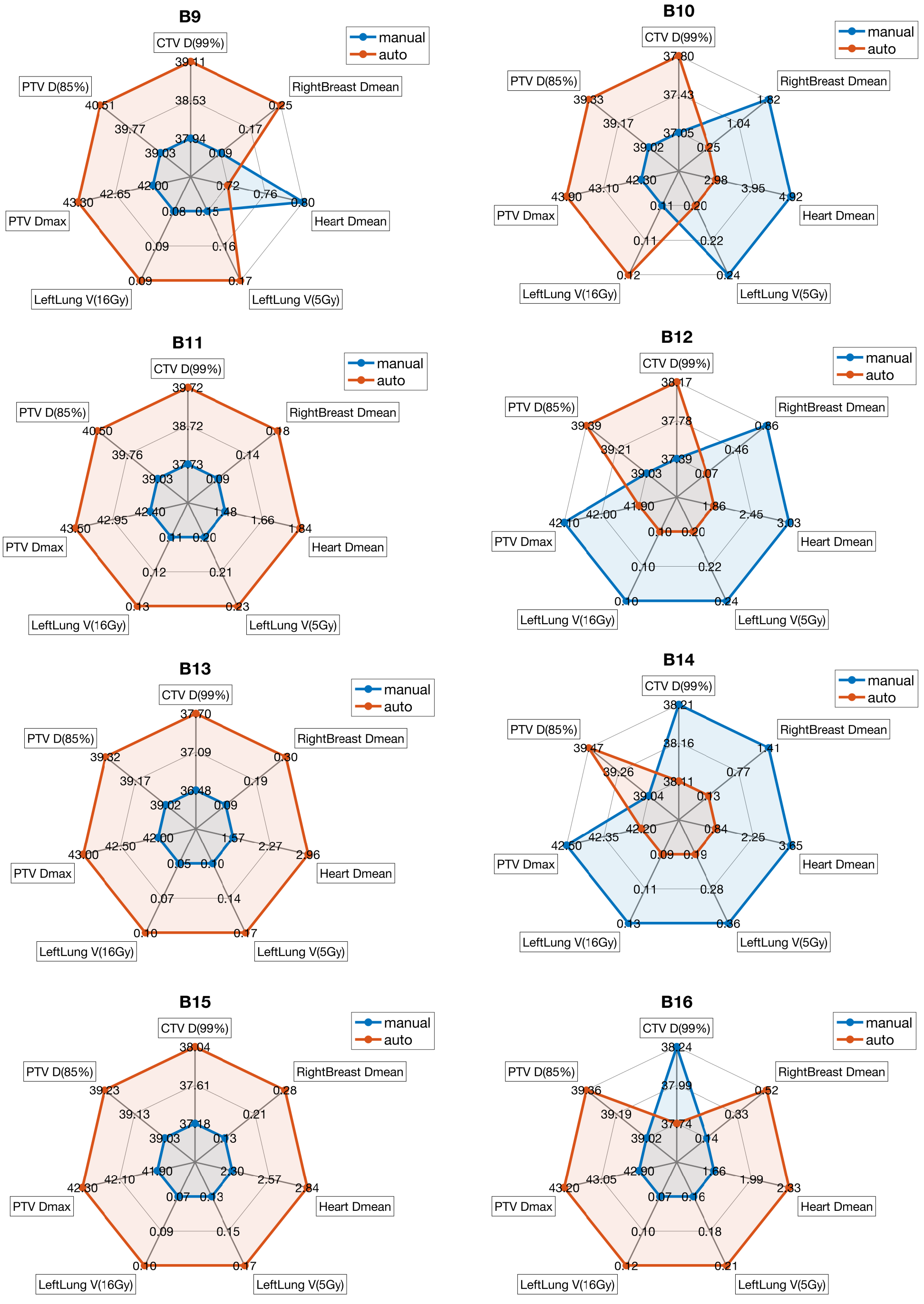


Figure 8.5: Main endpoints differences between Manual and Auto plans divided by patient (from patient#7(B7) to patient#16(B16)) in term of spider plot.

		Clinical Goal	Manual Plans	Automated Plans	p-value
			Mean ± SD	Mean ± SD	
CTV	*D(99%) [Gy] - Scheme1*	*38.5*	**39.12 ± 0.46**	**38.30 ± 0.73**	p<0.01
	D(99%) [Gy] - Scheme2	*37.1*	**37.94 ± 1.06**	**37.53 ± 0.61**	p<0.01
PTV	*D(85%) [Gy] - Scheme1*	*40.5*	**40.10 ± 0.67**	**40.13 ± 0.71**	p<0.01
	D(85%) [Gy] - Scheme2	*39*	**39.03 ± 0.01**	**39.64 ± 0.54**	p<0.01
	Dmax [Gy] - Scheme1	*44.5*	**43.23 ± 0.69**	**43.69 ± 1.12**	p<0.01
	Dmax [Gy] - Scheme2	*42.9*	**42.26 ± 0.33**	**42.91 ± 0.70**	p<0.01
	CI	-	0.96 ± 0.04	0.94 ± 0.04	p=0.25
Body	*Dmax [Gy] - Scheme1*	*44.5*	43.23 ± 0.69	43.25 ± 1.21	p=0.22
	Dmax [Gy] - Scheme2	*42.9*	42.26 ± 0.34	42.44 ± 0.64	p=0.22
Heart	*Dmean [Gy]*	*5 (mandatory)- 3(optimal)*	2.54 ± 1.36	2.29 ± 1.27	p=0.80
	V(5Gy) %	*15%(mandatory) -5%(optimal)*	7% ± 6%	9% ± 6%	p=0.08
	V(20Gy) %	*10%(mandatory)*	2% ± 2%	3% ± 3%	p=0.14
Left Lung	*Dmean [Gy]*	*10 (mandatory)- 6(optimal)*	4.75 ± 1.32	4.73 ± 1.26	p=0.68
	V(5Gy) %	*50%*	19% ± 6%	20% ± 4%	p=0.38
	V(16Gy) %	*20%*	**9% ± 3%**	**11% ± 4%**	p=0.01
Right Breast	*Dmean [Gy]*	*3 (mandatory)- 1(optimal)*	0.54 ± 0.59	0.32 ± 0.28	p=0.92
LAD	*Dmean [Gy]*	*10 (if it is possible)*	12.94 ± 9.04	12.61 ± 9.67	-
Shell-PTV	*V(100%)*	*ALARA*	3.7% ± 7.9%	1.9% ± 1.6%	p=0.80
(auxilla)	*V(105%)*	*ALARA*	0.1% ± 0.2%	0.0% ± 0.1%	p=0.50
	Dmax [%]	*<107%*	104.8% ± 1.7%	105.6% ± 1.5%	p=0.15

Table 8.3: For all 16 patients, mean values for automatically generated plans (Auto plans) and manually generated (Manual plans). Bold values represent the statistically significant differences as $p<0.05$ with the Wilcoxon's signed-rank test. CI = Conformity Index. Shell-PTV is a dummy structure create to reduce the dose to auxilia (see Figure 8.2) and material and method section.

(coverage endpoint), (i.e. the dose received by the structure to 85% and 99% of total volume, respectively), the higher value means the better results, thus more distant from the centre of the graph.

The differences in evaluations between Manual and Auto plans performed with PlanIQ software in terms of PQM% are shown in Figure 8.5. The median values of PQM% were 91.7 and 91.9, for Manual and Auto plans respectively, with a SD of 9.5 and 9.1. This means that auto plans obtained slightly better scores with smaller variation among plans. However, the difference was statistically insignificant with a p-value of 0.92 (Student's t-test).

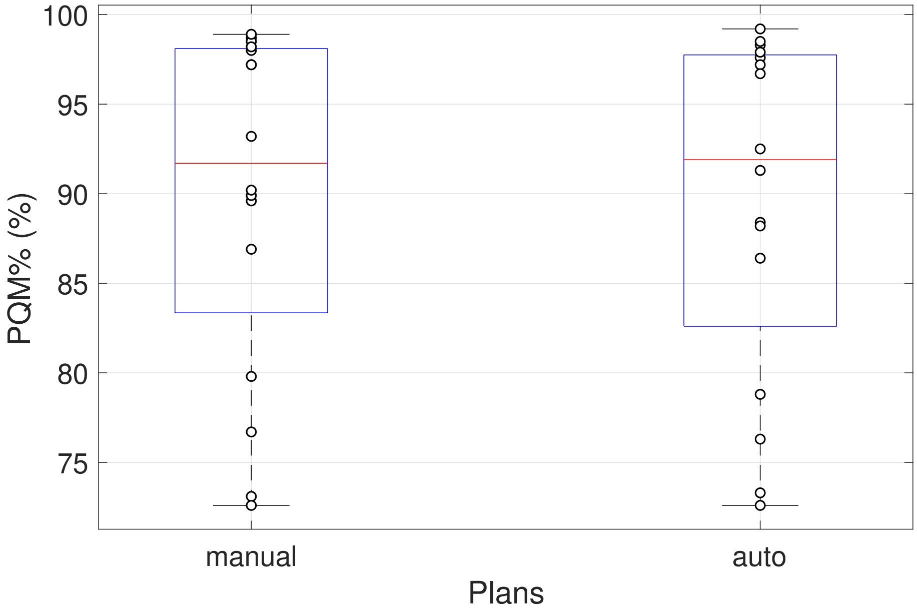


Figure 8.6: Boxplot of PQM% analysis with PlanIQ for Auto and Manual plans. For each box, the central mark represents the median value, while the bottom and top edges of the box are the 25th and 75th percentiles, respectively. The 'whiskers' represent the range of values. The circles represent individual plan PQM%.

8.5 DISCUSSION

The goal of the study presented in this chapter was to configure, test, and implement Erasmus-iCycle for automated, multicriterial IMRT treatment planning for left breast patients. For 18 consecutively treated patients, auto plans were compared to manually generated, clinically delivered plans using various dosimetric indices. The Manual plans and Auto plans were comparable in term of PTV and CTV coverage, homogeneity and OAR sparing. Moreover, IMRT auto plans showed similar results in terms of low dose bath compared to manual TF plans. Although the automated plans were of similar quality to the manual plans, the advantage of automation is in the consistency of plan generation and in a significant reduction of the time spent to generate the plans [7, 82, 101–104]. The automated treatment planning procedure produced plans that fulfilled all the clinical constraints in 100% of the cases. Similar results were achieved by Manual plans as shown in Table 8.3. Compared to the HN site, the breast site has more patient anatomy variation. The main OARs are located very close to the target in both cases, but in the latter (breast) low doses are also important as clinical endpoints. This changes the paradigm of the problem with respect to HN automatic planning, since it is more related to individual geometry (i.e. OAR distance from PTV, OAR volume and shape). For this reason, the dataset of patients to include in this comparison (Manual vs. Auto) should be enlarged to include several geometries. One hypothesis to be verified is that the observed differences between Auto and Manual plans were very small due to the limited cohort of breast patients. Also the plan quality variation measured by PlanIQ in terms of PQM% showed similar variability and quality between the two groups. This could be related to the choice of Manual plans, made by the same planner for 90% of cases. This condition does not represent the real clinical practice of in a medium size department, in which the plans are generally performed by 4-5 planners. The study presented in this chapter is part of a larger programme of work (see chapter 11) for further details). It is a prelude to a bigger study to use the methods developed in chapter 6 for HN, to perform a similar study on inter-user variability in plan quality assessment for the breast site, involving a larger number of patients and more than one institute. The project's long term main aim is to quantify the differences between ROs and MPs intra and inter-institute in perceived quality of breast treatment plans. As discussed in chapter 6, broad and inter-departmental discussions on plan requirements, aiming at a broadly shared, and well defined view on plan quality, could improve the inter-observer variation in plan quality assessments. The hypothesis is that, as found for the HN case, the introduction of automation techniques into breast radiotherapy planning practice can reduce inter-planner and inter-evaluator variability. A second aim is to use the automated plans as reference data to implement an independent template for breast radiotherapy planning using more complex techniques. Pareto optimal plans, generated with the Erasmus-iCycle module, presented in this chapter, will be used as reference plans to develop and define a more complex in-house technique (IMRT or VMAT) using the AUSL-IRCCS clinical treatment planning system (TPS), Eclipse (Varian Medical Systems).

8.6 CONCLUSIONS

Compared to conventional planning, Auto plan generation for left breast for the cohort considered in this study led to similar PTV and CTV coverage while maintaining similar OAR dose, and while keeping the low dose outside targets comparable. Both Auto and Manual treatment planning produced plans that fulfilled all the clinical constraints in 100% of the cases. Auto plans were quicker to generate than Manual, with a factor depending on the plan complexity, a factor that makes the Manual solution harder and thus more time consuming. Moreover, Auto showed slightly higher consistency in plan quality measured with PlanIQ tool. Further work will be done to confirm these preliminary results performed with a limited cohort of patients by including more cases with different kinds of anatomy.

Part IV

APPLICATION TO ADAPTIVE RADIOTHERAPY

9

EVALUATING THE QUALITY OF DEFORMABLE IMAGE REGISTRATION IN ADAPTIVE RADIOTHERAPY USING A DIGITALLY ENHANCED PHANTOM.

> “If you knew Time as well as I do,”said the Hatter, “you wouldn’t talk about wasting IT. It’s HIM.”
>
> *Lewis Carroll-Alice in Wonderland*

9.1 PREVIEW

Work from this chapter was published in Cagni *et al.* (Appl.Science 2022) [134]. The figures and tables that are shown in this chapter are drawn from that published work. Work from this chapter is in preparation for submission as an original article in a peer-reviewed journal. It presents a registration-based method for DIR quality assurance for ART using digitally post-processed head and neck anthropomorphic phantom image datasets. The investigation of certain methods that allow evolution towards real-time ART were performed. The development of automated tools that enable high-quality ART is likely to lead to significant efficiency gains and foster wider clinical uptake. The goal of this study was to assess a verification method to contribute to the automation and standardisation of and standardize the DIR verification phase.

9.2 INTRODUCTION

External beam radiotherapy is gradually evolving towards real-time adaptive radiotherapy (ART) [135], which is being developed as a new paradigm in radiation oncology [136, 137]. However, ART is, at present, not standardised or widely employed [138]. This is due not only to the time-consuming processes of the delineation of targets and organs at risk on daily computed tomography (CT) images and treatment re-planning, but also to the limited quality of the daily on-board cone beam computed tomography (CBCT) images. Such limitations add practical constraints to the ability to delineate structures manually and estimate the daily delivered dose accurately and also contribute to the limited accuracy of registration between daily CBCT and planning CT images [139]. The development of automated tools that enable high-quality ART is likely to lead to significant efficiency gains and foster wider clinical uptake [139]. In the process of delivering automated ART solutions, a challenging task is the validation of image registration algorithms for clinical use, as their performance depends on the complexity and quality of the images used in the registration task [25]. In the ART workflow, the deformable vector field (DVF), which is generated during deformable image registration (DIR), could also be applied for dose mapping and dose accumulation purposes. In this process, any uncertainty in DVF generation would be propagated directly to the calculated dose map. Considerable research has been done in the investigation of DIR accuracy, primarily through the creation of quality assurance (QA) metrics. Several QA methods for validating image registration have been proposed [140]. The report of the American Association of Physicists in Medicine (AAPM) Task Group 132 (TG132) provided an essential set of guidelines for quality assurance (QA) and quality control of image registration operations for the overall clinical process. The TG132 report recommends a series of tests and corresponding metrics that should be evaluated and reported during image registration software verification. AAPM TG 132 suggests the use of a known DVF to test deformable image registration [140]; this method is generally used with digital phantoms that consist of two image sets linked via a known DVF (ground-truth). Several phantom data sets have been generated by AAPM TG 132 [140] for use in QA programmes for this purpose. However, a few groups found incompatibility of certain digital phantoms provided by the TG132 report with commercial software [141, 142] and/or they found some of the provided phantom images still limited in their utility. For example, only a pelvic anatomical phantom was made publicly available for DIR verification and only a single known transformation (refDVF) was provided for test purposes [140]. Pukala *et al.* [143] reported differences observed between 10 virtual head and neck phantoms for DIR verification over 5 different commercial software systems; these results emphasized the need to assess DIR accuracy on each considered clinical DIR. In this chapter the work is focused on DIR verification for head and neck adaptive radiotherapy. Practical guidelines and dedicated phantoms for clinical implementation are still needed to test the accuracy of deformable image registration (DIR) in various clinical situations. Registration between CT and CBCT image sets can be considered a multimodal registration problem. The consistency of pixel intensity values between the two modalities is restricted because the image quality of CBCT is poorer than that of CT [144, 145] and CBCT images are commonly affected by artefacts due to scatter contamination and truncated CBCT patient volumes [146–150]. The use of digital phantoms for CT-CBCT DIR QA is limited by the fact that the applied transformations are not patient-specific but are based on pre-defined DVFs. Furthermore, the deformed images of digital phantoms do not have any representative noise variation inherent in the CBCT images of an actual patient [25]. It is possible to overcome such limitations using images of a physical phantom acquired with systems and acquisition protocols used in clinical practice. Physical phantoms do not reflect the complexity of real human anatomy which is affected by many factors, such as setup variation, organ motion and organ deformation. These factors contribute to the inherent differences between CT and CBCT images, even if both image sets are obtained on the same day. However, the use of a rigid anthropomorphic phantom allows the acquisition of image datasets with identical anatomy for both CT and CBCT scans. This ensures robust and reliable comparisons by providing a reference, as any discrepancies between the primary CT and CBCT images would be purely related to the performance of the DIR algorithm.

In this chapter, we have processed the images of a commercial anthropomorphic phantom to include common OARs as recommended by international guidelines [151]. The aim of this study was to build and test a novel workflow for the verification of patient-based DIR in HN ART using the developed phantom images.

9.3 MATERIALS AND METHODS

9.3.1 *Digitally enhanced phantom*

The study used the ATOM Max Dental and Diagnostic Head Phantom Model-711 (CIRS, Norfolk, VA) (HN phantom). This HN phantom approximates to the average male human head in both size and structure. The original phantom includes detailed 3D anthropomorphic anatomy, such as the brain, bone, larynx, trachea, sinus, nasal cavities and teeth. The bones contain both cortical and trabecular components. The teeth include distinct dentine, enamel, and root structures, including nerves. The sinus cavities are fully open. The HN phantom was scanned with both CT and CBCT at AUSL-IRCCS hospital of Reggio Emilia. The CT image datasets were acquired using a GE Hi-Speed (General Electric Company, Boston, USA) scanner with acquisition parameters set in agreement with local clinical HN protocol (120 KV, 100 mA, 2 mm slice thickness). As a second step, CBCT image sets were acquired on a TrueBeam STx linear accelerator (Varian Medical Systems, Palo Alto, CA) using local HN image acquisition protocols (HN modality). To acquire the entire head of the phantom, a multi-scan CBCT was performed, and the data were combined using the TrueBeam STx advanced reconstruction module. Both CT and CBCT image sets were post-processed to digitally insert several OARs not present in the original phantom, including the brainstem, oral cavity, left and right parotid glands, larynx and eyes. Firstly, a radiation oncologist drew outlines of these anatomical structures on the phantom CT and CBCT images, following international guidelines [151] for their standard shape, anatomical position, and volume. Subsequently, the phantom image Hounsfield Unit (HU) values were replaced with surrogate HU values for all the above-mentioned structures using an in-house developed script based on MATLAB code (The Mathworks, Natick, MA) and Computational Environment for Radiological Research tool (CERR, http://www.github.com/cerr/CERR) [152]. The surrogate HU value in each voxel i of a given organ OAR, $\mathrm{HU}^{\mathrm{phantom}}_{\mathrm{processedOAR_i}}$, was derived using the following equation:

$$\mathrm{HU}^{\mathrm{phantom}}_{\mathrm{processedOAR_i}} = (\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{OAR}}} - \overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{t}}}) + \mathrm{HU}^{\mathrm{phantom}}_{\mathrm{originalOAR_i}} \tag{9.1}$$

Where $\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{OAR}}}$ and $\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{t}}}$ are the mean HU values over the voxels included in a ROI belonging to a specific organ ($\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{OAR}}}$) and tissue around that organ ($\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{t}}}$) CT or CBCT images respectively; $\mathrm{HU}^{\mathrm{phantom}}_{\mathrm{originalOAR_i}}$ is the HU value in the original HN phantom CT or CBCT image for each voxel. The $\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{OAR}}}$ and $\overline{\mathrm{HU}^{\mathrm{pt}}_{\mathrm{t}}}$ were measured by taking the mean value of several (typically 4) square ROIs of 10x10 pixels placed in different positions inside the OAR and its surrounding tissue, to reflect typical HU variation in real patient images. These mean values were then averaged over a set of 10 pairs of actual patient CT and CBCT images. This process was done separately for CT and CBCT image sets (i.e. for each patient 1 CT and 1 CBCT image was considered). Using this method, the intrinsic noise of each image type, CT and CBCT, was maintained in the digitally modified phantom image sets. In the rest of the chapter, the phantom image set relates to the post-processed enhanced phantom CT.

9.3.2 *General workflow*

The method used to validate the DIR applied to the CT-CBCT image pairs was based on a known transformation for each voxel [25]. This method requires an independent 'third party' image registration software system to generate the reference DVF. The third-party software is referred to 'reference DIR software', as it will be used to generate a reference DVF. The only requirement for the reference and the clinical DIR software systems should be the independence of the two systems in the DIR procedure to avoid biasing the results. As a guarantee of the correctness of the procedure it is not essential that the reference software should be any more accurate in the DIR process than the clinical one.

The set of clinical patient images (planning CT and CBCT set acquired at particular fraction) used to produce the clinical DIR to be tested, have to be registered using the reference software (see Figure 9.1a, spatial analysis workflow). This registration is considered as the reference for the patient CT-CBCT data sets ($\mathrm{refDVF_{pt}}$). As a second step, the HN phantom CT images were artificially deformed into 'warped' phantom CT (wCT) images by applying the $\mathrm{refDVF_{pt}}$ transformation. Using the clinical DIR software, the phantom CBCT image was registered with the wCT phantom image, generating a test DVF to be evalu-

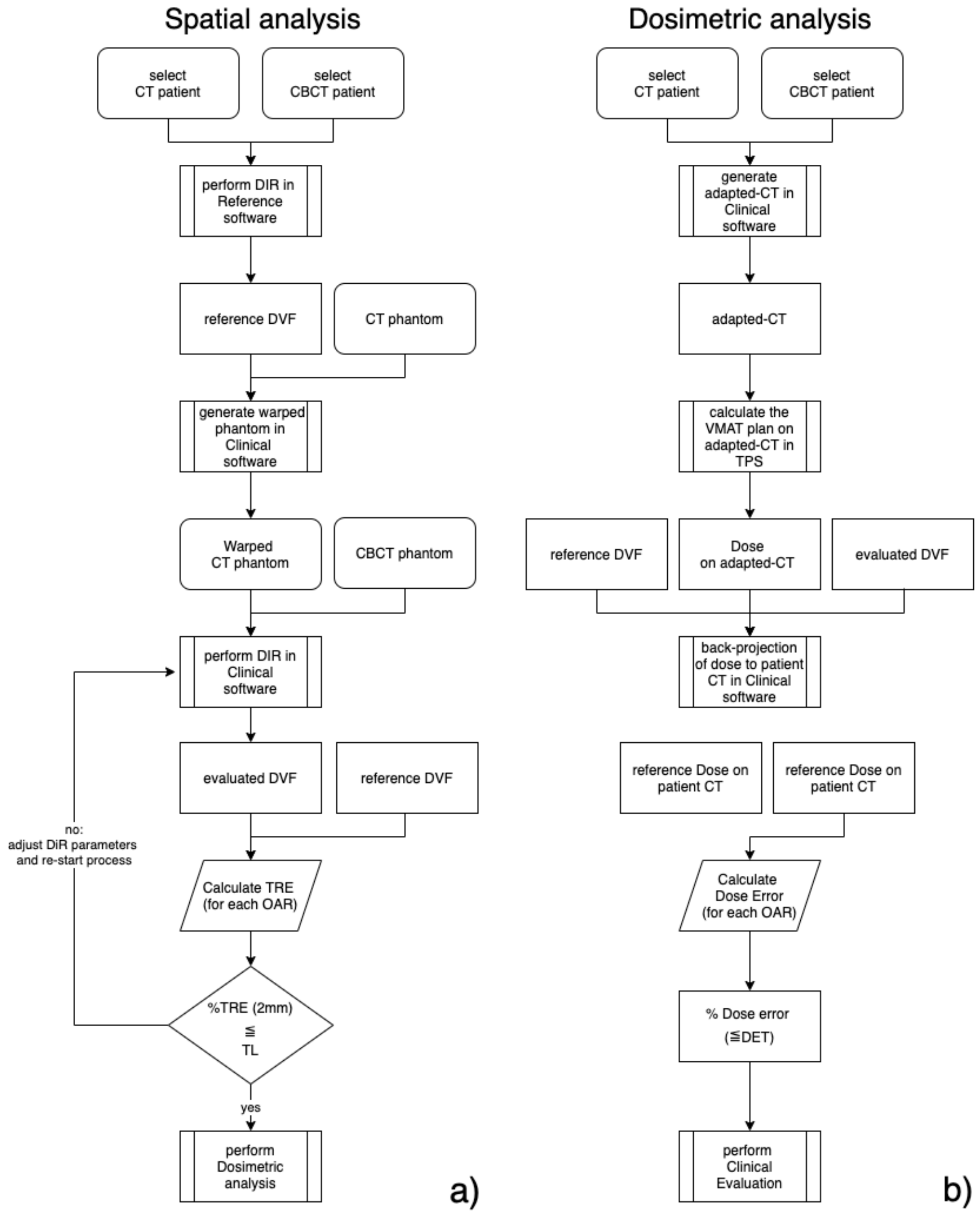

Spatial analysis
select CT patient
select CBCT patient
perform DIR in Reference software
reference DVF
CT phantom
generate warped phantom in Clinical software
Warped CT phantom
CBCT phantom
perform DIR in Clinical software
evaluated DVF
reference DVF
no: adjust DiR parameters and re-start process
Calculate TRE (for each OAR)
%TRE (2mm) ≦ TL
yes
perform Dosimetric analysis
a)
Dosimetric analysis
select CT patient
select CBCT patient
generate adapted-CT in Clinical software
adapted-CT
calculate the VMAT plan on adapted-CT in TPS
reference DVF
Dose on adapted-CT
evaluated DVF
back-projection of dose to patient CT in Clinical software
reference Dose on patient CT
reference Dose on patient CT
Calculate Dose Error (for each OAR)
% Dose error (≦DET)
perform Clinical Evaluation
b)

Figure 9.1: Schematic workflow of CT-CBCT DIR validation method for spatial and dosimetric analysis. a) Spatial Analysis. The HN patient CBCT-CT DIR was considered as reference DVF (refDVFpt). The refDVF$_{pt}$was generating using third-party reference software. Post-processed phantom CT images were digitally warped using refDVF$_{pt}$resulting in warped phantom CT. The phantom CBCT and warped CT images were then registered in the clinically DIR software, generating an evaluated DVF (testDVF$_{ph}$). The testDVF$_{ph}$was compared with the refDVF$_{pt}$by using a TRE metric and an operative tolerance level (OTL) analysis performed to quantify the accuracy of the DIR. b) Dosimetric Analysis. Treatment plans were calculated in the TPS on each patient's adapted-CT image (patient CT image warped into CBCT image) resulting in a delivered dose. The delivered dose was transferred back to the patient's planning CT image via both the refDVF$_{pt}$(reference dose) and testDVF$_{ph}$(test dose). The dose difference between the two was calculated with a clinical assessment required for a final DIR validation.

ated (testDVF$_{ph}$). Note that the term refDVF$_{pt}$relates to patient image registration throughout, whereas the term testDVF$_{ph}$always refers to the warped phantom CT and phantom CBCT DIR.

The quality of the registration was assessed as the ability of the clinical software to replicate the refDVFpt, artificially introduced into phantom wCT images. Target registration error (TRE) was computed as a metric to evaluate the difference between refDVF$_{pt}$ and testDVF$_{ph}$(subsection 9.3.4). As recommended by TG132, the percentage of voxels with TRE within 2 mm was compared with a designated OperativeTolerance Level (OTL) value (subsection 9.3.5). If the testDVF$_{ph}$does not overcome the OTL analysis, the evaluated registration process should be restarted adjusting certain registration parameters, such as the volume of interest.

The propagation of dosimetric errors due to differences in DIR maps is schematized in Figure 9.1b, labelled as dosimetric analysis workflow. VMAT clinical treatment plans were recomputed in the clinical TPS for each patient's CBCT adapted-CT images to avoid bias in dose calculation (see subsection 9.3.4) and back-projected to the corresponding patient CT images via both the refDVF$_{pt}$ and testDVF$_{ph}$, generating reference dose (refDose) and evalutated dose (testDose) arrays, respectively. Dosimetric differences between refDose and testDose should be evaluated in conjunction with clinical radiation oncologists.

9.3.3 *Clinical data*

Ten clinical patients with advanced oropharyngeal cancer who underwent radiotherapy at AUSL-IRCCS were randomly selected from the departmental database. All the plans used volumetric modulated arc therapy (VMAT, RapidArc) to treat three targets at dose levels of 69.96, 59.40 and 54.45 Gy in 33 fractions with simultaneous integrated boost, generated with an Eclipse Treatment Planning System (TPS) v.13.6 (Varian Medical Systems, Palo Alto, CA). Bilateral neck irradiation was delivered to all patients, with involved high dose PTV on the right side for 6 patients, left side for 2 patients and centrally for the remaining 2 patients. For each treatment fraction, a CBCT scan was acquired using the clinical HN protocol before delivering treatment to assess patient setup and anatomical variation. All CBCT images were automatically saved in an ARIA database (Varian Medical

Systems, Palo Alto, CA). Published studies have shown that anatomic changes in HN cancer patients are more pronounced in the first half of treatment [30], and based on the clinical institute experience that re-planning requests are often performed in the second half of treatment. Hence, for each patient, the CBCT scans performed around the middle (fraction #16) and latter parts of the treatment course (fraction #26) were extracted from the imaging database for further analysis.

9.3.4 *Demonstration of clinical application*

The extracted pairs of CBCT and planning CT images for the 10 clinical patients selected were imported into MICE Toolkit v.1.1.0 (NONPI Medical AB, Sweden) which was used in this study as reference software. MICE Toolkit was used to produce the refDVF$_{\text{pt}}$, being the only independent DIR software available in the institute for comparison with the clinical system (VelocityAI, described below). Within MICE, the open-source Elastix software module [153] was employed. The registration method used in this case is of the same type as the one in the clinical platform but implemented independently. It consists of a B-spline method with an adaptive stochastic gradient descent optimiser, with the number of resolutions equal to 3, the maximum number of iterations of 1000 and an interpolator order of 3, similar to those parameters reported in the literature [154, 155]. The DVF from each DIR (in total, 20 DVFs) was subsequently exported in MATLAB format, converted to binary format, and used as refDVF$_{\text{pt}}$.

The VelocityAI Oncology Imaging Informatics System v.4.0 (Varian Medical System, Palo Alto, CA) was the DIR module used in the institute in clinical practice. The post-processed phantom scans were imported into VelocityAI and linked to each of the 10 patients selected for the study. Since the clinical refDVF$_{\text{pt}}$ has the same frame of reference as the patient CT image, both phantom CT and CBCT images were rigidly registered to the CT image of the patient. This was done to align the phantom CT image with the clinical refDVF$_{\text{pt}}$ to produce a more clinically realistic warped phantom CT image. It is important that the warped phantom is created, applying as much as possible the deformation magnitude consistent with the spatial deformation of the patient images (same OARs). Moreover, the lack of an initial deformation or shift between phantom CT and CBCT (owing to the rigidity of the phantom) was the basis for the reference condition. Following the workflow described in Figure 9.1 using VelocityAI, the HN phantom CT images were artificially deformed into wCT images using each of the imported refDVF$_{pt}$. Then, the phantom CBCT image was registered with the phantom wCT image, generating a testDVF$_{ph}$ obtained using the VelocityAI algorithm. The 'CBCT corrected multi-pass deformable' modality in VelocityAI was used to generate the testDVF$_{ph}$. This applies a 'fade correction' prior to the registration which enhances low signal regions in the CBCT image [39]. VelocityAI's primary registration algorithm uses a multi-resolution approach based on mutual information. The transform is a cubic B-spline, the interpolator is a bi-linear interpolation function, and the optimiser is based on the steepest gradient descent method [39]. The testDVF$_{ph}$ was then exported from VelocityAI in binary format. This workflow was repeated for all 20 registrations for the 10 patients. To ensure that the simulated deformation could be used as a reference to validate the image registration algorithm, for each case the warped phantom CT images were carefully inspected to verify that the changes induced in the phantom scans produced by refDVF$_{pt}$ were realistic. Each testDVF$_{ph}$ was compared to the refDVF$_{pt}$ using MATLAB in-house scripts. The quality of the registration was assessed as the ability of the testDVF$_{ph}$ to replicate the artificially generated refDVF$_{pt}$. For each patient and treatment fraction considered, an adapted-CT image of each refDVF$_{pt}$ was generated in VelocityAI. Adapted-CT images used the HU values of the patient CT image mapped onto the CBCT image to avoid bias in dose calculation [39]. The adapted-CT images were then imported into the Eclipse TPS and the VMAT clinical treatment plans were recomputed on each patient's adapted-CT images and sent back to VelocityAI. To assess dose propagation errors due to differences in DIR maps, the dose on the adapted-CT was back-projected to the corresponding patient CT images via both the refDVF$_{pt}$ and testDVF$_{ph}$.

9.3.5 *Data Analysis*

Spatial and dosimetric errors were calculated for all the voxels contained within the following OARs: brainstem, spinal cord, mandible, left parotid, right parotid, larynx, and oral cavity, as well as external body contours. In two of the 10 patients, the larynx was missing as an OAR because those patients were affected by laryngeal tumours which were included in the clinical target volumes. Spatial errors were calculated using the TRE [25], which describes the difference between co-located voxels once they have been transferred through the refDVF$_{\mathrm{pt}}$ and testDVF$_{\mathrm{ph}}$. TRE values were evaluated as a function of treatment fraction number and refDVF$_{\mathrm{pt}}$ magnitude on a voxel basis. The percentage of voxels with TRE less than or equal to 2 mm was considered in the analysis as a significant parameter for spatial errors, as was suggested by AAPM-TG132 [140]. Dose errors were calculated on patient CT images as being the difference between the back-projected doses using the two propagation methods (reference vs. evaluated). Dose errors were evaluated on a voxel basis. As a significant metric for dose errors, the percentage of voxels within a specific dose error threshold (DET) was calculated. In this study a significant value for DET was considered as 5% of the prescribed dose of 70 Gy, thus corresponding to 3.5 Gy.

9.3.6 *Tolerance level based on AAPM-TG132*

AAPM TG132 Report [140] provided digital phantoms for use in commissioning and quality assurance programs for image registration accuracy tests. For rigid registration, TG-132 proposed 2 data sets created using ImSimQA (Oncology Systems Limited, UK) (basic phantom data set and anatomic data set (a pelvic phantom)) for various modalities (CT, CBCT, PET, MRI-T1, and MRI-T2) for rigid registration. For DIR accuracy assessment, TG-132 provides a dataset of anatomical pelvic phantom images comprising:

1. a basic anatomical dataset (CT);

2. a basic deformation dataset, same as basic anatomical dataset with added gaussian noise variation to simulate CBCT image;

3. a ground truth DVF transformation (TG132-refDVF) file in dicom format. The basic anatomical phantom CT provided by the AAPM was deformed by using the provided TG132-refDVF. The registration between warped anatomical phantom CT and basic deformed phantom (CBCT) was performed in

Velocity AI, generating the TG132-testDVF. Differences between TG132-refDVF and TG132-testDVF were quantified in terms of target registration error (TRE). The results of this test were used to establish an OTL for DIR accuracy in the clinical dataset used in this study.

9.3.7 *Statistical analysis*

The Wilcoxon two-sided signed-rank tests and the Pearson's correlation coefficient metric were used to assess the statistical significance of the observed TRE differences between fractions and TRE as a function of refDVF$_{pt}$ magnitude, respectively. The correlation between dose error, TRE, dose gradients and the combination (scalar product) of these two (TRE and dose error) was assessed using Pearson's correlation. P-values$<$0.05 were considered significant.

9.4 RESULTS

9.4.1 *Digitally enhanced phantom*

Original and post-processed CT and CBCT phantom images in three different transverse sections are shown in Figure 9.2. Table 9.1 reports the HU values of OARs in the original and processed CT and CBCT phantom images.

9.4.2 *Reference DVFs for patient CT-CBCT registrations*

Visual inspection of deformed images and DVF magnitude was performed after each registration using the designated reference (MICE Toolkit) software to check the algorithm performance and the integrity of resulting DVF. The magnitudes of refDVF$_{pt}$ (mean ± SD) over all fractions and patients were: 7.56±2.33 mm, 4.44±3.26 mm, 5.17±4.94 mm, 3.67±1.5 mm, 3.70±1.33 mm, 3.77±2.44 mm, 12.57±8.94 mm and 5.52±2.54 mm for the body, left parotid gland, right parotid gland, larynx, oral cavity, mandible, brainstem, and spinal cord, respectively.

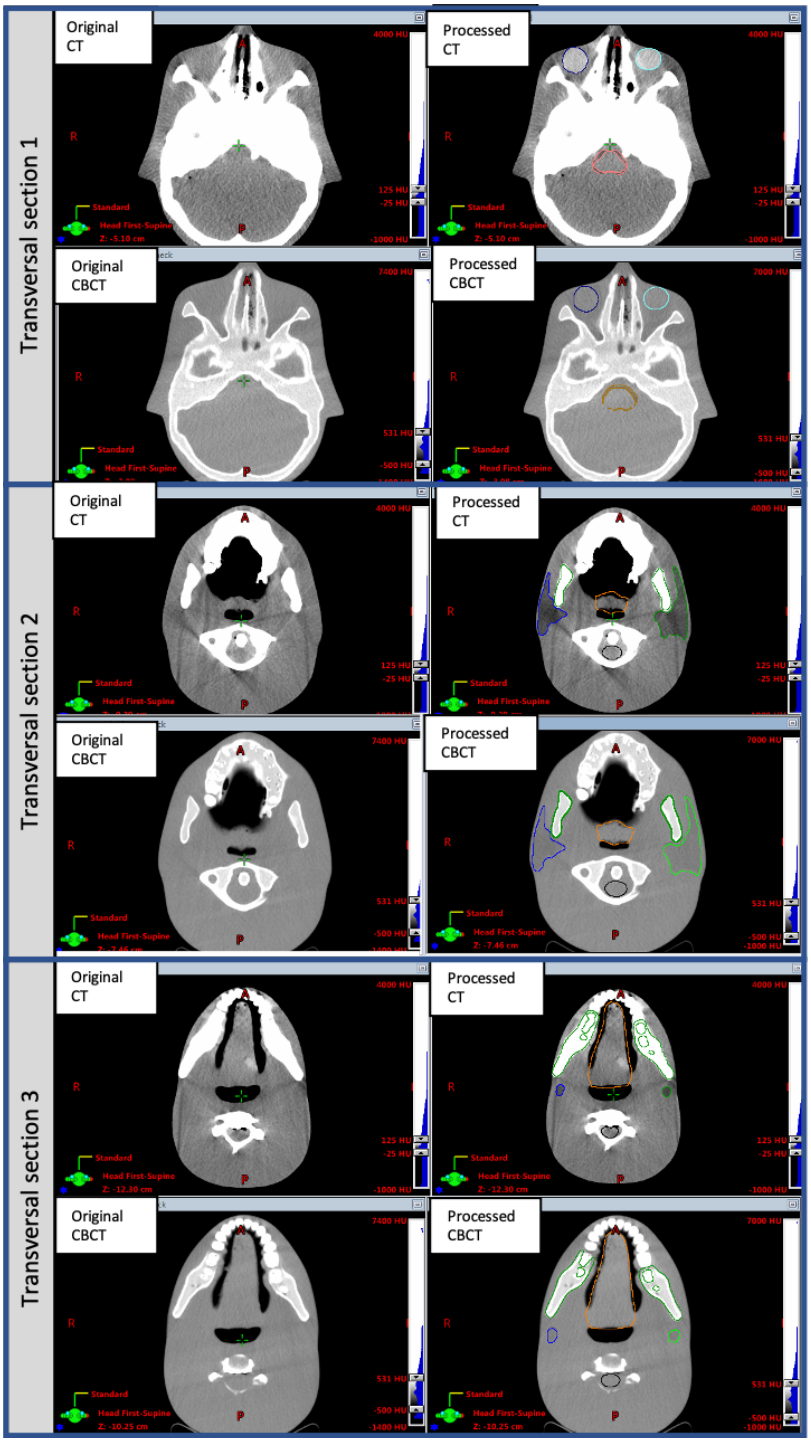
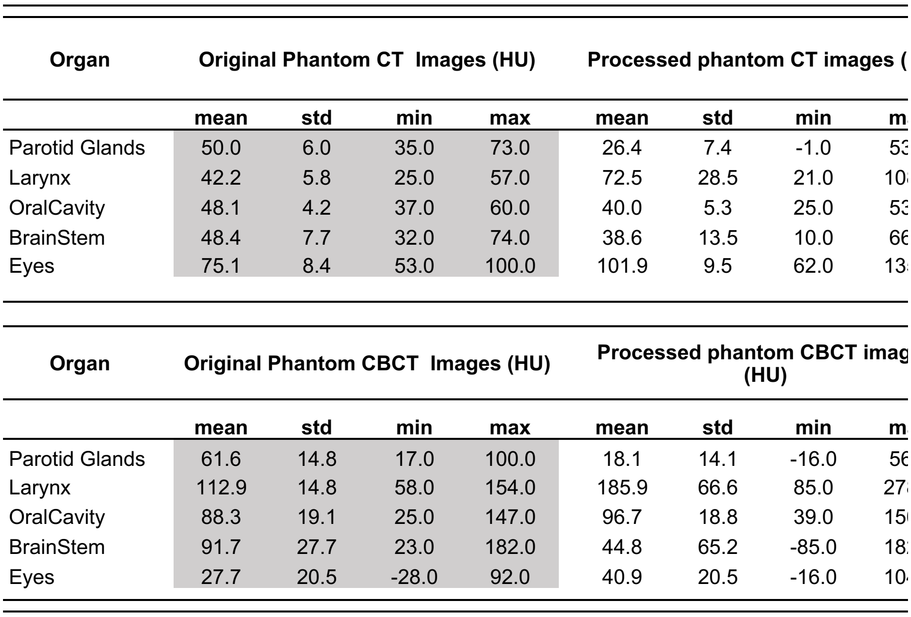


Figure 9.2: Original and post-processed CT and CBCT transverse sections of the CIRS ATOM Max Dental and Diagnostic Head Phantom Model-711 used in the method presented for DIR accuracy evaluation. Sections were chosen to visualize organs of interest. The HU values of parotid glands, larynx, oral cavity, brainstem, and eyes were digitally modified using Matlab and CERR tools as described in subsection 9.3.1



Organ	Original Phantom CT Images (HU)				Processed phantom CT images (			
	mean	std	min	max	mean	std	min	m
Parotid Glands	50.0	6.0	35.0	73.0	26.4	7.4	-1.0	53
Larynx	42.2	5.8	25.0	57.0	72.5	28.5	21.0	10
OralCavity	48.1	4.2	37.0	60.0	40.0	5.3	25.0	53
BrainStem	48.4	7.7	32.0	74.0	38.6	13.5	10.0	66
Eyes	75.1	8.4	53.0	100.0	101.9	9.5	62.0	13

Organ	Original Phantom CBCT Images (HU)				Processed phantom CBCT imag (HU)			
	mean	std	min	max	mean	std	min	m
Parotid Glands	61.6	14.8	17.0	100.0	18.1	14.1	-16.0	56
Larynx	112.9	14.8	58.0	154.0	185.9	66.6	85.0	27
OralCavity	88.3	19.1	25.0	147.0	96.7	18.8	39.0	15
BrainStem	91.7	27.7	23.0	182.0	44.8	65.2	-85.0	18
Eyes	27.7	20.5	-28.0	92.0	40.9	20.5	-16.0	10

Table 9.1: Original and post-processed CT and CBCT transverse sections of the CIRS ATOM Max Dental and Diagnostic Head Phantom Model-711 used in the method presented for QA of DIR. The HU values of parotid glands, larynx, oral cavity, brainstem, and eyes were digitally modified using Matlab and CERR tools as described in subsection 9.4.1. HU values for original and post-processed CT and CBCT phantom images reported in the table referred to a ROI of 20x20 pixels.

9.4.3 *Spatial analysis*

TRE values for each OAR and each registration are plotted in Figure 9.3a. The TRE was found to vary across the OARs of interest. The mean and standard deviation values (mean±SD) of the TRE over all fractions and patients were 4.6±4.6 mm, 2.6±1.4 mm, 2.7±1.5 mm, 2.1±1.0 mm, 2.0±1.2 mm, 1.8±1.2 mm, 5.6±3.3 mm and 2.5±2.0 mm for the body, left parotid gland, right parotid gland, larynx, oral cavity, mandible, brainstem and spinal cord, respectively. The large range of values of TRE for brainstem were due to the border of the ROI exceeding the CBCT boundary in the cranial direction in the MICE toolkit, when any mask was used. This could have induced unusually large shearing of the refDVF$_{\text{pt}}$ for brainstem in the cranial section of the CBCT boundary, as reported in the refDVF$_{\text{pt}}$ data in subsection 9.4.2. The testDVF$_{\text{ph}}$ generated with VelocityAI showed a smoother DVF in that area, resulting in consequently larger TRE values. Differences in the TREs between the two fractions, compared on a voxel basis, were found to be insignificant using the Wilcoxon signed-rank test with a p-value≈1.

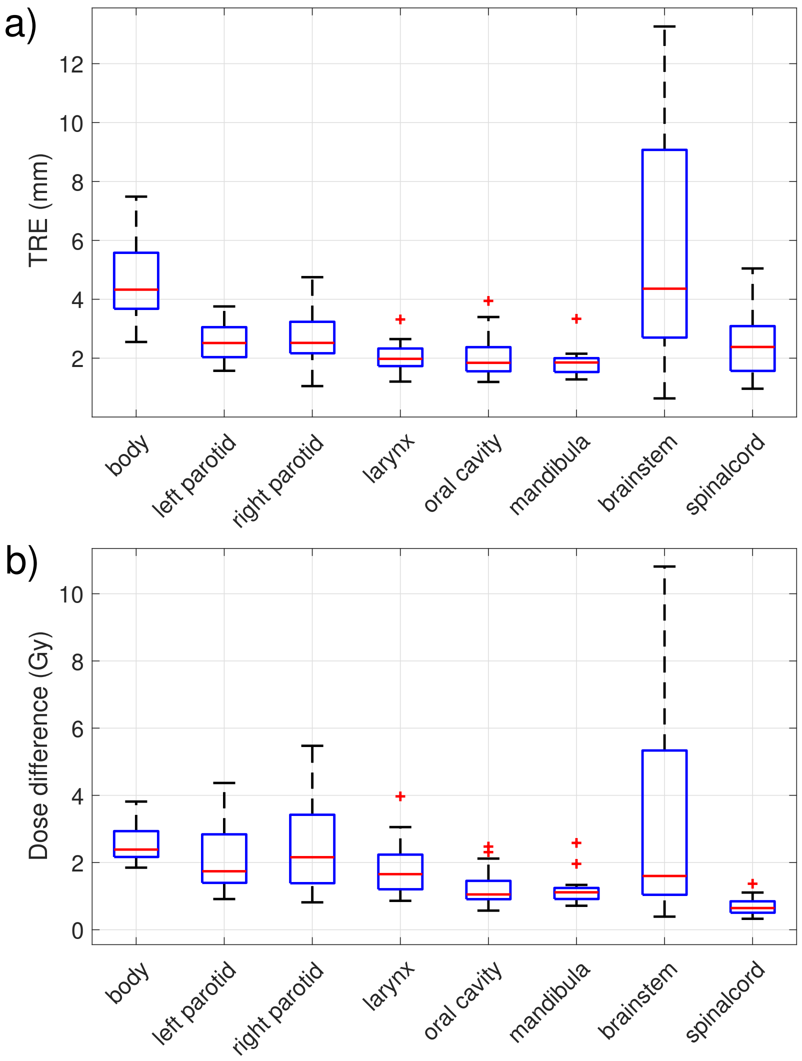


Figure 9.3: Boxplots of a) TRE difference and b) dose difference distribution sorted by OAR. For each box, the central mark represents the median value, while the bottom and top edges of the box are the 25th and 75th percentiles, respectively. The 'whiskers' represent the range of values. Observations beyond the whisker length are marked as outliers (+). By definition, an outlier is a value that is more than 1.5 times the interquartile range away from the bottom or top of the box. An outlier appears as a red + sign.

9.4.4 *Dosimetric analysis*

Dose error values for each OAR are plotted in Figure 9.3b. Similarly, to the TRE analysis, the median dose error value varied according to the OAR of interest. The mean ± SD dose difference values over all fractions and patients were 2.0 ± 4.3 Gy, 2.2 ± 2.4 Gy, 2.5 ± 2.7 Gy, 1.8 ± 1.8 Gy, 1.3 ± 1.6 Gy, 1.2 ± 1.8 Gy, 3.2 ± 2.6 Gy, 0.7 ± 0.7 Gy for the body, left parotid gland, right parotid gland, larynx, mandibula, brainstem and spinal cord, respectively.

9.4.5 *Baseline AAPM-132 image registration validation*

The mean TRE errors (mean ± SD) between the TG132-refDVF and TG132-testDVF, generated by VelocityAI, were 2.31 ± 1.13 mm for the phantom external contour (body). The overall % of voxels with TRE less than or equal to 2 mm was 45% (rounded to nearest 1%). This percentage value was considered as the OTL for the registrations evaluated in this study. Since the AAPM Report did not provide any dose error tolerance level information, the OTL could only be considered in terms of spatial accuracy.

9.4.6 *Tolerance levels analysis*

Figure 9.4 reports in the correlation graph below the percentage of voxels with TRE error $\leqslant$2mm versus the percentage of dose errors $\leqslant$ the DET (5% of prescribed dose in this analysis). The Figure 9.4 x-axes show the percentage of voxels with TRE within 2 mm (as suggested by AAPM-TG132) across the 20 registrations for each OAR. Significant variability was observed for all organs and registrations. The markers '+' indicate the case for which the percentage of voxels with TRE $\leqslant$ the 2mm OTL described in subsection 9.4.5, represented as a vertical red line in the graph. Percentages higher than OTL (indicated as a 'circle' marker) were found in 30% (body), 55% (left parotid), 40% (right parotid), 31% (larynx), 85% (oral cavity), 100% (mandibula), 25% (brainstem), 80% (spinal cord) of all cases (N=20) for all registrations.

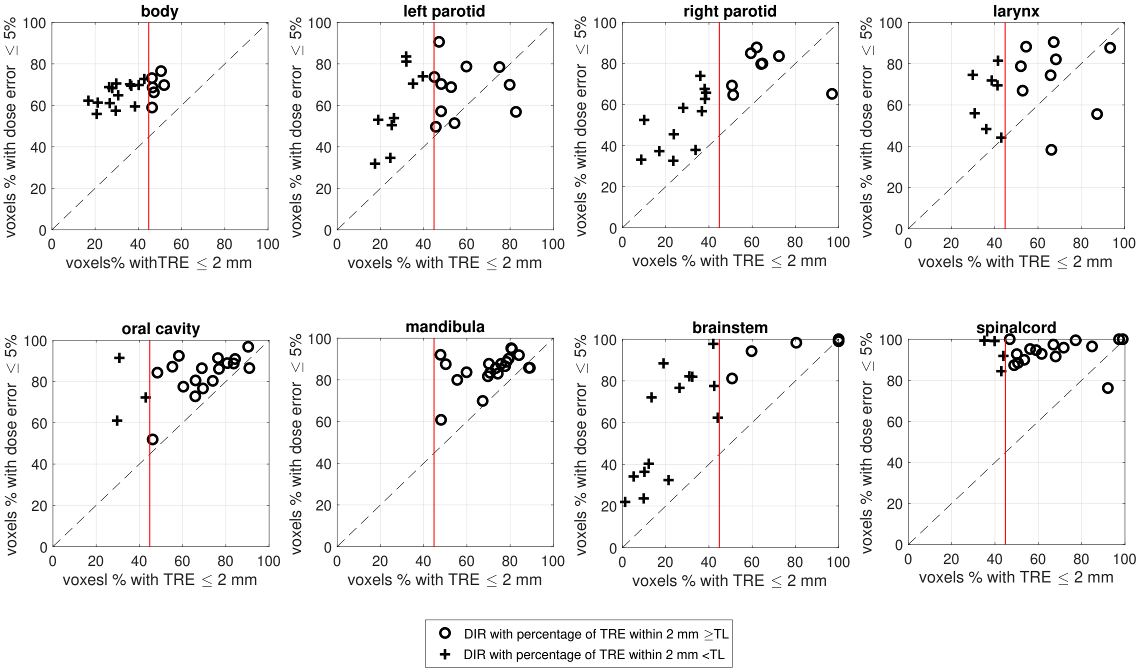


Figure 9.4: Correlation graph of the percentage of the TRE within 2 mm and percentage of voxels within 5% of dose error (3.5Gy), across the 20 registrations and for each OAR. The '+' markers indicate the cases for which the percentage of voxels was less than the 45% OTL, indicated as vertical red line. The OTL value was derived using TG132 datasets, as described in subsection 9.4.6

The Figure 9.4 y-axes report the corresponding dose error percentage within 3.5Gy (5% of prescribed dose), across every registration for each OAR. Percentages of 66.3±5.8%, 63.9± 16.1%, 62.0±17.5%, 69.3±16.4%, 82.2±11.1%, 85.1±8.0%, 66.7±28.2%, 93.6±6.2%, were found for body, parotids left and right, larynx, oral cavity, mandible, brainstem, and spinal cord, respectively. Overall, for the criteria used, the correlation graphs showed a generally higher proportion of voxels involving dose error less than 5% rather than TRE less than 2 mm (visualised as markers above diagonal bisector line of each panel). The relation between TRE and dose error percentage is quite complex and depends on a single registration (see also subsection 9.4.6). For almost all OARs, some registrations passed the spatial OTL analysis but showed quite low dose error $<$ 5% percentage and vice versa.

9.4.7 *Correlations*

The correlation between the TRE and the magnitude of the deformation in the reference DVF (refDVF$_{pt}$), was investigated on a voxel basis within the body contour, and a significant correlation was found (p-values$<$0.01 across all patients) with a Pearson correlation coefficient of 0.79± 0.14 (range 0.48-0.97), meaning that the DIR spatial errors are a function of the deformation magnitude of each registration. As reported in previously published studies [143, 156–159] the effect of the DVF on the dose map is complex. Theoretically, it is expected to depend on the local dose gradient and on the spatial registration error [159]. The correlation between the dose map error and TRE was also tested on an OAR-by-OAR basis and results shown in Figure 9.5. The upper panel shows in a correlation graph the median value of TRE versus median value of dose error for each OAR. Correlation values (indicated by the diameter of each marker size) varied significantly across the OARs. The Figure 9.5 lower panels show in correlation graphs TRE values versus dose errors for each of the 20 registrations for each OAR. Pearson coefficients, ranging from 0.21 (larynx) to 0.84 (right parotid), were reported for each panel. Significant correlations for body, right parotid, mandibula and brainstem were found.

The correlation between the dose error and TRE was also tested with a voxel-based approach inside the body. A statistically significant correlation was found (p-values$<$0.01) albeit with a low mean Pearson coefficient of 0.31±0.10 (range 0.15-0.53) across all registrations [30,31]. This confirmed the complex correlation between dose difference and TRE found on an OAR basis and reported above. The dose error showed a moderate but statistically significant correlation with the dose gradient, with a mean Pearson coefficient of 0.51±0.06 (range 0.37-0.61) and p-values $<$ 0.01 across all patients on a voxel basis, confirming the hypothesis that the local dose gradient influences dose errors related to DIR uncertainties [158, 159]. As suggested in the previous studies [143, 156–159] , the dose errors were also investigated as a function of the combination of dose gradient and TRE (here the scalar product between the two was considered). A moderate mean Pearson coefficient of 0.52±0.07 (range 0.41-0.66) was found, with a p-value $<$0.001, with respect to the single factor analysis (TRE and dose gradient).

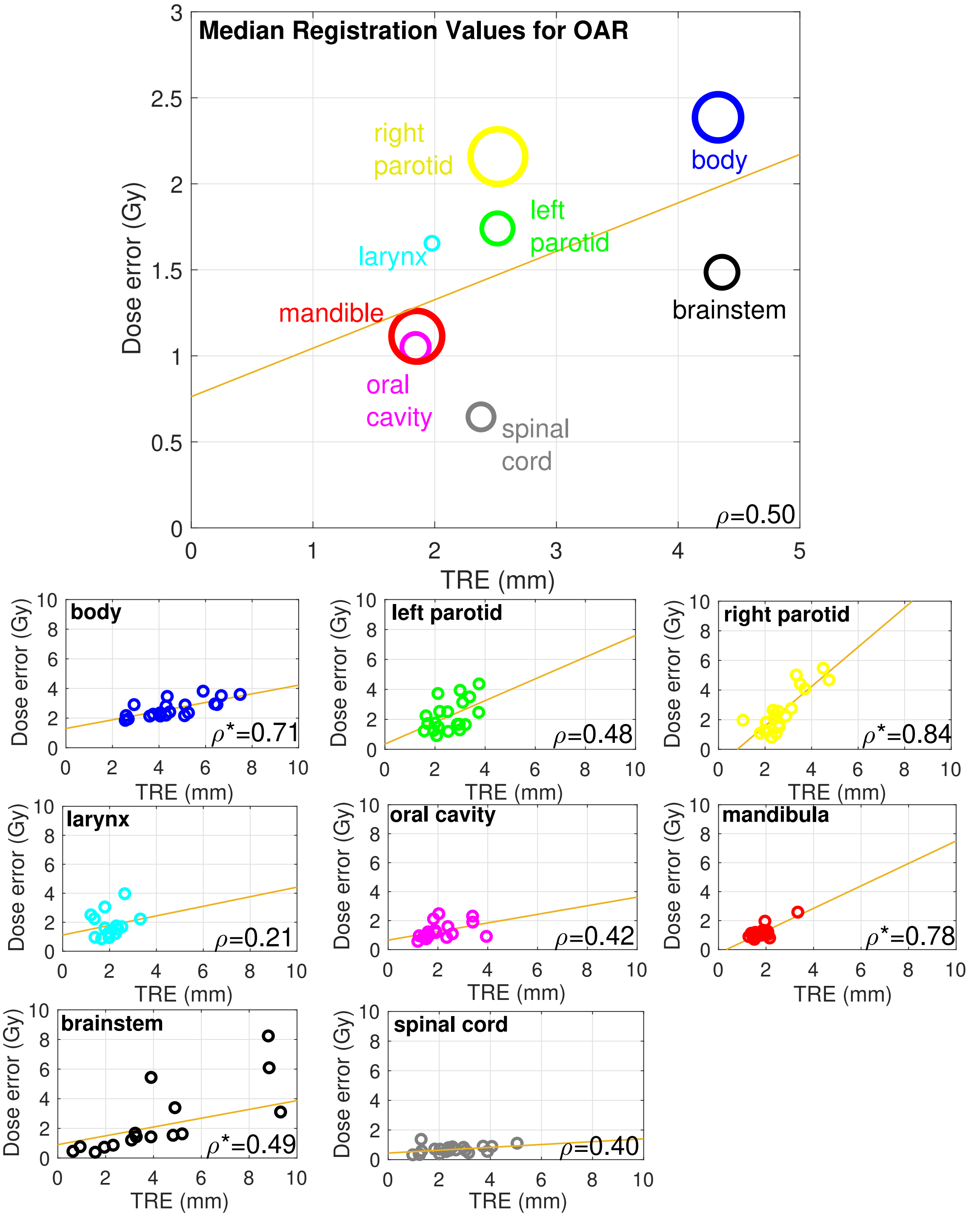


Figure 9.5: Upper panel: Correlation graph of the median TRE and dose errors values for each OAR. The diameter of each marker is proportional of the Pearson's correlation coefficient value (see lower panels for individual numeric values) A linear correlation fit line is also reported. Lower panels: Correlation graph of the TRE and dose errors, across 20 registrations divided for each OAR. A linear correlation fit line is also reported. In the lower right corner of all panels, the Pearson's coefficient was reported between TRE and dose error. A '*' indicates that Pearson test confirmed statistically significant correlation.

9.5 DISCUSSION

In this study, a novel method to evaluate the quality of patient-based multimodal CT-CBCT DIR using a post-processed, digitally enhanced anthropomorphic HN phantom has been presented. All post-processed ATOM Max Dental and Diagnostic Head Phantom image datasets (CT and CBCT) are in the DICOM format, derived using the procedure described in 9.3.1, and have been made publicly available as additional downloadable material for this paper as Supplementary Material. This allows the use of the proposed method for CT-CBCT HN DIR accuracy evaluation in other centres, by processing the image sets with two different image registration software systems (reference and clinical), as described in 9.3.1. The refDVF represents the known transformation map derived with third-party reference software based on the clinically observed anatomical changes, which is then applied to create the warped phantom image. Thus, the transformation of each voxel in the warped image was known because it was deliberately applied to each voxel. As a verification method, the refDVF data did not necessarily represent the real, individual deformations of the patients, but they served as a reference to check how much the clinical algorithm under investigation (in this study, the Velocity AI) was able to reproduce the known deformation for that patient case. However, care must be taken to avoid applying a deformation derived using the same third-party algorithm that the user employs for the clinical deformation, as this would create a testDVF that was biased to the results expected from the same third-party and clinical DIR algorithms [25]. The registration algorithms used in this work to generate the refDVFs and testDVFs were both of the same class (B-spline) but implemented independently (i.e., different optimiser, similarity measure and general parameters). Because of these differences, the registrations can be considered independent between the two software programmes, in line with the previous studies [151, 152]. The dose error variations were found to be related to both the registrations and the OAR of interest. This is in line with the findings of Qin *et al.* [156], who revealed that the impact of the DIR method on treatment dose warping is a function of the registration and is organ-specific. The OTL was derived from the material provided by the AAPM TG132 for quality assurance purposes. However, the TG132 utilised only pelvic anatomical data sets, processed using the ImSimQA, for the DIR accuracy assessment with a ground-truth transformation. Our findings confirmed that the registration accuracy is also a function of the selected organ at risk, underlying the importance of using an anatomical phantom for the DIR accuracy assessment, consistent with the clinical data considered. Thus, the OTL derived for this study, using the generic pelvic phantom proposed by the TG132, could have some limitations on the practical level. However, it was used here primarily as an exemplar to describe a suggested workflow for the DIR accuracy, as described in Figure 9.1a. In the published report on DIR validation by the AAPM's Task Group 132, the proposed ideal goal for the TRE is 95% of voxels within 2 mm, using the Report's provided image sets. Our results showed a considerably lower value, OTL = 45%, using the Velocity AI. The Velocity AI was previously benchmarked with different commercial DIR algorithms by Pukala *et al.* [143] and found to be very similar in accuracy with respect to the other commercial software systems included in the study. Based on this, it can be assumed that the OTL value found in this study could be considered a typical practical accuracy value using the commercial software, based on the B-spline method, for the CT-CBCT DIR registration using the TG-132 basic anatomical phantom. The difference between the OTL found in this study, and the goal suggested by the AAPM-TG132, underlines the pressing need for a universally available, comprehensive library of site-specific phantom datasets with the ground-truth deformation data to determine a more robust operative tolerance level to test the DIR software. The registrations that pass the OTL for spatial accuracy can still result in a low percentage passing the dose difference criterion (as reported in Figure 9.4 and Figure 9.5), indicating quite a complex relationship between the dose error and TRE accuracy. One limitation, shared by the spatial DIR metrics, is the indirect relationship between the quantification of the DIR error and the effect that error has on a given dose distribution. This is a complicated issue akin to invoking the gamma criterion for an IMRT QA analysis [160]. The TG-132 report did not suggest any tolerance goal for dose differences. For this reason, in the dosimetric analysis workflow reported in Figure 9.1, it has been suggested a clinical evaluation criterion for the dose difference results. A moderate, as classified in [161], but significant correlation on a voxel basis for both the dose gradient and the TRE with a dose difference was found. Veiga *et al.*

[159] observed a moderate correlation between the dose gradient and the resulting dose difference, in agreement with our findings. Murphy *et al.* have reported the effect of the DVF on dose-mapping and concluded that there is a complex correlation that depends on the TRE and the voxel location, relative to the dose gradient [158]. The Pearson correlation test results were always significant (p ¡ 0.05) for the voxel basis approach (9.4.7). As reported by Schober *et al.* [162], the p-value derived from the test provides no information on how strongly the two variables are related. With large datasets, as they are in the voxel-based analysis, very small correlation coefficients can be “statistically significant.” Therefore, a statistically significant correlation must not be confused with a clinically relevant correlation. It is important to underline that this study represents a Proof of Principle, and further work will be undertaken to explore the behaviour of the algorithm in other situations. In my experience, the clinical workflow is composed of several tasks and can be time-consuming.

9.5.1 *Range of Application of the Method*

The use of a phantom in the method has some intrinsic limitations. The lack of an initial deformation or shift between the CT and CBCT (owing to the rigidity of the phantom and not applicable with the real patient CT and CBCT images) was the basis for the reference condition of the validity of the method. Since the clinical reference DVF has the same frame of reference as the patient CT image, care must be taken to align the DICOM coordinate systems between the patient and HN phantom CT images before generating the warped phantom CT image based on the reference DVF. This was done to align the phantom CT image with the clinical refDVF to produce a more clinically realistic warped phantom CT image. Indeed, it is important that the warped phantom is created, applying as far as possible the deformation magnitude consistent with the spatial deformation of the patient images. From a mathematical point of view, the validity of the method is consistent for every size of patient anatomy, because the quality of the registration was assessed as the ability of the clinical software to replicate the reference DVF, artificially introduced into the deformed phantom CT images. However, the rigid *coupling* between the phantom anatomy and the reference DVF should be considered carefully to ensure the clinical relevance of the reference DVF. The more the anatomy is similar between the phantom and patient, the more the clinical value of the reference DVF will be maintained when it is applied to the phantom anatomy. The quality of this coupling should be tested visually, by checking the overlap between the patient and phantom structures. A potential method to overcome the mismatch between the patient and phantom anatomy is to introduce a baseline DIR between the patient and the phantom in the coupling phase, instead of a rigid one. However, introducing this step can increase the complexity of the evaluation method, therefore it was decided to avoid this point to avoid over-complicating the workflow. Indeed, it is important to underline that, as a verification method, the refDVF data did not necessarily represent the real, individual deformations of the patients. Moabbed *et al.* [163] used a physical anthropomorphic pelvic phantom for the CT and CBCT deformable image registration with the aim of validating a DIR evaluation method for prostate cancer cases. Even if their study focused on evaluating the accuracy of the dose calculation as performed directly on the CBCT images, they presented a similar workflow to the current study for their evaluation, in that the pelvic phantom images were digitally deformed using two realistic patient-based deformation fields, after a rigid alignment between the patient and the phantom's system of reference. Since the OAR insertion in the HN processed phantom was done in terms of HU, and not in physical density, the processed phantom images were not suited to studies of the dose calculation accuracy, but only for the DIR registration accuracy. Indeed, the dose calculation using this method is only performed on the patient adapted-CT.

9.5.2 *Similar DIR Accuracy Workflows to the Current Study*

Our proposed method quantified the effect of the DIR on the dose distribution, focusing on the DVF accuracy. As was reported in the introduction, for adaptive radiation therapy, the accurate dose deformation and summation/accumulation within each contour is determined by the DVF accuracy [164]. Other methods evaluated the dose deformation accuracy, focusing on the DVF. Recently, Lowther et al. [165] assessed the uncertainty in the dose accumulation procedure for the HN adaptive by comparing the DIR-facilitated dose accumulation using a commercial software system with the results of an in-silico model, based on the clinically observed deformations as the ground truth. HN patients CT and CBCTs images were used to generate in silico reference CBCTs and DVFs, as the ground truth, with a second reference B-spline DIR software. The in silico images, generated with an independent DIR software as the reference, were necessary to have a known reference DVF. Also, this method has the advantage of evaluating each patient-specific deformation and dose warping. However, Lowther's method works directly on the patients' images, making the clinical workflow simpler and reasonably quicker than the proposed one, avoiding the use of the HN phantom presented in this study, with its related inherent imitations (see 9.5.1). The complexity of the deformations in the human body, as well as its heterogeneity in terms of the HU values, are challenging to model accurately through a phantom. On the other hand, the method implemented in our work has the advantage of replicating the HU distribution found in the OARs on clinical images. In comparison, Lowther's method tested the DIR between the planning CT and the in silico model of the CBCT, which is different from the original CBCT in the HU and quality (the planning CT deformed to the anatomy of the CBCT of the considered fraction) and could not consider some issues of the CTBT-to-CT DIR, including the reduction in the accuracy due to the low image contrast of the real CBCT. Pukala et al. [[143] evaluated the performance of several DIR software systems using several HN digital phantoms, in terms of both the spatial and dosimetric accuracy at the voxel level. The study created publicly available phantoms based on the CT data from head and neck cancer patients. These phantoms provide a clinically based ground-truth model that encompasses the anatomical changes that occur over the course of a typical treatment. Also, this study assessed the deformation accuracy for each voxel by comparing the entire registration DVF to the ground-truth DVF, based on the clinical scenario. One key aspect of our study was the possibility to use real patient data (the DVF and dose distributions) to evaluate each patient-specific DIR, instead of commissioning specific DIR algorithms. Moreover, the study of Pukala was focused on CT-to-CT registrations, while our study referred to multimodal CBCT-to-CT registration. The latter presents more issues in the DIR accuracy, due to the limited quality of the CBCT images. Veiga et al. [159] investigated the transformations that mapped the anatomy between the planning

CT and CBCT using four different DIR approaches. DVFs were used to remap the dose of the day onto the planning CT. The data from five HN patients were used to evaluate the performance of each implementation, based on geometrical matching, the physical properties of the DVFs, and the similarity between the warped dose distributions. As reported in the Introduction, this approach can be useful to choose the optimal registration method, without conveying any information on the DIR uncertainties and related dose errors. Singhrao et al. [157] used a deformable HN phantom to test DIR algorithms, but, in practice, it is difficult to produce physical phantoms for every deformation scenario that occurs clinically. Essentially, the novelty of our approach consists of the use of an HN anthropomorphic phantom for the CT-CBCT registration, which was digitally enhanced to include the OARs which are typical of that anatomic region, (Figure 9.1), with a robust accuracy verification procedure taking into account the recommended OARs for the HN. Also, in this study, further recommendations in terms of a clinically feasible workflow and achievable thresholds for the CBCT-to-CT deformable image registration have been investigated. The findings of this study indicate that there is a need for standardised approaches and specific guidelines for different applications to supplement the overarching DIR guidelines such as the TG132.

9.6 CONCLUSIONS

A novel method to evaluate the quality of patient-specific multimodality CT-CBCT DIR for adaptive radiotherapy of head and neck patients was developed. The methodology described in this chapter allows direct testing of DIR algorithms for clinical registration, which can produce valuable insights into their clinical impact on the adapted dose distribution. The technical requirement for using this method is the availability of an independent image registration software platform (in addition to the clinical system). This work contributes to the study of standardisation and automation of quality assurance methods for deformable image registration in radiotherapy.

10

USE OF KNOWLEDGE BASED DVH PREDICTIONS TO ENHANCE AUTOMATED RE-PLANNING STRATEGIES IN HEAD AND NECK ADAPTIVE RADIOTHERAPY

For everything there is a season,
and a time for every purpose under heaven
Ecclesiastes

10.1 PREVIEW

Work from this chapter was published in Cagni et. al. (PMB 2021) [39]. The figures and tables that are shown in this chapter are drawn from that published work. Patient anatomical deformations, that often happen during radiotherapy treatment course, cannot always be corrected by simple couch shifts or patient repositioning, so these deformations should be managed by re-planning. Generating a new plan with the same planning goals as the original plan within a clinically acceptable time and with minimal user intervention is another important technical challenge and time consuming step in ART. As the patient anatomy varies, OARs are re-shaped and re-positioned with respect to the targets and change from day to day. The gain in OAR sparing with a new plan is *a priori* unknown, and re-planning decisions are often based on the clinician's practical experience. In this study the feasibility of using KBP iinformation as part of the ART process to estimate the potential gain given by OAR sparing during the treatment course for HN cases was investigated. Such relationships, if significant, could be used to establish the need for plan adaptation based on OAR sparing and to automate the process of re-planning itself.

10.2 INTRODUCTION

As discussed in section 9.2, external beam radiotherapy is gradually evolving towards real-time ART [135, 166], which is becoming a new paradigm in radiation oncology [136]. ART has the clinical rationale of reducing normal tissue and organ at risk (OAR) toxicity and/or improving tumour control through plan adaptation [137, 167]. The frequency of re-planning in patients with head and neck (HN) cancer was reported to vary from 32% to 70% depending on several criteria, such as weight loss, change in neck separation or poor immobilisation shell fit [139, 168]. However, at present, there are still a number of technical limitations in applying ART as an automated standardisation process. [138]. As a consequence, ART is not widely utilised [138, 168]. Indeed, in clinical practice, contouring and treatment planning processes are labour-intensive and use substantial resources. Some investigators are researching automated methods to predict the eventual need for re-planning, but more work is needed [169, 170]. One of the on-going technical issues is the challenging decision of choosing the appropriate time in the process to re-plan a patient [30, 171–177]. As anatomical deformation cannot be corrected by simple couch shifts these deformations should be managed by re-planning [138]. Generating a new plan with the same planning goals as the original plan within a clinically acceptable time and with minimal user intervention is another important technical challenge in ART. As the geometry of patient anatomy varies, OARs are shaped and positioned with respect to the targets and change from day to day. The gain in OAR sparing with a new plan is *a priori* unknown [171], and re-planning decisions are often based on the clinician's practical experience [30, 171]. The "trigger point" is the time at which significant dosimetric variation for a specific parameter is present as an indicator for ART re-planning [173]. However, the re-planning process is expensive in resource terms as it requires a new computed tomography (CT) scan, new contours, and a new optimisation. For this reason, the trigger should be carefully chosen, balancing the time-consuming procedures with the gain from re-planning. Knowledge-based planning (KBP) tools (previously described in chapter 4) can generate estimated dose-volume histograms (DVHs) based on previous patient anatomy and dose distributions [36, 178]. The KBP methods are generally equivalent to expert planners in terms of plan quality but preliminary results indicate that they are significantly more efficient timewise [36]. Currently, KBP is frequently used in clinical practice to drive new IMRT planning based on a database of prior clinical plan data and other sources of knowledge, such as treatment trade-off and clinician experience. Studies published in recent years have demonstrated that KBP allowed a general improvement in (a) inter-patient consistency of the treatment plans, (b) intrinsic quality and (c) efficiency (time and workflow) of the planning process applied to different anatomical sites using manually generated treatment plans [17–22], including the head and neck [23, 24]. Previously published studies have demonstrated that the efficacy of KBP was influenced by the quality of the data used for the training process, the regression applied to build the predictive models and the consistency between new cases and the population used for the training [178]. KBP training generally involves using a database of manually generated treatment plans that can suffer from plan quality variation and inconsistency [20]. To maximise the performance of KBPs, Pareto plan solutions can be used to train the model [81, 82, 179]. In the studies published to date [17–24], KBP was employed to assist the planner to achieve optimal dose distributions when new plans were created. The clinical implementation and use of KBP optimization models is a rapidly changing subject and previous studies have applied KBP for help in replanning for adaptive radiation therapy [180]. However none of the published studies have investigated if the accuracy of a KBP prediction is suitable for OAR sparing in an ART application. In this work, the possibility of using the KBP method as part of the ART process to estimate the potential gain given by OAR sparing during the treatment course for HN cases was investigated. Such relationships, if significant, could be used to establish the need for plan adaptation based on OAR sparing and to automate the process of re-planning itself.

10.3 METHODS AND MATERIALS

10.3.1 *Clinical Data*

In line with samples used in previous work [24, 81, 82], a dataset of 100 Head and Neck (HN) VMAT patients previously treated at AUSL-RCCS of Reggio Emilia was selected for this study. Treatment plans following two different fractionation schemes were included in this work, 69.96 Gy/59.4 Gy/54.12 Gy in 33 fractions (46 patients) and 66 Gy/60 Gy/54 Gy in 30 fractions (54 patients), both schemes using a simultaneous integrated boost (SIB) technique. For all plans, the goal was to deliver 100% of the prescribed dose to 95% of every PTV. Each plan was generated using a previously configured KBP model for HN patients trained on manually produced clinical plans. This model is discussed below (subsection 10.3.4). All plans were generated with the Eclipse Treatment Planning System (TPS) (Varian Medical Systems, Palo Alto, CA) using 3 fully coplanar arcs with collimator rotated to 30°, 315° and 90°, with 6 MV energy. For each treatment fraction, an on-board cone beam CT (CBCT) scan was acquired before treatment delivery to assess patient setup and anatomical variation. All CBCT images were automatically saved in an ARIA database (Varian Medical Systems, Palo Alto, CA).

10.3.2 *KBP RapidPlan tool*

RapidPlan is a knowledge-based automatic planning (KBP) solution integrated in the Eclipse TPS. For each new patient, RapidPlan predicts the most likely dose-volume histograms (DVHs) to occur based on the specific patient's anatomy in terms of structure set geometries. Predicted DVHs are then used to establish dose-volume objectives and weights for automated plan optimisation. DVH prediction in RapidPlan is based on a statistical model that is generated from the principal component analysis of anatomic and dosimetric features obtained from plans of previously treated patients. Therefore, the quality of RapidPlan DVH predictions depends on the quality of the plans used to train the model. Extensive descriptions of model configuration are provided in the Varian reference manual

[80], previous publications [81, 82, 179], in the studies published to date [17–24], and in chapter 4 of this thesis.

10.3.3 *Multicriteria Optimisation Trade-off module*

The Multicriteria Optimisation (MCO) approach is based on trade-off exploration modules and is implemented in the Eclipse TPS. In MCO, a range of different Pareto solution plans is generated based on a selection of optimisation objectives. The priority of each objective may vary from plan to plan but all plans belong to a *Pareto surface*, which means that is not possible to improve an objective without compromising another objective in the trade-off. The user can explore the trade-offs along the Pareto surface and select the plan that best fulfils the treatment goals. With the use of slider bars, dynamic DVHs and dynamic 3D dose distributions, the TPS allows users to visually review and evaluate plans along the Pareto surface in real-time [80, 81]. Extensive descriptions of model configuration are provided in the Varian reference manual [80] and in chapter 4 of this thesis.

10.3.4 *MCO-KBP model*

A KBP-model was configure using the procedure described in details in chapter 4. Here a brief summary is reported. The 100 manual plans in the database were first randomly divided into two groups: KBP training (80 plans) and KBP evaluation (20 plans). Using the training set, KBP DVH prediction models were created for the following OARs: brainstem, spinal cord, left parotid gland, right parotid gland, mandible, oral cavity, oesophagus, and larynx. The KBP model was built following the guidelines provided by the manufacturer [80]. Figure 10.1a outlines the workflow of the KBP model generation stage.

All 80 patients used in the training set were automatically re-optimised using the KBP model, the latter used as a starting plan in the MCO module. A wish list of objectives to fulfil was used to consistently select a solution on the Pareto surface in the MCO module. Due to the use of the same optimisation scheme for all patients, plan generation was highly consistent across the entire cohort, with no plans adjusted by the planning team. All plans generated with the KBP and

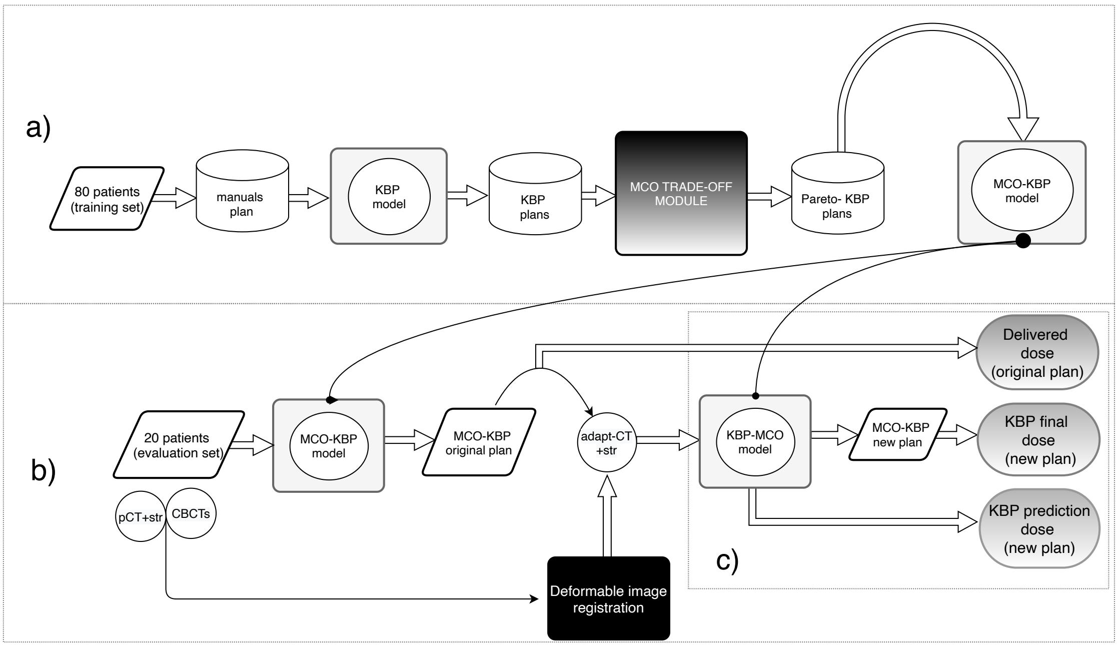


Figure 10.1: The workflow of the study was divided into 3 parts: MCO-KBP models creation a) and application of the gain of the re-planning estimation using the KBP tool b) and c). In detail: 100 HN patients were selected for the study. a) 80 patients were chosen to train the KBP model. To improve the quality of the model, the training set was automatically re-planned to use the KBP model and refined using the MCO tool based on a single wish list. The final MCO-KBP plans generated in this process were then used to train the model. b) 20 patients were considered in the evaluation set for the experiment. For each patient's planning CT (pCT), the structure set and CBCT of fractions 16 and 26 were considered. Using a deformable image registration tool, adapt-CT, deformed structures for both fractions were generated. c) An automated plan was created using the KBP-MCO model on pCT. The plan was recalculated using the adapt-CT of each fraction generating the delivered dose. The KBP-MCO model was used to generate a KBP prediction and final dose from the adapt-CT of each fraction if a new plan was created.

MCO combination were used to train a new, highly consistent MCO-KBP model (MCO-KBP) for the previously selected structures. The optimisation constraints for this model were unchanged with respect to the KBP model.

10.3.5 *Delivered DVHs for the evaluation set*

In the evaluation dataset, the MCO-KBP model was used for each patient to create an automatically generated plan following the same fractionation scheme as the original treatment. This procedure is outlined in Figure 10.1b. Two CBCT scans acquired before each treatment fraction were extracted for each patient, corresponding to the 16th fraction (around half-way through the treatment course) and the 26th fraction (about three quarters of the way through the treatment course). For each selected fraction, an adapted CT image [75] was generated in Velocity AI v.4.0 (Varian Medical Systems, Palo Alto, CA) through deformable registration of the planning CT (pCT) on the selected CBCT. The same HU values as the pCT were used, to avoid bias in dose calculations. The transformation used was a cubic B-spline, the interpolator was computed with a bi-linear function, and the optimiser was based on the steepest gradient descent [75]. For the selected set of patients, no artifacts were evident in the CT image. No image pre-processing was done before the deformable image registration. The OAR structures were automatically propagated from the planning CT to the adapted CT. The accuracy of registration was verified by means of a visual inspection of deformed CT and structure sets. This was combined with a visual check of deformable vector field (DVF) that was performed in Velocity AI after each registration. The adapted CT images with modified structure sets were then imported into the Eclipse TPS, and the plan created with the MCO-KBP model was recalculated for the adapted CT. The dose and DVH of the plan for this adapted CT will be referred to as the 'delivered DVH' ($\mathrm{DVH_d}$).

10.3.6 *MCO-KBP DVH model prediction validation*

For each patient in the evaluation dataset, KBP predictions were performed on both adapted CTs using the MCO-KBP model. This section of the workflow is outlined in Figure 10.1c. For each organ, RapidPlan presents two predicted DVH lines representing the ± 1 SD DVH confidence limits. For the validation group, the mean of these two predicted DVH curves, $\mathrm{DVH_{pKBP}}$, was compared with the corresponding DVH achieved after the optimisation, $\mathrm{DVH_{fKBP}}$ [23]. Following international guidelines [73, 74], specific DVH endpoints for validation were considered: maximum dose ($\mathrm{D_{max}}$) for spinal cord, brainstem and mandible and mean dose ($\mathrm{D_{mean}}$) for parotids, oral cavity, larynx and oesophagus, as used in the AUSL-IRCCS hospital's clinical practice.

10.3.7 *Gain from re-planning and KBP prediction uncertainty*

From the Varian Manual [80], the normal prediction bounds shown by the KBP tool indicated 68% probability (±1 SD) that the final DVH should fall between the bounds. Based on this assumption, we can consider the predicted KBP uncertainty as 1 SD, the distance between the $\mathrm{DVH_{pKBP}}$ (subsection 10.3.6) and one of the prediction bounds. Thus, for each OAR j and case i:

$$\mathrm{pKBPuncertainty_{ij}} = \mathrm{ASR}(\mathrm{upperboundDVH_{KBP_{ij}}} - \mathrm{DVH_{pKBP_{ij}}}) \tag{10.1}$$

ASR was the absolute sum of residuals (ASR) and was used to quantify the distance between DVHs:

$$\mathrm{ASR} = (\sum_{D=0}^{\infty} |\mathrm{DVH_1}(D) - \mathrm{DVH_2}(D)| \cdot \Delta D) \tag{10.2}$$

where $\mathrm{DVH_1}(D)$ and $\mathrm{DVH_2}(D)$ refer to the DVHs for which the distance was quantified. After the KBP prediction was obtained, the optimisation was performed and the dose calculated without manual intervention, generating a final KBP DVH ($\mathrm{DVH_{fKBP}}$).

In line with equation 10.1 and equation 10.2, the final KBP uncertainty was defined as the difference between predicted and final DVH:

$$\mathrm{fKBPuncertainty_{ij}} = \mathrm{ASR}(\mathrm{DVH_{fKBP_{ij}}} - \mathrm{DVH_{pKBP_{ij}}}) \tag{10.3}$$

where for each OAR j and case i, $\mathrm{DVH_{fKBP}}$ refers to the final KBP DVH after optimisation, with ASR and $\mathrm{DVH_{pKBP}}$ as defined above. The gain from re-planning was calculated in term of sum of residuals (SR), between delivered DVH (subsection 10.3.5), $\mathrm{DVH_{dij}}(D)$, and predicted KBP, $\mathrm{DVH_{pKBPij}}(D)$ or final DVH after optimisation, $\mathrm{DVH_{fKBPij}}(D)$, when the new plan is created:

$$\mathrm{SR_{ij}} = (\sum_{D=0}^{\infty} (\mathrm{DVH_{d_{ij}}}(D) - \mathrm{DVH_{KBP_{ij}}}(D)) \cdot \Delta D) \tag{10.4}$$

for each OAR j and case i. It is referred respectively to predicted gain values (pSRs) or to final gain values (fSRs), when $\mathrm{DVH_{pKBP}}$ or $\mathrm{DVH_{fKBP}}$ was considered in equation 10.4. The approach to use sum of residuals to estimate the difference of DVHs is similar to the one proposed by Appenzoller *et al.* [181]. However, in that original study, only positive differences between DVHs were considered for detection of suboptimal plans (restricted sum of residuals). This work accounted for both positive and negative differences, as both are important to detect improvement/detriment in the re-planning phase. To automate the gain from re-planning, predicted gain values (pSR) were plotted against the final optimisation gain (fSR). To evaluate the feasibility of the current method, the KBP uncertainty (equation 10.1 and equation 10.3) for each OAR and case was compared against the gain from re-planning. To use the same metric in the comparison, ASR, was used to quantify the distance between DVHs in the gain of replanning, using equation 10.2. Thus, for each OAR j and each case i:

$$\mathrm{ASR}_{ij} = (\sum_{D=0}^{\infty} |\mathrm{DVH}_{d_{ij}}(D) - \mathrm{DVH}_{\mathrm{KBP}_{i,j}}(D)| \cdot \Delta D) \tag{10.5}$$

It is referred respectively to ASR predicted gain values (pASRs) or to final gain values (fASRs), when $\mathrm{DVH}_{\mathrm{pKBP}}$ or $\mathrm{DVH}_{\mathrm{fKBP}}$ was considered in equation 10.5.

10.3.8 *Discriminant analysis*

The receiver operating characteristic (ROC) analysis curve (referred to in this chapter as *discriminant analysis*) was used to quantify predicted KBP performance measures in this experiment. With these analyses, four domains were highlighted:

- TP: positive pSR values predict significant OAR sparing, and re-planning demonstrates improved OAR DVH with positive fSR re-planning values.
- True negatives (TN): negative or null pSR values predict no OAR sparing, and re-planning demonstrates no improvements in OAR DVH with negative or null fSR re-planning values.
- FP: positive pSR values predict significant OAR sparing, but re-planning demonstrates no improvements in OAR DVH with negative or null fSR re-planning values.

- False negatives (FN): negative pSR values predict no OAR sparing, but re-planning demonstrates improved OAR DVH with positive fSR re-planning values.

The ROC curve was used to choose the best predicted operating point (OP) that gives the best trade-off between the sensitivity, or TP rate, and specificity, or 1-FP rate, of KBP predictions [118].

10.3.9 *Statistical Analysis*

Wilcoxon two-sided signed rank tests were used to assess the statistical significance of the observed differences between KBP predictions and final endpoints and SRs, to test the difference in gain between the two fractions and to test differences between ASR values and KBP uncertainty values. The differences were considered significant when $p<0.05$.

10.4 RESULTS

10.4.1 *Quality of the MCO-KBP model*

The quality of the generated MCO-KBP model was evaluated by checking the model's goodness-of-fit statistics for each structure such as the coefficient of determination (R^2 (between 0 and 1: the larger, the better)) and the average Pearson's chi-square (χ^2 (the closer to 1, the better)), as suggested by RapidPlan guidelines [80]. These parameters, together with the number of potential outliers (also known as influential points), are reported in Table 10.1 for all models. No particular trends were observed for χ^2 and R^2. A mean χ^2 of 1.11±0.05 and a mean R^2 of 0.83±0.10 were found.

The potential outliers identified by the Varian Model Analytic (MA) tool [80, 82] were evaluated on a case-by-case basis. Since all plans inserted in the model were Pareto plans, these cases were regarded as not actually outliers, and it was verified that there were no anomalous anatomical differences compared to the rest of the population in the model to cause such categorisation. The plans iden-

Structure	R^2	χ^2	# Outliers (MA)
Brainstem	0.84	1.05	3
Oesophagus	0.83	1.06	0
Larynx	0.82	1.18	3
Mandible	0.83	1.08	5
Oral Cavity	0.87	1.07	2
Parotids	0.82	1.05	0
Spinal Cord	0.54	1.10	0

Table 10.1: Goodness of the prediction models in terms of coefficient of determination, R^2, average Pearson's chi square, χ^2, and number of potential outliers (model analytics, MA, suggested plans to be removed and plans to be checked).

tified as potential outliers by MA, were representative of an *uncommon* patient anatomy with respect to the training set population, but these plans remained clinically suitable as they were created using the MCO module. The differences between predicted and final KBP endpoints for the OARs considered are reported in Table 10.2. The differences were calculated for every predicted and final endpoint of each organ and patient, then the mean and SD values over the patient population were derived. Mean difference ±1SD between the pKBP and fKBP endpoints were 1.8±4.5 Gy, 1.2±5.3 Gy. 0.8±1.1 Gy, 0.6 ±0.9 Gy, -1.3±1.8 Gy. 1.4±2.4 Gy, 0.9±2.1 Gy and -3.4±4.7 Gy for spinal cord, brainstem, right and left parotid, oral cavity, oesophagus, larynx and mandibula, respectively. Wilcoxon signed rank tests confirmed the differences across the range of endpoints (pKBP and fKBP) were not statistically significant ($p > 0.05$).

10.4.2 *Gain from re-planning*

Figure 10.2 shows a comparison of the DVHs across an individual case (patient #7 fraction 16), considered as a patient example. The predicted DVH_{pKBP}, the bounds of the predicted DVH_{pKBP} (1SD of the KBP), the final DVH_{fKBP} and the DVH_d are represented for the following organs: spinal cord, brainstem, left parotid, mandible, larynx, oesophagus, oral cavity and right parotid.

	Spinal cord	Brainstem	Right parotid	Left parotid	Oral cavity	Oesophagus	Larynx	Mandibula
	Dmax [Gy]	Dmax [Gy]	Dmean [Gy]	Dmean [Gy]	Dmean [Gy]	Dmean [Gy]	Dmean [Gy]	Dmax [Gy]
DVH_{pKBP}	28.5 ± 4.7	28.6 ± 9.0	21.5 ± 8.6	23.0 ± 6.4	34.3 ± 10.7	16.5 ± 7.1	32.0 ± 5.2	64.9±6.2
DVH_{fKBP}	26.7 ± 4.6	27.3 ± 8.5	20.7 ± 9.0	22.4 ± 6.6	35.5 ± 11.7	15.0 ± 6.3	31.1 ± 6.2	68.2±8.1
difference	1.8 ± 4.5	1.2 ± 5.3	0.8 ± 1.1	0.6 ± 0.9	-1.3 ± 1.8	1.4 ± 2.4	0.9 ± 2.1	3.4±4.7

Table 10.2: Endpoint differences (D_{mean} and D_{max}) between prediction (pKBP) and final DVH (fKBP) for all 20 patients (2 CBCTs for patient) of the evaluation set in term of mean value±SD.

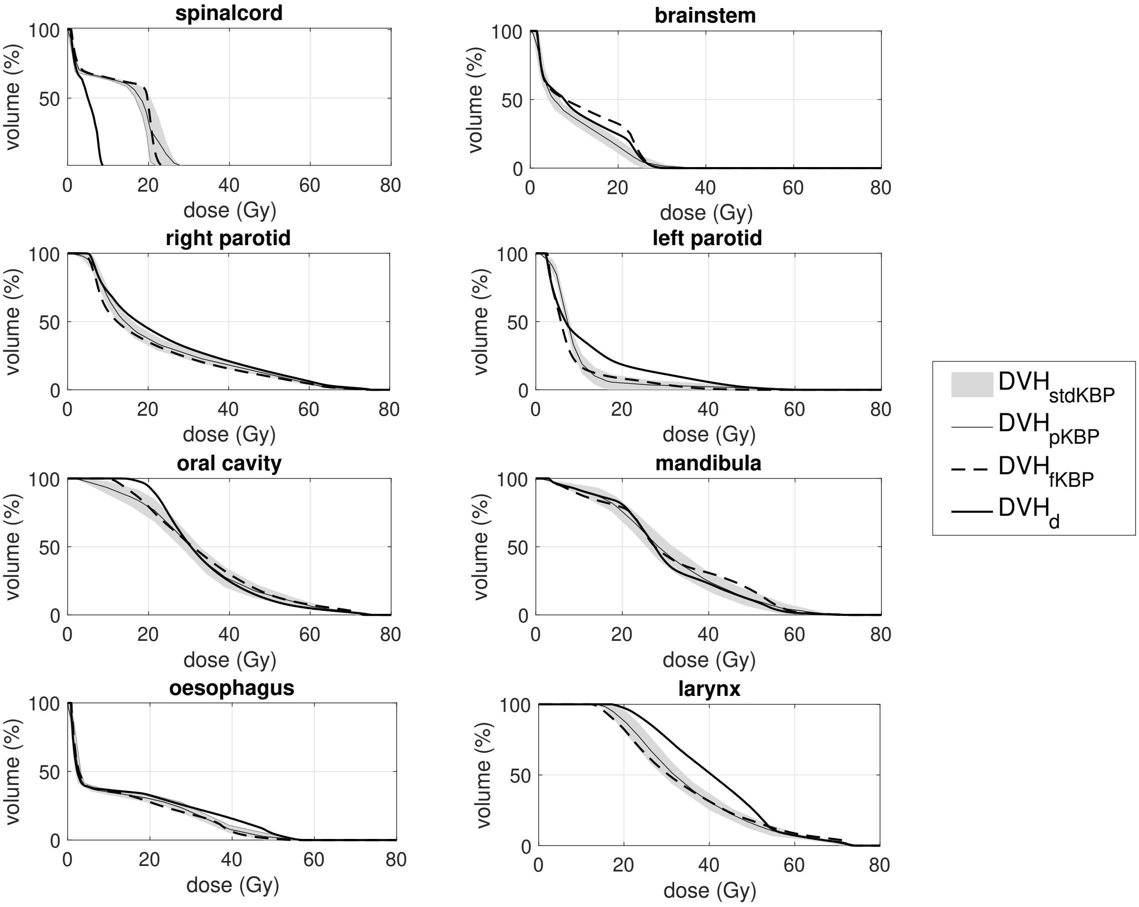


Figure 10.2: KBP-predicted DVH bounds (stdKBP), mid KBP-predicted DVH (DVH_{pKBP}), final KBP DVH (DVH_{fKBP}) and delivered DVH (DVH_d) for each organ for patient #7 fraction 16.

The overall fSR and pSR (mean±1SD) for all patient OARs and the fractions were 0.07±2.73 and 0.08±2.98, respectively. A significant difference was found between pSR and effective fSR (p=0.03). For each OAR, the mean and standard deviation of fSR and pSR values (the latter in brackets) were -1.34±2.15 (-0.97±2.30), 0.10±1.13 (0.07±1.20), 1.54±4.04 (0.70±3.81), 0.59± 2.96 (0.03

±2.96), -0.20 ±2.72 (1.07±2.61), -0.55±1.47 (1.48 ±3.00), 0.59±2.17 (-0.85± 3.13) and -0.27±3.08 (-1.18±3.04) for spinal cord, brainstem, left parotid, right parotid, mandible, oesophagus, oral cavity and larynx, respectively. The fSR and pSR values were similar between the two fractions. The Wilcoxon two-sided signed rank test confirmed significant differences for fSR values of the 16th and 26th fractions at p<0.01. No significant difference was observed in the pSR distribution of the two fractions. For 48% of the cases (N=310), KBP predicted a positive gain from re-planning (pSR>0). Final DVHs confirmed the effective gain fSR>0 in a similar percentage, 47%, of cases. Figure 3 shows a boxplot of the comparison between pSRs and fSRs for each organ for all 40 cases. Overall, there was no observed trend between median fSRs and median pSRs. For brainstem, parotids, oesophagus, and larynx, median fSRs were higher than their respective pSRs, as reported in Figure 10.3.

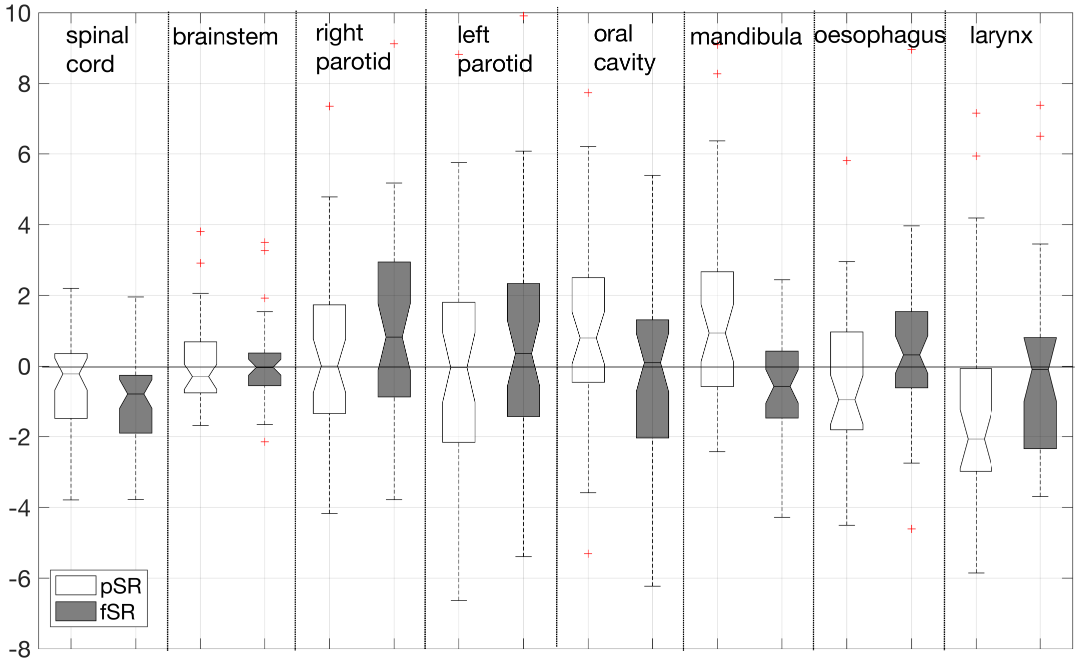


Figure 10.3: Boxplot of predicted and final gain in terms of pSRs (white) and fSRs (shaded) for each organ for all 20 delivered adapted-CBCT and KBP plans. The central mark represents the median value, whereas the edges of the boxes are the 25th and 75th percentiles, respectively. The whiskers extend to the maximum distance, and the crosses represent individual outliers.

10.4.3 *KBP model uncertainties*

Ideally KBP prediction uncertainties should be significantly smaller than, or at least comparable to the predicted gain. However, even if the model was trained with Pareto optimal plans, a relevant KBP prediction uncertainty is still present. For each OAR and each case, this study compared the pKBP uncertanty values versus the pASR, the latter representing the predicted gain of replanning quantified using ASR (subsection 10.3.7). These values are plotted in the correlation graph reported in Figure 10.4a. Each marker represents a plan of the evaluation set for which the pASR was higher (circle) or lower (cross) than the pKBP uncertainty. Overall, the mean pASR values were similar to KBP uncertainties and the pASR SD slightly higher than KBP uncertainties, with values of 2.78±2.13 and 2.38±0.79 respectively.

A Wilcoxon two-sided signed rank statistical test confirmed that there was not a significant difference between the two groups (p=0.56). From Figure 10.4, it is possible to observe that the correlation between KBP uncertainties and pASR was strictly related to OAR and a single item. The ideal situation in Figure 10.4a would be to see all points well above the diagonal line of unity, indicating a pASR value significantly larger than the KBP uncertainty (for each case). However, many items in Figure 10.4a fall below the diagonal. This finding means that in those cases KBP uncertainties were higher than predicted gains in terms of ASR. It was found that 15/40, 20/40, 19/40, 20/40, 20/40, 15/40, 23/40 and 15/30 points for pASR were higher than the corresponding pKBP uncertainty, for spinal cord, brainstem, left and right parotid, oral cavity, mandible, oesophagus, and larynx respectively. This is represented in Figure 10.4 as a square in the lower right corner. When the gain was lower/higher than uncertainties for more than half of the items the square is filled with black/gray colour; in the case of equality the square was white.

In Figure 10.4b the correlation between fKBP uncertainties and final gains, quantified as fASR, is shown. In comparison with Figure 10.4a, where pKBP uncertainties were considered, fewer points fall below the bisector. This means that the fKBP uncertainties were smaller than final gain, in terms of ASR. Overall, mean and SD values of 2.12±2.02 and 1.40±.61 were obtained for fASR and

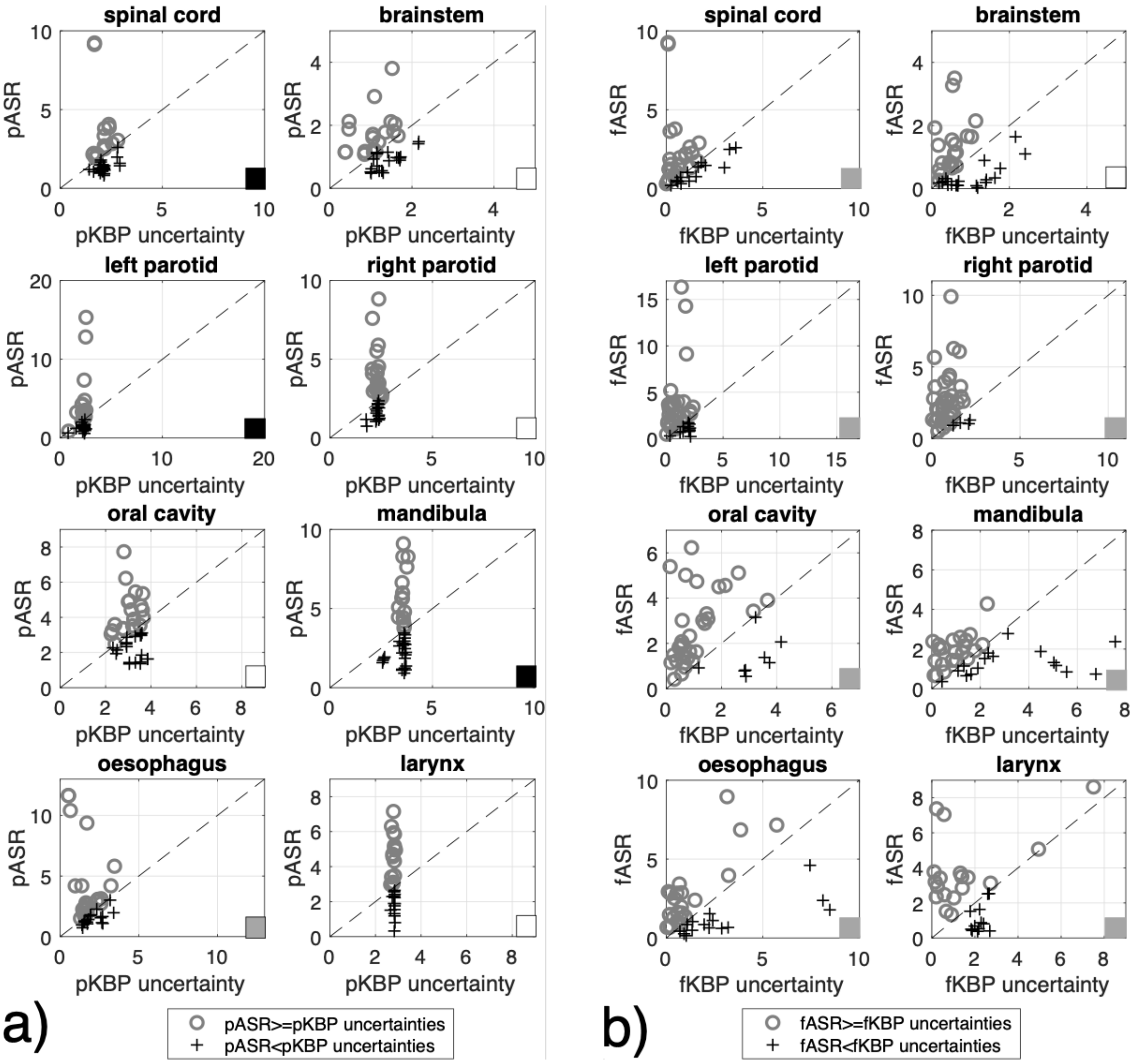


Figure 10.4: a) Correlation graphs of 1 SD predictions of pKBP uncertainties versus pASR from re-planning, for each OAR model. b) Correlation graphs of fKBP uncertainties versus fASR from re-planning, for each OAR model. Each marker represents a plan for the evaluation set for which the gain is higher (circle) or lower (cross) than the KBP uncertainty. This is represented in each panel of the figure as a square in the lower right corner. When the gain resulted lower/higher that uncertainties for more than half of the items the square is filled black/gray colour; in the case of equality the square was filled in white.

fKBP uncertainties respectively, with p-values $\ll$0.01. This confirms that the difference between the two distributions is statistically significant. In this case it was found that 26/40, 20/40, 28/40, 36/40, 31/40, 22/40, 25/40 and 18/30 cases for predicted gain were higher than the corresponding KBP uncertainties, for spinal cord, brainstem, left and right parotid, oral cavity, mandible, oesophagus, and larynx respectively.

10.4.4 *Discriminant analysis*

Figure 10.5 shows the correlation graph between pSR and fSR for the 20 patients and 2 fractions considered (N=40 cases). Overall, the two groups were deemed to be correlated with a coefficient of 0.72 (p$<$0.01). Single OAR correlation coefficients and p-values were 0.90 (p$<$0.01), 0.91 (p$<$0.01), 0.57 (p=0.18), 0.76 (p=0.04), 0.71 (p=0.04), 0.83 (p=0.01), 0.72 (p=0.04) and 0.42 (p=0.30) for the spinal cord, brainstem, parotids (left and right), oral cavity, mandible, oesophagus, and larynx, respectively. From discriminant analysis, the OP values for each of the OARs were 0.20, 0.09, -0.61, 0.70, 0.78, 1.67, 0.03 and -2.15. These OPs represent guidelines for clinical decision making using pSR of KBP values to obtain a real gain after plan optimisation (fSR$>$0). Figure 10.5 reports, for each OAR, the pSR and fSR. The OPs are represented by a vertical line in each graph for each organ. It is possible to observe that, for various OARs, many fSRs are positive for pSR values greater than OPs (marked as circles).

In Figure 10.6, ROC curves are shown separately for each OAR. The corresponding areas under the curve (AUCs) give an effective measure of the accuracy of the pSR prediction and were 0.824, 0.704, 0.974, 0.863, 0.772, 0.821 and 0.886 for the spinal cord, brainstem, parotids, oral cavity, mandible, oesophagus, and larynx, respectively. Overall, among the 310 cases (in 10 cases the larynx was missing), TP were 140, TN were 104, FP were 36 and FN were 30. From the ROC analysis, the sensitivity and specificity of the overall model were 0.69 and 0.78, respectively. The associated accuracy was 0.74, and the estimation error was 0.26.

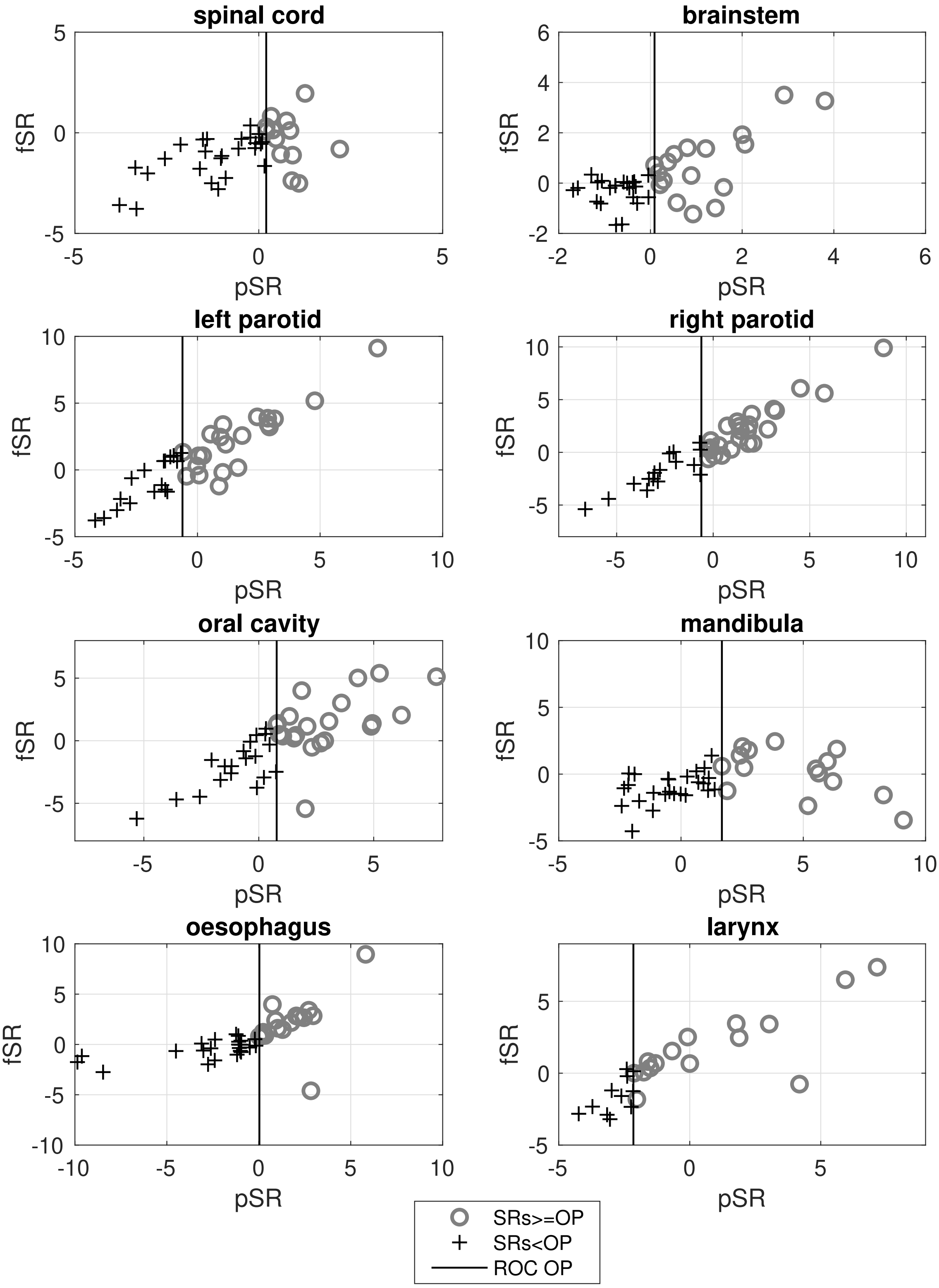


Figure 10.5: Correlation graph of predicted KBP gain pSR and fSR values for the 20 patients, where both fractions are considered (40 adapted CTs). OPs found from the ROC analysis are represented by a vertical line in each graph for each organ. Each marker represents a plan for the evaluation set for which the pSR were higher/right (circle) or lower/left (cross) of the OP value (vertical line).

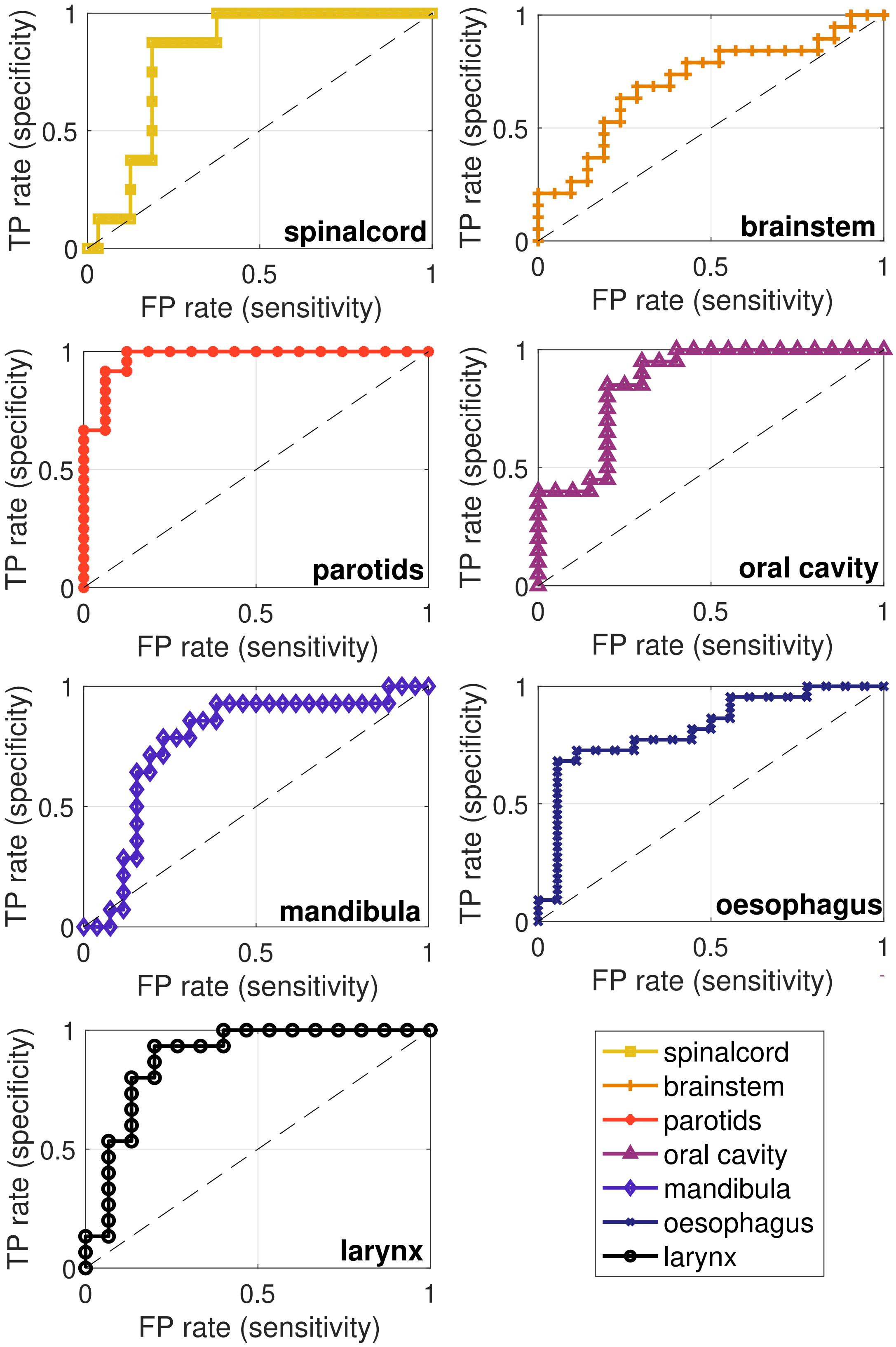


Figure 10.6: ROC curve for each OAR. The corresponding AUCs were 0.824, 0.704, 0.974, 0.863, 0.772, 0.821 and 0.886 for the spinal cord, brainstem, parotids, oral cavity, mandible, oesophagus, and larynx, respectively.

10.5 DISCUSSION

In this study, the use of a KBP tool for ART and evaluated the predicted gain from re-planning for OARs has been investigated. The KBP model was trained using Pareto-optimal treatment plans that were all automatically generated using the Eclipse MCO module with a uniquely prioritised, objective optimisation list, avoiding the use of manually generated plans to train the model as these generally suffer from variation in quality and inconsistencies among different planners. To assess the inherent accuracy of RapidPlan predictions for ART re-planning, the gain from re-planning was quantified by comparing DVHs of the original plan recalculated on patient anatomy, imaged at two different time points during treatment, with predicted and final KBP DVH obtained with a new plan on the same image set. Discriminant analysis performed in this study allowed an estimation of the KBP predictive power for the effective gain from re-planning, with an AUC value greater than 0.7 (Figure 10.6) for all OARs, confirming that mid-line DVH could be used as a good surrogate for prediction values as previously suggested in other works [23, 24]. ROC curve analysis established the best cut-off for predicted values for clinical purposes (OPs (subsection 10.3.8)). These OPs were found to be positive (range: [0.03, 1.67]) for 5 OAR models, except for the parotids (-0.61) and larynx (-2.15). Adapted-CTs were used, based on deformable image registration between planning CT and CBCT respectively, for fraction 16 and fraction 26. The validation of image registration algorithm for clinical use remains a challenging task in ART, as their accuracy depends on the complexity and quality of the images used in the registration task [182, 183]. In the absence of standardized tools, the accuracy of registration in this study was verified by means of a visual inspection. Deformed images (adapted-CTs), deformed structures on adapted-CTs and magnitude of the deformable vector fields were all carefully assessed. Since the investigations were performed directly on adapted-CTs by comparing predicted and final KBP DVH with the original plan recalculated on adapted-CT, this analysis was not directly influenced by the registration uncertainties. An optimal prediction model needs a prediction range (predicted uncertainty) that is as small as possible. Said range is defined in RapidPlan as 1 SD from the predicted result. A larger prediction range means that the model has larger uncertainties. In the MCO-KBP model training, consistently generated Pareto-optimal plans were used to build models. However, a significant statistical deviation from predicted DVH bounds is still present, as reported in Figure 10.2 for an example patient from the evaluation set. Since manual plans were not used to build the model, the deviation is not related to the quality of the suboptimal plans but only to the intrinsic limitations of the model. The KBP prediction error bounds led to significantly higher pKBP uncertainty, as reported in Figure 10.4a compared with the predicted gain in term of ASR for several cases. Both quantities (pASR and KBP uncertainties) were dependent on the OAR KBP model (Figure 10.4). Overall, pASR of pKBP uncertainties and the ASR of predicted gains when a new plan was used were very close in average values and Wilcoxon signed rank tests confirmed no significant differences between the two groups ($p>0.05$). When considering the final values after optimisation this relationship improved, showing fASR values higher than corresponding fKBP uncertainty values for the majority of cases (Figure 10.4b). Overall, predicted, and final gains, pSR and fSR, were correlated ($p<0.01$) and showed a similar proportion of cases with positive gain predicted (48%) with respect to positive actual gain (47%). In the process of building the KBP prediction model, outliers are defined as training plans that could result in undesirable bias in the models. In the absence of general rules, some criteria to better identify and manage these possible outliers have been previously published [24, 80, 82]. In this study, all plans used to build the model were Pareto-optimal with consistent trade-offs between all treatment objectives. Therefore, any dosimetric outliers were not expect in the training set. However, the Varian MA tool still identified some instances of outliers in the current training dataset (Table 10.1). Those potential outliers could have arisen due to the limited range of geometric information in the input data and/or intrinsic limits of the models, as reported in a previous study by Cagni *et al.* [82]. Varian's RapidPlan guide indicates a minimum number of 20 plans to train a model. However, Boutilier et al [81] have shown that for prostate cancer DVH prediction, RapidPlan needs at least 75 plans to achieve good prediction accuracy. Fogliata *et al.* [24] considered 83 patients for building a KBP model with RapidPlan and the models were validated on 20 HN patients. In line with this approach, our model was built on 80 cases for training and 20 cases for validation. The model quality was evaluated by checking the goodness of fit statistics for each structure, with the coefficient of determination R^2, with observed values higher than 0.8 for all structures except spinal cord, and the average Pearson's chi square χ^2, with all observed values less than 1.1 (Table 10.1). These results are in line with the KBP models of goodness of fit reported in other studies [18, 23, 24]. Moreover, differences in terms of model predicted and final endpoints (D_{max} and D_{mean}) were not statistically significant (Table 10.2), confirming the goodness of model prediction. Sum of residuals (SR) between entire DVHs were used as the metric to quantify predicted and final OAR gain. From a radiobiological perspective, SR gives interesting information about global DVH and mean dose variation, generally considered for parallel organs, such as the parotids, larynx, oral cavity, or oesophagus [182, 183]. On the other hand, for serial organs such as the spinal cord, only the high dose region of the DVH is correlated with radiation-induced complications [182, 183]. In this work, the spinal cord had the worst (negative) gain among all OARs with a mean final SR of -1.34, but it did not provide information about the radiobiological integrity of the organ since there was no indication of how its maximum delivered dose changed. For this reason, specific DVH endpoints were considered in order to summarize the RapidPlan prediction performance, as described in Table 10.2.

10.6 CONCLUSIONS

This work has demonstrated the feasibility of using knowledge-based tools to establish the need for plan adaptation based on OAR sparing to help automate the process of re-planning. This study has shown that the prediction uncertainties of knowledge-based planning tools trained with Pareto-optimal plans are sufficiently low for such tools to be used in adaptive radiotherapy. The approach described in this work, combined with the previous chapter, chapter 9, where a standardized method to test DIR in ART process was presented, has the potential to be implemented in an on-line adaptive radiotherapy process; and, in more generally, the work of chapter 9 and chapter 10 can speed up and standardize clinical ART process for HN patients.

Part V

ON-GOING RESEARCH

11

GENERAL CONCLUSIONS, CURRENT AND FUTURE WORK

> Human beings set out to encounter other worlds, other civilizations, without having fully gotten to know their own hidden recesses, their blind alleys, well shafts, dark barricaded doors.
>
> *Stanisław Lem - Solaris*

11.1 SUMMARY OF SIGNIFICANT CONTRIBUTIONS

In this thesis, automated treatment planning tools have been developed and investigated, focusing on head and neck primarily and also on breast cancer. Quality improvement related to manual planning was one key aspects within the research scope. The impact of automated planning (knowledge based tools) and standardization methods for quality assurance in adaptive radiotherapy was also investigated. The performed research has produced a published paper in an important journal in the field. Another area of research developed in this thesis concerned the choice of the best plan for clinical treatment. The results presented emphasise not only the necessity for standardisation using automation, but also the importance of the plan quality assessment step of the clinical radiotherapy workflow. This is a novel aspect not previously investigated, to best of current knowledge. In this area, the performed research also produced a published paper in an important journal in the field. There is now a new project already approved by ethical committee on breast planning evaluations over several institutes (see subsection 11.2.2 of this chapter for details). In this final chapter, the focus will be on challenges that automated planning and tools bring within clinics and the benefits of automated planning in treatment technique comparisons and plan quality evaluation. To conclude, possibilities for current and future research are discussed.

11.1.1 *Automatic Planning for plan quality and plan consistency*

Ideally, auto-planning results in all patients having a final acceptable high-quality plan, which should also be Pareto-optimal. For the vast majority of patients there should at least be no need for manual fine-tuning of the auto-plan. To reach this point, proper configuration of the auto-planning algorithm is crucial. In multicriteria optimisation algorithms, configuration means generation of a wish-list [65] to obtain plans of good quality and consistency. The process of wish-list tuning was described in chapter 3 for head and neck and chapter 8 for breast cancer. A well constructed wish-list can ensure that plans are generated in line with scientific knowledge, as well as local treatment traditions, e.g. regarding the required level of overall high-dose conformality relative to sparing of specific OARs. The configurations performed for the studies in this thesis have resulted in the following observations.

i. Each wish-list determines the plan quality for an entire patient group. If the configuration is sub-optimal, the quality of all plans will be sub-optimal, effectively introducing a systematic problem. There is never a guarantee that the best possible wish-list will be found, because the best possible plan quality is generally not well defined. Comparison with manually generated, clinically delivered plans is a basic measure to ensure that at least the equivalent clinical quality is obtained. However, as demonstrated in this thesis, in chapters chapter 3, chapter 6, and chapter 8, extensive tuning generally results in a wish-list that can exceed standard clinical plan quality. Tuning of wish-lists is however a complex, interactive procedure. As for manual planning for individual patients, it is not always clear when to stop. Previous experience in generation of wish-lists is likely to facilitate the task and may result in a better outcome. It is also important that enough time is reserved to obtain the best possible result. This observation is not new but is a confirmation of previous studies on automatic planning.

ii. A well-established treatment planning protocol, agreed upon by the treating clinicians, is crucial for proper wish-list tuning. However, these protocols can generally only partly describe how optimal plans should appear, as it is virtually impossible to fully quantify requirements for conformality, dose spikes, the dose bath, hot/cold spots, and balances between all treatment objectives. Chapter 3 and chapter 8 reported on the large differences found between clinical protocol and final wish-list used in Erasmus-iCycle. This observation is not knew but it is a confirm of previously studies on automatic planning.

iii. Automated planning showed comparable quality to manual planning but had higher consistency between plans relative to manual plans. This has been shown previously in the literature [7, 82, 101–104] and is also demonstrated in this thesis, with data presented in chapter 6 (Figure 6.4 panels f, g and h). There it was found that there was higher agreement between evaluators when automated plans using wish-lists were considered compared to manual plan based judgement.
Figure 11.1 reports for the different plan types described in chapter 6, i.e. CLIN (manual plans), MCOa (automated plan generated with a consistent single best wish-list) and MCOx (automated plan generated using suboptimal wish-list with different endpoint priorities) the absolute scores between 9 evaluators studied in chapter 6. MCOa showed the best absolute scores with the minimum spread over all evaluators compared to CLIN and MCOx. Moreover for 8/9 evaluators MCOa was the plan type which showed highest absolute scores. The finding that automatic planning can reduce the inter observer variability in plan quality assessment is new respect to the background presented in the literature (more details in the next section).

11.1.2 *Automatic tools to support plan quality assessment variability*

In most centres, treatment plans are prepared by MPs, and evaluated for final clinical approval by the treating ROs. The process, often denoted as manual planning or trial-and-error planning, may have several iterations in which the planner adjusts intermediate plans, based on feedback from the RO. Limited common understanding or agreement between planners and ROs on what constitutes a 'good plan' can result in sub-optimal dose distributions, even with iteration loops.

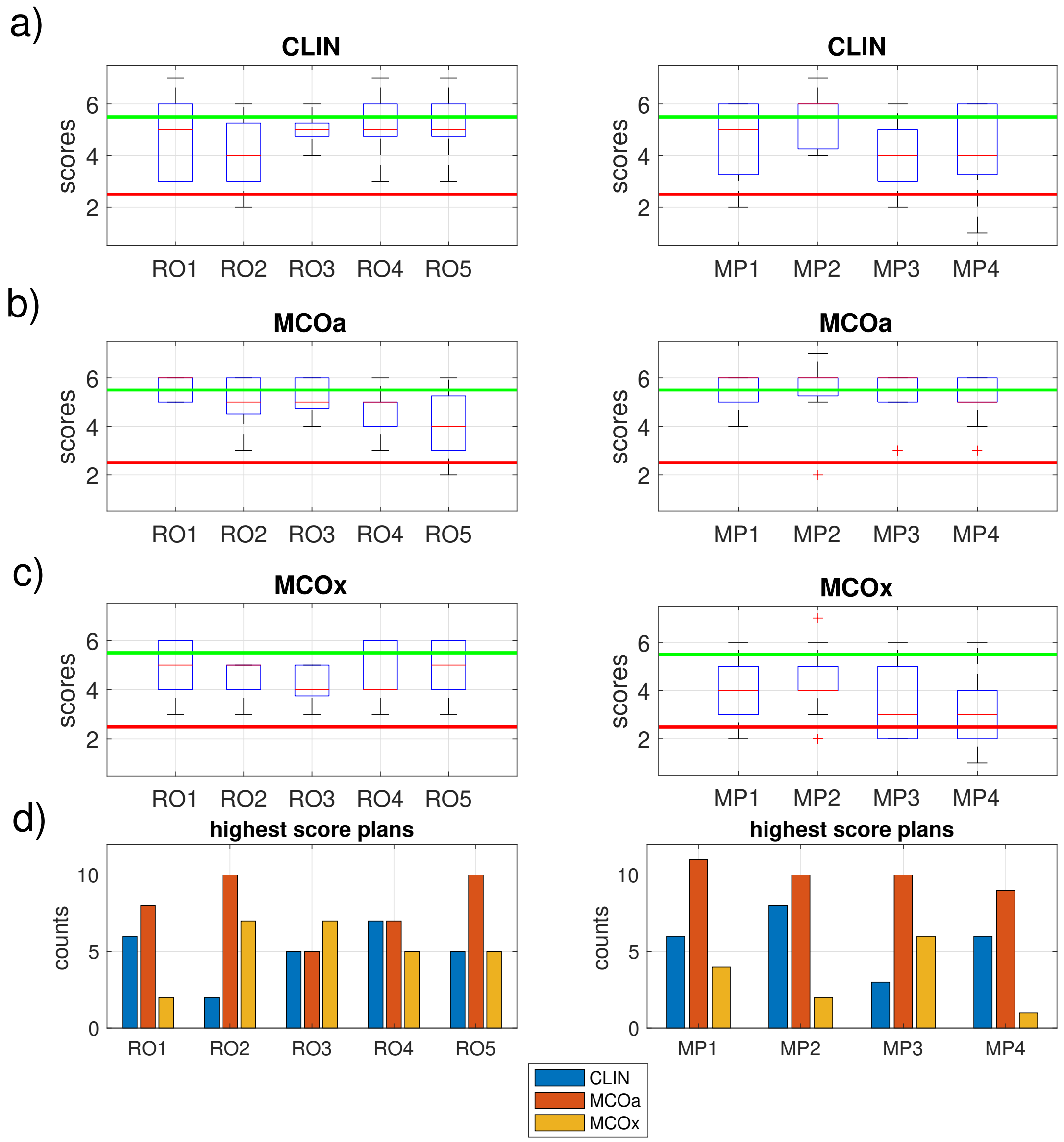


Figure 11.1: Raw a-b-c, boxplot of absolute scores over 65 plans for each evaluator (5 radiation oncologists (ROs) and 4 medical physicists (MPs) considered in the study reported in chapter 6 divided by plan type, CLIN (manual), MCOa (automatically generated using optimal wish-list) and MCOx(automatically generated using sub-optimal wish-list). Raw d, histogram of the highest score frequency divided by plan type for each evaluator.

In this thesis, the differences between groups of ROs and planning MPs in a single radiotherapy department were investigated in terms of perceived quality of oropharyngeal cancer plans. To the best of current knowledge, this is the first study (presented in chapter 6) that systematically investigates variations in subjective plan quality assessment among ROs and MPs working in a particular department.

i. The results of the current work could stimulate similar studies in other departments as they seem to indicate an important weak link in the radiotherapy planning chain. It is commonly recognized that variations between ROs in delineating targets is a major concern in clinical radiotherapy. This study suggests that large inter-observer variations in plan quality assessments (even in a single department), could be another 'Achilles heel' inhibiting successful treatment.

ii. It is possible that broad departmental discussions on plan requirements, aiming at a generally shared but precisely defined view on plan quality, could reduce the currently large inter-observer variations in plan quality assessment. In chapter 7, a method to better understand the treatment plan evaluation process in radiotherapy through the use of machine learning (ML) methods and an idealised dose model (called gUIDE) was presented. This tool was shown to improve the accuracy of using ML tools to model plan quality evaluation process for several users. A high degree of variability among users was found in terms of the features of importance considered by each evaluator with the ML approach considered for the evaluation. Future application of such tools could possibly contribute to enhanced plan quality consistency. One approach could be to use these for a training purposes, in order to reduce the variability in the pattern of feature importance revealed by ML and gUIDE.

iii. Automated planning could also result in reduction of plan quality variability. This has been shown earlier in this chapter (Figure 11.1). Enhanced plan quality with automated planning compared to manual planning has been observed previously [44, 54, 97–100], but the results presented in this thesis are the first showing reduced inter-observer variations in subjective plan scores for the automated plans compared to corresponding manual plans. Other studies have identified the use of numerical plan quality assessment tools to enhance treatment plan quality [112].

11.1.3 *Automatic tools for adaptive radiotherapy*

Within this thesis several aspects of the automation process in adaptive radiotherapy have been investigated (chapter 9 and chapter 10).

i. One current issue in ART is the lack of standardisation methods to test the quality of deformable image registrations. In this thesis a registration-based method for deformable image registration quality assurance for adaptive radiotherapy, using digitally post-processed head and neck anthropomorphic phantom image datasets was developed and analysed. One of the main findings of this work was that spatial and dose errors are a function of the magnitude of the deformation and of the gradient of the dose distribution. This emphasizes the importance of performing patient specific DIR verification and consequently, the need to develop and make available tools that are fit for this purpose. A novel method to evaluate the quality of patient-specific multimodality CT-CBCT DIR for adaptive radiotherapy of head and neck patients has been investigated in the thesis. The methodology described allows direct testing of DIR algorithms for clinical registration, which can produce valuable insights into their clinical impact on the adapted dose distribution. The key technical requirement for using this method is the availability of an independent image registration software platform (in addition to the clinical system). This work, reported in chapter 9 contributes to the study of standardisation and automation of quality assurance methods for deformable image registration in radiotherapy.

ii. The technique of knowledge-based planning used generally for automation of plan generation, was applied in this thesis for the adaptive radiotherapy process, using it for potential organ at risk sparing estimation in the replanning strategy for head and neck ART. KBP tools can generate estimated DVHs based on previous patient anatomy and dose distributions. Previously published studies have demonstrated the efficacy of KBP to create a new plan for several cancer sites. KBP training generally involves using a database of manually generated treatment plans that can suffer from plan quality variation and inconsistency. DVH predictions shown by the KBP tool consisted of two DVH bounds indicating 68% probability that the final DVH should fall between the bounds. To maximise the performance of KBPs (i.e. reducing the prediction DVH bounds), Pareto plan solutions, created with the Varian MCO Trade-Off module, were used to train the KBP model. However, even if the model was trained with Pareto optimal plans, a relevant KBP prediction uncertainty was still present. To be used effectively in ART, KBP prediction uncertainties should be significantly smaller than, or at least comparable to, the predicted gain in OAR sparing if a new plan is performed. Chapter 10 demonstrated the feasibility of using KBP tools to establish the need for plan adaptation based on OAR sparing to help automate the process of re-planning. This study has shown that the prediction uncertainties of KBP tools trained with Pareto-optimal plans are sufficiently low for such tools to be used in ART. A systematic workflow for identifying effective OAR sparing in replanning strategies based on KBP prediction is presented. It was concluded that this method could provide an important KBP application for adaptive radiotherapy and give feasible estimation of OAR sparing. The approach described in this work informs development of the automation of the adaptive radiotherapy processes.

11.2 CURRENT AND FUTURE WORK

11.2.1 *gUIDE and ML as tool to reduce the inter-user variability in plan quality assessment*

The study presented in Chapter 8 is the subject of current study. The following aspects are work in progress.

i. Improving the AUC of ML-gUIDE models considering other endpoints more related to dose distribution rather than DVH data. This is done to better simulate the real judgement of the clinician during plan approval. A key aspect of the plan quality assessment involves looking at the isodose distribution on planning CT slices and structure sets. To include this, a cumulative quality volume histogram (QVH) is considered. QVH is simply the histogram of ratio between plan dose and ideal dose, given in this study by gUIDE, defined for each voxel. The ideal QVH is a step function at Q = 1 [184]. The hypothesis is that by considering features more related to the real plan approval process, the ML model will give better results.

ii. Using the results from ML-gUIDE found in point i. to build a training process for plan quality assessment for a particular department. This includes recognition of a specific feature pattern as the favoured one (i.e. the most chosen one). By discussing and sharing the feature importance variations quantified by ML-gUIDE to all the evaluators, aiming at a widely shared, and precisely defined view on plan quality, could improve the current large inter-observer variation in plan quality assessments. A second round of plan quality assessment may then be done after this training to check if the agreement has improved among users.

11.2.2 *AIRPLAN B project*

A study called ‘Automation In Radiotherapy treatment PLANs: optimisation and evaluation processes in Breast cancer treatment’ (AIRPLAN B) was presented and approved by ethical committee in August 2021 (689/2021/OSS/IRCCSRE - AIRPLANE B) and with thesis author as principal investigator. This is based on the study design and methods developed in the research study for head-neck cancer treatment, described in chapter 6 and chapter 7 and recently published [95]. Initial results have been presented in chapter 8 of this thesis. The central hypothesis of this study is that it is possible to quantify the differences between ROs and MPs intra and inter-institute in perceived quality of breast cancer plans. Broad departmental and inter-departmental discussions on plan requirements, aiming at a broadly shared, and precisely defined view on plan quality, could improve the current large inter-observer variation in plan quality assessments. The data could be used to guide automated RT planning evaluation using machine learning tools able to assist departments for plan quality assurance, quality evaluation and reducing plan quality variation [95, 185]. A second hypothesis is that the introduction of automation techniques into breast radiotherapy planning practice can improve the plan quality and reduce the planners' variability. The study is a retrospective and observational study and it is divided into two main parts.

1. Generation of Pareto optimal plans to improve complex breast treatment planning (this is initially reported in chapter 8 of this thesis). About 60/80 left breast cancer patients who underwent radiotherapy treatment at AUSL-IRCCS are included. Two physicians will check all OARs and target contouring to be consistent within the patient selected group. The OARs considered were the ipsilateral and contralateral lungs, the heart, the contralateral breast, the spinal cord and the left arteria descending (LAD). An internal clinical protocol (wish-list) is defined for breast inside AUSL-IRCCS Reggio Emilia hospital, based on international guidelines (chapter 8). Pareto optimal plans are generated using the Erasums-iCycle module. Plan generation is based on a wish-list, describing hard planning constraints and planning objectives used at AUSL-IRCCS of Reggio Emilia. Each objective in the wish-list will have an assigned priority. Inside the optimizer, the objective priorities will be used for multi-criterial plan generation choice, aiming at clinically favourable balances between several, often competing objectives. These balances will be found in a procedure that is identical for all patients, guaranteeing consistent plan quality (see chapter 8). Pareto optimal plans, generated with the Erasmus-iCycle module, will be used as reference information to develop and define an in-house complex technique (IMRT or VMAT) using the AUSL-IRCCS clinical treatment planning system Eclipse. Comparison between Pareto optimal Erasmus-iCycle module plans and complex Eclipse plans will be performed in terms of dosimetric endpoint and clinical evaluation process by ROs and MPs of AUSL-IRCCS institute.

2. Evaluation of variation in plan quality assessment between radiation oncologists and medical physicists in both single and multiple radiotherapy departments. For the same group of patients considered in 1) apart from the clinical 3D-CRT (CLIN) plans used for the treatment, from 2 to 4 additional IMRT plans will be produced for this study for each patient. All these ad-

ditional plans will be generated using the Erasmus-iCycle module. One of these will be generated with the wish-list defined by AUSL-IRCCS of Reggio Emilia (MCOa plan) in Part 1 of the study and will be the reference plan. The extra plans will have variable plan quality and will be generated with automated planning by varying some geometrical parameters or wish-list priorities with respect to the reference (MCOx plans). All plans will be evaluated in this study, resulting in a total of 150/250 evaluable plans. All patients and plans will be anonymized and each of the available plans will be evaluated by three departmental ROs and three MPs. For each patient, every observer independently will give a score to each of the 3 or 5 available plans in a single session. Scoring will be blinded, i.e. observers will not know how the plans were generated. Evaluators will know only some information concerning patients: age, tumour stage and if the patient has any heart disease. Apart from giving a quality score to each plan, observers will be also asked what change they considered most desirable for improvement of the plan (without knowing whether this would be feasible or not) and to choose the plan for each patient that theoretically will go to the treatment. To assess intra-observer variability in quality scoring, 1 RO and 1MP will perform the entire scoring process a second time, with a delay of at least a month. Previous results will be blinded. The same procedures (inter-observers and intra-observer), considering the same group of patients/and plans of AUSL-IRCCS of Reggio Emilia, will be performed in other 2 institutes, Firenze and Piacenza, by a group of ROs and MPs for each institute. Inter and intra evaluator variability will be quantified for each individual department and over all departments. In this second part, the investigators will use both statistical and artificial ML methods to handle information contained in dose distributions and to determine training methods that can be used to support the evaluation of breast radiotherapy treatment plans and improve consistency both within and between departments.

BIBLIOGRAPHY

[1] Torre, L. A., Siegel, R. L., Ward, E. M., and Jemal, A. (2016). "Global cancer incidence and mortality rates and trends—an update". In: *Cancer Epidemiology and Prevention Biomarkers* 25.1, pp. 16–27.

[2] Todua, F., Gagua, R., Maglakelidze, M., and Maglakelidze, D. (2015). "Cancer incidence and mortality-Major patterns in GLOBOCAN 2012, worldwide and Georgia". In: *Bull Georg Natl Acad Sci* 9.1, pp. 168–173.

[3] Good, D., Lo, J., Lee, W. R., Wu, Q. J., Yin, F.-F., and Das, S. K. (2013). "A knowledge-based approach to improving and homogenizing intensity modulated radiation therapy planning quality among treatment centers: an example application to prostate cancer planning". In: *International Journal of Radiation Oncology* Biology* Physics* 87.1, pp. 176–181.

[4] Cherry, P. and Duxbury, A. M. (2019). *Practical radiotherapy: physics and equipment*. John Wiley & Sons.

[5] Breedveld, S., Craft, D., Van Haveren, R., and Heijmen, B. (2019). "Multi-criteria optimization and decision-making in radiotherapy". In: *European Journal of Operational Research* 277.1, pp. 1–19.

[6] Sanchez-Nieto, B and Nahum, A. (2000). "BIOPLAN: software for the biological evaluation of radiotherapy treatment plans". In: *Medical Dosimetry* 25.2, pp. 71–76.

[7] Hussein, M., Heijmen, B. J., Verellen, D., and Nisbet, A. (2018). "Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations". In: *The British journal of radiology* 91.1092, p. 20180270.

[8] Monz, M., Küfer, K.-H., Bortfeld, T. R., and Thieke, C. (2008). "Pareto navigation—algorithmic foundation of interactive multi-criteria IMRT planning". In: *Physics in Medicine & Biology* 53.4, p. 985.

[9] Munter, J. S. and Sjölund, J. (2015). "Dose-volume histogram prediction using density estimation". In: *Physics in Medicine & Biology* 60.17, p. 6923.

[10] Yang, Y. and Xing, L. (2004). “Clinical knowledge-based inverse treatment planning”. In: *Physics in Medicine & Biology* 49.22, p. 5101.

[11] Yang, Y., Ford, E. C., Wu, B., Pinkawa, M., Van Triest, B., Campbell, P., Song, D. Y., and McNutt, T. R. (2013). “An overlap-volume-histogram based method for rectal dose prediction and automated treatment planning in the external beam prostate radiotherapy following hydrogel injection”. In: *Medical physics* 40.1, p. 011709.

[12] Petit, S. F. and Elmpt, W. van (2015). “Accurate prediction of target dose-escalation and organ-at-risk dose levels for non-small cell lung cancer patients”. In: *Radiotherapy and Oncology* 117.3, pp. 453–458.

[13] Ahmed, S., Nelms, B., Gintz, D., Caudell, J., Zhang, G., Moros, E. G., and Feygelman, V. (2017). “A method for a priori estimation of best feasible DVH for organs-at-risk: Validation for head and neck VMAT planning”. In: *Medical physics* 44.10, pp. 5486–5497.

[14] Sheng, Y., Ge, Y., Yuan, L., Li, T., Yin, F.-F., and Wu, Q. J. (2017). “Outlier identification in radiation therapy knowledge-based planning: a study of pelvic cases”. In: *Medical physics* 44.11, pp. 5617–5626.

[15] Mayo, C. S., Yao, J., Eisbruch, A., Balter, J. M., Litzenberg, D. W., Matuszak, M. M., Kessler, M. L., Weyburn, G., Anderson, C. J., Owen, D., et al. (2017). “Incorporating big data into treatment plan evaluation: Development of statistical DVH metrics and visualization dashboards”. In: *Advances in radiation oncology* 2.3, pp. 503–514.

[16] Zhu, X., Ge, Y., Li, T., Thongphiew, D., Yin, F.-F., and Wu, Q. J. (2011). “A planning quality evaluation tool for prostate adaptive IMRT based on machine learning”. In: *Medical physics* 38.2, pp. 719–726.

[17] Fogliata, A., Wang, P.-M., Belosi, F., Clivio, A., Nicolini, G., Vanetti, E., and Cozzi, L. (2014). “Assessment of a model based optimization engine for volumetric modulated arc therapy for patients with advanced hepatocellular cancer”. In: *Radiation Oncology* 9.1, pp. 1–13.

[18] Hussein, M., South, C. P., Barry, M. A., Adams, E. J., Jordan, T. J., Stewart, A. J., and Nisbet, A. (2016). “Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy”. In: *Radiotherapy and Oncology* 120.3, pp. 473–479.

[19] Fogliata, A., Nicolini, G., Clivio, A., Vanetti, E., Laksar, S., Tozzi, A., Scorsetti, M., and Cozzi, L. (2015b). “A broad scope knowledge based model for optimization of VMAT in esophageal cancer: validation and assessment of plan quality among different treatment centers”. In: *Radiation Oncology* 10.1, pp. 1–11.

[20] Fogliata, A., Nicolini, G., Bourgier, C., Clivio, A., De Rose, F., Fenoglietto, P., Lobefalo, F., Mancosu, P., Tomatis, S., Vanetti, E., et al. (2015a). “Performance of a knowledge-based model for optimization of volumetric modulated arc therapy plans for single and bilateral breast irradiation”. In: *PLoS One* 10.12, e0145137.

[21] Chin Snyder, K., Kim, J., Reding, A., Fraser, C., Gordon, J., Ajlouni, M., Movsas, B., and Chetty, I. J. (2016). “Development and evaluation of a clinical model for lung cancer patients using stereotactic body radiotherapy (SBRT) within a knowledge-based algorithm for treatment planning”. In: *Journal of applied clinical medical physics* 17.6, pp. 263–275.

[22] Foy, J. J., Marsh, R., Ten Haken, R. K., Younge, K. C., Schipper, M., Sun, Y., Owen, D., and Matuszak, M. M. (2017). “An analysis of knowledge-based planning for stereotactic body radiation therapy of the spine”. In: *Practical radiation oncology* 7.5, e355–e360.

[23] Tol, J. P., Delaney, A. R., Dahele, M., Slotman, B. J., and Verbakel, W. F. (2015). “Evaluation of a knowledge-based planning solution for head and neck cancer”. In: *International Journal of Radiation Oncology* Biology* Physics* 91.3, pp. 612–620.

[24] Fogliata, A, Reggiori, G, Stravato, A, Lobefalo, F, Franzese, C, Franceschini, D, Tomatis, S, Mancosu, P, Scorsetti, M, and Cozzi, L (2017a). “RapidPlan head and neck model: the objectives and possible clinical benefit”. In: *Radiation Oncology* 12.1, pp. 1–12.

[25] Brock, K. K. (2013). *Image processing in radiation therapy*. CRC Press.

[26] Morgan, H. E. and Sher, D. J. (2020). “Adaptive radiotherapy for head and neck cancer”. In: *Cancers of the head & neck* 5.1, pp. 1–16.

[27] Castelli, J, Simon, A, Lafond, C, Perichon, N, Rigaud, B, Chajon, E, De Bari, B, Ozsahin, M, Bourhis, J, and Crevoisier, R de (2018). “Adaptive ra-

diotherapy for head and neck cancer". In: *Acta Oncologica* 57.10, pp. 1284–1292.

[28] Barker Jr, J. L., Garden, A. S., Ang, K. K., O'Daniel, J. C., Wang, H., Court, L. E., Morrison, W. H., Rosenthal, D. I., Chao, K. C., Tucker, S. L., et al. (2004). "Quantification of volumetric and geometric changes occurring during fractionated radiotherapy for head-and-neck cancer using an integrated CT/linear accelerator system". In: *International Journal of Radiation Oncology* Biology* Physics* 59.4, pp. 960–970.

[29] Yan, D., Lockman, D., Martinez, A., Wong, J., Brabbins, D., Vicini, F., Liang, J., and Kestin, L. (2005). "Computed tomography guided management of interfractional patient variation". In: *Seminars in radiation oncology*. Vol. 15. 3. Elsevier, pp. 168–179.

[30] Brouwer, C. L., Steenbakkers, R. J., Langendijk, J. A., and Sijtsema, N. M. (2015b). "Identifying patients who may benefit from adaptive radiotherapy: Does the literature on anatomic and dosimetric changes in head and neck organs at risk during radiotherapy provide information to help?" In: *Radiotherapy and Oncology* 115.3, pp. 285–294.

[31] Kessler, M. L. (2006). "Image registration and data fusion in radiation therapy". In: *The British journal of radiology* 79.special_issue_1, S99–S108.

[32] Lu, W., Olivera, G. H., Chen, Q., Chen, M.-L., and Ruchala, K. J. (2006). "Automatic re-contouring in 4D radiotherapy". In: *Physics in Medicine & Biology* 51.5, p. 1077.

[33] Sarrut, D. (2006). "Deformable registration for image-guided radiation therapy". In: *Zeitschrift für medizinische Physik* 16.4, pp. 285–297.

[34] Xing, L., Thorndyke, B., Schreibmann, E., Yang, Y., Li, T.-F., Kim, G.-Y., Luxton, G., and Koong, A. (2006). "Overview of image-guided radiation therapy". In: *Medical Dosimetry* 31.2, pp. 91–112.

[35] Kaus, M. R. and Brock, K. K. (2007). "Deformable image registration for radiation therapy planning: algorithms and applications". In: *Biomechanical Systems Technology: Volume 1: Computational Methods*, pp. 1–28.

[36] Ge, Y. and Wu, Q. J. (2019). "Knowledge-based planning for intensity-modulated radiation therapy: a review of data-driven approaches". In: *Medical physics* 46.6, pp. 2760–2775.

[37] Krayenbuehl, J, Zamburlini, M, Ghandour, S, Pachoud, M, Tanadini-Lang, S, Tol, J, Guckenberger, M, and Verbakel, W. (2018). “Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer”. In: *Radiation Oncology* 13.1, pp. 1–8.

[38] Amaloo, C., Hayes, L., Manning, M., Liu, H., and Wiant, D. (2019). “Can automated treatment plans gain traction in the clinic?” In: *Journal of applied clinical medical physics* 20.8, pp. 29–35.

[39] Cagni, E., Botti, A., Chendi, A., Iori, M., and Spezi, E. (2021a). “Use of knowledge based DVH predictions to enhance automated re-planning strategies in head and neck adaptive radiotherapy”. In: *Physics in Medicine & Biology*.

[40] Beyzadeoglu, M., Ozyigit, G., and Ebruli, C. (2010). *Basic radiation oncology*. Springer Science & Business Media.

[41] Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R. L., Torre, L. A., and Jemal, A. (2018). “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. In: *CA: a cancer journal for clinicians* 68.6, pp. 394–424.

[42] Batumalai, V., Jameson, M. G., Forstner, D. F., Vial, P., and Holloway, L. C. (2013). “How important is dosimetrist experience for intensity modulated radiation therapy? A comparative analysis of a head and neck case”. In: *Practical radiation oncology* 3.3, e99–e106.

[43] Moore, K. L., Brame, R. S., Low, D. A., and Mutic, S. (2011). “Experience-based quality control of clinical intensity-modulated radiotherapy planning”. In: *International Journal of Radiation Oncology* Biology* Physics* 81.2, pp. 545–551.

[44] Nelms, B. E., Robinson, G., Markham, J., Velasco, K., Boyd, S., Narayan, S., Wheeler, J., and Sobczak, M. L. (2012). “Variation in external beam treatment plan quality: an inter-institutional study of planners and planning systems”. In: *Practical radiation oncology* 2.4, pp. 296–305.

[45] Naghavi, M., Abajobir, A. A., Abbafati, C., Abbas, K. M., Abd-Allah, F., Abera, S. F., Aboyans, V., Adetokunboh, O., Afshin, A., Agrawal, A., et al. (2017). “Global, regional, and national age-sex specific mortality for 264

causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016". In: *The Lancet* 390.10100, pp. 1151–1210.

[46] Fisher, B., Anderson, S., Bryant, J., Margolese, R. G., Deutsch, M., Fisher, E. R., Jeong, J.-H., and Wolmark, N. (2002). "Twenty-year follow-up of a randomized trial comparing total mastectomy, lumpectomy, and lumpectomy plus irradiation for the treatment of invasive breast cancer". In: *New England Journal of Medicine* 347.16, pp. 1233–1241.

[47] Mayo, C. S., Urie, M. M., and Fitzgerald, T. J. (2005). "Hybrid IMRT plans—concurrently treating conventional and IMRT beams for improved breast irradiation and reduced planning time". In: *International Journal of Radiation Oncology* Biology* Physics* 61.3, pp. 922–932.

[48] Kestin, L. L., Sharpe, M. B., Frazier, R. C., Vicini, F. A., Yan, D., Matter, R. C., Martinez, A. A., and Wong, J. W. (2000). "Intensity modulation to improve dose uniformity with tangential breast radiotherapy: initial clinical experience". In: *International Journal of Radiation Oncology* Biology* Physics* 48.5, pp. 1559–1568.

[49] Dayes, I., Rumble, R., Bowen, J, Dixon, P, Warde, P, Panel, I. I. E., et al. (2012). "Intensity-modulated radiotherapy in the treatment of breast cancer". In: *Clinical Oncology* 24.7, pp. 488–498.

[50] Fogliata, A., Seppälä, J., Reggiori, G., Lobefalo, F., Palumbo, V., De Rose, F., Franceschini, D., Scorsetti, M., and Cozzi, L. (2017b). "Dosimetric tradeoffs in breast treatment with VMAT technique". In: *The British journal of radiology* 90.1070, p. 20160701.

[51] Pignol, J.-P., Olivotto, I., Rakovitch, E., Gardner, S., Sixel, K., Beckham, W., Vu, T. T. T., Truong, P., Ackerman, I., and Paszat, L. (2008). "A multicenter randomized trial of breast intensity-modulated radiation therapy to reduce acute radiation dermatitis". In: *Journal of Clinical Oncology* 26.13, pp. 2085–2092.

[52] Donovan, E., Bleakley, N., Denholm, E., Evans, P., Gothard, L., Hanson, J., Peckitt, C., Reise, S., Ross, G., Sharp, G., et al. (2007). "Randomised trial of standard 2D radiotherapy (RT) versus intensity modulated radiotherapy (IMRT) in patients prescribed breast radiotherapy". In: *Radiotherapy and Oncology* 82.3, pp. 254–264.

[53] Harsolia, A., Kestin, L., Grills, I., Wallace, M., Jolly, S., Jones, C., Lala, M., Martinez, A., Schell, S., and Vicini, F. A. (2007). "Intensity-modulated radiotherapy results in significant decrease in clinical toxicities compared with conventional wedge-based breast radiotherapy". In: *International Journal of Radiation Oncology* Biology* Physics* 68.5, pp. 1375–1380.

[54] Mukesh, M. B., Barnett, G. C., Wilkinson, J. S., Moody, A. M., Wilson, C., Dorling, L., Chan Wah Hak, C., Qian, W., Twyman, N., Burnet, N. G., et al. (2013). "Randomized controlled trial of intensity-modulated radiotherapy for early breast cancer: 5-year results confirm superior overall cosmesis". In: *Journal of clinical oncology* 31.36, pp. 4488–4495.

[55] Group, E. B. C. T. C. et al. (2011). "Effect of radiotherapy after breast-conserving surgery on 10-year recurrence and 15-year breast cancer death: meta-analysis of individual patient data for 10 801 women in 17 randomised trials". In: *The Lancet* 378.9804, pp. 1707–1716.

[56] Grantzau, T., Thomsen, M. S., Væth, M., and Overgaard, J. (2014). "Risk of second primary lung cancer in women after radiotherapy for breast cancer". In: *Radiotherapy and Oncology* 111.3, pp. 366–373.

[57] Mast, M. E., Kempen-Harteveld, L. van, Heijenbrok, M. W., Kalidien, Y., Rozema, H., Jansen, W. P., Petoukhova, A. L., and Struikmans, H. (2013). "Left-sided breast cancer radiotherapy with and without breath-hold: does IMRT reduce the cardiac dose even further?" In: *Radiotherapy and Oncology* 108.2, pp. 248–253.

[58] Zagar, T., Tang, X. I., Jones, E., Matney, J., Das, S., Green, R., Sheikh, A., Khandani, A., McCartney, W., Wong, T., et al. (2015). "Prospective assessment of deep inspiration breath hold to prevent radiation associated cardiac perfusion defects in patients with left-sided breast cancer". In: *International Journal of Radiation Oncology, Biology, Physics* 93.3, E11.

[59] Virén, T., Heikkilä, J., Myllyoja, K., Koskela, K., Lahtinen, T., and Seppälä, J. (2015). "Tangential volumetric modulated arc therapy technique for left-sided breast cancer radiotherapy". In: *Radiation Oncology* 10.1, pp. 1–8.

[60] Zhao, H., He, M., Cheng, G., Han, D., Wu, N., Shi, D., Zhao, Z., and Jin, J. (2015). "A comparative dosimetric study of left sided breast cancer after

breast-conserving surgery treated with VMAT and IMRT”. In: *Radiation Oncology* 10.1, pp. 1–10.

[61] Vikström, J., Hjelstuen, M. H., Wasbø, E., Mjaaland, I., and Dybvik, K. I. (2018). “A comparison of conventional and dynamic radiotherapy planning techniques for early-stage breast cancer utilizing deep inspiration breath-hold”. In: *Acta Oncologica* 57.10, pp. 1325–1330.

[62] Yu, P.-C., Wu, C.-J., Tsai, Y.-L., Shaw, S., Sung, S.-Y., Lui, L. T., and Nien, H.-H. (2018). “Dosimetric analysis of tangent-based volumetric modulated arc therapy with deep inspiration breath-hold technique for left breast cancer patients”. In: *Radiation Oncology* 13.1, pp. 1–10.

[63] Supakalin, N., Pesee, M., Thamronganantasakul, K., Promsensa, K., Supaadirek, C., and Krusun, S. (2018). “Comparision of different radiotherapy planning techniques for breast cancer after breast conserving surgery”. In: *Asian Pacific journal of cancer prevention: APJCP* 19.10, p. 2929.

[64] Redapi, L, Rossi, L, Marrazzo, L, Penninkhof, J., Pallotta, S, and Heijmen, B (2021). “Comparison of volumetric modulated arc therapy and intensity-modulated radiotherapy for left-sided whole-breast irradiation using automated planning”. In: *Strahlentherapie und Onkologie*, pp. 1–11.

[65] Breedveld, S., Storchi, P. R., Voet, P. W., and Heijmen, B. J. (2012). “iCycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar IMRT plans”. In: *Medical physics* 39.2, pp. 951–963.

[66] Breedveld, S., Storchi, P. R., Keijzer, M., Heemink, A. W., and Heijmen, B. J. (2007). “A novel approach to multi-criteria inverse planning for IMRT”. In: *Physics in Medicine & Biology* 52.20, p. 6339.

[67] Romeijn, H. E., Dempsey, J. F., and Li, J. G. (2004). “A unifying framework for multi-criteria fluence map optimization models”. In: *Physics in Medicine & Biology* 49.10, p. 1991.

[68] Alber, M. and Reemtsen, R. (2007). “Intensity modulated radiotherapy treatment planning by use of a barrier-penalty multiplier method”. In: *Optimisation Methods and Software* 22.3, pp. 391–411.

[69] Voet, P. W., Dirkx, M. L., Breedveld, S., Fransen, D., Levendag, P. C., and Heijmen, B. J. (2013). “Toward fully automated multicriterial plan gener-

ation: a prospective clinical study”. In: *International Journal of Radiation Oncology* Biology* Physics* 85.3, pp. 866–872.

[70] Haimes, Y. (1971). “On a bicriterion formulation of the problems of integrated system identification and system optimization”. In: *IEEE transactions on systems, man, and cybernetics* 1.3, pp. 296–297.

[71] Clements, M., Schupp, N., Tattersall, M., Brown, A., and Larson, R. (2018). “Monaco treatment planning system tools and optimization processes”. In: *Medical Dosimetry* 43.2, pp. 106–117.

[72] Voet, P. W., Dirkx, M. L., Breedveld, S., Al-Mamgani, A., Incrocci, L., and Heijmen, B. J. (2014). “Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients”. In: *International Journal of Radiation Oncology* Biology* Physics* 88.5, pp. 1175–1179.

[73] Grégoire, V., Evans, M., Le, Q.-T., Bourhis, J., Budach, V., Chen, A., Eisbruch, A., Feng, M., Giralt, J., Gupta, T., et al. (2018). “Delineation of the primary tumour clinical target volumes (ctv-p) in laryngeal, hypopharyngeal, oropharyngeal and oral cavity squamous cell carcinoma: Airo, caca, dahanca, eortc, georcc, gortec, hknpcsg, hncig, iag-kht, lprhht, ncic ctg, ncri, nrg oncology, phns, sbrt, somera, sro, sshno, trog consensus guidelines”. In: *Radiotherapy and Oncology* 126.1, pp. 3–24.

[74] Dinshaw, K. A., Agarwal, J. P., Laskar, S. G., Gupta, T., Shrivastava, S. K., and Cruz, A. D. (2005). “Head and neck squamous cell carcinoma: The role of post-operative adjuvant radiotherapy”. In: *Journal of surgical oncology* 91.1, pp. 48–55.

[75] Yu, Y. and Lee, N. Y. (2019). “JAVELIN head and neck 100: a phase III trial of avelumab and chemoradiation for locally advanced head and neck cancer”. In: *Future Oncology* 15.7, pp. 687–694.

[76] Gupta, T., Agarwal, J., Jain, S., Phurailatpam, R., Kannan, S., Ghosh-Laskar, S., Murthy, V., Budrukkar, A., Dinshaw, K., Prabhash, K., et al. (2012). “Three-dimensional conformal radiotherapy (3D-CRT) versus intensity modulated radiation therapy (IMRT) in squamous cell carcinoma of the head and neck: a randomized controlled trial”. In: *Radiotherapy and Oncology* 104.3, pp. 343–348.

[77] Sharfo, A. W. M., Breedveld, S., Voet, P. W., Heijkoop, S. T., Mens, J.-W. M., Hoogeman, M. S., and Heijmen, B. J. (2016). “Validation of fully automated VMAT plan generation for library-based plan-of-the-day cervical cancer radiotherapy”. In: *PloS one* 11.12, e0169202.

[78] Della Gala, G., Dirkx, M. L., Hoekstra, N., Fransen, D., Lanconelli, N., Pol, M. van de, Heijmen, B. J., and Petit, S. F. (2017). “Fully automated VMAT treatment planning for advanced-stage NSCLC patients”. In: *Strahlentherapie und Onkologie* 193.5, p. 402.

[79] Buschmann, M., Sharfo, A. W. M., Penninkhof, J., Seppenwoolde, Y., Goldner, G., Georg, D., Breedveld, S., and Heijmen, B. J. (2018). “Automated volumetric modulated arc therapy planning for whole pelvic prostate radiotherapy”. In: *Strahlentherapie und Onkologie* 194.4, pp. 333–342.

[80] Systems, V. M. (2015). *Eclipse Photon and Electron Algorithms Reference Guide*.

[81] Boutilier, J. J., Craig, T., Sharpe, M. B., and Chan, T. C. (2016). “Sample size requirements for knowledge-based treatment planning”. In: *Medical physics* 43.3, pp. 1212–1221.

[82] Cagni, E., Botti, A., Wang, Y., Iori, M., Petit, S. F., and Heijmen, B. J. (2018). “Pareto-optimal plans as ground truth for validation of a commercial system for knowledge-based DVH-prediction”. In: *Physica Medica* 55, pp. 98–106.

[83] Alpuche Aviles, J. E., Cordero Marcos, M. I., Sasaki, D., Sutherland, K., Kane, B., and Kuusela, E. (2018). “Creation of knowledge-based planning models intended for large scale distribution: Minimizing the effect of outlier plans”. In: *Journal of applied clinical medical physics* 19.3, pp. 215–226.

[84] Castriconi, R, Fiorino, C, Broggi, S, Cozzarini, C, Di Muzio, N, Calandrino, R, and Cattaneo, G. (2019). “Comprehensive Intra-Institution stepping validation of knowledge-based models for automatic plan optimization”. In: *Physica Medica* 57, pp. 231–237.

[85] Delaney, A. R., Tol, J. P., Dahele, M., Cuijpers, J., Slotman, B. J., and Verbakel, W. F. (2016). “Effect of dosimetric outliers on the performance of a commercial knowledge-based planning solution”. In: *International Journal of Radiation Oncology* Biology* Physics* 94.3, pp. 469–477.

[86] Yuan, L., Ge, Y., Lee, W. R., Yin, F. F., Kirkpatrick, J. P., and Wu, Q. J. (2012). "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in IMRT plans". In: *Medical physics* 39.11, pp. 6868–6878.

[87] Nwankwo, O., Mekdash, H., Sihono, D. S. K., Wenz, F., and Glatting, G. (2015). "Knowledge-based radiation therapy (KBRT) treatment planning versus planning by experts: validation of a KBRT algorithm for prostate cancer treatment planning". In: *Radiation oncology* 10.1, pp. 1–5.

[88] Fusella, M., Scaggion, A., Pivato, N., Rossato, M. A., Zorz, A., and Paiusco, M. (2018). "Efficiently train and validate a RapidPlan model through APQM scoring". In: *Medical physics* 45.6, pp. 2611–2619.

[89] Tamura, M., Monzen, H., Matsumoto, K., Kubo, K., Otsuka, M., Inada, M., Ishikawa, K., Nakamatsu, K., Sumida, I., Mizuno, H., et al. (2018). "Mechanical performance of a commercial knowledge-based VMAT planning for prostate cancer". In: *Radiation Oncology* 13.1, pp. 1–7.

[90] Scaggion, A., Fusella, M., Roggio, A., Bacco, S., Pivato, N., Rossato, M. A., Peña, L. M. A., and Paiusco, M. (2018). "Reducing inter-and intra-planner variability in radiotherapy plan output with a commercial knowledge-based planning solution". In: *Physica Medica* 53, pp. 86–93.

[91] Chatterjee, A., Serban, M., Abdulkarim, B., Panet-Raymond, V., Souhami, L., Shenouda, G., Sabri, S., Jean-Claude, B., and Seuntjens, J. (2017). "Performance of knowledge-based radiation therapy planning for the glioblastoma disease site". In: *International Journal of Radiation Oncology* Biology* Physics* 99.4, pp. 1021–1028.

[92] Chang, A. T., Hung, A. W., Cheung, F. W., Lee, M. C., Chan, O. S., Philips, H., Cheng, Y.-T., and Ng, W.-T. (2016). "Comparison of planning quality and efficiency between conventional and knowledge-based algorithms in nasopharyngeal cancer patients using intensity modulated radiation therapy". In: *International Journal of Radiation Oncology* Biology* Physics* 95.3, pp. 981–990.

[93] Wu, H., Jiang, F., Yue, H., Li, S., and Zhang, Y. (2016). "A dosimetric evaluation of knowledge-based VMAT planning with simultaneous integrated

boosting for rectal cancer patients". In: *Journal of applied clinical medical physics* 17.6, pp. 78–85.

[94] Fried, D. V., Chera, B. S., and Das, S. K. (2017). "Assessment of Plan IQ Feasibility DVH for head and neck treatment planning". In: *Journal of applied clinical medical physics* 18.5, pp. 245–250.

[95] Cagni, E., Botti, A., Rossi, L., Iotti, C., Iori, M., Cozzi, S., Galaverni, M., Rosca, A., Sghedoni, R., Timon, G., et al. (2021b). "Variations in head and neck treatment plan quality assessment among radiation oncologists and medical physicists in a single radiotherapy department". In: *Frontiers in oncology*, p. 3714.

[96] Nutting, C. M., Morden, J. P., Harrington, K. J., Urbano, T. G., Bhide, S. A., Clark, C., Miles, E. A., Miah, A. B., Newbold, K., Tanay, M., et al. (2011). "Parotid-sparing intensity modulated versus conventional radiotherapy in head and neck cancer (PARSPORT): a phase 3 multicentre randomised controlled trial". In: *The lancet oncology* 12.2, pp. 127–136.

[97] Chun, S. G., Hu, C., Choy, H., Komaki, R. U., Timmerman, R. D., Schild, S. E., Bogart, J. A., Dobelbower, M. C., Bosch, W., Galvin, J. M., et al. (2017). "Impact of intensity-modulated radiation therapy technique for locally advanced non–small-cell lung cancer: a secondary analysis of the NRG oncology RTOG 0617 randomized clinical trial". In: *Journal of Clinical Oncology* 35.1, p. 56.

[98] Viani, G. A., Viana, B. S., Martin, J. E. C., Rossi, B. T., Zuliani, G., and Stefano, E. J. (2016). "Intensity-modulated radiotherapy reduces toxicity with similar biochemical control compared with 3-dimensional conformal radiotherapy for prostate cancer: A randomized clinical trial". In: *Cancer* 122.13, pp. 2004–2011.

[99] Staffurth, J et al. (2010). "A review of the clinical evidence for intensity-modulated radiotherapy". In: *Clinical oncology* 22.8, pp. 643–657.

[100] Berry, S. L., Boczkowski, A., Ma, R., Mechalakos, J., and Hunt, M. (2016). "Interobserver variability in radiation therapy plan output: results of a single-institution study". In: *Practical radiation oncology* 6.6, pp. 442–449.

[101] Hansen, C. R., Nielsen, M., Bertelsen, A. S., Hazell, I., Holtved, E., Zukauskaite, R., Bjerregaard, J. K., Brink, C., and Bernchou, U. (2017). "Automatic treat-

ment planning facilitates fast generation of high-quality treatment plans for esophageal cancer”. In: *Acta Oncologica* 56.11, pp. 1495–1500.

[102] Rossi, L., Sharfo, A. W., Aluwini, S., Dirkx, M., Breedveld, S., and Heijmen, B. (2018). “First fully automated planning solution for robotic radiosurgery–comparison with automatically planned volumetric arc therapy for prostate cancer”. In: *Acta Oncologica* 57.11, pp. 1490–1498.

[103] Heijmen, B., Voet, P., Fransen, D., Penninkhof, J., Milder, M., Akhiat, H., Bonomo, P., Casati, M., Georg, D., Goldner, G., et al. (2018). “Fully automated, multi-criterial planning for Volumetric Modulated Arc Therapy–An international multi-center validation for prostate cancer”. In: *Radiotherapy and Oncology* 128.2, pp. 343–348.

[104] Marrazzo, L., Meattini, I., Arilli, C., Calusi, S., Casati, M., Talamonti, C., Livi, L., and Pallotta, S. (2019). “Auto-planning for VMAT accelerated partial breast irradiation”. In: *Radiotherapy and Oncology* 132, pp. 85–92.

[105] Edge, S. B. and Compton, C. C. (2010). “The American Joint Committee on Cancer: the 7th edition of the AJCC cancer staging manual and the future of TNM”. In: *Annals of surgical oncology* 17.6, pp. 1471–1474.

[106] Taherdoost, H. (2019). “What is the best response scale for survey and questionnaire design; review of different lengths of rating scale/attitude scale/Likert scale”. In: *Hamed Taherdoost*, pp. 1–10.

[107] Colman, A. M., Norris, C. E., and Preston, C. C. (1997). “Comparing rating scales of different lengths: Equivalence of scores from 5-point and 7-point scales”. In: *Psychological Reports* 80.2, pp. 355–362.

[108] Preston, C. C. and Colman, A. M. (2000). “Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences”. In: *Acta psychologica* 104.1, pp. 1–15.

[109] Cohen, J. (1960). “A coefficient of agreement for nominal scales”. In: *Educational and psychological measurement* 20.1, pp. 37–46.

[110] Landis, J. R. and Koch, G. G. (1977). “The measurement of observer agreement for categorical data”. In: *biometrics*, pp. 159–174.

[111] Watson, P. and Petrie, A (2010). “Method agreement analysis: a review of correct methodology”. In: *Theriogenology* 73.9, pp. 1167–1179.

[112] Wang, Y., Heijmen, B. J., and Petit, S. F. (2017). "Prospective clinical validation of independent DVH prediction for plan QA in automatic treatment planning for prostate cancer patients". In: *Radiotherapy and Oncology* 125.3, pp. 500–506.

[113] El Naqa, I. and Murphy, M. J. (2015). "What is machine learning?" In: *machine learning in radiation oncology*. Springer, pp. 3–11.

[114] Sidey-Gibbons, J. A. and Sidey-Gibbons, C. J. (2019). "Machine learning in medicine: a practical introduction". In: *BMC medical research methodology* 19.1, pp. 1–18.

[115] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., and Brown, S. D. (2004). "An introduction to decision tree modeling". In: *Journal of Chemometrics: A Journal of the Chemometrics Society* 18.6, pp. 275–285.

[116] Hamza, M. and Larocque, D. (2005). "An empirical comparison of ensemble methods based on classification trees". In: *Journal of Statistical Computation and Simulation* 75.8, pp. 629–643.

[117] Tu, M. C., Shin, D., and Shin, D. (2009). "A comparative study of medical data classification methods based on decision tree and bagging algorithms". In: *2009 Eighth IEEE International Conference on Dependable, Autonomic and Secure Computing*. IEEE, pp. 183–187.

[118] Bradley, A. P. (1997). "The use of the area under the ROC curve in the evaluation of machine learning algorithms". In: *Pattern Recognition* 30.7, pp. 1145–1159. ISSN: 0031-3203.

[119] Organization, W. H. et al. (2020). "WHO report on cancer: setting priorities, investing wisely and providing care for all". In.

[120] Chui, C.-S., Hong, L., Hunt, M., and McCormick, B. (2002). "A simplified intensity modulated radiation therapy technique for the breast". In: *Medical physics* 29.4, pp. 522–529.

[121] Smith, W., Menon, G., Wolfe, N., Ploquin, N., Trotter, T., and Pudney, D. (2010). "IMRT for the breast: a comparison of tangential planning techniques". In: *Physics in Medicine & Biology* 55.4, p. 1231.

[122] Donovan, E., Yarnold, J., Adams, E., Morgan, A, Warrington, A., and Evans, P. (2008). "An investigation into methods of IMRT planning applied to

breast radiotherapy". In: *The British journal of radiology* 81.964, pp. 311–322.

[123] Johansen, S., Cozzi, L., and Olsen, D. R. (2009). "A planning comparison of dose patterns in organs at risk and predicted risk for radiation induced malignancy in the contralateral breast following radiation therapy of primary breast using conventional, IMRT and volumetric modulated arc treatment techniques". In: *Acta Oncologica* 48.4, pp. 495–503.

[124] Nichols, G. P., Fontenot, J. D., Gibbons, J. P., and Sanders, M. E. (2014). "Evaluation of volumetric modulated arc therapy for postmastectomy treatment". In: *Radiation Oncology* 9.1, pp. 1–8.

[125] Johansen, H., Kaae, S., Johansen, H., Kaae, S., Jensen, M.-B., and Mouridsen, H. T. (2008). "Extended radical mastectomy versus simple mastectomy followed by radiotherapy in primary breast cancer. A fifty-year follow-up to the Copenhagen Breast Cancer randomised study". In: *Acta Oncologica* 47.4, pp. 633–638.

[126] Darby, S. C., Ewertz, M., McGale, P., Bennet, A. M., Blom-Goldman, U., Brønnum, D., Correa, C., Cutter, D., Gagliardi, G., Gigante, B., et al. (2013). "Risk of ischemic heart disease in women after radiotherapy for breast cancer". In: *New England Journal of Medicine* 368.11, pp. 987–998.

[127] Penninkhof, J., Spadola, S., Breedveld, S., Baaijens, M., Lanconelli, N., and Heijmen, B. (2017). "Individualized selection of beam angles and treatment isocenter in tangential breast intensity modulated radiation therapy". In: *International Journal of Radiation Oncology* Biology* Physics* 98.2, pp. 447–453.

[128] Zhao, X., Kong, D., Jozsef, G., Chang, J., Wong, E. K., Formenti, S. C., and Wang, Y. (2012). "Automated beam placement for breast radiotherapy using a support vector machine based algorithm". In: *Medical physics* 39.5, pp. 2536–2543.

[129] Wang, W., Purdie, T. G., Rahman, M., Marshall, A., Liu, F.-F., and Fyles, A. (2012). "Rapid automated treatment planning process to select breast cancer patients for active breathing control to achieve cardiac dose reduction". In: *International Journal of Radiation Oncology* Biology* Physics* 82.1, pp. 386–393.

[130] Purdie, T. G., Dinniwell, R. E., Fyles, A., and Sharpe, M. B. (2014). "Automation and intensity modulated radiation therapy for individualized high-quality tangent breast treatment plans". In: *International Journal of Radiation Oncology* Biology* Physics* 90.3, pp. 688–695.

[131] Chen, G.-P., Liu, F., White, J., Vicini, F. A., Freedman, G. M., Arthur, D. W., and Li, X. A. (2015). "A planning comparison of 7 irradiation options allowed in RTOG 1005 for early-stage breast cancer". In: *Medical Dosimetry* 40.1, pp. 21–25.

[132] Lee, B. M., Chang, J. S., Kim, S. Y., Keum, K. C., Suh, C.-O., and Kim, Y. B. (2020). "Hypofractionated radiotherapy dose scheme and application of new techniques are associated to a lower incidence of radiation pneumonitis in breast cancer patients". In: *Frontiers in oncology* 10, p. 124.

[133] Trialists' Group, T. S. (2008). "The UK Standardisation of Breast Radiotherapy (START) Trial B of radiotherapy hypofractionation for treatment of early breast cancer: a randomised trial". In: *The Lancet* 371.9618, pp. 1098–1107.

[134] Cagni, E., Botti, A., Orlandi, M., Galaverni, M., Iotti, C., Iori, M., Lewis, G., and Spezi, E. (2022). "Evaluating the quality of patient-specific deformable image registration in adaptive radiotherapy using a digitally enhanced head and neck phantom". In: *Applied Sciences* 12.19, p. 9493.

[135] Keall, P., Poulsen, P., and Booth, J. T. (2019). "See, think, and act: real-time adaptive radiotherapy". In: *Seminars in radiation oncology*. Vol. 29. 3. Elsevier, pp. 228–235.

[136] Jaffray, D. A. (2012). "Image-guided radiotherapy: from current concept to future perspectives". In: *Nature reviews Clinical oncology* 9.12, pp. 688–699.

[137] Langen, K. M. and Jones, D. T. (2001). "Organ motion and its management". In: *International Journal of Radiation Oncology* Biology* Physics* 50.1, pp. 265–278.

[138] Liu, C., Kim, J., Kumarasiri, A., Mayyas, E., Brown, S. L., Wen, N., Siddiqui, F., and Chetty, I. J. (2018). "An automated dose tracking system for adaptive radiation therapy". In: *Computer methods and programs in biomedicine* 154, pp. 1–8.

[139] Heukelom, J. and Fuller, C. D. (2019). “Head and neck cancer adaptive radiation therapy (ART): conceptual considerations for the informed clinician”. In: *Seminars in radiation oncology*. Vol. 29. 3. Elsevier, pp. 258–273.

[140] Brock, K. K., Mutic, S., McNutt, T. R., Li, H., and Kessler, M. L. (2017). “Use of image registration and fusion algorithms and techniques in radiotherapy: Report of the AAPM Radiation Therapy Committee Task Group No. 132”. In: *Medical physics* 44.7, e43–e76.

[141] Rong, Y., Rosu-Bubulac, M., Benedict, S. H., Cui, Y., Ruo, R., Connell, T., Kashani, R., Latifi, K., Chen, Q., Geng, H., et al. (2021). “Rigid and Deformable Image Registration for Radiation Therapy: A Self-Study Evaluation Guide for NRG Oncology Clinical Trial Participation”. In: *Practical Radiation Oncology*.

[142] Latifi, K., Caudell, J., Zhang, G., Hunt, D., Moros, E. G., and Feygelman, V. (2018). “Practical quantification of image registration accuracy following the AAPM TG-132 report framework”. In: *Journal of applied clinical medical physics* 19.4, pp. 125–133.

[143] Pukala, J., Johnson, P. B., Shah, A. P., Langen, K. M., Bova, F. J., Staton, R. J., Mañon, R. R., Kelly, P., and Meeks, S. L. (2016). “Benchmarking of five commercial deformable image registration algorithms for head and neck patients”. In: *Journal of applied clinical medical physics* 17.3, pp. 25–40.

[144] Feldkamp, L. A., Davis, L. C., and Kress, J. W. (1984). “Practical cone-beam algorithm”. In: *Josa a* 1.6, pp. 612–619.

[145] Hatton, J., McCurdy, B., and Greer, P. B. (2009). “Cone beam computerized tomography: the effect of calibration of the Hounsfield unit number to electron density on dose calculation accuracy for adaptive radiation therapy”. In: *Physics in Medicine & Biology* 54.15, N329.

[146] Schulze, R, Heil, U, Groβ, D, Bruellmann, D., Dranischnikow, E, Schwanecke, U., and Schoemer, E (2011). “Artefacts in CBCT: a review”. In: *Dentomaxillofacial Radiology* 40.5, pp. 265–273.

[147] Stock, M., Pasler, M., Birkfellner, W., Homolka, P., Poetter, R., and Georg, D. (2009). “Image quality and stability of image-guided radiotherapy (IGRT)

devices: A comparative study". In: *Radiotherapy and Oncology* 93.1, pp. 1–7.

[148] Park, S., Plishker, W., Quon, H., Wong, J., Shekhar, R., and Lee, J. (2017). "Deformable registration of CT and cone-beam CT with local intensity matching". In: *Physics in Medicine & Biology* 62.3, p. 927.

[149] Zhen, X., Yan, H., Zhou, L., Jia, X., and Jiang, S. B. (2013). "Deformable image registration of CT and truncated cone-beam CT for adaptive radiation therapy". In: *Physics in Medicine & Biology* 58.22, p. 7979.

[150] Nithiananthan, S., Schafer, S., Uneri, A., Mirota, D. J., Stayman, J. W., Zbijewski, W., Brock, K. K., Daly, M. J., Chan, H., Irish, J. C., et al. (2011). "Demons deformable registration of CT and cone-beam CT using an iterative intensity matching approach". In: *Medical physics* 38.4, pp. 1785–1798.

[151] Brouwer, C. L., Steenbakkers, R. J., Bourhis, J., Budach, W., Grau, C., Grégoire, V., Van Herk, M., Lee, A., Maingon, P., Nutting, C., et al. (2015a). "CT-based delineation of organs at risk in the head and neck region: DAHANCA, EORTC, GORTEC, HKNPCSG, NCIC CTG, NCRI, NRG Oncology and TROG consensus guidelines". In: *Radiotherapy and Oncology* 117.1, pp. 83–90.

[152] Deasy, J. O., Blanco, A. I., and Clark, V. H. (2003). "CERR: a computational environment for radiotherapy research". In: *Medical physics* 30.5, pp. 979–985.

[153] Klein, S., Staring, M., Murphy, K., Viergever, M. A., and Pluim, J. P. (2009). "Elastix: a toolbox for intensity-based medical image registration". In: *IEEE transactions on medical imaging* 29.1, pp. 196–205.

[154] Broggi, S., Scalco, E., Belli, M. L., Logghe, G., Verellen, D., Moriconi, S., Chiara, A., Palmisano, A., Mellone, R., Fiorino, C., et al. (2017). "A comparative evaluation of 3 different free-form deformable image registration and contour propagation methods for head and neck MRI: the case of parotid changes during radiotherapy". In: *Technology in cancer research & treatment* 16.3, pp. 373–381.

[155] Brouwer, C. L., Kierkels, R. G., Veld, A. A. van't, Sijtsema, N. M., and Meertens, H. (2014). "The effects of computed tomography image charac-

teristics and knot spacing on the spatial accuracy of B-spline deformable image registration in the head and neck geometry". In: *Radiation Oncology* 9.1, pp. 1–10.

[156] Qin, A., Liang, J., Han, X., O'Connell, N., and Yan, D. (2018). "The impact of deformable image registration methods on dose warping". In: *Medical physics* 45.3, pp. 1287–1294.

[157] Singhrao, K., Kirby, N., and Pouliot, J. (2014). "A three-dimensional head-and-neck phantom for validation of multimodality deformable image registration for adaptive radiotherapy". In: *Medical physics* 41.12, p. 121709.

[158] Murphy, M. J., Salguero, F. J., Siebers, J. V., Staub, D., and Vaman, C. (2012). "A method to estimate the effect of deformable image registration uncertainties on daily dose mapping". In: *Medical physics* 39.2, pp. 573–580.

[159] Veiga, C., Lourenço, A. M., Mouinuddin, S., Van Herk, M., Modat, M., Ourselin, S., Royle, G., and McClelland, J. R. (2015). "Toward adaptive radiotherapy for head and neck patients: uncertainties in dose warping due to the choice of deformable registration algorithm". In: *Medical Physics* 42.2, pp. 760–769.

[160] Low, D. A., Harms, W. B., Mutic, S., and Purdy, J. A. (1998). "A technique for the quantitative evaluation of dose distributions". In: *Medical physics* 25.5, pp. 656–661.

[161] Asuero, A. G., Sayago, A., and Gonzalez, A. (2006). "The correlation coefficient: An overview". In: *Critical reviews in analytical chemistry* 36.1, pp. 41–59.

[162] Schober, P., Boer, C., and Schwarte, L. A. (2018). "Correlation coefficients: appropriate use and interpretation". In: *Anesthesia & Analgesia* 126.5, pp. 1763–1768.

[163] Moteabbed, M, Sharp, G., Wang, Y., Trofimov, A, Efstathiou, J. A., and Lu, H.-M. (2015). "Validation of a deformable image registration technique for cone beam CT-based dose verification". In: *Medical physics* 42.1, pp. 196–205.

[164] Shi, L., Chen, Q., Barley, S., Cui, Y., Shang, L., Qiu, J., and Rong, Y. (2021). "Benchmarking of deformable image registration for multiple anatomic sites

using digital data sets with ground-truth deformation vector fields”. In: *Practical Radiation Oncology* 11.5, pp. 404–414.

[165] Lowther, N. J., Marsh, S. H., and Louwe, R. J. (2020). “Quantifying the dose accumulation uncertainty after deformable image registration in head-and-neck radiotherapy”. In: *Radiotherapy and Oncology* 143, pp. 117–125.

[166] Green, O. L., Henke, L. E., and Hugo, G. D. (2019). “Practical clinical workflows for online and offline adaptive radiation therapy”. In: *Seminars in radiation oncology*. Vol. 29. 3. Elsevier, pp. 219–227.

[167] Yan, D., Vicini, F., Wong, J., and Martinez, A. (1997). “Adaptive radiation therapy”. In: *Physics in Medicine & Biology* 42.1, p. 123.

[168] Ahunbay, E. E., Peng, C., Godley, A., Schultz, C., and Li, X. A. (2009). “An on-line replanning method for head and neck adaptive radiotherapy a”. In: *Medical physics* 36.10, pp. 4776–4790.

[169] Gensheimer, M. F. and Le, Q.-T. (2018). “Adaptive radiotherapy for head and neck cancer: are we ready to put it into routine clinical practice?” In: *Oral oncology* 86, pp. 19–24.

[170] Guidi, G, Maffei, N, Meduri, B, D’Angelo, E, Mistretta, G., Ceroni, P, Ciarmatori, A, Bernabei, A, Maggi, S, Cardinali, M, et al. (2016). “A machine learning tool for re-planning and adaptive RT: a multicenter cohort investigation”. In: *Physica Medica* 32.12, pp. 1659–1666.

[171] Zhang, P., Simon, A., Rigaud, B., Castelli, J., Arango, J. D. O., Nassef, M., Henry, O., Zhu, J., Haigron, P., Li, B., et al. (2016). “Optimal adaptive IMRT strategy to spare the parotid glands in oropharyngeal cancer”. In: *Radiotherapy and Oncology* 120.1, pp. 41–47.

[172] Stoll, M., Giske, K., Debus, J., Bendl, R., and Stoiber, E. M. (2014). “The frequency of re-planning and its variability dependent on the modification of the re-planning criteria and IGRT correction strategy in head and neck IMRT”. In: *Radiation Oncology* 9.1, pp. 1–8.

[173] Huang, H., Lu, H., Feng, G., Jiang, H., Chen, J., Cheng, J., Pang, Q., Lu, Z., Gu, J., Peng, L., et al. (2015). “Determining appropriate timing of adaptive radiation therapy for nasopharyngeal carcinoma during intensity-modulated radiation therapy”. In: *Radiation Oncology* 10.1, pp. 1–9.

[174] Ahn, P. H., Chen, C.-C., Ahn, A. I., Hong, L., Scripes, P. G., Shen, J., Lee, C.-C., Miller, E., Kalnicki, S., and Garg, M. K. (2011). “Adaptive planning in intensity-modulated radiation therapy for head and neck cancers: single-institution experience and clinical implications”. In: *International Journal of Radiation Oncology* Biology* Physics* 80.3, pp. 677–685.

[175] Lee, C., Langen, K. M., Lu, W., Haimerl, J., Schnarr, E., Ruchala, K. J., Olivera, G. H., Meeks, S. L., Kupelian, P. A., Shellenberger, T. D., et al. (2008). “Assessment of parotid gland dose changes during head and neck cancer radiotherapy using daily megavoltage computed tomography and deformable image registration”. In: *International Journal of Radiation Oncology* Biology* Physics* 71.5, pp. 1563–1571.

[176] Stoiber, E. M., Schwarz, M., Huber, P. E., Debus, J., Bendl, R., and Giske, K. (2013). “Comparison of two IGRT correction strategies in postoperative head-and-neck IMRT patients”. In: *Acta Oncologica* 52.1, pp. 183–186.

[177] Graff, P., Kirby, N., Weinberg, V., Chen, J., Yom, S. S., Lambert, L., and Pouliot, J. (2013). “The residual setup errors of different IGRT alignment procedures for head and neck IMRT and the resulting dosimetric impact”. In: *International Journal of Radiation Oncology* Biology* Physics* 86.1, pp. 170–176.

[178] Fogliata, A, Cozzi, L, Reggiori, G, Stravato, A, Lobefalo, F, Franzese, C, Franceschini, D, Tomatis, S, and Scorsetti, M (2019). “RapidPlan knowledge based planning: iterative learning process and model ability to steer planning strategies”. In: *Radiation Oncology* 14.1, pp. 1–12.

[179] Miguel-Chumacero, E., Currie, G., Johnston, A., and Currie, S. (2018). “Effectiveness of Multi-Criteria Optimization-based Trade-Off exploration in combination with RapidPlan for head & neck radiotherapy planning”. In: *Radiation oncology* 13.1, pp. 1–13.

[180] Castriconi, R., Fiorino, C., Passoni, P., Broggi, S., Di Muzio, N. G., Cattaneo, G. M., and Calandrino, R. (2020). “Knowledge-based automatic optimization of adaptive early-regression-guided VMAT for rectal cancer”. In: *Physica Medica* 70, pp. 58–64.

[181] Appenzoller, L. M., Michalski, J. M., Thorstad, W. L., Mutic, S., and Moore, K. L. (2012). “Predicting dose-volume histograms for organs-at-risk in IMRT planning”. In: *Medical physics* 39.12, pp. 7446–7461.

[182] Bentzen, S. M., Constine, L. S., Deasy, J. O., Eisbruch, A., Jackson, A., Marks, L. B., Ten Haken, R. K., and Yorke, E. D. (2010). “Quantitative Analyses of Normal Tissue Effects in the Clinic (QUANTEC): an introduction to the scientific issues”. In: *International Journal of Radiation Oncology* Biology* Physics* 76.3, S3–S9.

[183] Beltran, M., Ramos, M., Rovira, J. J., Perez-Hoyos, S., Sancho, M., Puertas, E., Benavente, S., Ginjaume, M., and Giralt, J. (2012). “Dose variations in tumor volumes and organs at risk during IMRT for head-and-neck cancer”. In: *Journal of Applied Clinical Medical Physics* 13.6, pp. 101–111.

[184] Bowen, S. R., Flynn, R. T., Bentzen, S. M., and Jeraj, R. (2009). “On the sensitivity of IMRT dose optimization to the mathematical form of a biological imaging-based prescription function”. In: *Physics in Medicine & Biology* 54.6, p. 1483.

[185] Guida, S. (2021). “Approvate da EMA le raccomandazioni dell’ICMRA sulla regolamentazione dell’Intelligenza Artificiale in medicina-EMA approved the ICMRA recommendations on regulation of artificial intelligence in medicine”. In: *European Journal of Privacy Law & Technologies* 2021.1.