Machine Learning Applications for Personalised Automated Radiotherapy Planning

by

Iona Foster

A thesis submitted to Cardiff University for the degree of Doctor of Philosophy

December 2022

CARDIFF
UNIVERSITY
PRIFYSGOL
CAERDYDD

School of Engineering
Ysgol Peirianneg

Dedicated to

Albertina “Aunty Alma” Byron

Thesis Abstract

Automated radiotherapy planning is characterised by reduction in manual planning due to an increase in computerised planning. Current methods can produce plans suitable for clinical use. However, every case is unique and manual intervention is often needed. The goal of this work was to determine whether it is feasible to develop a fully automated planning system producing clinically optimal plans, and if so, to begin developing it. This work explored relationships between automated planning parameters and anatomical features with respect to dosimetric outcomes. A rules-based automated planning technique was used, an algorithm requiring calibration of input parameters prior to use. This calibration determines the target objectives the algorithm will optimise to. Existing calibration methods use a single set of calibrated parameters per treatment site and are applied to all patients. This approach is considered sufficient to meet clinical goals but may not be sufficient for development of optimal personalised planning due to anatomical variance between patients. Using a validated rules-based planning methodology and obtaining patient bespoke expert-driven calibrated parameters as the optimal gold standard and validation benchmark, two machine learning techniques were explored for apriori configuration of parameters for the delivery of personalised treatment planning. The main objective was to train models to predict gold standard parameters hence generating expert planning automatically. A secondary objective was to determine dosimetric differences between plans generated via machine learned parameters and a traditional single set of parameters applied to all cases. Preliminary studies were carried out to define what will be considered gold standard and to identify anatomical features for inclusion in the main study as well as their relationships to calibrated parameters. The research presented here was applied to three sites: prostate, rectum and lung. Findings are also expected to provide heuristics for research to be carried out on other treatment sites.

Acknowledgements

I would like to thank my primary Cardiff university supervisor Prof Emiliano Spezi. Emiliano, I have valued your guidance and direction. I often diverged deep in the mathematics of things and you have helped me to put together the pieces of work into a coherent structure and provided me with the lens through which to view the project. Thank you for helping me to see the bigger picture and to develop a project valuable to community.

I would also like to thank my secondary supervisor Prof John Staffurth. John, I never had a meeting with you where I did not leave with a new perspective. You always challenged me to answer the important questions and have given me ways of thinking that I will carry with me in all my work going forward.

But I would most of all like to show my a heartfelt appreciation to my Velindre Cancer Centre supervisory Dr Philip Wheeler. Your investment of time and interest in this work has been greatly appreciated. I have learned more from you in these few years than people I have known much longer. You've provided advice, knowledge, insights and new skills. But most of all you have been a pleasure to work with and a valued friend.

Of course a big thank you to my all of the people in my life. You helped me through the daily toil and loneliness and helped me see how blessed I am to have so many people in my life. Special thank you to Leti, Daniel, Alex and Martin for laughs. Taqmina, Abbie, Rhianna, Ibbi, Lottie, Natalie and Gemma for the continued sisterhood even with all the miles between us. I would also like to thank my family especially my parents Derrick and Jacqueline, step parents Neville and Zaarah and my brothers and sisters Marvin, Catherine, Daniel and Grace for your continued support. Without it, I will have struggled a lot more especially during lockdown. Also, a very special thank you to Dan. You joined me on this journey toward the end and quickly became my rock. I know you have been on a *rollercoaster* too so thank you for always listening, caring, helping and loving. You have been more valuable to me than I could have imagined. I only hope you know how much.

Funding

This work was supported by Velindre NHS Trust Charitable Funds

Publications and Output

Publications

I. D. Foster, E. Spezi, P. A. Wheeler, "Evaluating the Use of Machine Learning to Predict Expert-Driven Pareto-Navigated Calibrations for Personalised Automated Radiotherapy Planning", *Applied Sciences*, 2023

Posters

I. D. Foster, P. A. Wheeler, E. Spezi, J. Staffurth, A. Millin, "PD-0822 Bespoke vs machine learned: can expert Pareto navigated treatment planning be modelled?", *Radiotherapy and Oncology*, vol. 161, pp. S655–S656, 2021

I. D. Foster, P. A. Wheeler, E. Spezi, J. Staffurth, A. Millin, "PO-1504: Pareto navigation guided automated planning: is a single patient enough to calibrate a solution?", *Radiotherapy and Oncology*, vol. 161, pp. S811–S812, 2020

Oral presentations

I. D. Foster, E. Spezi, P. A. Wheeler, "Inter-Planner Variability in Expert-Driven Pareto-Guided Automated Planning Solutions", *Cardiff University School of Engineering Conference*, Cardiff, UK, 2023

I. D. Foster "Radiotherapy planning: past, present and future", *Cardiff University School of Engineering Gregynog Conference*, Gregynog, UK, 2019

I. D. Foster "Radiotherapy planning: what is it and why is it important?", *Cardiff University Society for Women Graduates*, Cardiff, UK, 2019

Contributions

Chapters include published material listed in the previous section. This content and the contributions of this author correspond to chapters in the following way:

- sections of chapter 3 contain material presented as part of the *Cardiff University School of Engineering Gregynog Conference* and *Cardiff University Society for Women Graduates* titled “Radiotherapy planning: past, present and future” and “Radiotherapy planning: what is it and why is it important?” respectively. This work included the authors critical appraisal of existing automated planning techniques and identification of literature gaps in which further research could contribute to the body of knowledge. Conducting a literature review, the key weaknesses of current standard manual planning were identified, the main automated planning approaches in the literature were highlighted and areas of development suggested and outlined.

- chapters 3 and 4 contain work published in *Radiotherapy and Oncology* titled “Pareto navigation guided automated planning: is a single patient enough to calibrate a solution?” Research into the strength of protocol-based automatic iterative optimisers as automated planning approaches was addressed regarding classic calibration methods. This author quantitatively compared numerical calibration parameters of an automated planning system and dosimetric characteristics for resulting plans. This author anonymised a cohort of patients and generated two plans for each patient using the RayStation treatment planning system via an existing automated planning system. Plans generated were based on: (1) parameters tailored only to one of the patients in the cohort, (2) parameters tailored to each individual patient in the cohort. This author assessed dosimetric differences between the planning methods statistically.

- a section of chapter 4 is based on material presented at the *Cardiff University*

School of Engineering Conference titled “Inter-Planner Variability in Expert-Driven Pareto-Guided Automated Planning Solutions”. This author anonymised a new cohort of patients, generated plans using the Raystation treatment planning system and recruited four participants for the study. This author also, prepared the test conditions and notes each participants would use as well as analysing the results of the study.

- chapters 5 and 6 contain developments of the work published in *Radiotherapy and Oncology* titled “Bespoke vs machine learned: can expert Pareto navigated treatment planning be modelled?” Personal contributions of this author included development of machine learned parameters for automated planning configuration. The numerical parameters were machine learned via regression built using code written in python and plans were generated in the RayStation treatment planning system via an existing automated planning system. Resulting plans were then dosimetrically assessed by this author

- chapter 6 contains a development of the work published in *Applied Sciences* titled “Evaluating the use of machine learning to predict expert-driven Pareto-navigated calibrations for personalised automated radiotherapy planning.” Python scripts and planning comparison were completed by this author. Assessments included differences in numerical automated planning parameters and dosimetric differences for a prostate seminal vesicles cohort of patients.

Contents

List of Figures

List of Tables

Nomenclature and acronyms

Nomenclature

Automated planning	Automated radiotherapy planning. A class of planning methods that minimise human manual planning
Bremsstrahlung	Radiation emitted by electron liberated from an atom
Clustering	Automatic clustering. A class of unsupervised machine learning methods characterised by grouping data points such that points within a cluster have greater similarities to points outside of the cluster
Conformality index	A value between zero and one indicating the level to which the planning dose conforms to the prescribed dose
Cross validation	A resampling method to train and validate models using a single dataset
Degree	Degree of a polynomial model e.g., a quadratic model has two degrees
Delineation	Contour of the outline of a regions-of-interest
Feature	Anatomical variable that defines a geometric characteristic relating to regions-of-interest. May be used in raw form or as Principal Components
Feature set	Set of Features. May be a subset of FeatureDS2 or a subset of Principal Components of FeatureDS1
FeatureDS1	Database of all raw Features used for generating Principal Components. Contains no variables with missing data or low variance

FeatureDS2	A subset of FeatureDS1. No pair of Features has a correlation coefficient greater than 0.85
Homogeneity index	A value between zero and one indicating a level of uniformity of dose over the target region
Knowledge-based planning	A class of automated planning approaches that utilise a knowledge-base for plan generation
Machine learning	Forms of artificial intelligence trained on data to produce outputs
Multi-criteria optimisation	Mathematical optimisation techniques involving optimisation of more than one objective function
Organ-at-risk	Radiosensitive regions-of-interest at increased risk of radiation induced side effects due to treatment
Pareto optimisation	Optimisation approaches focused on solutions along the Pareto front, a subset of feasible solutions such that no objectives can be improved further without leading to a detriment for at least one other
Planning	An external beam radiotherapy treatment stage in which treatment delivery is planner
Planning goal	Derived from clinical goals defined by an oncologist or local practice for use in planning. Used to help define optimisation objectives used by a treatment planning system for plan optimisation
Planning parameter	A weighting factor or other input of an automated planning system. It is often used to reference a numeric value between zero and infinity that determines the level priority a planning goal should receive during optimisation.
Protocol based automatic iterative optimisation	An automated planning solution using an iterative update approach
Pareto-guided automated planning	Automated planning method that incorporates Pareto-navigation techniques
Radiotherapy	Also known as radiation therapy. A class of medical treatment modalities involving ionising radiation

Region-of-interest	A region to be considered during planning. These will often be whole organs or sub-sections of organs.
Regression	A class of supervised machine learning methods in which inputs are related to outputs using mathematical functions
Rules-based planning	A class of automated planning approaches utilising algorithms to generate plans
Slope	Rate of change i.e. change in $y \div$ change in x

List of Acronyms

3D-CRT	Three dimensional conformal radiotherapy
ANOVA	Analysis of variance
AP	Automated radiotherapy planning
CI	Conformality index
CT	Computerised tomography
CTV	Clinical treatment volume
DiceC	Sørensen–Dice coefficient
DICOM	Digital Imaging and Communications in Medicine
$\mathrm{D}_{\mathrm{max}}$	Dose volume histogram max dose metric in Gray
$\mathrm{D}_{\mathrm{mean}}$	Mean average dose metric in Gray
DMPO	Direct machine parameter optimisation
DNA	Deoxyribonucleic acid
DVH	Dose volume histogram
ϵc	Epsilon constrained rules-based automated planning
EdgeVcc	Experience Driven plan Generation Engine by Velindre Cancer Centre
EBRT	External beam radiotherapy
FMO	Fluence map optimisation
GED	Geometry-based expected dose
Gy	Gray
GTV	Gross tumour volume
HI	Homogeneity index
IMRT	Intensity modulated radiotherapy
ITV	Internal target volume
KBP	Knowledge-based planning
LOOCV	Leave-one-out cross validation
MCO	Multi-criteria optimisation
$\mathrm{MCO}_{\mathrm{gs}}$	Expert-driven Pareto-guided protocol-base automatic iterative optimisation calibrations and plans
ML	Machine learning

$\mathrm{ML}_{\mathrm{clus}}$	Machine learned calibrations and plans generated using parameters learned via clustering
$\mathrm{ML}_{\mathrm{reg}}$	Machine learned calibrations and plans generated using parameters learned via regression
MLC	Multi-leaf collimator
MSE	Mean squared error
NTCP	Normal tissue control probability
OAR	Organ-at-risk
OLS	Ordinary least squares
$\mathrm{OV}_{ROI1,ROI2}$	Sub-region defined by the overlap of two regions-of-interest (i.e. ROI1 and ROI2). Measured in cm^3
PBAIO	Protocol-based automated iterative optimisation
PC	Principal component
PCA	Principal component analysis
PDD	Percentage depth dose
PG	Planning goal
PG_{H}	Serial organ (P1) and target (P2) planning goals i.e. higher planning goals that are not trade-off (P3) planning goals
PGAP	Pareto-guided automated planning
PRV	Planning organ at risk
PSV	Prostate and base seminal vesicles
PTV	Planning target volume. A delineated volume used for treatment planning. The prescribed dose to be achieved is often denoted in the suffix e.g. PTV48 for a prescribed dose of 48 Gray
$\mathrm{PTV}_{x\mathrm{cm}}$	The planning target volume expanded isoptropically by $x$cm e.g. $\mathrm{PTV48}_{0.02\mathrm{cm}}$ is PTV48+0.02cm
RBP	Rules-based planning
ROI	Region-of-interest
ROI1 VOF_{PTV1}	Total volume of a region-of-interest (ROI1) above the most superior slice and below the most inferior computed tomography slice of a PTV (PTV1). Measured in cm^3
TCP	Tumour control probability
TPS	Treatment planning system

VIF	Volume-in-field. The volume of a region-of-interest in field of the PTV on the treatment plane
VMAT	Volumetric modulated arc therapy
WF	Weighting factor

Chapter 1

Introduction

1.1 Thesis aims

The field of automated planning (AP) was developed to reduce reliance on human interaction during the planning process. The possibility of developing a fully automated planning application with zero hands-on time is still to be realised especially those achieving clinically desirable planning equivalent to or even surpassing manual human planning.

The goal of this work was to determine whether it is feasible to develop a fully automated planning system. The hypothesis is: if expert-driven AP calibration can be modelled using anatomical features as predictive variables, not only will this aid in uncovering the underlying relationships between anatomy, planning parameters and dosimetry, but may also facilitate a full AP method that produces clinically desirable plans. The main objective of the original research presented in this thesis involves the generation and testing of machine learning (ML) techniques to predict expert-driven AP calibration for an in-house built AP method. Additionally it was to determine dosimetric differences between machine learned parameters and traditional site-specific parameters applied to all cases.

1.2 Thesis outline

Radiotherapy background (chapter 2) This chapter presented the context of the project with respect to the field of oncology and cancer research. The various types of treatments that are available are discussed including the prevalence of radiotherapy as a treatment modality. General principles of radiotherapy are introduced such as radiobiological interactions and dose deposition. Methods of dose calculation are discussed in light of varying densities within anatomy and the importance of treatment planning systems is explained. Finally, aspects of planning are defined including factors affecting inverse planning with the concept of *automated planning* presented for further discussion in later chapters.

Automated planning (chapter 3)

This literature review presents the range of AP solutions that have been proposed and applied in clinical practice. Characteristics of each are described and some examples of use and success rates discussed. The approaches are compared and critiqued with advantages and pitfalls outlined including possible areas of development. The AP approach focused on in this work will be introduced including details of the algorithm.

With respect to the literature review, developments to the current AP method are proposed particularly for protocol-based automatic iterative optimisation (PBAIO) approaches. In the latter section of the chapter, the clinical sites considered in this thesis are outlined and discussed including why they were chosen and patient inclusion-exclusion criteria.

Hypothesis generation (chapter 4)

This chapter presents some of the preliminary work undertaken to gain knowledge about the relationships between anatomy, planning parameters and dose distribution to aid in the main ML study. It is split into three sections:

1. intra-planner study - a study that assesses discrepancies in planning choices by the same expert planning professional when planning the same patients on different days. Results of this study help in defining the gold standard

2. inter-planner study - a study to compare planning choices and prioritisation of planning goals. The results of this study help in the definition and justification of gold standard planning

3. anatomy simulation study - a study to methodically manipulate anatomy and produce new planning parameters. For each anatomy, new planning parameters were obtained such that they result in a comparable dose distribution to the original plan. Controlled augmentations of ROIs were created in order to simulate anatomical variance and better understand the underlying relationships between anatomy

and planning parameters.

From these studies, a definition for gold standard planning was obtained and heuristics were generated for used as a guide in the ML solutions for the main study.

Modelling and cross validation (chapter 5)

In this chapter, ML approaches are discussed with detailed descriptions of regression including multiple polynomial regression equations and descriptions of the range of automatic clustering algorithms with a detailed outlining of K-means. The ML approaches chosen are outlined and their parameters explained. Their appropriateness is justified and includes a discussion of why they were chosen over alternative ML methods. The concept of goodness-of-fit metrics is introduced and explained with descriptions of each. Also, outlined in this chapter is the concept of variable standardisation including the importance of doing so and choosing the correct method.

Cross validation is also outlined in detail with a focus on the leave-one-out cross validation approach. Cross validation is discussed with respect to planning parameter prediction including a detailed example of the methods used in this thesis and why they were chosen over alternative methods.

Regression modelling and cluster modelling (chapter 6)

The results of all cross validation models and final models are presented and discussed. The various modelling methodologies used for the prediction of gold standard calibration parameters towards a full AP system are presented including:

1. multi-polynomial regression using standardised raw features values
2. multi-polynomial regression using Principal Components in a reduced dimension space
3. K-means clustering over standardised raw features
4. K-means clustering over Principal Components

Conclusion (chapter 7)

This chapter discusses the outcomes of the work and final thoughts on the application of ML to AP calibration in general. Future work is suggested including developments of this work and other related approaches.

Chapter 2

Radiotherapy Background

2.1 About radiotherapy

Cancer statistics for the UK indicate survival rates have doubled in the past 40 years. Of those diagnosed, 50% live for 10 years or more and these rates are expected to improve. Nevertheless, 375,000 new cancer cases are reported each year[14]. Radiotherapy is one of the three main treatment modalities for cancer along with chemotherapy and surgery. Of the patients receiving at least one of these three main treatment types, at least 27% are treated with radiotherapy[15,16]. Given it is common for patients to receive more than one form of treatment, up to 50% will benefit from radiotherapy as part of their course of treatment[17].

Matter that absorbs high energy electromagnetic radiation (such as the radiation associated with radiotherapy) are referred to as *absorbers* and interactions of radiation in such matter occurs at the atomic level during radiotherapy. Human tissue contains absorbing matter and atomic level interactions with radiation affect tissue function at the cellular level. This is usually due to alterations in the function of the deoxyribonucleic acid (DNA). DNA molecules are polynucleotides present in the nucleus of every cell and encode information including cell renewal and programmed cell death.

The radiation delivered during radiotherapy is "ionsing" as interactions with matter can cause atoms to detach electrons resulting in charged particles known as ions. Ions with unpaired electrons form free radicals that are chemically unstable. The production of free radicals leads to more than one kind of chemical reaction but alterations in DNA structure are dependent on the interactions that occur with water molecules[18] as illustrate in Figure 2.1. Alterations in DNA structure can occur due to base protein damage or can

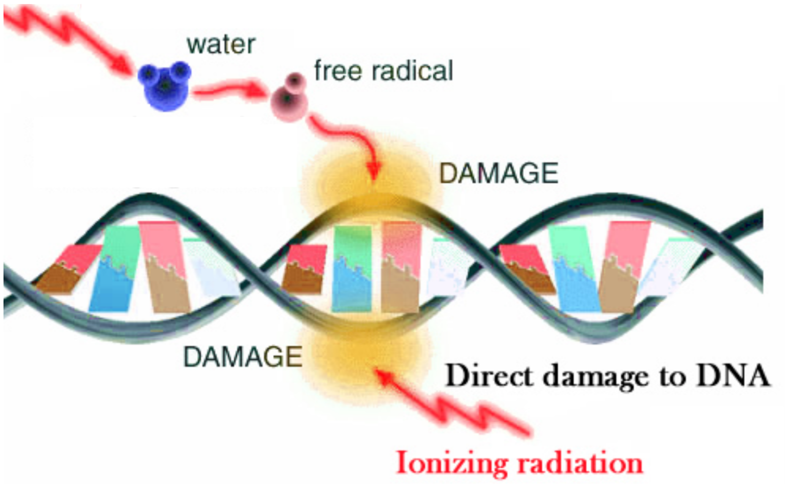


Figure 2.1: An illustration of DNA damage due to radiation. Adapted from RF Safe[1]

be due to single-stand or double-strand breaks in polynucleotide bonds.

Well designed radiotherapy treatment will usually cause the cell to follow a degradation pathway, an outcome that is highly correlated with an accumulation in the number of double strand breaks[19]. Alternatively, treatment can lead to DNA adjustments that encourage damaged cells to begin repairing themselves or simply to the loss of reproductive integrity. When there remain no clonogenic cells that maintain reproductive integrity in a volume of tissue, mitosis is therefore inhibited. The tissue mass will then shrink as cells begin to die given no further proliferation. Well designed radiotherapy treatment is governed by the 5R's of radiotherapy and these are outline below.

2.2 The five R's of radiotherapy

The rationale behind many principles used in clinical radiotherapy treatments are based on the five R's of radiotherapy: repair, repopulation, redistribution, reoxygenation, and radiosensitivity[20,21] and these will be discussed in the following sections 2.2.1-2.2.5.

2.2.1 Radiosensitivity

Radiosensitivity refers to the level of resistance cells have to radiation-induced damage. It can vary based on cells ability to repair damage, hypoxia, cell cycle position, and growth fraction[22]. These will be discussed in more detail later. Radiosensitivity may also be based on genetics of the individual.

2.2.2 Repair

When DNA is damaged, the cell will usual repair itself in a process known as DNA repair of which there are many known pathways[23,24]. Cell repair is less likely to occurs when a cell is in what is known as the mitotic phase and treatments are designed to take advantage of the differential in cell cycles between cancerous cells and health cells. This will be discussed in more detail in section 2.3.2.

2.2.3 Repopulation

Repopulation refers to the proliferation of clonogenic cells. Repopulation of cells following treatment occurs at an increased rate than normal[25] and research shows that treatments designed to last an extended period of time can lead to detrimental outcomes due to the increased rate of population in cancerous cells due to treatment[26]. Therefore, treatments are designed to occur in short window to account for the increase in repopulation rates due to treatment.

2.2.4 Redistribution

Cells enter different cycles at different times and are more radiosensitve during certain cycles than others. Therefore, designing treatments such that cancerous cells are targeted during the time they are expected to entering into their most radiosensitive phase has therapeutic benefits. Research also shows that even sublethal damage can accumulate hence continued treatment is advisable[27].

2.2.5 Reoxygenation

After a dose of treatment, cells become *hypoxic*. That is an insufficient amount of oxygen is available at the cell level to maintain *homeostasis*, the cells ability to self-regulate[28]. Hypoxic cells are comparatively radioresistant[29] and this leads to a therapeutic disadvantage. Reoxygenation is the process in which cells that are hypoxic become oxygenated again hence less radioresistant. Treatment is therefore designed to account for the need of time between dose deliveries for cells to become reoxygenated and to take advantage of this therapeutic benefit.

2.3 Radiotherapy as a treatment for cancer

Cancer treatment is bespoke to each patient with treatment considerations made on a patient-by-patient basis by a multidisciplinary team of professions. Radiotherapy treatments are highly versatile and suitable in an array of cases including cancers of the soft and connective tissue (sarcoma)[30], the lymphatic system (lymphoma)[31] and the bone marrow (myeloma)[32]. However, radiotherapy is particularly well suited to treating carcinoma: cancers originating in the surface or skin of organs that typically form tumours. Treatment of carcinoma is well documented and this thesis will focus on these therapies.

2.3.1 Cell survival

The term *cell survival* refers to the number of clonogenic cells remaining following treatment and is known to depend on a number of factors[33]. Nevertheless, there is a particularly strong documented relationship between the radiation energy absorbed by matter and cell survival and this relationship can be modelled. A common model used for this is the Linear-Quadratic model[34], an exponential decay function of the form:

$$S(D) = e^{-(\alpha D + \beta D^2)}, \tag{2.1}$$

where cell survival S is a function of *radiation dose*, D, which is the amount of absorbed radiation energy. The constants α and β refers to the relative radiosensitivity of the cell due to the number of double strand breaks that are expected. A large $\frac{\alpha}{\beta}$ ratio indicates the relatively greater importance of the linear coefficient and suggests a comparatively constant rate of cell death as radiation energy increases. Smaller $\frac{\alpha}{\beta}$ ratios suggest cell death will occur at a faster rate as radiation energy increases.

2.3.2 Cell Cycles

Cell cycle repetition in tumorous cells occur more frequently than in normal tissue meaning there are a larger number of tumour cells entering the mitotic phase of the cell cycle (Figure 2.2) than the surrounding tissue. This links closely to redistribution mentioned in section 2.2.4. Double strand breaks are achieved at an increased rate when cells are in this phase, hence an increased rate of cell death is likely to occur in cancer cells when treated[35]. However, not only the disease is subject to treatment. When healthy tissue is irradiated, cell damage can occur and lead to radiation induced toxicity. Toxicity has

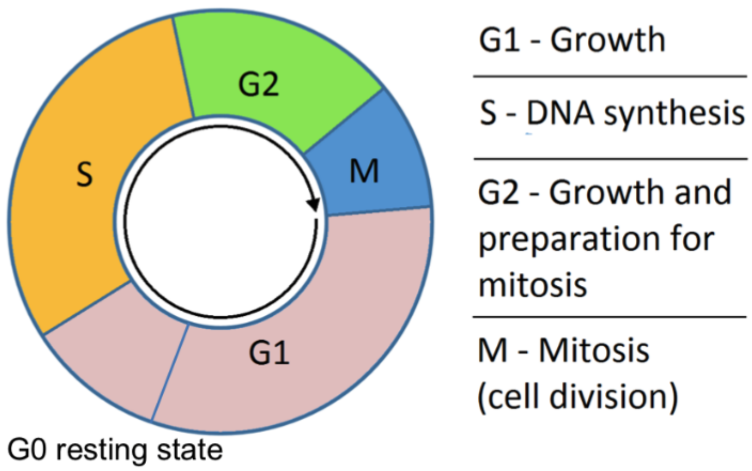

G2
S
M
G1
G0 resting state
G1 - Growth
S - DNA synthesis
G2 - Growth and preparation for mitosis
M - Mitosis (cell division)

Figure 2.2: Illustration showing the various stages of the cell cycle broken down by respective duration for the average healthy cell[2]

been associated with radiation induced secondary malignancies in cells that were previously healthy[36–38] and can also lead to acute radiation syndrome causing the patient to go through four stages of symptoms before ultimately recovering or dying. Targeting treatment appropriately is therefore vital, not only for ensuring effective treatment of the disease but also to minimising the likelihood of radiation induced side-effects. Radiotherapy is administered using different methods, each with benefits and pitfalls. These will be discussed in more detail.

2.3.3 Types of radiotherapy treatments

Delivery methods include molecular radiotherapy[39], brachytherapy[40] and external beam radiotherapy (EBRT). Molecular radiotherapy refers to the use of radiopharmaceuticals administered intravenously or orally to be taken up by the disease. This method therefore has the advantage of accurate targeting and it is also useful for treating any disseminated disease. However, it is not as widely used as other treatments[41] as it is most effective in specific cases. The radiopharmaceutical is chosen based on properties of the cancer and the propensity of the associated region to take it up and will therefore not be applicable in all cases.

Brachytherapy refers to the use of radioactive seeds or metal pellets that are inserted into or near to the target area. This treatment is widely used today and is especially useful for achieving local control but less useful for treating the disseminated disease. EBRT is among the most common forms of radiotherapy in use today and refers to any radiotherapy treatment involving radiation beams directed towards the treatment site from outside the body. Although it is not as effective at obtaining local control as other methods, a key advantage is the non-invasive nature of the procedure. Advanced forms of EBRT will be the focus in this work, specifically intensity-modulated radiotherapy (IMRT) and volumetric modulated radiotherapy (VMAT) and these will be discussed further in the following sections.

EBRT refers to a range of treatments[42] most of which utilise photons and almost all deliver dose using a linear accelerator (linac). The linac comprises a couch upon which the patient is immobilised and a moving gantry head that delivers the therapeutic dose. A schematic of linac gantry head composition can be found in Figure 2.3 and it contains a few key elements including:

- The gantry head target - a metal plate (usually tungsten) that interacts with accelerated electrons and is used to produce photons
- Primary collimator - used to define the primary field size by absorbing scatter photons outside of a specified field
- Flattening filter - used to flatten the beam and create a uniform field
- Ion chamber - measures the delivered dose and used to terminate the beam when the specified dose has been achieved
- Mulit-leaf collimator - used to achieve conformal shaping of the beam to match the treatment target contours at the angle being treated

Within a modern photon linac, free electrons are created in the cathode through a process known as thermionic emission and accelerated towards the gantry head target through an electron waveguide using two sets of steering coils. When accelerated electrons hit the gantry head target, bremsstrahlung interactions with atoms in the target lead to the production of photons that are scattered in all directions. Forward travelling photons pass through the primary collimator and due to scatter form a cone beam. Photon *fluence* is defined as the number of photons incident on a surface per unit area of the surface per second. As the distribution of photons exiting the primary collimator are more concentrated at the center of the beam, a flattening filter is placed in the beam path and absorbs energy close to the center producing a more uniform fluence across the beam. The secondary collimator or jaws, define the maximum beam field with the multi-leaf collimator

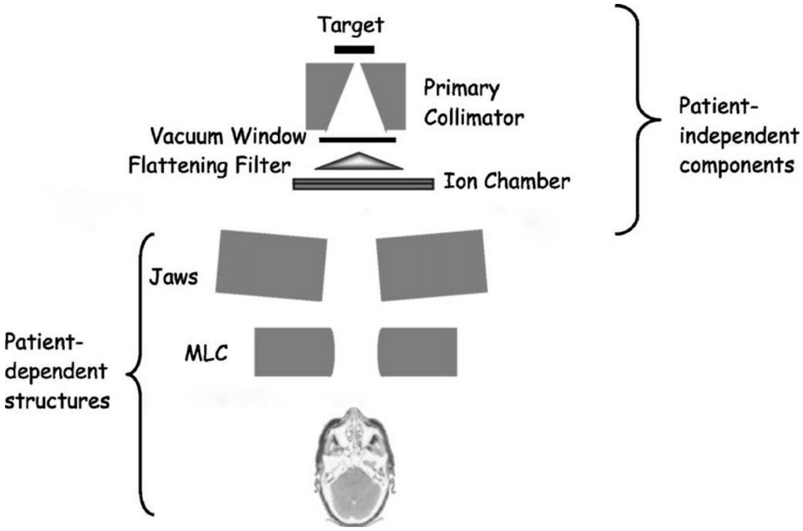


Figure 2.3: Illustration of the key components with the linac gantry head. Adapted from Chetty et al. (2007)[3]

(MLC) used to form specific aperture shapes to conform to treatment target contours and modulate dose.

During three-dimensional conformal radiotherapy (3D-CRT) for example, a 3D digital simulation of patient anatomy is created based on prior imaging (usually computerised tomography or CT image slices) that have been manually delineated to contour regions-of-interest (ROIs). These 3D images are used to plan aperture shapes that match the treatment volume at defined angles in a single plane in 3D space and these can then be used to define the position of the linac jaws and MLCs at various angles. This form of EBRT is delivered from set coplanar beam angles around the patient with each beam delivering uniform fluence.

However, IMRT is considered the standard today. Multiple modulated beam apertures enable non-uniform beam fluence ensuring not only for an appropriate dose delivery to target structures, but greater sparing of healthy tissue[43]. IMRT can be implemented using step-and-shoot in which MLCs form aperture shapes whilst the beam is off, or a sliding window delivery method in which aperture shapes are formed with the beam on. Research suggests therapeutic results of each implementation are greatly comparable[44,45]. However, a weakness of IMRT is the increase in treatment delivery times compared to other EBRT delivery methods.

Volumetric modulated arc therapy (VMAT) is a novel implementation of IMRT where aperture shapes are altered dynamically whilst the beam is on and dose is delivered as the gantry moves continuously in a 360° motion. A finite number of coplanar beam fields are defined and the machine parameters between defined fields are interpolated. It has been found that although VMAT is likely to result in a low dose bath to the patient outside of the treatment volume due to all angles of the patient being exposed to dose, it allows for greater sparing to organs-at-risk (OARs) and is still expected to correlate well with a reduction in dose to OARs and radiation induced side-effects[46–48].

For small tumours, tumours in areas sensitive to radiation and non-localised cancers such as myeloma and leukemia, alternative forms of EBRT are available and include stereotactic radiation therapy, proton therapy and total body irradiation respectively. Also, in some institutions, non-coplanar VMAT is also possible[49,50]. Although these other forms of EBRT are used in current practice, for the remainder of this thesis the focus will be on coplanar IMRT and VMAT only.

2.4 Considerations prior to implementing IMRT or VMAT

Prior to clinical use, quality assurance practices must be carried out including a commissioning stage of the linac and corresponding tools to ensure they are aligned and configured appropriately. This involves measuring intended dose delivery against actual delivered dose and adjusting machine parameters accordingly. This often involves the use of *phantoms* and these will be discussed in more detail in the next paragraph. However, IMRT and VMAT still present some risk of leading to radiation induced side effects and this can be thought of an an unavoidable consequence of these EBRT techniques. Nevertheless, there are additional procedures for mitigating random and systematic error for clinical effectiveness and better protection for patients. These include the use of immobilisation, fractionation and planning.

The term *phantom* refers to any object or device used in place of a human and there are many types of phantom. For example, the Delta^4 phantom (Scandidos, Uppsala, Sweden) comprises two intersecting perpendicular planes of polymethylmethacrylate containing dose sensors. It is used to measure dose delivered by the linac and can be used to determine whether a plan is deliverable or not. Dose to the phantom is compared against intended dose to the phantom with machine parameter adjustments made when significant discrepancies are found. The computerised XCAT phantom aids with realistic patient modelling in 4-dimensions[51] and can be useful for estimating dose delivery to treatment regions under motion such as the lungs. A water phantom can be used to measure dose at depth in homogeneous matter and produce estimates for dose in inhomogeneous matter such as human tissue. Phantoms are therefore safe measurement tools and verification mechanisms for machine configuration and help to facilitate in silico machine configuration prior to patient treatment.

Immobilisation refers to the fixed position patients are asked to assume during treatment sessions. The purpose is minimisation of systematic error following machine configuration and patient treatment targets. A common standard immobilisation is the head-first supine position in which a patient lays on their back with their head at the top of the couch and feet at the bottom. There are many others, some of which include the use of immobilisation devices including personalised molds and supports. Immobilisation is useful for: (i) initial treatment planning which will be discussed in a later paragraph, (ii) adaptive radiotherapy which relates to changes made to the initial treatment plan to account for anatomical variance that has occurred since the original plan was made, and (iii) a reduction in systematic errors during treatment due to consistent positioning.

Fractionation refers to the practice of delivering the prescribed dose over a series of treatment sessions. The full prescribed dose is usually not delivered to a patient in one session. Instead it is split into a series of fractional doses delivered in succession (e.g., daily). This has a therapeutic benefits for the healthy tissue, given cell repair and recovery is achieved at a higher rate in non-cancerous tissue and will commonly result in manageable acute side effects such as nausea. For this reason, it is expected that with effective targeting, most of the treated healthy tissue will be repaired between fractions. Fractionation also links closely to the $\frac{\alpha}{\beta}$ ratio of the coefficients of equation 2.3.1 and is sometimes known as the *therapeutic ratio* which will be discussed later in section 2.10.1. But fractionation is not only beneficial in terms of cell repair. Short fractions help to manage the repopulation of clonogenic tumour cells, it ensures an accumulation of sublethal dose to cells as they redistribute through the cell cycle, and it allows time for reoxygenation of hypoxic cells. However, all other pre-treatment considerations depend upon the achievement of a clinically applicable *treatment plan*.

Treatment cases may be comparable but each is unique and must be considered on a case-by-case basis. Achievement of effective treatment is strongly dependent on the production of an appropriate plan and planning comprises a number of stages that will be outlined in section 2.10. Firstly, in order to better understand planning, the treatment pipeline will be outlined. Following this, the background of radiotherapy will be discussed including understanding imagining, dose deposition and how dose maps are calculated and modelled. Also, as previously mentioned, radiotherapy planning is closely related to adaptive radiotherapy. Although applications of adaptive radiotherapy are outside of the scope of this work, it should be noted that treatment planning approaches applicable to the original plan are often applicable in adaptive planning also.

2.5 Radiotherapy treatment pipeline

There are four main stages of the radiotherapy treatment pipeline:

1. Diagnosis and patient consent to treatment
2. Pre-treatment preparation
3. Treatment
4. Post-treatment follow-up

Following consultation with an oncologist, the disease will be diagnosed and as mentioned, a multidisciplinary team of professions will identify radiotherapy as a viable form of treatment (or otherwise). If the patient consents, the pre-treatment stage will commence. The pre-treatment stage itself contains a number of phases including imagining for the visualisation of anatomy, delineation of key ROIs and treatment planning. Following this, the treatment stage of the pipeline commences in which the prescribed dose is delivered to the patient in a series of fractional doses over a defined period (e.g., everyday for a week). The patient will then follow-up with the oncologist to reassess the prognosis. This thesis will focus on the pre-treatment stage of the pipeline, in particular the treatment planning phase of pre-treatment. All stages of pre-treatment are important to planning and all will be outlined in the following sections. This includes outlining the fundamentals of radiation as pertain to radiotherapy and the background of radiotherapy planning techniques.

2.6 Imaging

More than one imaging modality is considered appropriate for clinical use due the ability of various modalities to produce a 3D snapshot of patient anatomy. CT is the predominant imaging modality within radiotherapy. A series of cross-sectional images through the body are captured via non-ionising radiation, usually up to 0.015Gy. Regions of high density (such as bone) will appear white, regions containing air will appear black and other regions will appear grey relative to their density. A core advantage of CT is the images can be used to determine the Hounsfield Units (HU) of different structures. From HU values, it is possible to directly calculate electron density and this detail is an essential requirement for subsequent dose prediction and optimisation systems, the details on which will be address in section 2.10.2. However soft tissue contrast is poor with this imaging modality and it provides no “functional image” information.

An alternative modality is magnetic resonance imaging (MRI) and this uses high energy magnetic fields to manipulate water molecules. It exploits the fact that water molecules in material of different densities move at different rates. Therefore, MRI is valuable for imaging areas of low density such as the brain and requires no radiation hence has the advantage of no dose deposition.

A further alternative is positron emission tomography (PET), a modality that makes use of gamma cameras to capture gamma ray production due to positron collisions of electrons. Given tumours often have a higher metabolic rate than surrounding tissue, administering a radioactive tracer results in a high rate of uptake by the tumour and gamma ray production is often well localised in this area. PET scans are therefore useful for functional imaging such as visualising metabolic activity and has the advantage of providing superior visualisation of the tumour when multi-modality imaging is applicable.

2.7 Delineation

Following imaging, organs and treatment volumes are delineated by qualified professionals and important ROIs identified. These include the sequentially determined volumes defined for treatment outlined by the the International Commission on Radiation Units and Measurements (ICRU)[52] and an illustrative example of these volumes can be found in Figure 2.4. These volumes include:

- Gross Tumor Volume (GTV) - delineates the gross visible malignant growth and contains the macroscopic disease. Multiple imaging modalities may be used to improve the delineation of this region.

- Clinical Target Volume (CTV) - accounts for the microscopic and/or subclincal invasion of the tumour to the surrounding area.

- Planning Target Volume (PTV) - a volume used to minimise the likelihood of under or over treating the CTV due to intra- and inter-fraction geometric changes. Planning utilises static imaging, although idiosyncratic geometric anatomical changes can occur at any time. Changes can be due to tumour shrinkage, unexpected organ motion, weight loss or other unexpected changes. Estimations of possible variations in ROI positioning are considered including variations in CTV positioning. The defined PTV is therefore applied as a treatment margin used to plan dose delivery and increase the likelihood the CTV will receive the prescribed dose upon treatment even when its position deviates from that of the original imaging.

- Internal volume target (ITV) - a volume accounting for known or expected changes in CTV position and used in specific planning cases such as lung radiotherapy. A CTV is considered prior to the PTV contour being delineated.

- Organs at Risk (OAR) - comparatively radiosensitive normal tissue proximal to treatment target volumes. These include regions at increased risk of long term and/or severe functional damage given certain treatment conditions. Avoidance of these structures are considerations made during planning.

- Planning Organ at Risk Volume (PRV) - similar to the PTV concept, PRVs considered OAR position variance given unexpected changes in anatomy following original imaging. A PRV can be used during planning to increase the likelihood of appropriate avoidance of dose to OAR during treatment.

- Treated volume - refers to the total region receiving the prescribed dose intended for the PTV. This volume is used to define conformality metrics that are useful for reviewing a plans clinical applicability.

- Irradiated volume - refers to all other dose received during treatment that is considered significant with respect to the normal tissue.

Achieving conformal fields about the PTVs is desirable in the avoidance of OAR treatment. However, given PTVs can overlap OARs, conflicts can occur during planning and management of these conflicts determines overall clinical desirability of the final plan. This will be discussed in more detail in section 2.16. First, the fundamentals of radiotherapy and planning will be discussed.

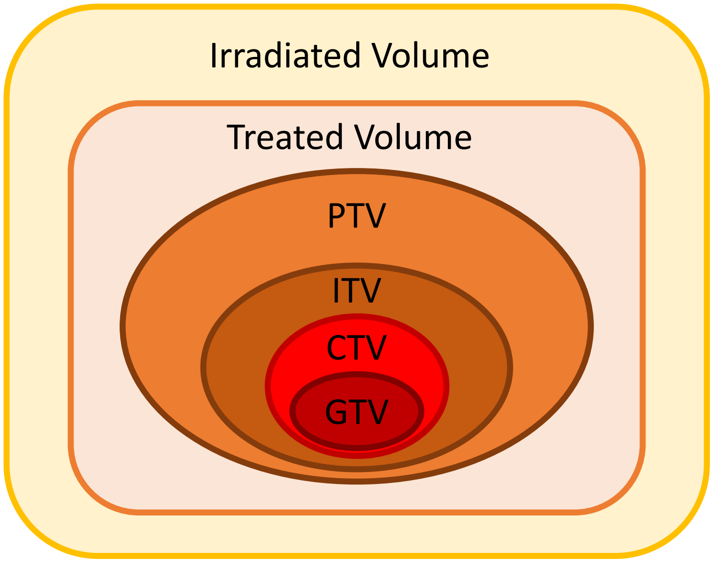


Figure 2.4: An example schematic of volume definition as defined by the ICRU. Adapted from ICRU report 62

2.8 Fundamentals of radiotherapy

2.8.1 Photon interaction with matter

Electromagnetic radiation moves as a wave and through a vacuum travels at the speed of light, c. Given wavelength λ, unit oscillation of an electromagnetic wave is defined as $v = c/\lambda$. Gamma rays contain uncharged photon particles and have the quality of short wavelengths, high energy and high penetration. Given unit oscillation of the gamma ray, v, by the unit-mass equivalency each photon carries energy hv where $h = 6.626\ J\ s^{-1}$ is Planck's universal constant and 1 J is defined as

$$1\ J = 1\ kg\ m^2\ s^{-2}. \tag{2.2}$$

Absorbed dose is defined as the amount of energy deposited within matter and is measured in Gray (Gy) where 1 Gy of absorbed dose is defined as 1 Joule (J) of energy absorbed per kilogram. The kinetic energy of accelerated particles is measures in mega-electron volts (MeV) where 1 MeV is defined

$$1\ MeV = 1.6 \times 10^{-13} J \tag{2.3}$$

The term *scatter* used here will refer to a change in direction of a particle once following a different trajectory. Photon attenuation is due to one of four main kinds of interaction[18,53]:

- Photo-electric effect - ejection of an electron (known as a photoelectron) by a photon incident to the atom. Given incoming photon energy of hv, the ejected photoelectron has kinetic energy $T = hv - E_B$ where E_B is the binding energy of the electron. Electrons liberated from lower shells have a higher E_B than outer shell electrons and this form of photo-electric scatter can result in liberation of an electron as well as a production of a photon. An outer shell electron dropping to replace the ejected electron produces a photon with energy equal to the difference between the energy level of the two shells

- Rayleigh scatter - refers to "coherent" scatter in which no energy is lost or absorbed but scatter occurs due to deflection

- Compton scatter - refers to "incoherent" scatter in which some energy is lost due to absorption and the rest converted to one of two types of scatter. The incident photon leads to liberation of an electron at an angle θ and deflection of the photon at angle ϕ and reduced energy. The kinetic energy of the ejected electron is $T = hv - hv'$ where hv' is the energy of the scattered photon and defined

  $$hv' = \frac{hv}{1 + \alpha(1 - \cos\theta)} \tag{2.4}$$

  where $\alpha = hv/m_0c^2$, c is the speed of light and m_0 is the rest mass of the electron. The relation between θ and ϕ is given by

  $$\cot\phi = (1 + \alpha)\tan\frac{\theta}{2}, \tag{2.5}$$

  and T is maximised when the photon is scattered directly backwards.

- Pair production - refers to creation of an electron-positron pair following collision with the nucleus with a combined kinetic energy of $T^- + T^+ = hv - 2m_0c^2$. Given $2m_0c^2 \approx 1.022$ MeV, this is the threshold for pair production.

Figure 2.5 illustrates how dominant forms of interaction vary with energy and atomic number. At lower energies, photo-electric scatter is dominant especially as atomic numbers increase and at higher energies, pair production. Compton scatter is the dominant interaction mechanism in EBRT discussed in this work.

2.8.2 Factors influencing dose deposition

Photon scatter is stochastic and therefore calculations for dose to patients are probabilistic. For any one beam energy, there are nominally three criteria affecting dose delivery

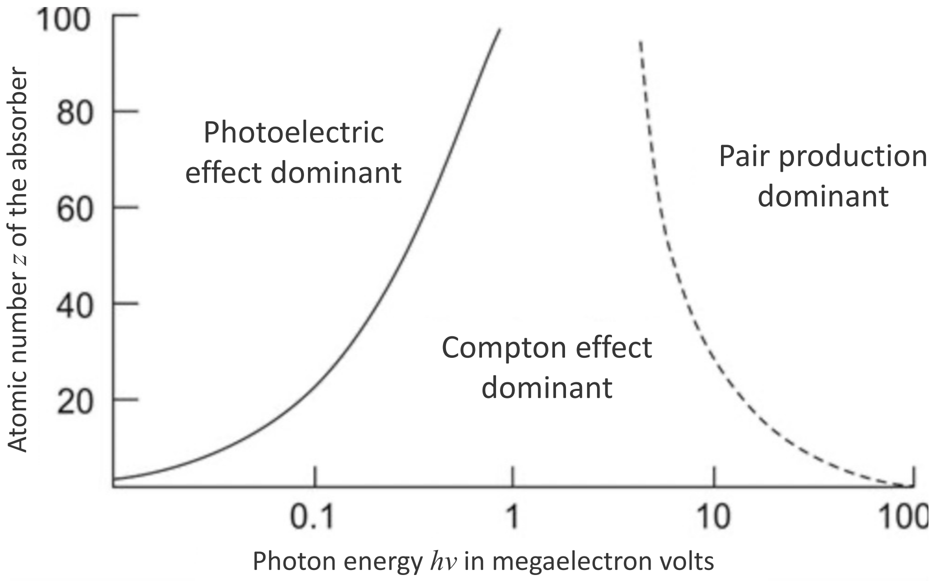


Figure 2.5: An illustration of dominant interaction mechanisms with respect to the atomic number of the absorber and beam energy. Adapted from Parajuli et al. (2022)[4].

in matter[18]:

- distance between the radiation source and the surface of the matter - the unattenuated dose rate is inversely proportional to the square of the distance between the source and point of measurement[18,54]. That is, expected dose at the point of measurement is proportional to r^2 where r is the distance between the radiation source and matter.

- beam energy - assuming uniform density of matter and a set skin-to-source distance, the energy of the beam determines the percentage of absorbed dose given the depth. This is known as the "percentage depth dose" or PDD. Maximum PDD ($\mathrm{D_{max}}$) occurs at some depth below the surface of the matter. This is due to the accumulation (or build-up) of energy due to the release of secondary electrons. With higher energy beams, the path length of secondary electrons increases meaning the build-up occurs at a greater length and $\mathrm{D_{max}}$ occurs at a lower depth.

- size and shape of the beam field - as the depth increases so too does the overall rate of scatter. Total scatter is minimised at depth for smaller field sizes. Field size also affects the build-up region with the maximum dose occurring closer to the surface over larger fields of the same beam intensity and source-to-skin distance.

As source-to-skin distance and/or beam size increases, the percentage of scatter increases. As beam energy increase, the percentage of scatter decrease. With these factors controlled, generation of computational dose models are possible. In particular, managing dose deposition by manipulating the beam shape and size is exploited in modern EBRT using MLCs to deliver non-uniform fluence to the target area. Initial dose calculation is generated assuming patient composition is equivalent to that of water given the large proportion of human tissue that is water. Corrections and models are therefore based on this fact.

2.9 Dose Calculation

Recognised dose calculation algorithms can be broadly categorised into three groups[55]:

- correction-based algorithms - measurement-based algorithms that rely on predictions made based point dose kernel and equivalent path length in water phantoms with corrections made for known inhomogeneities and variations in density. Correction-based algorithms are no longer widely used given comparably expense computations and inaccuracies
- model-based algorithms - dose calculations based directly on patient representation
- Monte Carlo simulations - models built based on calculated probabilities using historical data to model dose deposition given the stochastic scatter. Monte Carlo methods are currently considered the gold standard given comparably accurate dose calculation.

Dose deposition is dependent on where it hits the body and on attenuation of the beam through matter of varying densities. Therefore modelling accurate patient dose deposition is important. This is achieved with the help of imaging modalities such as CT.

2.10 Fundamentals of radiotherapy planning

Treatment planning refers to the process of establishing a dose delivery protocol. It is unique to each patient case with bespoke parameters tailored to suit the anatomy of the patient and match therapeutic requirements as defined by an oncologist. The planning of dose delivery protocols will be discussed in more detail in section 2.11 and beyond but the principle is for the treatment area to receive sufficient dose coverage with healthy tissue (especially radiosensitive tissue) spared as best as possible. Oncologists will therefore

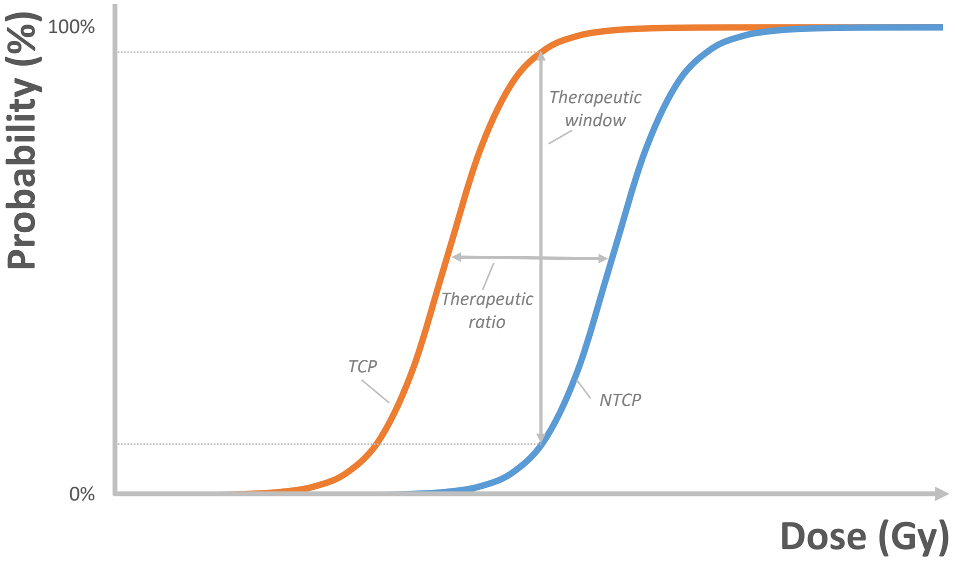


Figure 2.6: An illustration of the TCP-NTCP relationship with a strong therapeutic ratio. Adapted from Reda et al. (2020)[5]

outline their preferences taking into consideration dose coverage of treatment volumes and maximum dose tolerance of OARs.

To achieve a plan congruent with oncologist preferences, modulation of the dose field shape and energy is defined by a planner using computer simulation software to model patient anatomy, machine parameters and dose distribution. This software is known as a *treatment planning system* or TPS. Functionality and use of a TPS will be discussed in more detail in section 2.10.2, but with the use of imaging and dose calculation algorithms, a TPS can be used to model dose distribution given differences in clinical preferences such as increasing the relative importance of sparing one region over another.

However, despite state-of-art software, this process remains highly non-trivial requiring specialist expertise among planners and identifying the best plan can still be difficult given more than one plan may meet the oncologist's key treatment goals. Nevertheless, with respect to clinical outcomes, a "best possible" plan is achievable in each case and convergence on this plan is dependent on management of two over arching probabilities of control: maximising the tumour control probability and minimising the normal tissue complication probability.

2.10.1 Tumour control and normal tissue complication

The therapeutic outcome of treatment can be understood using an elegantly illustrated relationship between tumour control probability (TCP) and normal tissue complication probability (NTCP)[56]. TCP and NTCP are known as biological effect models. The relationship between biological effects and dose due to radiotherapy has been widely studied[57] with classic models taking a sigmoidal shape similar to those illustrated in Figure 2.6. TCP models represent the percentage of clonogenic tumour cells that are eliminated given the tumour receives a certain dose. The model for NTCP is the probability complications will occur given normal tissue is treated with a certain dose. The aim of effective planning is to maximise a theoretical *therapeutic ratio* (or index) between TCP and NTCP where this ratio is defined as the difference between TCP and NTCP at a given probability.

Biological effects link closely with fractionation. As mentioned, cell cycles tend to repeat more frequently in cancer cells making them more susceptible to cell damage due to the increased incidence of DNA double strand breaks. This scenario makes fractionation a desirable treatment mechanism especially when the $\frac{\alpha}{\beta}$ ratio of the Linear-Quadratic model (equation 2.3.1) is high for the tumour and low for nearby OARs. However, when treatment volumes are proximal to normal tissue that also have short cell cycles or are otherwise notable radiosensitive, the therapeutic ratio may be small regardless of the fractionation schedule. This is similarly true for aggressive tumours containing a high population of clonogenic cells where regrowth of the tumour may occur at a rate comparable to that of the healthy tissue. Therefore, increasing the fractions will in some circumstances enable effective dose delivery to targets while allowing healthy tissue recovery between fractions hence improving the therapeutic output. Given a similar $\frac{\alpha}{\beta}$ ratio for all regions this may not be the case. The fractionation schedule is determined by an oncologist on a case-by-case basis to suit the nature of the treatment region.

However, the TCP-NTCP relationship is particularly dependent on the optimisation of the dose distribution when a plan is defined. Managing planning priorities can be complex because approaches for maximising TCP may simultaneously have a negative impact on NTCP and vice versa. A clinically desirable approach is not always apparent and planning often uncovers trade-offs between these two biological effects. Management of NTCP itself can become complex given planning priorities for OARs may uncover trade-off relationships such as variances in proximity to treatment target volumes and varying radiosensitivity of different regions. Direct reporting of TCP and NTCP can be complicated[58–60] and in practice are not directly modelled for review during planning. Instead, dose delivery is managed entirely by planning parameters defined during the planning process.

Oncologist's clinical goals are defined using their own expertise along with ICRU standards and dose-volume constraints that have been locally defined by the institution. Planning goals (PGs) are then defined by the planning professional for use in the TPS during plan *optimisation* with the aim of meeting clinical goals. Optimisation will be discussed more in section 2.11 but examples of PGs include limiting the average dose delivered to an OAR or setting a minimum dose for a target. Chosen PGs are assigned to what is know as a *planning protocol* that lists and prioritises the PGs with respect to one another. The choice and configuration of PGs will have an impact on the final dose distribution. However, the relationship between dose distribution and PG configuration is unknown apriori and dose distribution is only established following full optimisation of PGs within the TPS when a 3D dose distribution is achieved. This is a barrier in planning and relates closely to the aims of this work outlined in section 1.1. First, principles relating to optimisations are outline and discussed to provide context.

2.10.2 Treatment planning systems

In modern EBRT planning, a dedicated TPS is often adopted within a clinic with common and commercially available systems including Pinnacle^3 by Philips Healthcare, Eclipse by Varian Medical Systems and Monaco by Elekta. A benefit of a modern TPS (as well as that of other planning machinery) is support of the Digital Imaging and Communications in Medicine (DICOM) Standard introduced by the National Electrical Manufacturers Association[61]. This Standard enables coherent file transfer of imaging and data between imaging machines, the TPS and linac without having to change the format.

Nevertheless, each TPS differs. For example, each will have its own dose calculation method, delineation tools, image processing tools and further optimisation management including *objective functions* which will be discussed later in section 2.13. An overview of plan optimisation management will now be addressed.

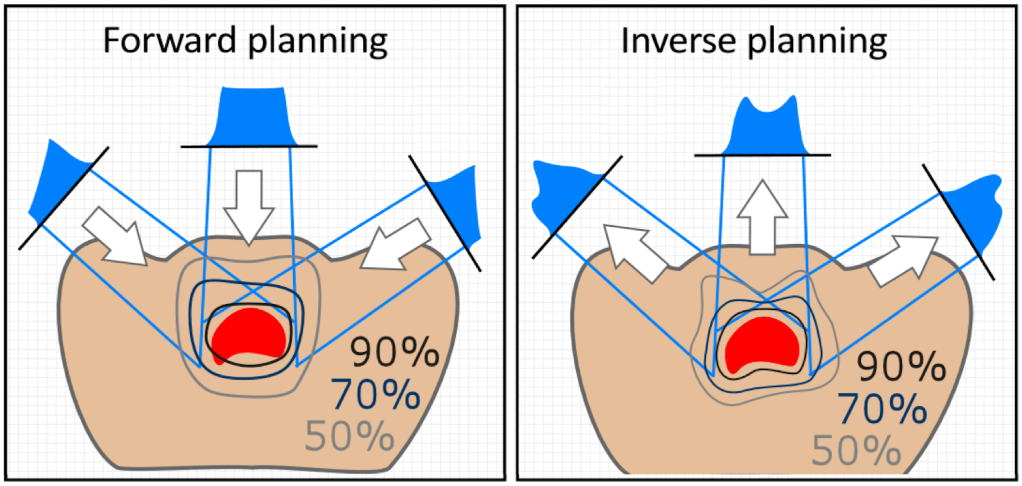


Figure 2.7: An illustration showing differences in forward and inverse planning methods. Adapted from Carlsson (2008)[6].

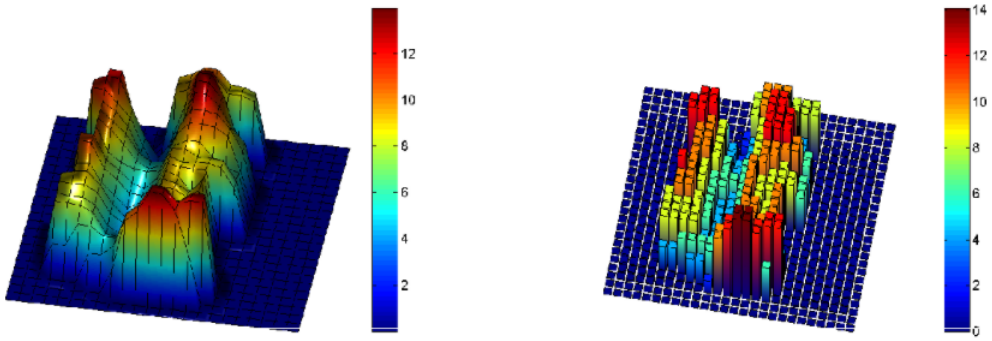
2.11 Plan optimisation

There are two main approaches to plan optimisation: forward planning and inverse planning.

2.11.1 Forward planning

Forward planning refers to a process of manually manipulating beam orientation (assuming a finite number of beams) and beam profiles to achieve the best dose distribution. Beam orientation is chosen by the planner based on beams-eye-view analyses. Beam modification is then determined using *wedges*, high density material that can attenuate the beam across the treatment region (Figure 2.7). All choices are based predominantly on the planner's knowledge of modifications that will yield changes congruent with clinical goals. This method is intuitive given modifications to the plan result in expected changes and the overall dose distribution is directly managed including hot spots. This method is also less computationally expensive when compared to inverse planning.

However, with this planning method, choices are dependent on the planner's experience and the process can be time consuming. It is also difficult to ascertain the influence of smaller changes in the plan on dose distribution. With the adoption of advanced delivery techniques such as IMRT and VMAT, the increased degrees of freedom enable improved modulation of the beam field and more effective delivery. However, due to the increased number of considerations such as the large number of aperture shapes to be



(a) (b)

Figure 2.8: An example fluence map for a single beam indicating the relative intensity across the field. (a) indicates the ideal theoretical fluence of the beam field and (b) indicates a deliverable fluence map when machine parameters are considered. Source: Rocha et al. (2011)[7].

chosen and the weight of individual beamlets given non-uniform fluence, the process is significantly more computationally expensive when using a forward planning approach.

2.11.2 Inverse planning

Inverse planning refers to the choice of feasible machine parameters given some desirable dose distribution. Typically choices are not made directly by the planner but using a TPS that explores various configurations. This method has the advantage of not requiring as much hands-on planning time and can be considered less subjective than forward planning. Nevertheless, the process is dependent on PGs and hot spots and non-standard areas of avoidance are not inherently managed. This thesis will focus on optimisation using inverse planning techniques and inverse planning will now be discussed in the more detail.

2.12 Inverse plan generation process

A TPS will commonly use one of two inverse planning approaches to produce deliverable plans: 1. fluence map optimisation (FMO) with subsequent leaf sequencing or 2. direct machine parameter optimisation (DMPO) also known as direct aperture optimisation. Technical details of plan generation will be discussed later in this section but a *fluence map* is a beam intensity profile that defines the variance in fluence across the beam field when the beam is modulated. Figure 2.8 shows an illustration of a fluence map with intensity varying across the field. Leaf sequencing defines MLC configurations necessary

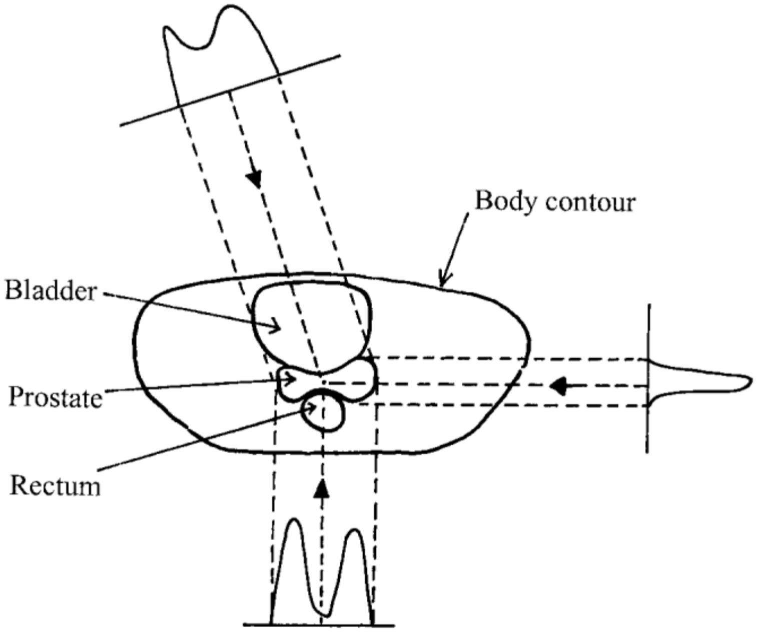


Figure 2.9: An illustration of three beam fluence maps for prostate treatment. Fluence maps are optimised to spare OARs and treat the target volume. Source: Webb et al. (2003)[8]

to achieve this fluence map and is achieved using either a “close in” method where leaves move towards each other or a “sweep across” method where all leaves move in one direction. The DMPO method considers MLC positioning during optimisation. Nevertheless regardless of the optimisation method used, all plans are generated and tailored with respect to beam angles. Beam angles refer to the predefined coplanar positions around the patient that are selected for dose delivery. In VMAT, all angles of the patient are subject to dose delivery and beam angles are usually not explicitly defined. With IMRT, a finite number of beam angles are defined and the selection of these will be discussed in more detail later.

For context, a simple IMRT plan can contain, for example, two parallel opposed beam fields each with uniform fluence. This is the simplest possible plan that contains two beams and such a plan has a number of advantages. It is firstly not very computationally expensive and can be be produced very quickly. It is also easy to ensure the treatment volumes obtain the necessary dose coverage. Nevertheless, opposing fields are not always advantageous for sparing regions outside of the treatment volume.

Prostate planning will be discussed in section 3.6 and chapter 4 but Figure 2.9 shows a schematic of a prostate patient treated with three intensity modulated beams. In contrast to Figure 2.8 which shows a full fluence map for a given beam, Figure 2.9 illustrates a flu-

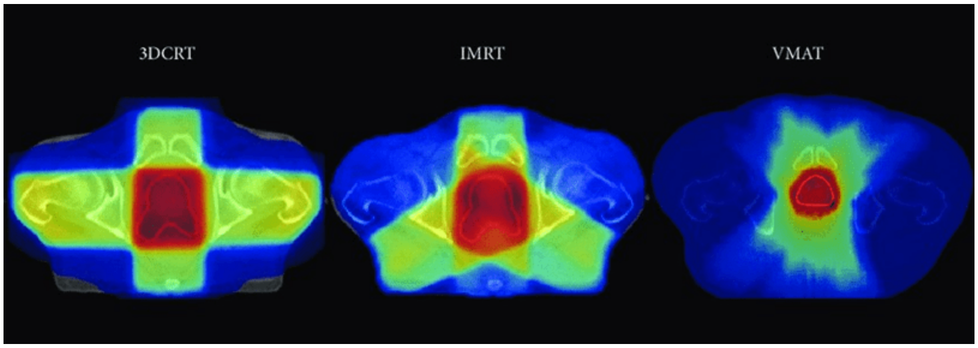


Figure 2.10: An example depiction of a planning dose distributions for the prostate treatment site using different treatment methods. In the example a transverse slice of the prostate is shown with 3D-CRT (left) ensuring dose coverage to the treatment region using opposed fields. IMRT (middle) applies dose modulation and an odd number of beam fields to spare radiosensitive regions such as the rectum and femoral heads. With modulated dose delivered at all angles, VMAT (right) achieves the greatest sparing and conformality of dose to the treatment region whilst ensuring appropriate treatment volumes dose coverage. Source: Mahatma Gandhi Cancer Hospital & Research Institute[9]

ence map profile in a sagittal plane. In this image, beam modulation is designed to spare the rectum and bladder whilst maximising dose to the treatment target (the prostate).

Figure 2.10 shows example prostate plan dose distributions for three EBRT methods. The left-hand pane shows a 3D-CRT plan with two sets of perpendicular parallel opposed beams, the middle pane shows an IMRT plan with three beams and the right-hand panes shows a VMAT plan. This illustration shows modified beam angles with IMRT can lead to comparable coverage of the treatment structure as seen with 3D-CRT whilst leading to greater sparing of radiosensitive regions including femoral heads to the left and right of the target and the rectum posterior to the target. Avoidance of these regions correlates inversely with incidence of radiation induced side effects such as reduced blood flow to the femoral heads (avascular necrosis) and rectal inflammation leading to discomfort and impaired function (radiation proctitis).

Therefore, to produce a plan congruent with oncologist preferences and maximise the therapeutic ratio, the plan generation process in IMRT and VMAT inverse planning involves the following general stages[62]:

1. Beam optimisation

    - Beam Orientation Optimisation for IMRT - beam orientation is usually predefined to be an odd number of equi-spaced coplanar beams determined by local practice but can be changed at the planners discretion using beams-eye-view analysis

- Control point optimisation for VMAT - a standard finite number of *control points* are defined along the gantry trajectory. One of the two inverse planning approaches is applied at each control point with MLC positions interpolated between control points[63]. The number of control points is often predefined and modified when needed by the TPS or amended manually by a planner. Typically a TPS will modify the number of control points by adding intermediate points when notable jumps in MLC sequences are observed between existing points.

2. Planning method

- Fluence map optimisation and leaf sequencing - fluence maps are defined by discretising beam fields into a series of beamlets each with its own fluence dependent on expected dose deposition to the patient. The MLC configuration necessary to achieve these fluence maps are usually determined via an optimisation algorithm known as *gradient descent*
- Direct Machine Parameter Optimisation - MLC configuration is usually determined via an optimisation algorithm known as *simulated annealing*

FMO mathematically accounts for the expected dose deposition of each beamlet in each patient voxel where a voxel is defined as a discretised unit volume within the patient. It can be defined:

$$z_{ij} = \sum_{i \in \mathcal{B}} D_{ij} x_i, \quad \forall s \in \mathcal{S}, \mathcal{V}_s \in s, j \in \mathcal{V}_s \tag{2.6}$$

where $s \in \mathcal{S}$ denote ROIs, $\mathcal{V}_s$ denotes the voxels in ROI s and z_{ij} is the dose deposited by beamlet $i \in \mathcal{B}$ of fluence x_i to voxel $j \in \mathcal{V}_f$. The matrix D_{ij} contains the dose deposition coefficients: the fraction of dose produced by beamlet i that reaches voxel j. Figure 2.11 illustrates the FMO process given nine MLCs and two voxels.

2.13 Inverse planning objective functions

Applying an inverse optimisation first requires definition of a target dose distribution. This can be defined mathematically with respect to voxels and beamlet intensity. To

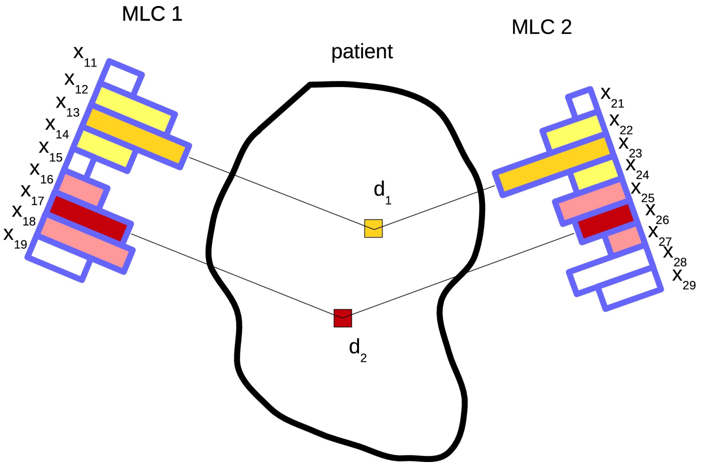


Figure 2.11: An illustration of a beam arrangement depicting the contribution of two beamlets to the resulting fluence map. Image sourced from Breedveld et al. (2019)[10]

optimise dose, PGs can be used as parameters to modify some target objective function (or cost function) to meet predefined target objective values.

Examples of standard optimisation objectives include:

- Maximum/minimum dose to the whole ROI
- Maximum/minimum mean dose to ROI
- Dose volume histogram (DVH) maximum/minimum dose to ROI

A dose volume histogram or DVH illustrates the minimum radiation dose received by a certain percentage of a volume. They are traditionally presented as line graphs where any one point on the line indicates that at least $y\%$ of the ROI volume receives x Gy or more. See Figure 2.12 for an example. They do not contain spatial information but elegantly summarise 3D dose distribution for individual ROIs and points on a DVH can be used as optimisation objectives.

Inverse planning algorithms formulate *objective functions* (also known as a cost functions) that can be used to solve for voxel level dose to ROIs and produce a model of dose distribution across the patient. A standard optimisation problem defines a scenario in which more than one solution is feasible but not all are necessarily desirable. In order to converge on desirable solutions, it is necessary to define the constraints and other objectives.

Constraints define characteristics of the solution space that must not be violated. Objectives define a hierarchy of desirable characteristics that may be violated but only in

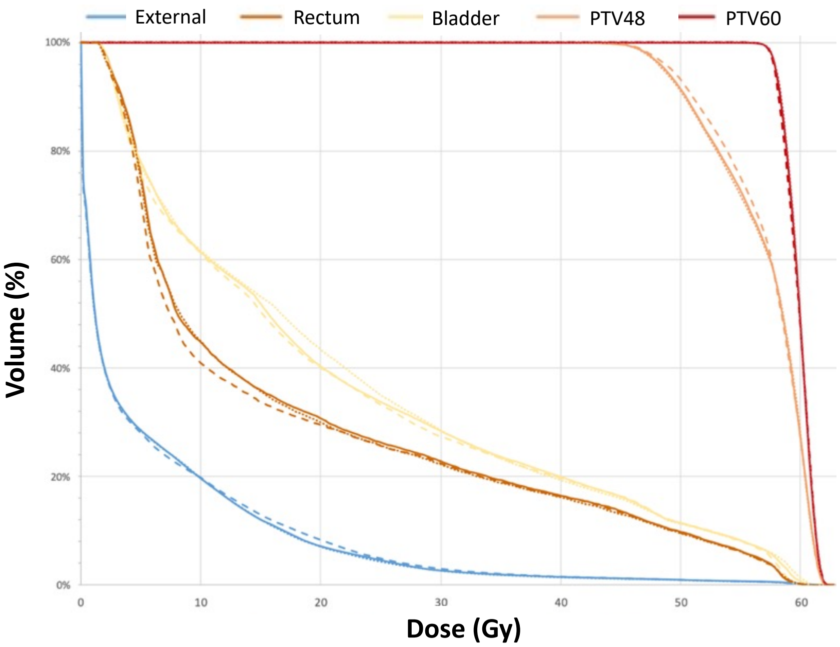


Figure 2.12: An example plot showing DVHs for five ROIs (colours) for three different plans (lines).

favour of meeting constraints and higher priority objectives. The purpose of an objective function is to provide the planner with a mechanism to minimise a *target objective value* for each ROI. A target objective value is a decision variable defined by the planner (or standard local practice) for each ROI that determines the dosimetric "cost" of any one solution where cost is defined as the difference between the actual outcome and the desired outcome. Target objectives represent the supposed most clinically desirable prioritisation of objectives and constraints such that any trade-off relationships between PGs are managed. In practice, objectives and constraints are defined using PGs in the planning protocol. Treatment volume target objective functions are optimised to maximise the prescribed dose to the volume. That is, actual dose close to the prescribed dose results in a low target objective value. The converse is the case for OARs. To obtain an overall performance metric of any one optimised plan, a *composite objective value* is obtained by adding together all target objective values for individual ROIs.

The objective function is defined by the TPS and will usually be a *convex function*. That is, a function that has one and only one solution that will minimise the objective function for any set of PGs. For each configuration of parameters, the algorithm returns a single composite objective value and the aim is to find the configuration of parameters that minimises this value. A widely used composite objective function is a quadratic function of the form:

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} w_i \Big(D(x_i) - D^P(x_i) \Big)^2 \tag{2.7}$$

where N is the number of PGs and D and D^P are the delivered dose and prescribed dose respectively[64,65]. Values for D^P are determined by the target objective function and D are determined by the actual achieved dose distribution. The composite objective function, $f(\boldsymbol{x})$, applies a stricter penalty the further D deviate from D^P.

For example, consider a planning scenario containing the following three optimisation objectives only: PG1 - minimum dose of 58Gy to the PTV1, PG2 - maximum dose of 62Gy to PTV1 and PG3 - maximum dose of 5Gy to OAR1. The set of D^P values for this example are (58,62,5). Given PTV1 receives 59Gy and OAR1 receives 7Gy and assuming $w_i = 1$ for all i, the composite objective function value will be $1^2 + (-3)^2 + 2^2 = 14$. However, the influence of any one PG on the composite objective function can be managed by w_i, the importance factor or optimisation weight. Both optimisation objectives and optimisation weights can be modified to influence dose distribution and these modifications will be explored in more detail in the following sections.

Minimisation of the objective function is achieved through implementation of one of the following three types of optimisation algorithm:

- **Emuneration methods**: exhaustive search technique. When the solution space contains a countable number of solutions, it is possible to consider each individually to identify the optimal solution

- **Gradient descent methods**: sample points are taken on either side of the current chosen point and the gradient size and directions is used to determine the choice of the next point

- **Random search methods**: a stochastic strategy is used to search the solution space and converges based on some probabilistic model e.g., particle swarm, simulated annealing and genetic algorithms.

2.14 Multi-criteria optimisation

As mentioned, planning goals lead to conflicts and trade-off relationships to be managed and prioritised. This can be parsed as a multi-criteria optimisation problem. The most clinical desirable plan is a feasible plan that has congruence with oncologist preference and maximises the therapeutic ratio. This is achieved by manipulating PGs and PG optimisation weights. However, manipulation of objective and/or their weights can influence the resulting plan in different ways.

2.15 Pareto optimality

A plan is said to be Pareto optimal when no further dosimetric improvements can be made by manipulating any one PG except to simultaneously lead to a detriment for another and the most clinically desirable plan is considered to be Pareto optimal. Planners therefore aim to converge on these plans when adapting PGs prior to inverse planning.

Traditionally, PGs are updated by the planner between iterations of the TPS inverse optimisation algorithm. Beginning with a standard set of PGs, planners will present stricter and stricter PGs to the optimiser until eventually no further changes can be made that will improve the dose distribution. This trial-and-error process can be considered the clinical standard. However, Pareto optimality is not guaranteed with this method and even achieving a Pareto optimal solution does not imply clinical desirability.

The general form of the MCO problem is given by[66]:

$$\begin{aligned} \min_{\boldsymbol{x}} \quad & f(\boldsymbol{x}) = [f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), ..., f_n(\boldsymbol{x})], && (2.8)\\ \text{s.t.} \quad & g(\boldsymbol{x}) = [g_1(\boldsymbol{x}), g_2(\boldsymbol{x}), ..., g_m(\boldsymbol{x})] \leq 0, && (2.9)\\ & \boldsymbol{x} > \boldsymbol{0}, \\ & n, m \in \mathbb{N}. \end{aligned}$$

where $\boldsymbol{x}$ is a feasible solution in the optimisation space, the $f(\boldsymbol{x})$ are objective functions for each optimisation objective and the $g(\boldsymbol{x})$ are constraints on those objectives.

2.16 Plan calibration

Given a Pareto optimal solution in which all values of w_i in equation 2.7 are equal, no one PG takes precedence over another. Given the dosimetric characteristics of a plan is not known apriori, the best configuration of w_i to deliver an acceptable distribution is notably difficult to define. When incorrectly balanced, these values can lead to clinically subpar planning likely to lead to complications if delivered. Therefore management of this during planning is vital.

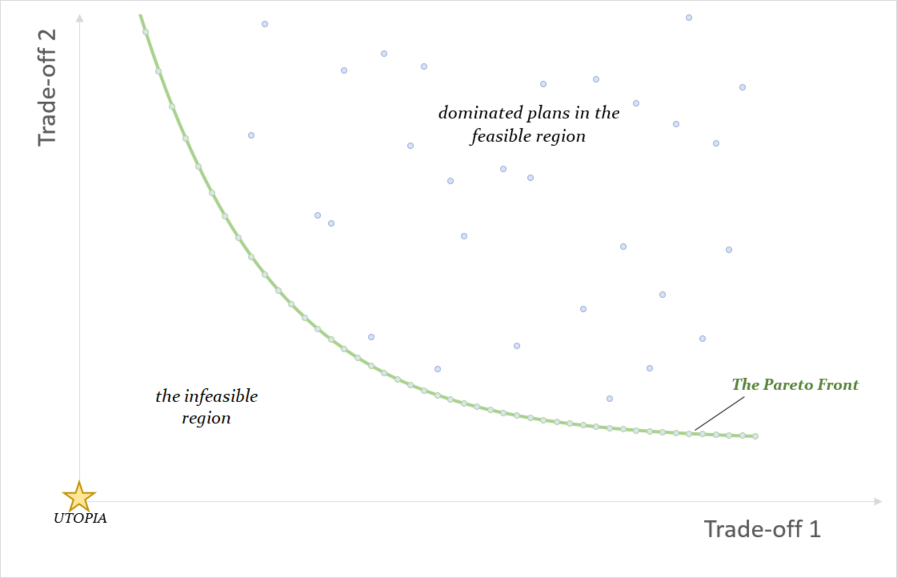


Figure 2.13: An example of a typical Pareto front relationship defined by dose in Gy of two competing trade-offs e.g., two organs-at-risk. Adapted from Rebello et al. (2021)[11]

Consider Figure 2.13. Illustrated is a bi-variate case (two PGs) showing the feasible planning space bounded by what is known as the "Pareto Front". This is the set of plans for which no further improvements can be made. Dependent on the choice of optimisation weights, Trade-off 1 may be spared to a lesser or greater amount with respect to Trade-off 2. The most desirable solution will lie somewhere on this front with the chosen plan dependent on the calibration of optimisation weights and the clinical goals. It is common for optimisation weights to be assigned via a locally defined planning protocol based on previously planned cases. The implication is that patient geometry is similar on average and this approach has been adopted with some success for a number of sites. Moreover, it has been suggested that dose distribution is geometry based[67,68] and therefore the ideal calibration of optimisation weights may differ greatly from patient-to-patient. Therefore, the application of a universal or site-specific approach is not appropriate and weights must be modified on a per-patient basis to ensure clinically acceptable and optimal plans.

2.17 Automated planning

Given the reliance of planning on manual techniques, truly objective planning cannot be guaranteed with standard methods. Therefore, there may be a margin of planning error due to standard methods and this is difficult to measure especially in the absence of a clinically relevant alternative approach. In addition, planning comes with a number of inherent challenges including the need for expert practitioners with knowledge of the treatment site in question and the TPS, as well as appreciation of oncologists preferences.

Automated planning (AP) is a new innovation of radiotherapy planning in which plans are not configured manually via human planning but automatically. All AP approaches can be considered algorithms and have the same basic requirement: to generate plans non-inferior to manual planning. In this regard, the success rate of existing techniques has lead the field of AP to gain traction with a boost in literature and research in recent years. However, existing AP technology is still being developed and researched with many applications still requiring expert-driven calibration or subsequent improvement by a human planner prior to clinical use. The ultimate goal is to establish a full AP system that consistently and reliably delivers clinically applicable planning. This will provide a benchmark for the comparison of standard methods and a foundation for improvements in these methods.

Chapter 3

Automated planning

3.1 Issues related to manual planning

The state-of-the-art in radiotherapy treatment enables clinically effective planning in ways previously not possible. Nevertheless, even with modern technology and inverse optimisation algorithms, planning relies heavily on manual manipulation of optimisation parameters. Issues include:

- **Expert human time that could be utilised differently**: Manual inverse planning always results in patient-tailored plans and expert-driven planning with appropriate quality assurance measures ensure sure plans are fit for treatment. However, manual methods follow a trial-and-error process that can be time consuming especially when planners are presented with uncommon cases. If plans could be produced without the need for expert knowledge, planning duties could be delegated to a wider team hence a more efficient use of resources

- **Planning time and efficiency**: Irrespective of the time taken to produce a plan, it can only be performed for one patient at a time. If the knowledge used to produce a single plan could be leveraged to produce plans for other cases, overall planning time for the subsequent patients could be reduced and human planner time used more efficiently

- **No guarantee of Pareto optimality**: Convergence on a Pareto optimal solution can be difficult to ascertain with a trail-and-error method. Given a means of limiting the solution space to plans adjacent to the Pareto front only, convergence on a Pareto optimal solution becomes more likely

- **Subjectivity, inconsistency and/or variability in planning methods**: Manual methods are planner dependent and subject to discrepancies due to the variability of human choices. Also, the trial-and-error process requires patience and familiarity with the treatment site for the production of a plan congruent with oncologists preferences. Furthermore, the standard method of comparison for one plan over another is the judgement of the oncologist and this can be a difficult choice when both plans meet the clinical goals. It can also be unclear whether the final plan is among the most optimal set of plans or can be improved further. Even when a plan is produced with no time limitations, an optimal result is not guaranteed. An objective planning method would lead to consistency with all patients receiving treatments planned with the same integrity

Some of these issues can be resolved with increased automation. This chapter will discuss current AP methods with respect to these issues as well as developments in existing methods. The literature contains a variety of AP techniques that have been gaining traction with most falling into at least one of two main categories: knowledge-based planning (KBP) and rules-based planning (RBP). These methods will now be outlined in more detail.

3.2 Knowledge-based planning

KBP techniques are developed using information obtained from previous clinical planning and itself falls into two categories[69]: model-based and atlas-based KBP. Atlas-based KBP involves matching the current case to a case in the knowledge-base. Matching is often dependent on the similarity of key attributes of the patient case such as OAR sizes and positioning. The case in the knowledge-base is then used to inform planning of the new case. Model-based techniques, however, are generated using multiple cases. This can be done in a number of ways including the use of statistical and ML techniques. Documented approaches to these methods will now be discussed.

3.2.1 Atlas-based KBP

Atlas-based KBP has a few stages. First, key attributes of plans are defined. These are chosen based on assessment of planning characteristics that are influential to dose distribution. Then, a method for comparing cases to each other is defined including some measure of similarity. Lastly, the existing solution is applied to the new matched case.

Key attributes are often anatomical variables that have been locally defined by an institution. As planning is a strongly geometry-based task, the selection of attributes found in the literature are primarily spatial. Studies have included features such as the distance between ROIs, angle of incidence between ROIs and the beams-eye/in-field view of the overlap between ROIs[70–73].

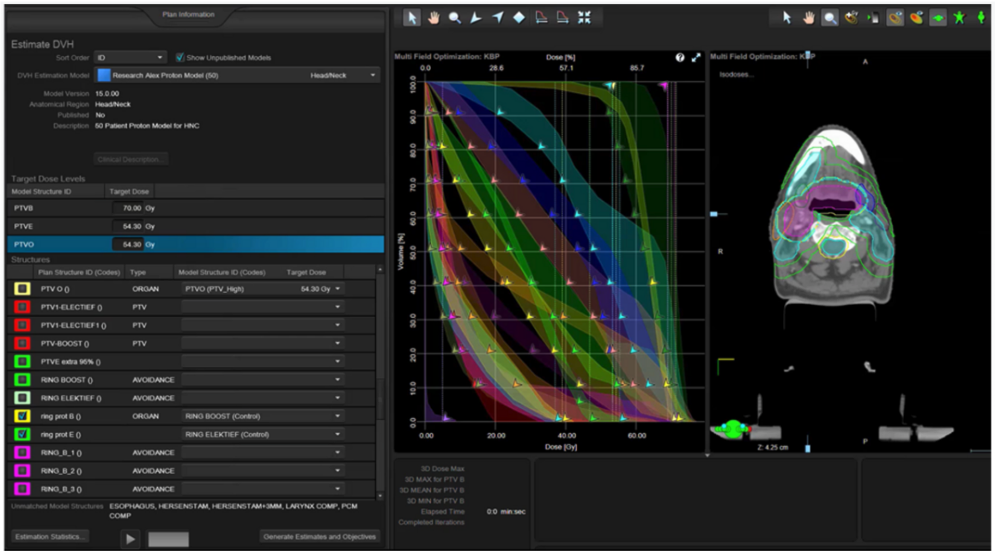
Once key attributes have been defined, new cases outside of the knowledge-base are assigned to an already-planned case within the knowledge-base. Petrovic et al. (2016) used a weighted nearest neighbours approach where each attribute was an empirically weighted continuous variable and new cases were assigned based on the smallest sum of the weighted differences[70]. Chanyavanich et al. (2011) defined similarity based on 2D beams-eye view images used to calculate a similarity metric in a manner akin to a conformality index[73]. Sheng et al. (2015) based similarities on PTV shape and size alone[72].

Once a reference case has been identified within the knowledge-base, there are a few ways this can be used to inform planning for new cases. Typical examples include the use of a DVH, use of a 3D dose distribution[74,75] or application of an existing planning protocol. For example, a DVH or 3D dose can be used to define PGs for the new patients and objective weights defined in the original planning protocol can be adopted. Regarding 3D dose distribution, one study considered the comparison of three deformable image registration methods to the development of KBP 3D dose distribution prediction for left-sided breast cancer[76] and concluded all methods to be appropriate for clinical use. Atlas-based KBP has many advantages and these will be discussed in section 3.4.

3.2.2 Model-based KBP

Some of the most well established KBP approaches apply model-based methods. Potentially the most well-known KBP method in the literature is DVH prediction such as that found in the commercially available Varian RapidPlan™ software (Varian Medical Systems, Palo Alto, USA). This KBP method is available within the Varian Eclipse TPS and has been widely used and researched showing its planning efficacy in a clinical setting[77–81].

Such applications require a training phase for the ML model to be generated with respect to all data points in the knowledge-base. With RapidPlan for example, in each



Plan Information

Estimate DVH

Sort Order: ID

☑ Show Unpublished Models

DVH Estimation Model: Research Alex Proton Model (50) Head/Neck

Model Version: 15.0.00

Anatomical Region: Head/Neck

Published: No

Description: 50 Patient Proton Model for HNC

Clinical Description...

Target Dose Levels

Model Structure ID	Target Dose
PTVB	70.00 Gy
PTVE	54.30 Gy
PTVO	54.30 Gy

Structures

	Plan Structure ID (Codes)	Type	Model Structure ID (Codes)	Target Dose
☐	PTV O ()	ORGAN	PTVO (PTV_High)	54.30 Gy
☐	PTV1-ELECTIEF ()	PTV		
☐	PTV1-ELECTIEF1 ()	PTV		
☐	PTV-BOOST ()	PTV		
☐	PTVE extra 95% ()			
☐	RING BOOST ()	AVOIDANCE		
☐	RING ELEKTIEF ()	AVOIDANCE		
☑	ring prot B ()	ORGAN	RING BOOST (Control)	
☑	ring prot E ()		RING ELEKTIEF (Control)	
☐	RING_B_1 ()	AVOIDANCE		
☐	RING_B_2 ()	AVOIDANCE		
☐	RING_B_3 ()	AVOIDANCE		

Unmatched Model Structures: ESOPHAGUS, HERSENSTAM, HERSENSTAM+3MM, LARYNX COMP, PCM COMP

Estimation Statistics... | Generate Estimates and Objectives

3D Dose Max

3D MAX for PTV B

3D MEAN for PTV B

3D MIN for PTV B

Elapsed Time: 0:0 min:sec

Completed Iterations

Figure 3.1: An example of a Varian RapidPlan predictive DVH plots a head and neck case with 12 ROIs. Sourced from Varian[12]

case the ROIs are segmented into four different regions based on their relation to the PTV. The software then generates a series of geometry-based expected dose (GED) histograms defined as the dose received by a portion of an OAR based on its distance from the target. Using a combination of principal component analysis and regression, the software uses GED values and the combined DVHs of the OAR sub-segments to generate DVH predictions when presented with new cases[82] (Figure 3.1). Using the DVH prediction, the PGs of the current case are overlayed and assessed. If the planner finds further adjustment is required for the case at hand, further manual planning can follow.

Other DVH-based KBP models have been proposed such as using a probability density function to estimate points along the DVH curve given the value of the point preceding it[83]. Others generated DVH predictions using a dose-distance relation that implies dose to an OAR will diminish the further away it is from the target[84].

The DVH approach is prominent in the literature, however, DVHs do not provide spatial information and 3D dose prediction models may be advantageous. For example, studies have shown the use of artificial neural networks trained to predict dose matrices. A proof of concept study trained voxel geometry to voxel dose for prostate and stereotactic radiosurgery cases. For comparison, DVHs were derived from the 3D models and they found 3D dose prediction improved on existing DVH-based predictions with respect to overall deviation from prescribed doses[85]. There are also newer approaches that use deep learning for improved 3D dose prediction. For example, a study has considered

IMRT beam configuration in addition to anatomical features and has indicated a substantial benefit of bespoke beam configurations regarding reduced dose outside of the target region[86]. Studies have also demonstrated the use of reinforcement learning such as adapting existing reinforcement learning models used in other fields (e.g., gaming) to help define machine parameters based on patient contours. Results show comparable dosimetric outcomes to clinical planning[87].

3.2.3 Implementation issues for KBP

A major barrier to KBP planning is the acquisition of a repository or database of planning examples to draw from. In addition, ensuring the diversity of cases (such as a range of anatomical variations) can be difficult and may lead the underlying model to be skewed by outliers. Alternatively an approach could be to choose only what is considered high quality planning. That could be, for example, cases that have been specially considered by the planning team and oncologist. Nevertheless, the composition of the knowledge-base is a key consideration for the efficacy of a KBP approach and this will also be discussed in section 3.4.

3.3 Rules-based planning

RBP methods use logic to converge on a solution and different approaches can be found in the literature. All RBP methods rely on some predefined processes and the two most well documented are epsilon constrained methods (also known as constrained hierarchical optimisation) and protocol-based automatic iterative optimisation methods.

3.3.1 Epsilon-constrained RBP

Epsilon constrained (ϵc) optimisation refers to methods of converting unconstrained optimisation problems into constrained scenarios that are more easily managed. In radiotherapy planning this refers to the constraint of certain PGs.

A key ϵc approach is known as a lexicographic ordering. With this approach, PGs are optimised sequentially in an order given by some predefined lexicon of PGs. The lexicon of PGs is locally agreed and designed to correspond with the preferences of the oncologist. The process begins with the single most high priority PG and the TPS optimisation run to generate a plan for this PG irrespective of any others (i.e., a univariate optimisation). Once complete, it is set as a constraint and the next PG considered. All

PGs are optimised in this way until all have been optimised. The logic is, the process should result in a single Pareto optimal plan that takes into account the preferences of the oncologist and versions of this AP method have been implemented[88,89]. A well-known lexicographic ordering is found within the Erasmus-iCycle system[90], a fully automated planning solution.

3.3.2 Protocol-based automatic iterative optimisation

Protocol-based automatic iterative optimisation (PBAIO) methods are algorithms that adapt and update planning parameters during optimisation. These iterative optimisers will often operate predominantly by mimicking a human planning adjustment style of PGs between runs of the inverse optimiser. PBAIO is therefore characterised by dynamic objective adjustment.

Some research has aimed to emulate this directly using recorded scripts of the logic applied during manual planning[91]. Manual planners would follow a standard iterative planning method and the process recorded using C# scripts. Using a standard coordinate system applied overlaid on a CT, the same logic can be applied to new patients by scaling the coordinate space to fit the new anatomy.

Most PBAIO approaches manipulate PGs using a standard approach. Early research explored the concept of a fast monotonic-descent algorithm coupled with a “fuzzy weight function”[92]. PG weighting factors were assigned a value between zero and one that was explicitly defined based on the value of the prescribed dose of the PG. The closer the prescribed dose of the dynamic PG is to the upper limit of the prescribed dose for the ROI, the higher the value of the weighting factor. Prescribed doses were then iteratively updated using a script that adjusted them by some arbitrarily defined delta value (small number) between runs of the inverse optimiser with a cost function assessed for each iteration to determine the optimal solution. This approach aimed not only to achieve Pareto optimality but also to manage the balance of trade-offs between PGs to converge on clinically desirable plans. This approach had theoretical benefits that directly deal with some of the issues identified for manual planning such as convergence on Pareto optimality.

However, all aspects relating to dose distribution cannot always be directly handled by manipulating standard PGs. For example, even Pareto optimal plans may contain extreme hot spots that are undesirable for clinical solutions. Researchers have developed similar iterative optimisers to those mentioned that automatically segment regions containing hot and cold spots between runs of the inverse optimiser[93]. The new volumes are then automatically assigned a PG and an objective weight to help manage the overall dose distribution and maintain dose uniformity across regions.

Some PBAIO approaches have been developed using scripts. For example, the AIO was developed using the Lazarus Pascal compilers. It functions by manipulating dose-volume objectives, moving them a specified increment at the start of every new pass[94]. This is a similar approach to manual human planning but adjustment of PGs is performed consistently and automatically. Other similar approaches to this include the commercially available Auto-Planning within the Philips Pinnacle TPS[3][95] and the Experience Driven Plan Generation Engine (EdgeVcc) by Velindre Cancer Centre developed using RaySearch's Raystation TPS[96]. Auto-planning is a software built with reference to the penalty scheme developed by Cotrutz and Xing[97]. PGs related to OAR are managed by the user who categorises them into one of three groups: high priority, medium priority and low priority. Although this system contains a proprietary algorithm, it is known to automatically generate new contours during optimisation to help meet clinical goals[95]. For example, given a large overlap of a high priority OAR with a target treatment volume, the algorithm will adjust the relative prioritisation of the overlap region to ensure dose to that region is limited for great sparing of the OAR. That is, given the original list of PGs, additional structures are iteratively added to the planning protocol and automatically assigned an objective function weight. This is done to help meet the highest priority PGs. However, this thesis will be focused on the EdgeVcc[96] and this approach will be discussed in section 3.3.2.1.

3.3.2.1 EdgeVcc

For a full and comprehensive description of the Experience-Driven plan Generation Engine by Velindre Cancer Centre (or EdgeVcc), refer to Wheeler et al. (2019)[96]. Plan generation is dependent upon a base site-specific "AutoPlan protocol" containing a set of PGs and this will be used by the PBAIO system to interact with the TPS native optimiser. Within the AutoPlan protocol, PGs are assigned to one of three priority levels: primary normal tissue goals (P1), target goals (P2) and trade-off goals (P3). Target PGs (P2) ensure target volume dose objectives are met including PTV coverage and hot spots. All other planning objectives are known as trade-off PGs (P3). Each PG is assigned a numeric weighting factor that the PBAIO AP solution will use to determine prioritisation of each objective during plan generation. The algorithm explicitly defines weighting factors for P1 and P2 goals and a dynamic objective algorithm applied to help balance P3 goals. Essentially, EdgeVcc is a PBAIO process that applies a dynamic adjustment of TPS optimisation weights used to defined the target objective function of the optimisation and therefore sets the level of priority for each PG whilst maintaining an appropriate hierarchy of PGs in the underlying algorithm. Weighting factors can be thought of as relative values and PGs assigned high values receive a high relative priority over those with lower values. This method also incorporates a novel Pareto navigation based calibration process for defining the weighting factors for P3 PGs that will be discussed more in section 3.5.3.

The user-defined AutoPlan protocol denotes P1, P2 and P3 PGs with respect to ROI names as found in the patient case in the TPS. For geometric specificity, auxiliary optimisation ROIs (AuxROIs) are generated for use in the algorithm and each applied a TPS objective function optimisation weight for use in the native inverse planning algorithm. These weights for each AuxROI are derived of the weighting factors in the following way:

$$\mathrm{w_{TPS}} = \mathrm{w_{nom}}\ \mathrm{F_V}\ \mathrm{F_T}\ \mathrm{F_C}\ \mathrm{F_N}, \tag{3.1}$$

where $\mathrm{w_{nom}}$ is the weighting factors and $\mathrm{F_V}$, $\mathrm{F_T}$, $\mathrm{F_C}$ and $\mathrm{F_N}$ are scaling factors. Planning experience indicated optimal objective function weights were dependant on ROI volumes therefore $\mathrm{F_V}$ is derived of the volume of the ROI. $\mathrm{F_T}$ is a correction scaler to offset strict penalties related with the Raystation TPS objective function algorithm. $\mathrm{F_C}$ is a hard coded constant relating to PTV AuROIs to ensure the necessary balance in priorities between P1 and P2 PGs. $\mathrm{F_N}$ is a normalisation factor relating to P3 goal volumes and is usually set to 1.

PTVs are each subdivided into three AuxROIs:

- $\mathrm{PTV_{SV\text{-}1}}$ - retracted from the skin and proximal P1 OARs
- $\mathrm{PTV_{SV\text{-}2}}$ - PTV within the skin or extending into air
- $\mathrm{PTV_{SV\text{-}3}}$ - volume not covered by PTVSV-1 or PTVSV-2 i.e. PTV proximal to primary OARs

For P1 and P2, weighting factors values (w_{nom}) are defined using hard-coded algorithms because clinical preference between P1 and P2 is considered well defined across all tumour sites. Therefore, the balance of conflicting P1 and P2 PGs is a fixed relationship with P2 goals compromised in favour of P1 goals.

P3 weighting factors values can be derived in one of two ways. Either via an already defined set of weighting factors such as values defined for a previous patient or, a bespoke weighting factor set derived via Pareto navigation. The details of this will be discussed in more detail in section 3.5.3.

Once weighting factors have been assigned to all PGs, in the PBAIO framework P3 PGs prioritisation is updated dynamical. That is, their target objective value within the inverse optimiser is altered after FMO. For each P3 PG, target objective values are defined:

$$T = D_c - 0.35D_{pres}, \tag{3.2}$$

where D_c and D_{pres} are the current and prescribed dose parameters corresponding to the PG respectively. After each pass, the target objective value is reassessed using the follow equation:

$$\Delta = \frac{D_c - T}{D_{Pres}}. \tag{3.3}$$

The criteria for PBAIO termination is $\Delta \in [0.15, 0.5]$ or $\Delta \in [0, 0.5]$ if $T = 0$. Otherwise, values of T are updated and a new pass run. For termination, the aim is to ensure D_c is within an acceptable tolerance of D_{Pres} for all P3 PGs. Values for Δ have been defined empirically and were derived based on experience from manual planning. The lower bound where $T \neq 0$ (i.e., 0.15) is considered to be a strict tolerance criteria in which the PGs target objective value indicates a high priority to meet the prescribed dose, 0.5 is the lowest tolerance and 0.35 is midway between the two.

The resultant DVH where $\Delta = 0.35$ presents a starting point to begin updating objective positioning. Upon definition of these tolerances, it was observed that between these values, a similar organ-at-risk DVH is obtained and this method was used for manual class solutions also[98].

3.3.3 Implementation issues for RBP

Calibration of PGs is an existing issue for many RBP approaches given the trade-off relationships must be defined prior to running the algorithm. The classic approach to managing this is to use trial-and-error[99–102]. With this approach an arbitrary weighting factor solution set is defined and manually manipulated until an appropriate solution is agreed upon. Trial-and-error will often involve the use of a small group of test patients on whom the configuration is based. This is an effective approach to RBP calibration but can be time consuming. Also, given the arbitrary starting position and the fact further improvements are made only with respect to previously tried examples, final solutions may be subjective. Newer approaches to RBP calibration have been proposed and these will be discussed in section 3.5.3.

3.4 Review and comparisons of automated planning techniques

AP has great potential for relieving some of the stress in the treatment pipeline and even improve on conventional methods. It has a number of general benefits including time and resource efficiency, and planning consistency.

Given all or part of the planning process is automated, the amount or time an expert would have been needed to complete the whole task is reduced. This frees expert planning time to be used elsewhere. Given multiple AP planning units with appropriate processing power, there is also the potential for more than one case to be considered at a time and faster than manual planning. Studies have shown planning times to reduce by 50%[103], 64%[104] and even up to 94%[105] when compared with conventional manual planning. Clinical planning is dependent on expert planning knowledge but conventional methods can be tedious for the planner. Automated solutions have the potential to surpass human planning by producing clinically desirable plans in lieu of hands on expert knowledge[106,107]. The consistency it brings is desirable for an institution as it can be difficult to obtain with conventional methods due to intra-and inter-planner variability. Standardised approaches help mitigate variability and ensure a structure for consistent logic to be applied.

KBP will always require a knowledge-base in which resulting solutions will be as reliable as the logic that was applied during original planning. It will also be limited by the homogeneity of the cases in the knowledge-base. In this way, there may be some bias in the KBP modelling process related to the composition of the knowledge-base as well as the objective quality of the plans within. Also, given advanced and convoluted modelling, KBP can fast lead to a black box scenario with the final choices and outcomes unknown to the user leading to ambiguity in plan analysis.

Despite the advantages, planning is not an exact science and all AP methods have pros and cons[108]. RBP methods drive the optimiser towards a Pareto optimal solution and KBP methods determine the best trade-off relationships. Therefore, when it comes to clinical desirability, convergence on a Pareto optimal plan is implied with most RBP methods but is not necessarily the case with KBP methods. Also, KBP methods may result in an acceptable balance of trade-offs but may not always converge towards the Pareto optimal solution[109] even when the knowledge-base is chosen to contain only what are considered highly desirable examples of Pareto plans[110].

The appropriate balancing and prioritisation of trade-off PGs is implied with KBP methods with no need to be explicitly defined by the user because this is determined by the previous planning in the knowledge-base. The application of the knowledge used in previous planning to novel cases implies a comparable balancing of trade-offs reflecting expert knowledge and oncologist preferences. RBP however, requires calibration that explicitly manages the relative importance of PGs. Calibrations are often developed empirically and are refined using a small set of up to 10 patients before application to the wider patient base[100–102]. This task is especially difficult given calibration parameters do not necessarily have direct dosimetric relationships and the result of any one calibration set is only realised once a plan has been generated. As a result, this task can take some time to perform and may even result in calibrations biased to the patient group it was refined for.

Given the outcomes of KBP are limited by the cases in the knowledge-base and resulting solutions dependent on the logic that was applied during original planning, a shortcoming of KBP includes the fact it will always require a substantial knowledge-base of clinical planning to be useful. This can be resource intensive and is still no guarantee of unbiasedness. Also, given advanced and convoluted modelling, KBP can fast lead to a black box scenario with the final choices and outcomes unknown to the user and causing some ambiguity in plan analysis.

However, knowledge of how a plan was generated is arguably irrelevant given the plan is ultimately approved and black box KBP may not be a wholly negative development. It is also a good way of leveraging the time and expertise that have already been employed in the clinic. Even given AP that is deemed “improvable” due to not being Pareto optimal, these solutions are still useful for relieving pressure in the clinic by removing much of the tedious trial-and-error of planning and providing a more useful starting point.

Both KBP and RBP methods are being used in clinics with no approach currently being categorically deemed superior over another. This is due partly to the inherent difficulty in generating what is known to be a reliable AP solution. There have been many studies comparing AP to manual planning and showing the notable benefits. However, comparison between AP solutions tend to show comparable outcomes with only incremental gains[96,110,111]. Nevertheless, in light of some of the existing pitfalls, it is known that existing methods of AP can be improved upon. In the next section, some of the recent developments on classic approaches are discussed.

3.5 Advances in AP

The increase in the number of AP studies in recent years shows the level of interest in the research community to fill gaps in the literature[108]. In fact this increased interest called for review and standardisation of practice to ensure quality and validity of findings and ensure further development in this area is reliable[102,112–114].

Given the dichotomous benefits of using either a KBP or RBP planning method, it can be hypothesised that an approach combining the two may mitigate some of the pitfalls of employing one on its own. As a result, hybrid approaches have emerged. Additionally, other ways of converging on clinically desirable planning is being explored. These are methods that have consonance with oncologists preferences and relate to intuitive ways MCO approaches. Some of these advances in AP will be discussed and compared here.

3.5.1 Advances in KBP

Given KBP methods often employ statistics and machine learning, the field of KBP for AP is becoming extensive. Research highlights potential developments on existing methods including the use of less conventional predictive variable and use of advanced modelling techniques. In addition to anatomical and clinical data, predictive features for use in KBP have emerged and include the addition of radiomic features[115], dosimetric features[116,117], objective function features[118–121] and neural network generated features[122–125]. Especially with the increase in readily available patient-specific data and outcomes, patient specific planning is considered a vital area of exploration of research[125].

3.5.2 KBP-RBP composite approaches

Currently, there are no novel end-to-end hybrid approaches that incorporate both KBP and RBP methods. However, examples of hybrid approaches are found in the literature that incorporate separate AP systems. For example, a composite approach was developed using two commercial packages in conjunction. The result is a solution that takes advantage of the benefits of both approaches to achieve a fully automated solution[126]. First, Varian's RapidPlan™ is used to generate a DVH prediction based KBP model. Using the resulting DVH prediction, PGs are derived and applied to a planning protocol for use in the Pinnacle Auto-Planning RBP system. This approach leverages and benefits from previous planning and the implied trade-off will reflect clinical preferences in local practice and converge towards Pareto optimal solutions. Such an approach not only benefits from the advantages of both, they each also mitigate some of the disadvantages of each and lead to a fully automated and clinically desirable solution.

As mentioned in Section 3.3.2, one PBAIO approach was founded on the recorded behaviour of expert planners that translated their actions into a scripts[91]. More recently, a similar PBAIO approach has been developed again founded on the recorded actions of expert planners but with a KBP overlay. It uses artificial neural networks for a deep-reinforcement learning approach to predict the iterative PG updates[127]. This approach combines the benefits of PBAIO with the power of deep learning and can be thought to not only deliver plans that converge to Pareto optimality but reflect clinical preference using an innovative approach.

However, the most popular KBP-RBP composite approaches involve using existing RBP methods calibrated using KBP. One example of this uses a probabilistic KBP method known as *kernel density estimation* to calibrate a PBAIO solution[128]. The KBP method assigns priorities to PGs by considering the conditional dependencies of voxels in related organs-at-risk to their distance from the surface of PTVs. The researchers used this method in conjunction with Pinnacle Auto-Planning and the method was successfully used to improve the quality and consistency of breast and rectal plans. Another approach clustered PG weighing factors of the training database into five weighting factor sets and used these for RBP planning of novel cases[121]. They found that some patients are more sensitive to weighting factor perturbations than others and that small number of weight sets.

3.5.3 MCO approaches

Pareto navigation has been widely identified as a beneficial approach to planning. Such *a posteriori* (retrospective) approaches allow for explicit exploration of the trade-off relationships between PGs and has been shown to enable the identification of highly clinically applicable planning[67,129,130]. Studies have emerged showing various methods incorporating Pareto navigation techniques into the planning process with successful results. These AP systems will be referred to as Pareto-guided automated planning (PGAP) systems.

Commercial MCO software has become available within TPS packages. For example, RaySearch Laboratories's Raystation® TPS research version 2.4.11 and Varian's Eclipse™ treatment planning system version 15.5 each started providing MCO planning functionality[131–134]. Varian's RapidPlan™ can now be used in conjunction with Varian's Eclipse™ MCO for a KBP-MCO hybrid approach also[135,136].

Outside of commercially available applications, research has been developed to show newer approaches for utilising MCO navigation techniques. For example, researchers have developed a technique called the Pareto optimal projection search (POPS) algorithm that defines the parameters of an ideal Pareto plan using a knowledge-base of previous plans and automatically generates and then navigates along the Pareto front to obtain a plan that best fits some predefined planning parameters[137]. This research applies a KBP-MCO approach for automated navigation of the Pareto front. This is desirable as the final plan will converge to a Pareto optimal plan as well as taking advantage of previous clinical planning for trade-off management of PGs.

A similar study used Pareto front exploration and previous planning to configure plans. The work built upon successful implementation of an ϵcapproach to prostate, known as the *two-phase ϵ-constraint method* (2pϵc), by applying a KBP configuration of the trade-offs[138]. The researchers generated linear approximations of the Pareto front and used the trade-off relationships found in the training database to automatically generate a configuration for prostate and identify a Pareto optimal plan. The results showed favourable median performance and smaller outliers over the previous version of the AP solution.

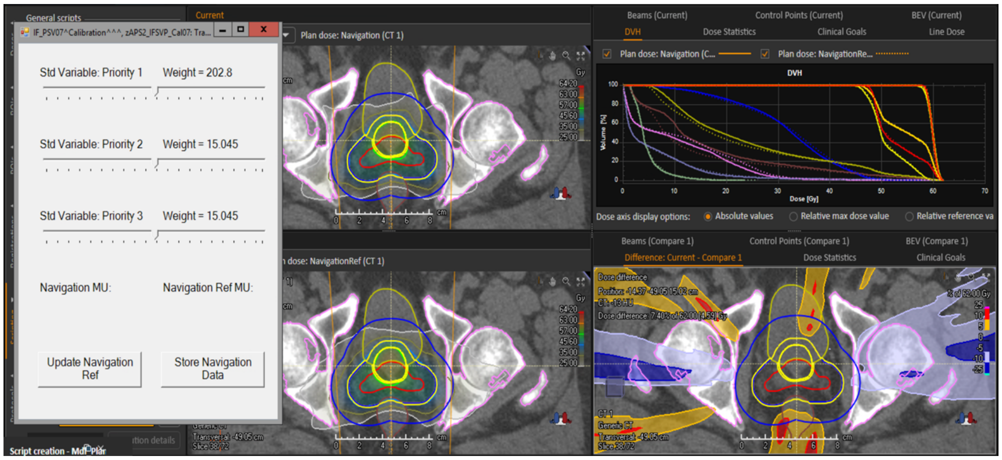
Other researchers have attempted to produce solutions that account for patient-specific trade-offs in the planning process. A new approach builds on an ϵcapproach[90] by showing the benefit of patient-specific wish-lists[102] as well as proposing an advanced technique called *NovelAPT* that uses weighted-sum cost functions to deliver plans of comparable quality to the aforementioned patient-specific planning[139]. This research shows MCO navigation can be used to help identify clinically desirable plans but that it is possible for *a priori* calibrated RBP methods to produce plans of comparable quality given advanced enough rules.

3.5.3.1 MCO approaches: EdgeVcc

The PBAIO planning system within EdgeVcc includes fully incorporated PGAP functionality allowing for an exploration of the wider planning space. The clinically applicable region of the Pareto front is sampled using a finite set of unique weighting factors and plans generated using PBAIO. A sliding interface is then used to navigate through the samples and select the weighing factor set resulting the in the most clinically desirable balance of trade-offs. Expert-driven PGAP decisions are made using 3D dose distributions, DVHs and numerical dose statistics all of which are updated on screen in the TPS in real-time as the navigation takes place. Figure 3.2 shows an example of the EdgeVcc sliding interface.

Weighting factors are user defined but are often chosen to follow a geometric progression to approximately follow the trend of the Pareto front. For navigation, regions of the Pareto front that have not been explicitly generated can be approximated using convex combinations of adjacent Pareto plans. These approximate solutions can aid the navigator in making a selection, and weighting factors can be navigated for any set of PGs at once. For example, weighting factors for a single PG can be navigated (a univariate case) or a combination of PGs can be considered together. Given PGs trade-off against each other, it is considered advisable to navigate PGs together to better understand their trade-off relationships and make a more informed decision. Once the navigator has made a choice, their weighting factors are saved within a JSON file readable by the Raystation TPS and in an easily accessibly CSV file.

However, a key issue with this methodology is the number of necessary Pareto plans necessary for navigation. The number of Pareto plans grows exponentially with the number of PGs that are considered at once. Given n PGs with weighting factors to calibrate and m weighing factor levels for each, n^m Pareto plans will be generated for navigation. As the number of weighing factor levels increase, so too does the navigation accuracy due to improved convex combination approximations but this contains a computational and resource cost given the space needed to store the plans. A larger issue in the increase in Pareto plans due to consideration of more PGs at once usually leading to a compromise navigation accuracy when fewer weighing factor level are considered in order to



IF_PSV07^Calibration^^^, zAPS2_IFSVP_Cal07: Tra...

Std Variable: Priority 1 Weight = 202.8

Std Variable: Priority 2 Weight = 15.045

Std Variable: Priority 3 Weight = 15.045

Navigation MU:

Navigation Ref MU:

Dose axis display options: ● Absolute values ○ Relative max dose value ○ Relative reference va

Dose difference
Position: -14.37 -49.05 15.02 cm
CT: -13 HU
Dose difference: 7.40% of 62.00 [4.59] Gy

Figure 3.2: An example of an EdgeVcc sliding interface containing three PGs. In the background can be see two identical transverse CT slices for a PSV patient on the left plane. The top slice shows a representation of isodose lines given the navigation. The bottom box contains a static image of the top box when the Store Navigation Data button is pressed. On the top right hand pane can be seen DVH plots of all delineated regions and the bottom right pane contains dose percentage differences between the two navigation panes on the left side.

minimised resource costs.

3.5.4 AP for advanced EBRT

In addition to improvements in standard modern EBRT, more advanced delivery methods have shown to result in dosimetric improvements[140,141] nevertheless the search space increases substantially making the identification of the most desirable plan even more elusive. Researchers have begun to develop methods of Pareto navigation for 4π (non-coplanar) delivery more readily facilitate the use of advanced methods of treatment. This leads to large potential for dosimetric benefits whilst maintaining reasonable planning efficiency[142].

3.6 Treatment sites considered in this work

The results of this work will be best realised for treatment sites most prevalent for tumorous cancers including breast, prostate, lung and colorectal tumours[143]. Therefore, it was pertinent to the development of the work that solutions be defined within this subset of treatment sites. Findings would potentially have far reaching implications in the short terms and generalised methods could then be adapted for other sites within this treatment institution. Breast was excluded from this work because EdgeVcc AutoPlan protocols

Site	PTV Prescribed Dose	Type	Prescribed Dose (%)
PSV	60	Min Dose	96.5
		Max Dose	105
		Min Median Dose	96.5
		Max Median Dose	105
	48	Min Dose	100
		Max Dose	100
Rectum	45	Min Dose	97
		Max Dose	102.5
		Max Median Dose	99.5
Lung	55	Min Dose	97.5
		Max Dose	102.5
		Min Median Dose	100
		Max Median Dose	100

Table 3.1: Summary of EdgeVcc PGs for PTVs of each site. Information outlined here was sourced from Velindre's internal documentation.

for breast were still being developed at the time of this work. Hence, chosen sites were prostate, lung and rectum.

All work relates to current local practice with the exclusion of as few patient groups as possible. All patient were previously treated at Velindre Cancer Centre and chosen at random from their respective time windows. Patients were originally planned using computed tomography (CT) scans of 3mm slice thickness and treated in the head-first supine position. Patients with non-standard areas of avoidance such as hip prostheses or hernias were excluded from patient databases as well as patients with non-standard margins. Plans for each site were generated within the RayStation (Raysearch Laboratories, Stockholm, version 8B) TPS with identical methodologies applied across patients including treatment units and arc configurations. PTVs were created using ICRU standard definitions and PTV suffixes indicate the prescribed dose in Gy. See Table 3.1 for a breakdown of PTV PGs included for each site.

For prostate, the chosen dataset included prostate seminal vesicles (PSV) patients treated between January and June 2018 (inclusive). Delineated ROIs included the external body contour (referred to from here on simply as the external), prostate, seminal vesicles, rectum, bladder and bowel (volume up to 2cm superior of the prostate). Forty-five PSV patients were considered in total of which 5 were excluded for not meeting the criteria: three for having a non-standard area of avoidance and two for having non-standard margins. Two PTVs were derived: (1) PTV60 defined as prostate expanded 5mm isotropically (6mm craniocaudally), (2) PTV48 defined as prostate and base seminal vesicles expanded by 10mm isotropically. Patients were treated on a Varian TrueBeam STx (Varian Medical Systems, Palo Alto, CA, USA) linac in 20 fractions using a simultaneous integrated boost technique and PGs derived from local clinical goals defined following the UK PIVOTAL trial[144].

For rectum, the chosen dataset included patients treated between June 2016 and June 2020 (inclusive). Delineated ROIs included external, bowel bag, stoma and genitals. Bowel bag delineates the abdominal cavity one transverse slices above the PTV down to the pelvic symphysis. Sixty-four rectum patients were considered in total of which 4 were excluded for not meeting the criteria due to non-standard areas of avoidance. All patients had an external and bowel bag delineated. Regarding stoma and genitals, 13 patients had both stoma and genitals delineated, 43 had genitals but no stoma and 4 had neither. One PTV, PTV45, was derived for each patient. Patients had been treated on an Elekta Agility VMAT (Elekta Solutions AB, Stockholm, Sweden) linac or a Varian TrueBeam STx linac in 25 fractions and PGs derived from local clinical goals.

For lung, the chosen dataset include patients treated between June 2018 and June 2020 (inclusive). Delineated ROIs included contralateral lung, ipsilateral lung and combined lungs minus GTV, heart, cord, oesophagus, brachial plexus and liver. In total 68 patients were considered of which 8 were excluded for not meeting the criteria: four for metal works, three for having the kidneys delineated and for having the bronchus delineated. Of the brachial plexus or liver volumes, 11 had a brachial plexus delineated, 8 had liver and non have 42 had neither and none had both. One PTV, PTV55, was derived for each patient. One PTV, PTV55, was derived for each patient. Patients were treated on an Elekta Agility VMAT linac in 25 fractions and PGs derived from local clinical goals.

Prior to this work were three pilot studies for knowledge gain and corroboration of other researchers findings. All preliminary work pertains mainly to prostate given it is directly comparable to related research in this area[109,118–121] and previous research related to the AP approach discussed on this work[96,105].

Chapter 4

Hypothesis generation

The overall aim of the work presented in this thesis is to defined ML solutions for automated and patient-tailored RBP calibration reflecting the choices of qualified professionals. This is done with the intent of leading to a full AP approach in which zero human planning is required. To do this, an RBP technique is used as a base AP system. Classical RBP methods are reliant on *apriori* (upfront) calibration of PG priorities and this is an obstacle of its implementation because resulting plan quality cannot be assessed by this calibration alone. Plan quality is only assessed once a plan has been fully generated and the dose distribution is known. Therefore, the direct relationship between RBP calibration parameters and dose distribution is yet unknown[145] and calibration is difficult for this reason. Also, RBP planning is not always patient-tailored as the calibration task is often carried out once per treatment site with the same prioritisation of PGs then applied to all cases. Understanding the relationship between calibration parameters and anatomy with respect to dose distribution will better facilitate straight forward planning and be valuable when implementing automation. For example, this will make choosing appropriate features for use in ML modelling simpler and will be useful for development of heuristics used to analyse those models.

4.1 Aims and objectives

The aim of this section is to better understand what constitutes “clinical preference” and how to obtain it consistently via apriori calibration of RBP. Using a PGAP system that permits navigation along the Pareto front, the levels of consistency between different calibrations can be assessed. Parameters in need of calibration for this automated planning system are the weight factors. The difference between two weighting factors is a calculable metric and variances in choices can be understood in terms of these differences. Defining the clinically applicable domain of the Pareto front is necessary for modelling a ML solution that maps to this front. Moreover, given a reliable and consistent definition of clinical preference can be obtained, the underlying relationships between clinically relevant weighting factors and anatomy can be better understood also. In light of this, three studies were carried out:

(i) Intra-planner calibration study
(ii) Inter-planner calibration study
(iii) Anatomy simulation study

All studies in this chapter were carried out for PSV only and the patient database consisted of randomly selected PSV patients as outlined in section 3.6. Participants of these three studies were all qualified for plan creation or plan verification for the site in question and were trained in the use of the PGAP system. These studies help to lay the foundations for closing a gap in knowledge. The first two studies were closely related and have a number of similarities that will be outlined now.

4.2 Defining “gold standard” planning

Planning can be thought of as translation of oncologist dosimetric preferences into deliverable plans. Therefore, planners are required to have an adequate understanding of this preference to ensure theoretical consistency in planning choices. This is not always the case given sources of variability lead to differences in assessment of appropriate trade-off relationships[146]. Inter-planner variability may be due to differences in training, experience, interpretation of protocols or personal inclination. Intra-planner variability may be due to workflow order effects (order of patient cases), developments in knowledge over time or other daily experiences. Studies presented in this chapter aim to establish if there exists a discernible gold standard domain of plans given these variances or whether the translation of clinical preference is truly planner specific.

The aims of this thesis are conditional on a reliable and consistent ground truth database being obtained from qualified professionals. Assuming consistent planning is possible, this ground truth can be established in one of two ways, using a database of plans generated by: (1) a single person or (2) a collection of individuals. To explore (1), the consistency in planning choice of a single participant were observed (intra-planner variability). The intra-planner study provides information about the degree to which a single human planner is expected to see discrepancies in their own weighting factors. The value of an intra-planner study is isolation of behaviour of a single qualified professional with years of training and experience. The study highlights the level of deviation in choices of the highly qualified when calibrating RBP solutions and therefore, helps in identifying the domain of clinically applicable planning that can be observed when a single person is considered. Not only does the intra-planner variability study aid in the research of (1), but a single participant design has the benefit of being more convenient than a multi-participant design and obtaining a gold standard database with this approach may require less time and computation. Also, if a multi-participant design is found to be more appropriate, conducting an intra-planner study first is still useful as an experiment refinement method prior to the larger multi-participant design[147].

To explore (2), multiple qualified professionals were selected (inter-planner variability) and the consistency in planning choices between them was observed. The inter-planner study illustrates the level of consistency in planning observed between different qualified professions of the same institution. The true range of clinically applicable planning across planning professions is unknown but choices between planners is known to vary. This study provides a more general view of the clinically applicable region of the Pareto front.

In all cases, participants calibrated solutions for patients using the PGAP system and choices compared. Similarity between choices was measured using a Sørensen–Dice coefficient and statistically significant differences in dosimetry and weighing factor values. Weighting factors hold little intrinsic value on their own but are strongly relative to each other. Relative weighing factor values are closely linked to the target objective value in the TPS optimiser as the PGAP system uses them to set the targets. Therefore, for analysis purposes, weighing factor values were compared in relative form as well as their raw values. That is relative values are derived by dividing raw weighing factor values by the sum of all weighing factor values. Participants were also interviewed following planning sessions and qualitative aspects will be discussed as well as quantitative differences.

4.3 The relationship between calibration and anatomy

The aim of the anatomy simulation study was to generate hypotheses about the relationships between anatomy and weighting factors. The study explores the changes in weighting factors necessary to achieve a consistent dose distribution given anatomical variations. Therefore, this study helps to uncover what is necessary to enable consistent planning in the clinic. Weighting factors are calibrated for a single patient following which the anatomy of the patient was augmented in a controlled way. The findings of this study can also be used as a guide for judging the successful application of ML solutions given anatomical variations in real cases. This is due to the anatomic augmentation helping to explicitly define some of the underlying relationships between geometric anatomy, weighting factors and the resulting dose distribution.

4.3.1 AutoPlan protocol

The base AutopPlan protocol presented in Tables 4.1a-4.1d is based on a clinically approved and implemented solution for PSV. It was created in-line with local practice and similar PGs have been considered appropriate to manage dose distribution for this clinical site in other work[120,148]. The AutoPlan protocol contains seven P1 and P2 PGs which aim to control maximum bowel dose and PTV homogeneity by restricting them within fixed tolerances. It also contains seven trade-off (P3) PGs: (1) average dose to the rectum, (2) average dose to the bladder, (3) PTV dose conformality, (4) maximum dose to the rectum, (5) intra-PTV dose fall-off, (6) maximum dose to the bladder and (7) medium-high dose to the bowel ($V_{36.0Gy}$ and $V_{45.6Gy}$). Average dose refers to mean dose across the voxels in the ROI.

4.3.2 Planning procedure for inter- and intra-planner study

When using a PGAP system, the number of Pareto plans increases exponentially as the number of PGs considered during navigation increases. This occurs because the number of plans required for navigation is raised to the power of the number of PGs[149]. This means the process can become increasingly computationally expensive as the number of PGs increases. Five weighing factor values were selected for each PG to sample the clinically relevant span of Pareto plans in this work. That is 5^N Pareto plans generated for navigation where N is the number of PGs considered simultaneously. Five was chosen

Table 4.1: PGs for PSV AutoPlan protocol.

Abbreviations: $\%_{\text{Presc, PTV}}$ = % of individual PTV prescription dose; $\%_{\text{Presc}}$ = % of overall treatment prescription; $\%_{\text{Vol}}$ = % volume of ROI.

Notes: Priority 3 Weighting Factors are subject to change prior to optimisation if desired. Priority 3 Target = 0.0 by default but can be specified if desired. Target is dynamically adjusted for all during optimisation and therefore initial values have negligible impact on plan quality but may decrease planning time if correctly defined.

(a) Priority 1: Primary OAR Goals

ROI Name	Dose Parameter	Target (Gy)	Weighting Factor
Bowel	D_{max}	51.0	1000

(b) Priority 2: Target Goals

ROI Name	Dose Parameter	Target ($\%_{\text{Presc, PTV}}$)	Weighting Factor
PTV60	D_{min}	96.5	250
PTV60	D_{max}	102.5	250
PTV60	$D_{50\%}$ max	99.5	250
PTV48	D_{min}	96.5	250
PTV48	D_{max}	105.0	250

(c) Priority 3: Trade-off Goals (Standard)

ROI Name	Dose Parameter	Target (Gy or $\%V_{\text{Vol}}$)	PG Number	Weighting Factor
Rectum	D_{mean} (Gy)	5.0	1	21.3
Bladder	D_{mean} (Gy)	5.0	2	6.86
Rectum	D_{max} (Gy)	60.0	4	195
Bladder	D_{max} (Gy)	54.0	5	0.880
Bowel	$V_{36.0\text{Gy}}$	0.0	7	0.762
Bowel	$V_{45.6\text{Gy}}$	0.0	7	0.762

(d) Priority 3: Trade-off Goals (Dose Fall Off)

ROI Name	Fall Off Type	High Dose Level (Gy)	Low Dose Level (Gy)	Dose Gradient ($\%_{\text{Presc cm}^{-1}}$)	PG Number	Weighting Factor
PTV48	Falloff	57.0	40.8	50%	3	23.6
PTV48	Intra PTV Falloff	54.0	52.8	50%	6	1.47

empirically and considered highly appropriate given previous AP studies use as few as N+1 Pareto plans for navigation[67,150] and recent mathematical studies use as few as five Pareto plans in total[151].

Of the five weighing factor values selected to sample the clinically relevant span of the Pareto set, maximum and minimum weights were chosen empirically based on a clinically approved and implemented AP protocol. Values for the three intermediate weighting factors were chosen such that they follow a geometric progression to ensure an even spread across the Pareto set[152]. It was possible to modify weights on a patient-by-patient basis if they were not found to be sufficient. However, no participants requested this for any patients in these studies.

4.4 Statistical Testing

The statistical testing used in this thesis will be explained and justified. This is done not only to illustrate the integrity of the results but also importantly to engage in the discussion of the most appropriate statistical testing for true scientific validity within the community.

Parametric testing uses sample parameters to explain the data opposed to individual observations. They are computationally efficient and expected to approximate population relationships well when the data are appropriately large. However, parametric testing usually requires some assumptions be met prior. For example, parametric tests can be used to determine if two or more samples follow a similar distribution. Of these tests, some use mean and standard deviation related parameters such as the t-test and analysis of variance (ANOVA). In these cases, the data must be expected to follow a standard normal distribution (or not deviate significantly from a standard normal distribution) and the variance of each sample is expected to be approximately equal. In order to determine this with statistical tests, the sample size should be appropriately large.

However, when assumptions are not met by the data, non-parametric alternatives must be employed instead. Non-parametric tests often make fewer assumptions (or no assumptions) about the data structure and will yield more reliable results than their parametric equivalents in certain circumstances. However, non-parametric tests are often more computationally intensive to carry out as the data size increases and may not converge to population outcomes as well as parametric tests would.

In this work, studies were conducted such that different measures are taken for the same cohort for comparison against each other. These studies are known as *repeated measures* because the measures are repeated for the same subjects for within subject comparison and subject cohorts are not independent. When only two measures are taken, pair-wise testing is appropriate and such testing may include parametric t-tests or non-parametric Wilcoxon signed rank tests. However, when more than two measures have been taken, omnibus tests should be carried out to manage the family-wise *type I* error. A type I error is the probability of rejecting the null hypothesis in error and a type II error is the probability of accepting the null hypothesis in error.

Given the number of patients in this work, ANOVA omnibus testing and Tukey post hoc testing have been employed providing ANOVA assumptions are not violated. Otherwise, Friedman omnibus testing with Nemenyi post hoc testing were used. The justification for this will follow later in this section. Statistically significant differences are reported at the 5% level of significance (or 95% confidence level). To determine if samples followed a standard normal distribution, a statistical test of normality was carried out. To do this, a Shapiro–Wilk test were employed[153] using a significance level of 5%. Given the test for normality is not violated, a test for sphericity of variances was carried out. To do this, Mauchly's test[154] was used and judged at the 5% level of significance. ANOVA was applied given neither test indicated the data do not meet the assumptions. Given the Shapiro–Wilk test was not significant but Mauchly's test was, ANOVA was used with a Greenhouse-Geisser corrected p-value[155] to adjust for lack of sphericity.

Given ANOVA omnibus testing with a significant difference indicated, a suitable post hoc test was employed to interpret pairwise differences whilst managing the family-wise error. One way of doing this would have been to use t-tests with a p-value correction to adjust for the number of pairwise comparisons. For example, a popular way to adjust for the family-wise error is to use a Dunn or Bonferroni p-value correction[156]. In this work however, a Tukey test was used for the ANOVA post hoc test given this test inherently manages the family-wise error rate and is recognised as a traditional ANOVA post hoc test. All ANOVA and ANOVA post hoc testing was implemented in python using the Pingouin 0.5.2 library.

When the normality assumption was violated, a non-parametric test was used. Traditionally in this field of work, a paired Wilcoxon signed rank test will be employed. This is valid for paired comparison and similar approaches have been supported by recent statistical research as a valid method[157]. Also, for independent measures, one of the most appropriate and a well known non-parametric test is the Kruskal-Wallis test[158–160]. However, this work relates to repeated measures, and a more appropriate test would be to use a Friedman test[161,162] as has been used in a number of related studies recently[76,163,164]. Similarly to the pairwise scenario of a Wilcoxon test, the rank of values is used to compare groups opposed to the actual values. Wilcoxon testing does not inherently manage the family-wise error therefore using it increases the potential of committing a type I error. To get around this, one could employ a correction such as those mentioned earlier to adjust the p-value due to the number of pairwise comparisons and control the family-wise error. However a more standard post hoc test is the Nemenyi test given it inherently controls for the family-wise error. In this work, Friedman testing was implemented using the Pingouin 0.5.2 library in python and Nemenyi testing implemented using the scikit_posthocs 0.7.0 library.

4.5 Intra-planner study

4.5.1 Methods

Of the PSV patients defined in section 3.6, PSV patients 01-20 were included in this study. At this institution, there is a known heuristic in local PSV planning given traditional planning protocols used to develop the AutoPlan protocol seen in Tables 4.1a-4.1d. That is, given the generally large relative priority associated with PG 1-3, these PGs show the most significant and notable trade-off relationships when navigated together with negligible impact to other PGs and navigating all PGs simultaneously was not considered strictly necessary. Therefore, given the exponential relationship between the number of PGs and the number of Pareto plans, navigation was performed in two stages for resource and time efficiency. Fewer Pareto plans were generated therefore plan generation occurred in less time and less computing power and space was taken than would have been if all six PGs were to navigated together. Also, given this study was to purely assess navigation choices, generation of fully optimised plans was not required and was not done. Navigated plans were therefore compared based on navigated data only with fully optimised plans not generated via PBAIO.

It is also known that repeated measures studies can be highly skewed by order effects if the planning professional is able to anticipate the next case. To mitigate this, navigation sessions were not held close together but with a suitable interval between them. In this study the interval was at least 7 days from the end of the last session. Cases were also presented in a different random order each time.

4.5.1.1 Navigation Session

In addition to the EdgeVcc sliding interface, the TPS interface was available for the participant to interact with. It contained a CT scan of the patient in each case, a DVH, numerical dose statistics and clinical goals. The participant had the ability to select and deselect ROIs that would update the view in the CT, DVH and statistics. CT scans could be switched between transverse, sagittal or coronal plane view whenever needed to allow dose distributions to be viewed in all directions. A key characteristic of the EdgeVcc sliding interface is the option to freeze a navigation at a certain position (set of weighting factors) and copy it to a reference plan and use the reference as a guide to continue sliding. With EdgeVcc, navigators are able to use other plans as a reference (e.g. clinically approved plans), but the participants choices were based on their own knowledge and judgement and such references were not permitted as they may confound the results. The participant performed navigation under standard planning conditions.

PGs 1-3 were navigated simultaneously whilst the latter four were held constant at the level defined in the clinically approved AutoPlan protocol. Weighting factors were stored during each session. PGs 4-6 were then navigated in three more sessions in a similar way with PGs 1-3 held at their newly defined values. Qualitative information regarding planning choices was provided by the participant after the first session. PG 7 (high dose to the bowel) was not re-calibrated due to its low priority, negligible trade-off impact and low proximity of the related organ-at-risk (OAR) to PTVs. This PG was held constant at its clinically defined weighing factor in all cases. Five weight level were used for each navigated PG with three navigated at a time in each navigation. That is $2 \times (5^3) = 250$ Pareto plans per patient.

Also included for analysis was the *Average Session*. Values for the Average Session were derived of the values from the three actual session. For each patient and each PG, the mean weighing factor over the three session was calculated. A useful finding of this work will be to determine a single set of weighting factors that represent the participants overall choices and interpretation of the oncologist clinical preference. One way would be to choose the weighing factor sets from a single session but instead the Average Session was proposed as a non-biased choice that would best control for day to

Absolute Weight Metric	Average Session	Session 1	Session 2	Session 3	Test
Rectum D_{mean}	114 ± 28.3	129 ± 33.3	101 ± 33.0	112 ± 34.5	ANOVA
Bladder D_{mean}	32.2 ± 11.3	36.4 ± 19.8	25.5 ± 10.5	34.6 ± 22.1	Friedman
PTV Conformality	292 ± 92.4	328 ± 103	269 ± 98.1	280 ± 107	Friedman
Rectum D_{max}	2.78 ± 0.836	2.45 ± 1.06	2.71 ± 1.13	3.19 ± 1.47	Friedman
Intra-PTV dose falloff	18.9 ± 3.37	19.3 ± 7.24	17.9 ± 6.98	19.5 ± 4.33	ANOVA
Bladder D_{max}	12.5 ± 3.40	13.2 ± 4.68	11.0 ± 3.10	13.5 ± 4.60	Friedman
Bowel D_{medium}	1.52	1.52	1.52	1.52	None
PG_{H}	2250	2250	2250	2250	None

Table 4.2: The mean raw navigated weights over all patients by the navigator in the three sessions and the average weights across these sessions. Boldface indicates statistically significant difference (at the 95% level) within the omnibus test (ANOVA or Friedman test).

day variance.

4.5.2 Results

4.5.2.1 Absolute weights

In each case, the participant took between 4-5 minutes per navigation. See Table 4.2 for a summary of absolute weighing factor values including indications of statistically significant differences of omnibus tests.

Given absolute values were constant for bowel $V_{36.0\mathrm{Gy}}$ and $V_{45.6\mathrm{Gy}}$ (bowel D_{medium}) and higher priority PGs (PG_{H}), testing was not done for these PGs. Values for PG 2-4 and 6 violated normality assumptions and a Friedman test was applied. In terms of absolute values, the Average Session illustrated a high levels of consensus across the sessions. No statistically significant differences were observed between the Average Session and individual sessions. Notably, Sørensen–Dice (DiceC) coefficients between each of the three sessions with the Average Session were greater than 0.9 indicating a high degree of similarity for all PGs in all sessions with those of the Average Session. Figure 4.1 shows the distribution of individual session weighting factors about the Average Session. Session 3 was the most similar to the Average Session given it resulted in the highest DiceC and the lowest median difference with the Average Session with values of 0.977 and 0.888 respectively. Session 2 differed the most from the Average Session given it resulted in the lowest average DiceC across the PGs with a value of 0.967 and the largest median deviations from the Average Session for five of the six PG as seen in Figure 4.1.

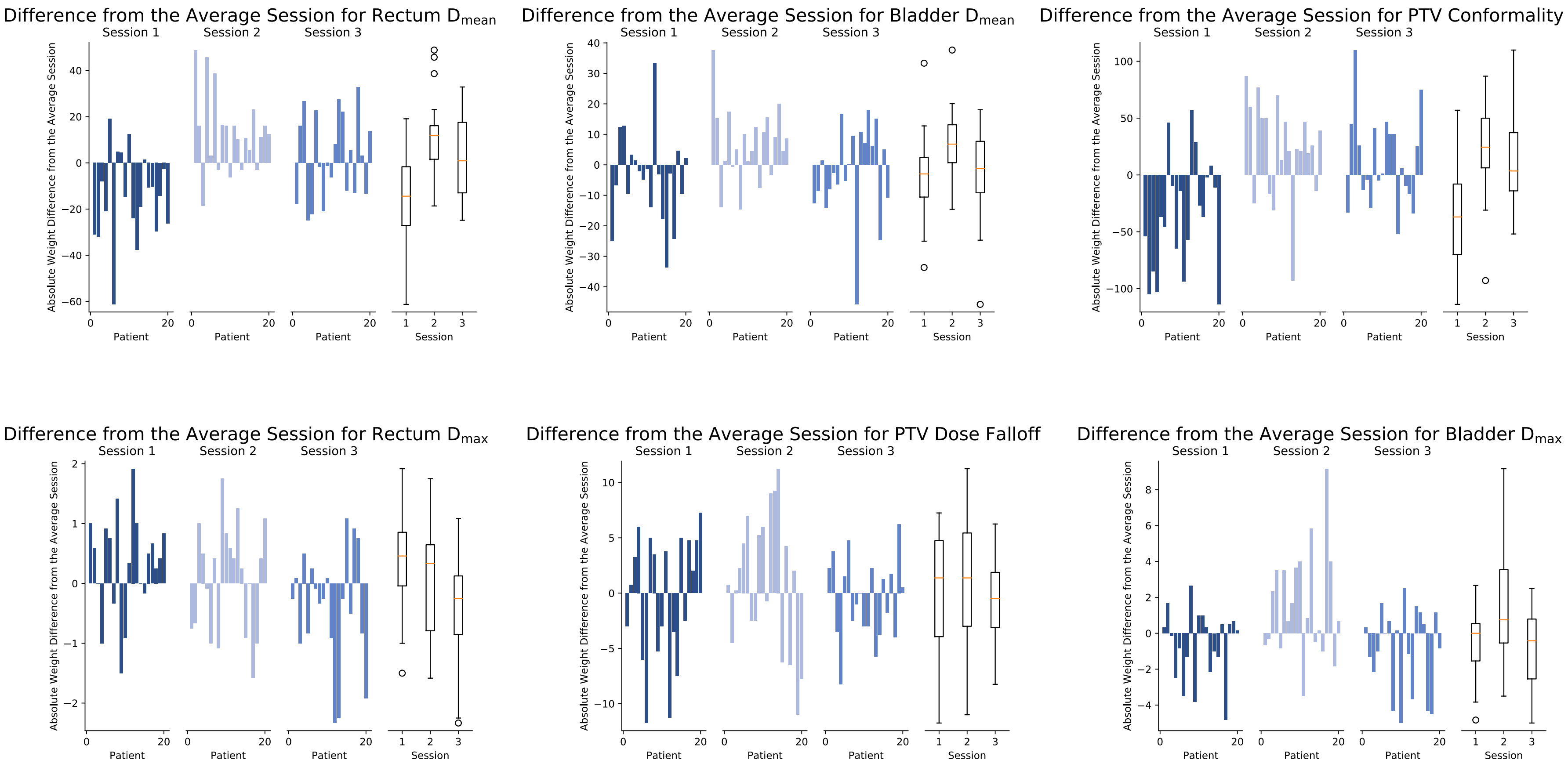


Figure 4.1: Summary of PG absolute weights for the intra-planner study. Values are differences from the Average Session

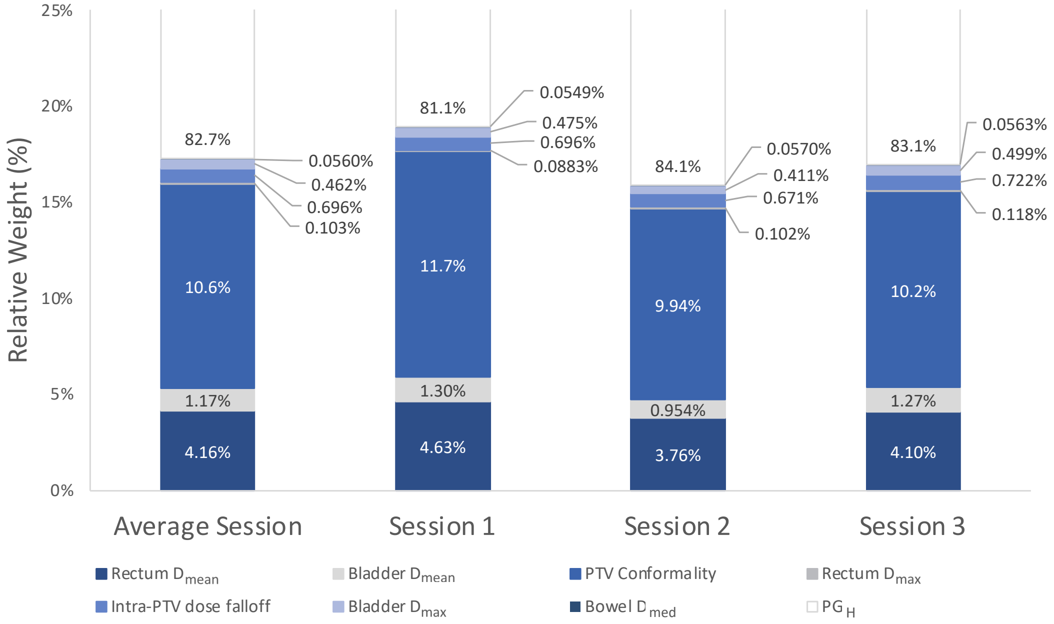


Figure 4.2: Intra-planner relative weight comparison between Session 1-3 and the Average Session. Values represent mean relative weighing factor values across all 20 patients.

There is some evidence that the average navigation (i.e. the Average Session) may be appropriate for representing the navigation behaviour of a planner overall given strong consensus with all sessions even when differences are observed between session. This may be considered a means of controlling for inconsistencies observed during individual sessions.

Differences between individual sessions were comparatively greater. Statistically significant differences were observed between Session 1 and 2 for rectum D_{mean} and PTV conformality only with Session 2 and 3 having no absolute weight significant differences for any PGs. However, the greatest similarity was observed between Session 1 and 3 with a DiceC of 0.933. This shows that Session 2 and 3 differed least on aggregate (at population level) and at patient-level Session 1 and 3 differed least. Nevertheless, there is evidence showing comparability of performance across the three session by this participant.

4.5.2.2 Relative weights

See Table 4.3 for a summary of relative weighting factors including indications of statistical significant difference in boldface. Of the eight PGs tested, three did not meet normality assumptions and were tested using Friedman. These PGs include conformal-

Relative Weight Metric	Average Session	Session 1	Session 2	Session 3	Test
Rectum D_{mean}	4.16% ± 1.02%	4.62% ± 1.15%	3.76% ± 1.19%	4.10% ± 1.21%	ANOVA
Bladder D_{mean}	1.18% ± 0.396%	1.30% ± 0.674%	0.954% ± 0.396%	1.27% ± 0.809%	Friedman
PTV Conformality	10.6% ± 2.98%	11.7% ± 3.21%	9.94% ± 3.22%	10.2% ± 3.45%	Friedman
Rectum D_{max}	0.103% ± 0.0312%	0.0883% ± 0.0378%	0.102% ± 0.0434%	0.118% ± 0.0555%	ANOVA
Intra-PTV dose falloff	0.696% ± 0.129%	0.696% ± 0.265%	0.671% ± 0.258%	0.722% ± 0.164%	ANOVA
Bladder D_{max}	0.462% ± 0.132%	0.475% ± 0.173%	0.411% ± 0.121%	0.499% ± 0.181%	Friedman
Bowel D_{medium}	0.0560% ± 0.0185%	0.0549% ± 0.00222%	0.0570% ± 0.0200%	0.0563% ± 0.00231%	Friedman
PG_H	82.7% ± 2.72%	81.1% ± 3.27%	84.1% ± 2.96%	83.1% ± 3.42%	Friedman

Table 4.3: Relative navigated weights over all 20 patients by the expert planning professional in the three navigation session and the average relative weights across these sessions. Statistical significance is measured at the 95% level.

ity and bladder D_{mean} and D_{max}. Similarly to the absolute weight scenario, a high level of similarity is observed between the Average Session and the three individual sessions given DiceC values of greater that 0.9 for all PGs. Additionally, no statistically significant differences were observed between the Average Session and any of the individual sessions and this supplements the evidence from the absolute value case above, that the Average Session is representative of this participants planning behaviour and sufficient for use instead of choices made during individual session. Figure 4.3 indicates deviations of navigating sessions from the Average Session. Consensus between individual sessions and the Average Session were strong given deviations ranges of between ∓ 0.04 or $\mp 4\%$. The PG that showed the largest deviations from the Average Session across the three navigation's sessions was rectum D_{mean} with DiceC values of 0.918, 0.918 and 0.916 for Session 1, 2 and 3 respectively.

Individual navigation session choices were also comparable on average (as seen in Figure 4.3) similarly to the absolute weighing factor case and statistically significant differences were observed between Session 1 and 2 only. Differences included three PG groups: PTV conformality, bowel D_{medium} and PG_H. A borderline significant difference was also observed for rectum D_{mean} also. Rectum D_{mean} and PTV conformality were prioritised lower on aggregate during planning Session 2 giving a higher relative priority to the two PGs that were not navigated i.e. bowel D_{medium} and PG_H.

A notable finding was that higher levels of similarity are observed between the Average Session and individual session for PG 1-3 when weights are relative compared

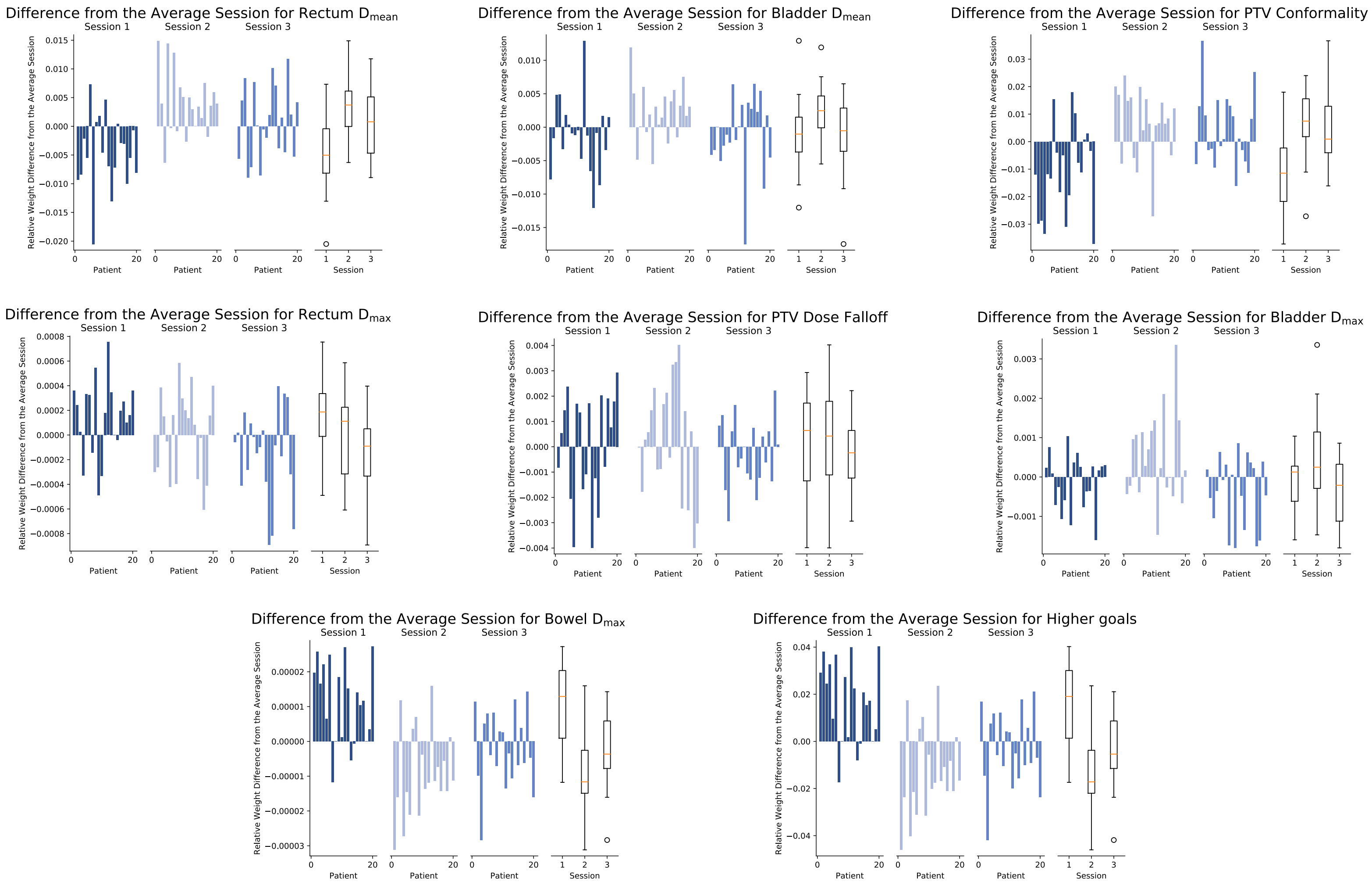


Figure 4.3: Summary of PG relative weights for the intra-planner study. Values are differences from the Average Session.

to when they are absolute. This is evidence that although planners may use a different region of the sliding scale during planning, their relative choice may still balance similarly in terms of relative prioritisation. This is a key finding given the relative PGs values are more applicable to weighing factor calibration and the dosimetry of the resulting plan. There is also strong evidence that static and non-patient specific weights (e.g. bowel D_{medium} or PG_H) can have a statistically significant impact on planning given other weights vary notably when calibration is patient-specific. Therefore, considering the relative relationships between weights is arguably more important than observing them in absolute terms.

4.5.2.3 Dosimetry

Table 4.4 summarises key dose-volume metrics for the three session and the Average Session. Of the 24 dose-volume metrics considered, eight violated the normality assumption for ANOVA and were tested using Friedman. These included PTV60 $D_{98\%}$ (Gy), PTV60 $D_{2\%}$ (Gy), homogeneity of PTV60, PTV48 $D_{50\%}$ (Gy), rectum V_{60Gy} (%), rectum $V_{60.8Gy}$ (%), bladder $V_{52.7Gy}$ (%) and bladder $V_{56.8Gy}$ (%). Statistically significant differences were observed for seven dose-volume metrics of which six related to the Average Session. All three navigation sessions differed from the Average Session for PTV60 $D_{98\%}$ (Gy), PTV60 $D_{50\%}$ (Gy), rectum V_{60Gy} (%) and bladder $V_{56.8Gy}$ (%). Smaller values were observed for all navigated sessions when compared with the Average Session for all four dose metrics showing a statistically significant difference.

Session 3 was dosimetrically most comparable to the Average Session given mean and median difference were minimised by this session for the majority of key metrics presented in Table 4.4. Most notably, Session 3 minimised median differences from the Average Session for dose to the rectum. A statistically significant difference was observed for high dose to the rectum (rectum $V_{60.8Gy}$ and V_{60Gy} mentioned earlier) between Session 3 and the Average Session however. This significant difference was observed for Session 2 against the Average Session also. This indicates good consensus between the Average Session and Session 3 except at the most extreme dose levels.

Session 1 and 2 are comparably further from the Average Session than Session 3. Statistically significant differences are observed for lower doses to the rectum with mean and median differences notably greater than for Session 1. Session 1 however, shows notable deviations from the Average Session in terms of higher doses to the rectum also,

DVH Statistic		Average Session	Session 1	Session 2	Session 3	Test
PTV60	**$D_{98\%}$ (Gy)**	57.3 ± 1.23	57.2 ± 1.23	57.2 ± 1.32	57.2 ± 1.25	Friedman
	$D_{50\%}$ (Gy)	60.0 ± 0.0567	59.9 ± 0.0425	59.9 ± 0.0709	59.9 ± 0.0460	ANOVA
	$D_{2\%}$ (Gy)	61.7 ± 0.103	61.7 ± 0.109	61.7 ± 0.0859	61.7 ± 0.0962	Friedman
	CI	0.848 ± 0.0177	0.851 ± 0.0201	0.845 ± 0.0196	0.848 ± 0.0192	ANOVA
	HI	0.0732 ± 0.0212	0.0758 ± 0.0213	0.0745 ± 0.0223	0.0753 ± 0.0213	Friedman
PTV48	$D_{98\%}$ (Gy)	46.2 ± 0.445	46.1 ± 0.451	46.3 ± 0.430	46.2 ± 0.399	ANOVA
	$D_{50\%}$ (Gy)	53.5 ± 1.51	53.4 ± 1.53	53.6 ± 1.52	53.6 ± 1.51	Friedman
	$D_{2\%}$ (Gy)	59.2 ± 0.261	59.2 ± 0.292	59.2 ± 0.250	59.2 ± 0.246	ANOVA
	CI	0.822 ± 0.0211	0.821 ± 0.0269	0.815 ± 0.0257	0.814 ± 0.0284	ANOVA
	HI	0.244 ± 0.0103	0.244 ± 0.0106	0.241 ± 0.0104	0.242 ± 0.0102	ANOVA
Rectum	$V_{24.3Gy}$ (%)	27.1% ± 8.00%	26.7% ± 7.86%	27.3% ± 7.87%	27.1% ± 7.74%	ANOVA
	$V_{32.4Gy}$ (%)	21.9% ± 7.00%	21.6% ± 6.91%	22.1% ± 6.91%	22.0% ± 6.80%	ANOVA
	$V_{40.5Gy}$ (%)	17.0% ± 5.82%	16.8% ± 5.84%	17.2% ± 5.84%	17.1% ± 5.75%	ANOVA
	$V_{48.6Gy}$ (%)	11.7% ± 4.27%	11.6% ± 4.31%	11.8% ± 4.38%	11.7% ± 4.31%	ANOVA
	$V_{52.7Gy}$ (%)	8.59% ± 3.26%	8.43% ± 3.25%	8.61% ± 3.34%	8.53% ± 3.26%	ANOVA
	$V_{56.8Gy}$ (%)	5.01% ± 1.92%	4.85% ± 1.96%	4.96% ± 2.08%	4.87% ± 2.00%	ANOVA
	V_{60Gy} (%)	0.571% ± 0.519%	0.221% ± 0.219%	0.267% ± 0.268%	0.186% ± 0.210%	Friedman
	$V_{60.8Gy}$ (%)	0.170% ± 0.228%	0.0407% ± 0.0479%	0.0303% ± 0.0398%	0.0280% ± 0.0448%	Friedman
	D_{mean}(Gy)	17.8 ± 3.82	17.5 ± 3.72	17.8 ± 3.76	17.8 ± 3.57	ANOVA
Bladder	$V4_{0.5Gy}$ (%)	20.4% ± 10.1%	20.4% ± 10.1%	20.7% ± 10.2%	20.6% ± 10.4%	ANOVA
	$V_{48.6Gy}$ (%)	14.2% ± 7.34%	14.3% ± 7.48%	14.4% ± 7.39%	14.4% ± 7.67%	ANOVA
	$V_{52.7Gy}$ (%)	11.2% ± 6.15%	11.2% ± 6.22%	11.4% ± 6.18%	11.3% ± 6.28%	Friedman
	$V_{56.8Gy}$ (%)	8.13% ± 4.74%	7.74% ± 4.63%	7.94% ± 4.74%	7.30% ± 4.62%	Friedman
	D_{mean}(Gy)	21.9 ± 7.65	21.7 ± 7.60	22.0 ± 7.73	21.9 ± 8.00	ANOVA

Table 4.4: Summary of key dose metrics. Values shown are Mean ± 1 Standard Deviation. Statistical difference at the 95% level of significance is indicated by boldface.

with Session 3 showing more notable deviations for dose to the bladder.

Significantly higher homogeneity indices were observed for Session 1 over the Average Session and differences were observed between Session 1 and 2 for two metrics PTV48 $D_{50\%}$ (Gy) and Bladder $V_{56.8Gy}$ (%). Regardless of statistical significance, all dose-volume metric differences were considered clinically negligible given deviation within ∓ 2 Gy, ∓ 1 percent volume and ∓ 0.04 units for dose, volume percentage and index units respectively.

PTV60 $D_{98\%}$ (Gy), bladder $V_{52.7Gy}$ (%), PTV60 $D_{2\%}$ (Gy), rectum $V_{60.8Gy}$ (%), homogeneity of PTV60, PTV48 $D_{50\%}$ (Gy), bladder $V_{56.8Gy}$ (%) and rectum V_{60Gy} (%)

4.5.3 Discussion

This study presents evidence that expert-driven PGAP is expected to result in consistent planning with minimal intra-planner variability (Figure 4.5). This is a marked improvement on the level of intra-planner variability that can be expected when there is a heavy reliance on manual planning[165]. This work gives credence to the idea expert-driven

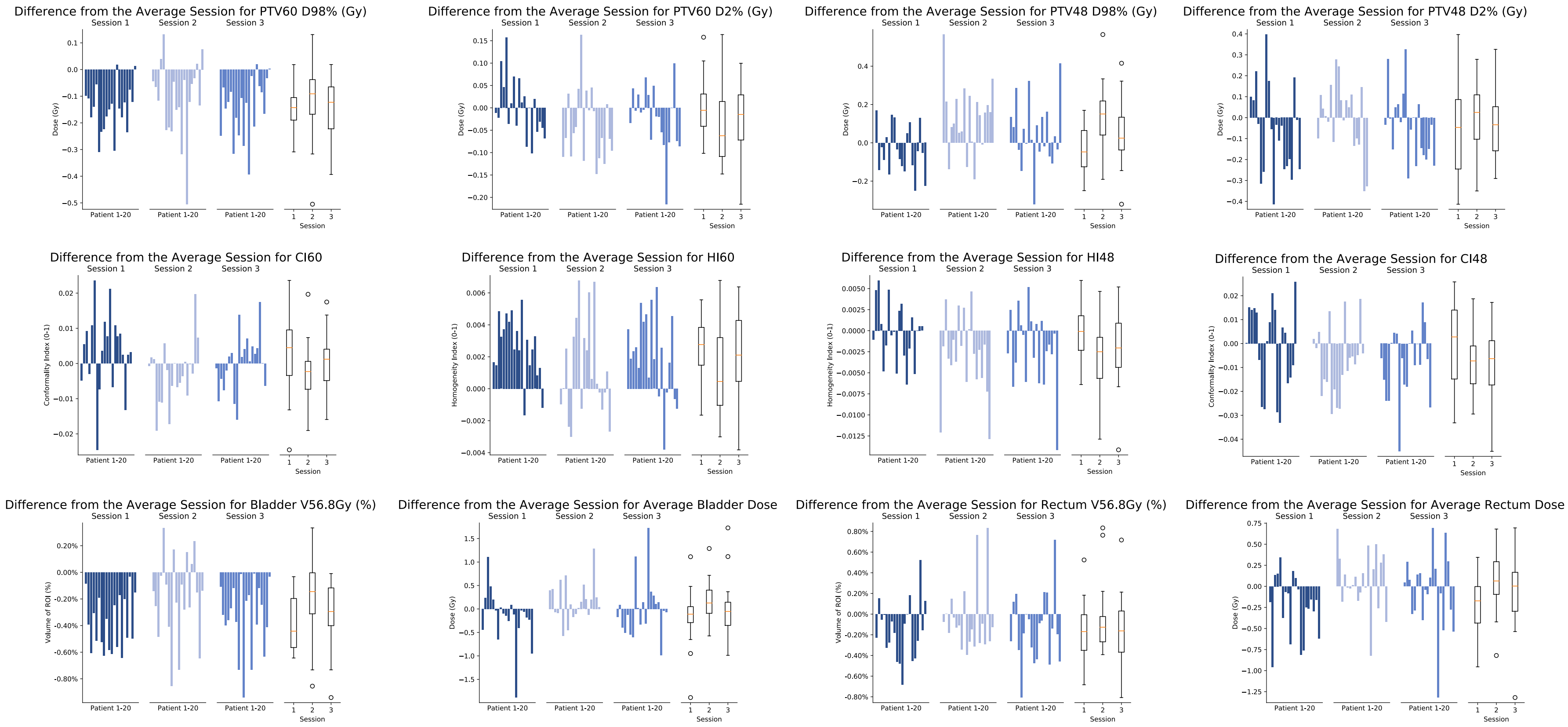


Figure 4.4: Plots showing absolute difference from the Average Session. Distributions are across the patient database for key dose related metrics for each of the three calibration session.

PGAP is useful for reflecting the preferences of the user hence can aid in the generation of plans that have consonance with their preferences[67] opposed to manual planning that can be tedious and cause the planner to stop as soon as clinical goals have been met[166].

This study also highlights the importance of understanding the relative relationships between weighting factors. Although there are no strictly imposed restrictions on weighting factors during the navigation process, assigned weighting factors for each PG are strongly relative within the PBAIO process and native TPS optimiser given they are necessary for defining the optimisation objective function. Such is the importance of navigating PGs simultaneously and the importance of understanding the relative relationships between navigated PG weights in terms of the dosimetry of the resulting plan.

Due to known intra-planner variance the feasibility of defining an appropriately representative planner and patient-specific gold standard was previously unknown. In this study it was shown that marked differences from session-to-session give lower Sørensen–Dice coefficients. However, given the intuitive use of weighting factors and their relationship to the PBAIO system using in this work, it was possible to define Average Session values that had a high degree of congruence with individual sessions and this approach is considered appropriate for defining a planner-specific gold standard. This approach not only showed consensus with individual planning but could help to mitigate some of the discrepancies observed during individual planning sessions.

This study helps to fill a gap in the body of literature given so little information is known regarding intra-planner variability and given newer planning methods such as these rely heavily on the ability to define a suitable ground truth.

4.5.4 Conclusion

The results indicates small differences can be expected between expert-driven PGAP planning sessions. Definition of an Average Session showed negligible clinically significant differences between navigated session metrics and high degrees of consistency. The Average Session is therefore considered an appropriate representative of planning behaviour over multiple sessions and can be used as the definition of gold standard planning with individual sessions comparable and gold standard adjacent.

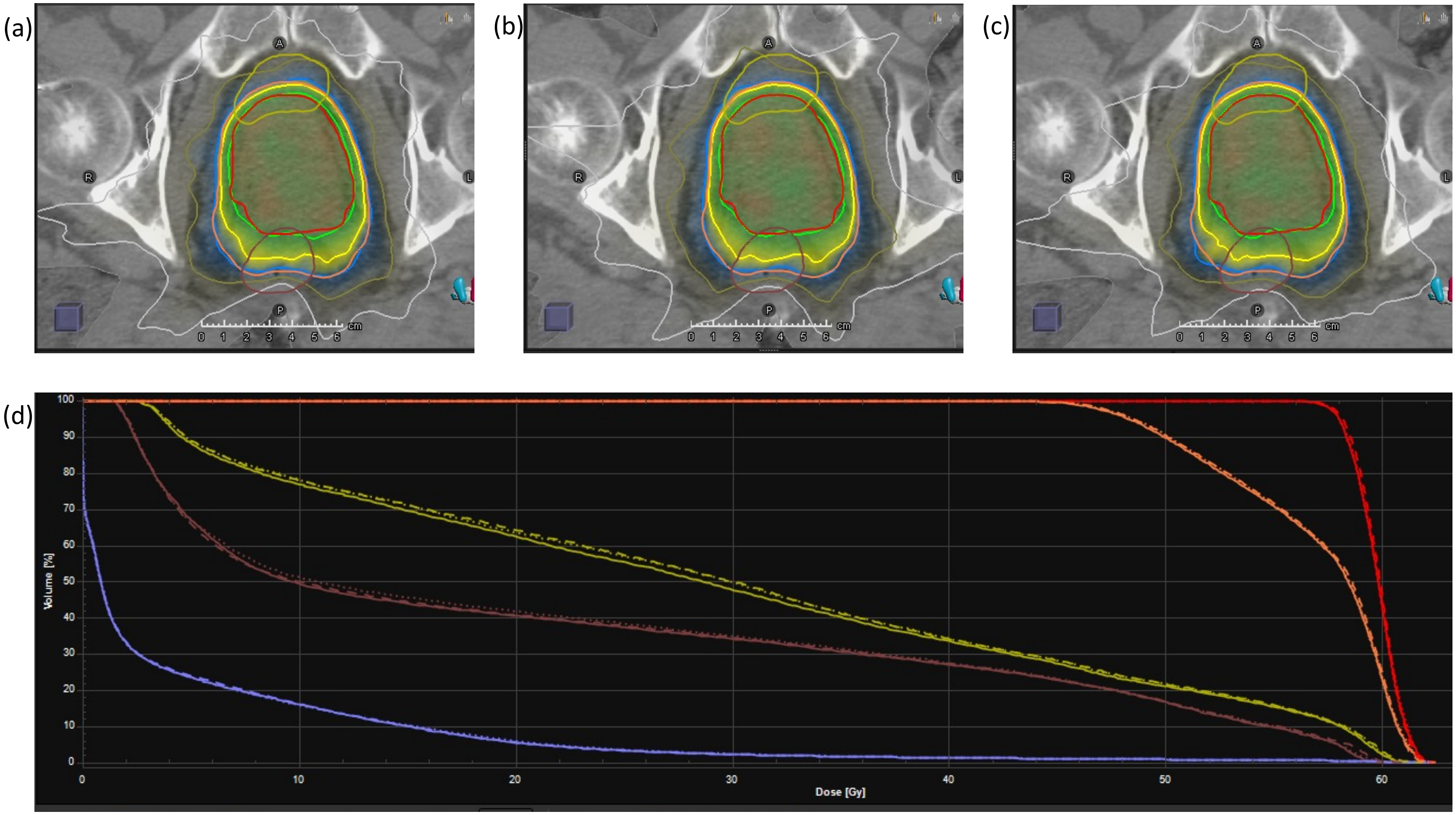


Figure 4.5: A single transverse slices of Patient 15: a patient that stood out as an outlier in for a number of key dose-volume metrics. Panels are: (a) Session 1, (b) Session 2, (c) Session 3 and (d) the DVH of these three sessions. DVHs are: Session 1 (solid line), Session 2 (dotted line) and Session 3 (dashed line). The five ROIs are: External (purple), rectum (brown), bladder (yellow), PTV48 (orange) and PTV60 (red)

4.6 Inter-planner study

4.6.1 Introduction

Even in highly regulated fields such as that of radiotherapy planning where qualified practitioner adhere to strict local and universal practices, it is not unreasonable to expect individuals will make different choices when presented with the same task. Controls are in place to encourage safe and consistent planning that has congruence with clinical preferences. Nevertheless, studies relating to inter-planner variability have already lead to some heuristics surrounding variability between the planning behaviours of different individuals.

van Beek (2018) conducted an inter-planner study to determine consistency in plan selection for rectal cancer. The study found that improvements in defined guidelines and an increase in experience of planners increased their level of agreement from 69% to 87% in five months[167]. Erkal (2022) wanted to assess inter-planner variability with respect to clinical preference and use the findings to define a planning protocol for the prostate treatment site by isolating key optimisation objectives. Four planners produced IMRT plans for 15 patients with plans assessed using dosimetric objectives. Notable differences were observed dosimetrically including as the number of monitor units per plan even when planners planned according to standardized protocol but found pre-determined optimization protocols enable a transfer of experience[168].

Therefore, studies have shown consistent planning among practitioners of the same institution is attainable with the experience of the planner being one of the most significant factors in deviations from clinically preferred planning. This study aims to explore discrepancies in choices made by different qualified practitioners when using a PGAP. Although all are qualified, practitioners were from a range of backgrounds and experience levels. Given a PGAP system is being used, it is hypothesised that the interactive and intuitive nature of this approach will enable planning choices reflecting planner-specific clinical preference. Hence, findings of this work will help to determine the clinically relevant region of the Pareto front as defined by a range of qualified individuals.

4.6.2 Methods

All sessions took place between 1st December 2019 and 28th February 2020 and four qualified professionals familiar with the PSV treatment site were selected to take part:

- medical physicist (participant A)
- two oncologists (participant B and D).
- clinical technologist (participant C)

All professionals were fully qualified, highly familiar with the PSV treatment site and had multiple years of experience. The clinical technologist will here and after be referred to simply as a professional planner.

Given this was an inter-planner study where the differences in weighting factors was the key outcome, it was not considered strictly necessary that a full clinical plan be calibrated. For this reason, only PGs 1-3 were calibrated with the remained PGs held constant at the level defined by the original AutoPlan protocol. In this way, only PGs with the most significant trade-off relationships were considered. Also, inter-planner choices could be assessed without unnecessary burden to institution and its clinical resources. Five weight levels were chosen for each PG with the middle value based on that given in the clinically defined AutoPlan protocol. The remaining four weights were chosen such that they followed a geometric progression.

Eight PSV patients were chosen for calibration from the set of patients defined in section 3.6 and correspond to Patient 11-18 from this set. In this study, patients have been label 1-8 for simplicity. A small set of patients was chosen such that it was considered large enough to observe a sufficient range of anatomies but small enough not to become a time consuming task for participants. Participants completed the task under similar conditions to the intra-planner case, in an environment fit for clinical planning. They had access to the clinical goals and could interact with the TPS however they desired. However, given not all participants were familiar with the PGAP system, they were all required to complete a practice case before completing the eight study cases. As in the intra-planner study, plans were compared in terms of absolute weighting factors, relative weighting factors and dosimetric features. The results of the practice case were not considered in this study.

4.6.3 Results

4.6.3.1 Weights

See Table 4.5 for a summary of navigated weighting factors. As expected, rectum D_{mean}, bladder D_{mean} and PTV conformality showed similar relationships in absolute form as in relative form. Differences were observed between participants for all PG groups except

Planning Goal		Participant A	Participant B	Participant C	Participant D
Absolute	**Rectum $\mathbf{D_{mean}}$**	94.1 ± 19.8	138 ± 59.3	53.6 ± 12.1	207 ± 80.7
	Bladder D_{mean}	35.6 ± 10.9	44.0 ± 25.7	59.2 ± 27.7	42.8 ± 20.9
	PTV Conformality	235 ± 71.3	118 ± 75.9	176 ± 25.2	251 ± 143
	Other PGs	2278	2278	2278	2278
Relative	**Rectum $\mathbf{D_{mean}}$**	3.55% ± 0.659%	5.26% ± 2.11%	2.09% ± 0.462%	7.27% ± 2.66%
	Bladder D_{mean}	1.34% ± 0.392%	1.67% ± 0.908%	2.30% ± 1.07%	1.50% ± 0.725%
	PTV Conformality	8.84% ± 2.41%	4.48% ± 2.72%	6.85% ± 0.957%	8.71% ± 4.53%
	Other PGs	86.3% + 2.95%	88.6% + 4.96%	88.8% + 1.00%	82.5% + 7.27%

Table 4.5: Summary of navigated weights for each PG group. Values are mean ± standard deviation of navigated weights over all eight patients. Statistically significant difference differences are indicated at the 5% level.

bladder D_{mean}. See Figure 4.6 for the distribution of relative weights across the patient database for each participant.

ANOVA assumptions were not met by the rectum D_{mean} PG and a Friedman test of significance was used. All other PGs were tested using ANOVA. Statistically significant differences observed were the same for absolute weights and there relative weight counterparts. No significant difference were observed for bladder D_{mean} but were observed for all other PG groups.

Participant C prioritised rectum D_{mean} significantly lower than D and C and showed consistently lower prioritisation of this PG in all cases. Participant A prioritised PTV conformality higher than B overall. Borderline significant differences were observed for the PG4 and higher PG group given comparably low prioritisation within this group for participant D than any other group. Weighting factors for this group were not recalibrated from the original AutoPlan protocol but were prioritised lower for participant D due to high relative prioritisation of the calibrated PGs.

For rectum D_{mean}, the highest degree of similarity was observed between B and D and the lowest level of similarity observed between C and D with DiceC values 0.884 and 0.493 respectively. The PG with the highest degree of agreement between participants was bladder D_{mean}. The lowest DiceC value for bladder D_{mean} was observed for C and D with a value of 0.672. Similarity metrics indicate A and C prioritised PTV conformality similarly and the significant difference between B and D yielding the lowest DiceC with values of 0.950 and 0.625 respectively. All DiceC values are higher than 0.99 for the PG 4 and higher group. However, the most dissimilar pair was B and D with a DiceC value

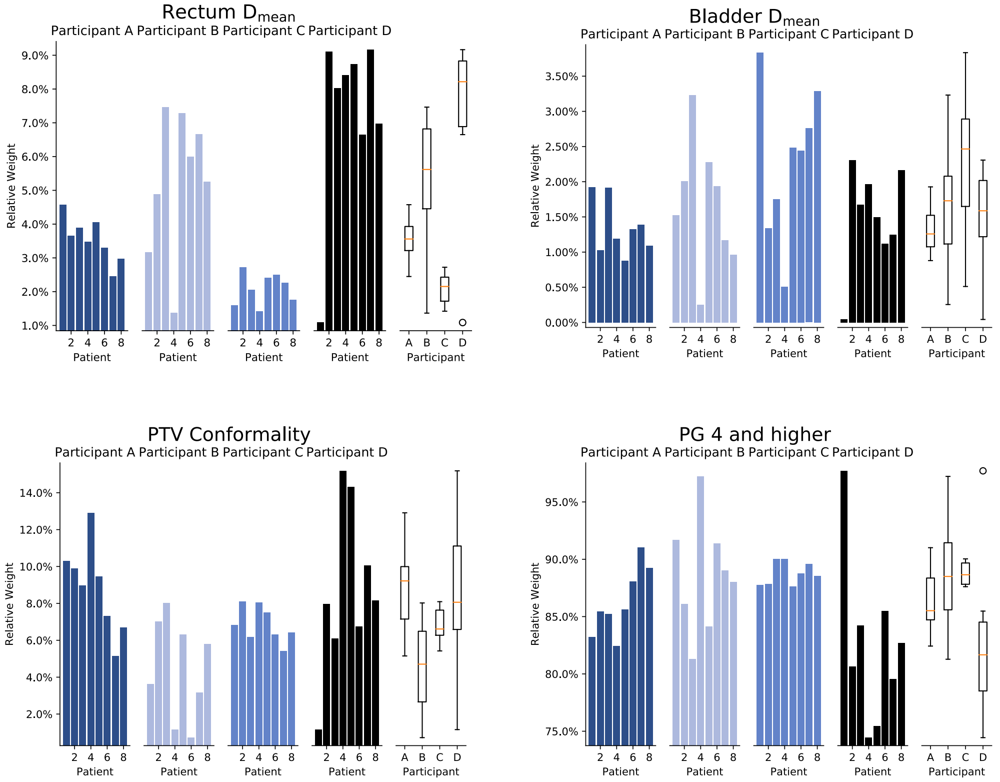


Figure 4.6: Relative weighting factor choices distributions.

of 0.9930.

Patient 4 stood out as an outlier given notable relative weighting factor inconsistencies when compared to other patients. Participant B and C assigned Patient 4 a comparably lower weight than other patients for rectum and bladder D_{mean}. Participants A and D assigned a higher than average weight to patient 4 for PTV conformality and participant B and D assigned this patient a comparably lower weight for the higher PG group. Figure 4.7 shows a sagittal slice of Patient 4. This patient has the largest PTV48 and PTV60 volume in the patient database with volumes 3.07 and 2.66 times the database median respectively. This patient also shows an atypical bowel position with notable PTV overlap. The PTV overlap meant prioritisation of the rectum and bladder PGs is likely to be lower than average to ensure appropriate dose coverage of the PTVs whilst avoiding undesirable compromising of the bowel. Also, patient 1 was a notable outlier for participant D only and resulted in weights for patient 1 that were outliers for rectum D_{mean} and the higher PG group for this patient.

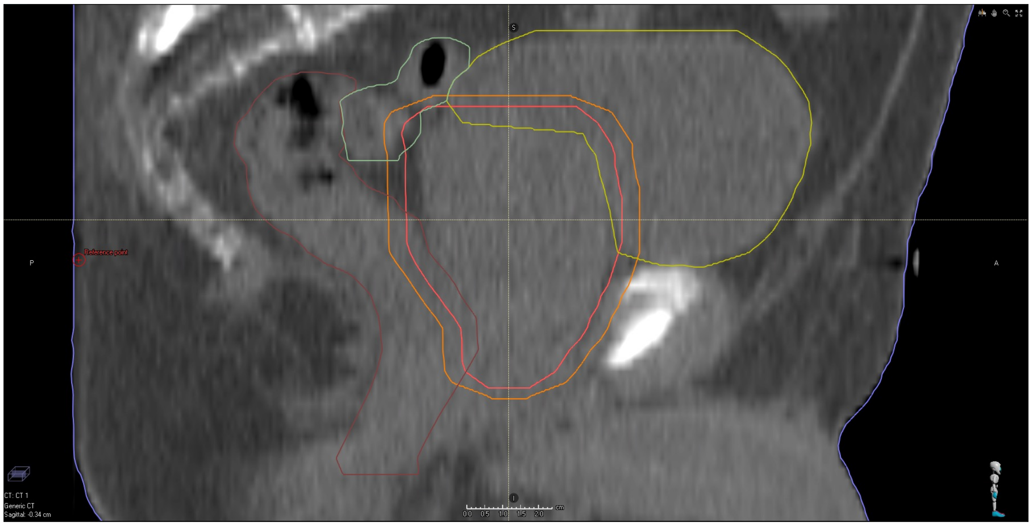


Figure 4.7: A sagittal slice of Patient 4 showing the overlap of the delineated bowel with the PTVs. ROIs include: rectum (brown) bowel (green), bladder (yellow), PTV60 (red) and PTV48 (orange)

4.6.3.2 Dosimetry

Normality assumptions were violated for four dose-volume metrics and a Friedman test was used. These included PTV60 $D_{98\%}$ (Gy), PTV48 $D_{50\%}$ (Gy), conformality indices of PTV60 and rectum $V_{60.8Gy}$ (%). See Table 4.6 for a summary of the key dose-volume metrics of interest and Figure 4.8 for an illustration of key metrics. Few statistically significant differences were observed with differences found for four metrics only: PTV60 $D_{98\%}$ (Gy), PTV48 $D_{50\%}$ (Gy) and conformality indices of PTV48 and PTV60.

Difference were observed between participant C and D for PTV60 $D_{98\%}$ (Gy). Higher doses were observed for C than D with a mean difference of 0.155 Gy. PTV48 difference observed for $D_{50\%}$ (Gy) related to participants A and B only. Observed dose was lower for A than B with a mean difference of 0.832 Gy. For CI60 observed difference relate to participant A and B with observed indices low for A on average given a mean difference of 0.0315 units. CI48 saw participant B observe lower indices than all other participants with deviations of 0.0973, 0.0899 and 0.0546 units for A, C and D respectively. All observed difference were considered clinically small indicating differences is planning decisions may be clinically negligible with this PGAP system.

The two most notably comparable participants in terms of dose-volume metrics are participant A and C. On average, these participants differ little across the metrics. Figure 4.8 also illustrates the comparability of these dosimetric outcome for these participants at

DVH Statistic		Participant A	Participant B	Participant C	Participant D
PTV60	**$D_{98\%}$ (Gy)**	57.6 ± 0.264	57.7 ± 0.148	57.6 ± 0.264	57.5 ± 0.303
	$D_{50\%}$ (Gy)	60.0 ± 0.0746	60.0 ± 0.0592	60.1 ± 0.0756	60.0 ± 0.0537
	$D_{2\%}$ (Gy)	61.7 ± 0.0816	61.6 ± 0.0592	61.7 ± 0.086	61.7 ± 0.104
	CI	0.850 ± 0.00965	0.820 ± 0.0235	0.845 ± 0.0111	0.838 ± 0.0187
	HI	0.0686 ± 0.00547	0.0658 ± 0.00305	0.0677 ± 0.0055	0.0698 ± 0.00668
PTV48	$D_{98\%}$ (Gy)	46.2 ± 0.197	46.4 ± 0.469	46.4 ± 0.238	46 ± 0.644
	$D_{50\%}$ (Gy)	53.8 ± 0.515	54.6 ± 0.907	54.0 ± 0.386	54.1 ± 0.681
	$D_{2\%}$ (Gy)	59.2 ± 0.202	59.4 ± 0.329	59.2 ± 0.118	59.3 ± 0.269
	CI	0.815 ± 0.0156	0.717 ± 0.0549	0.808 ± 0.0175	0.769 ± 0.0417
	HI	0.242 ± 0.00432	0.239 ± 0.00859	0.237 ± 0.00538	0.246 ± 0.0116
Rectum	$V_{24.3Gy}$ (%)	27.2% ± 5.83%	25.3% ± 5.24%	28.6% ± 5.86%	24.8% ± 4.70%
	$V_{32.4Gy}$ (%)	22.1% ± 5.23%	20.9% ± 4.57%	22.8% ± 5.25%	20.4% ± 4.09%
	$V_{40.5Gy}$ (%)	17.2% ± 4.36%	16.8% ± 3.89%	17.8% ± 4.38%	16.1% ± 3.43%
	$V_{48.6Gy}$ (%)	12.1% ± 2.99%	11.9% ± 2.88%	12.4% ± 3.08%	11.2% ± 2.39%
	$V_{52.7Gy}$ (%)	8.93% ± 2.07%	8.84% ± 2.12%	9.25% ± 2.20%	8.30% ± 1.71%
	$V_{56.8Gy}$ (%)	5.25% ± 1.38%	5.16% ± 1.46%	5.32% ± 1.54%	4.78% ± 1.13%
	$V_{60.8Gy}$ (%)	0.0792% ± 0.1050%	0.0731% ± 0.1420%	0.0470% ± 0.1150%	0.0514% ± 0.0761%
	D_{mean} (Gy)	17.9 ± 2.80	17.0 ± 2.80	18.7 ± 2.77	16.7 ± 2.26
Bladder	$V_{40.5Gy}$ (%)	12.6% ± 6.27%	12.9% ± 6.65%	12.5% ± 6.42%	12.8% ± 6.26%
	$V_{48.6Gy}$ (%)	8.53% ± 4.29%	9.17% ± 5.07%	8.57% ± 4.44%	8.71% ± 4.40%
	$V_{52.7Gy}$ (%)	6.83% ± 3.60%	7.27% ± 4.05%	6.83% ± 3.67%	6.96% ± 3.68%
	$V_{56.8Gy}$ (%)	4.81% ± 2.67%	5.11% ± 3.01%	4.79% ± 2.74%	4.94% ± 2.76%
	D_{mean} (Gy)	15.9 ± 6.44	15.2 ± 6.24	15.5 ± 6.70	15.7 ± 6.33

Table 4.6: Summary of key dose metrics. Values shown are mean ± 1 standard deviation and statistical difference at the 95% level of significance of non-parametric analysis of variance are indicated in boldface.

patient level with overall distributions showing high levels of similarity. This may suggest planners and medical physicists follow similar heuristic knowledge and interpretation of clinical desirability.

Patient 4 (Figure 4.7) also stood out for dosimetric irregularities showing the highest CI60 and HI60 values across all participants as well as higher doses to bladder and rectum than any other patient. Again, this was primarily due to the size and location of the delineated bowel that overlapped PTVs. Patient 1 has notably low doses to OARs and achieves a desirably high dose to PTVs irrespective of the participant. Patient 1 had the largest external delineation and largest rectum volume in the database. Within the inverse optimiser, this may have made the sparing of comparably large percentage volumes of the rectum easier whilst still maintaining reasonable PTV coverage.

4.6.4 Discussion

There still exists a gap in the literature for further inter-planner studies, but of those that do exist, there is evidence showing inconsistencies in participant choices[165,168–170]. Given this expectation, the aim of applying PGAP to mitigate discrepancies was explored here with a view of observing clinically significant differences.

It was observed that oncologists (participant B and D) applied a higher priority to

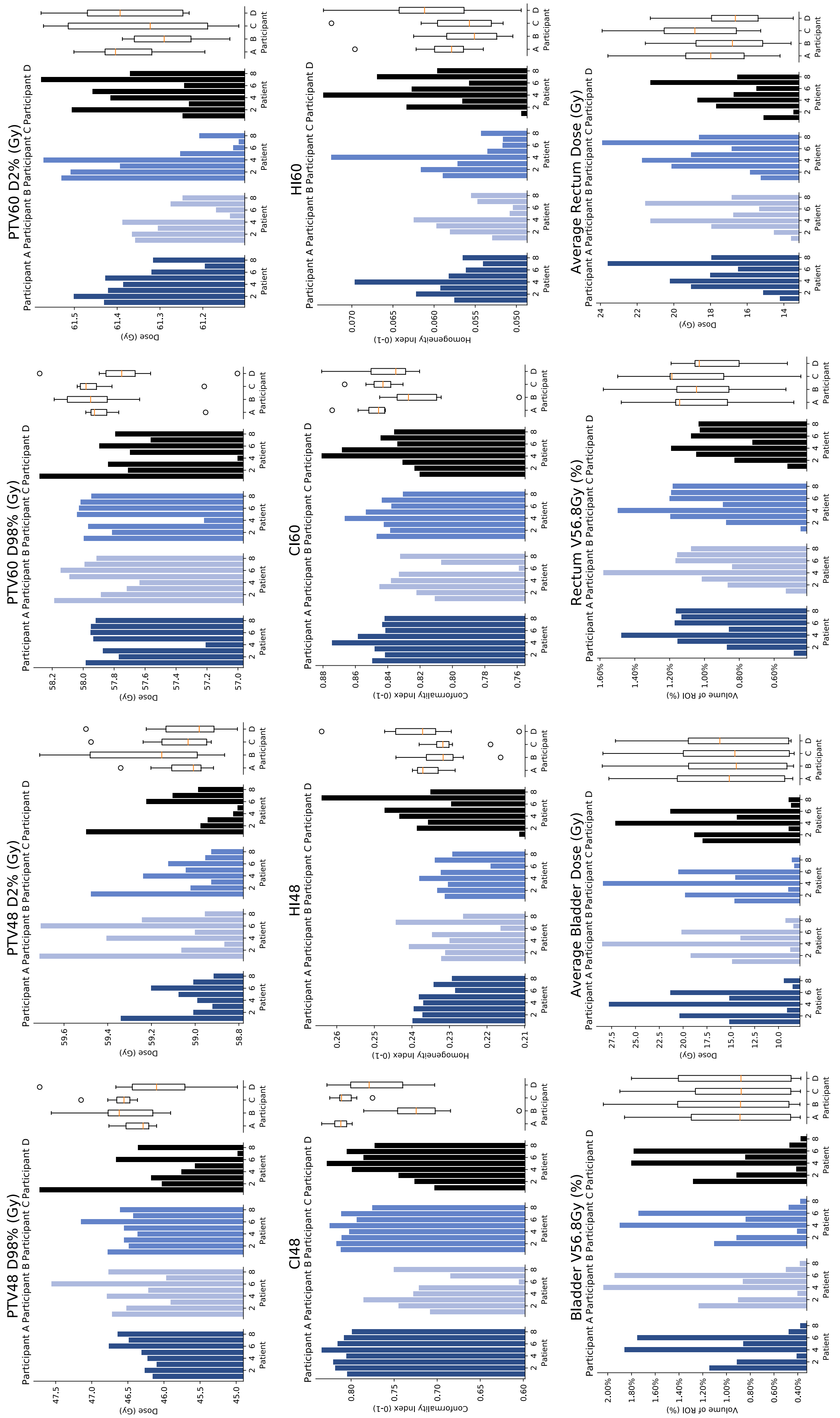


Figure 4.8: Distribution of values for some key dose-volume metrics by each participant.

sparing the rectum than do planners or physicists. Following interviews with the participants, participant B stated a preference to push dose in the anterior direction to help spare the rectum even at some cost to conformality or even increasing dose to the bladder. This participant considered the rectum a notably higher clinical priority over the bladder and would increase sparing of the rectum given a suitable dose distribution to PTVs was still achieved and all clinical goals were being met.

A key difference between the oncologists in this study was the tendency for participant D to use a higher range of the navigation scale than participant B. The tendency of participant D to use the higher end of the scale resulted in a generally higher priority to the navigated PGs over the other PGs than is seen for any other participant. However, interviews with participant D revealed simple preferences. Participant D wanted to ensure the achievement of the clinical goals but had fewer concerns about the planning details than some of the other participants. Although traditional IMRT planning methods have been criticised for being tedious and lacking an intuitive approaches that facilitates interaction of physicians[67], this work suggests clinical preference can at times be broad. The number of clinically applicable choices can be overwhelming for physicians even with the use of intuitive techniques such PGAP.

The physicist (participant A) and the planner (participant C) performed the most similarly by default with participant A in particular showing notably greater levels of consistency in planning choices between patients. Nevertheless, the PBAIO system was valuable in mitigated the majority of discrepancies in deviations at the calibration stage with few statistically significant differences dosimetric observed none of which were clinically significant.

4.6.5 Conclusion

When calibrating weighting factors with a PGAP for prostate seminal vesicle, individuals will make different choices even when all participants are considered expert-level qualified professions. Differences were considered partially due to background and experience. Oncologists prioritised the three PGs similarly but used different regions of the sliding scale indicating difference in clinical preference even among members of the same group and institution. The planner showed the greatest level of consistency between patients but dosimetric difference were negligible between participants and patients. There is evidence that expert-driven PGAP can be used to deliver consistent dosi-

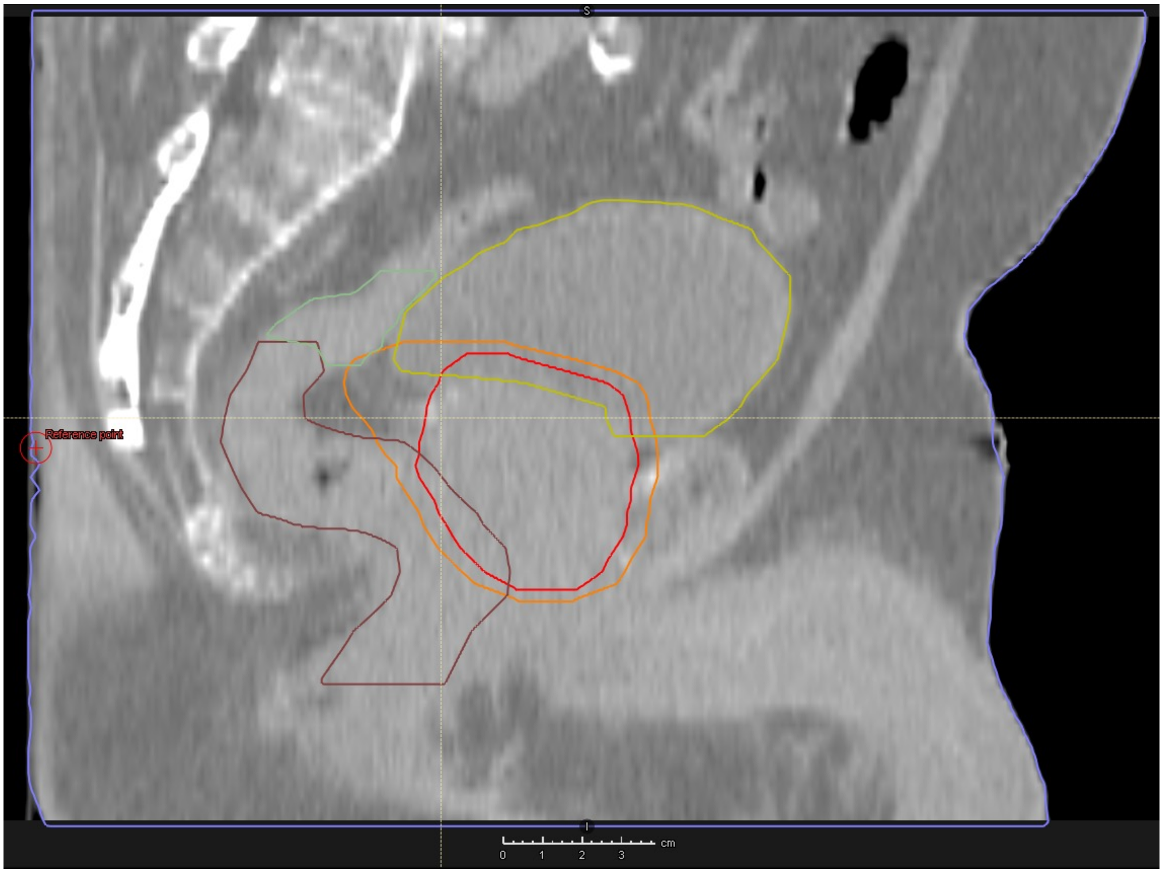


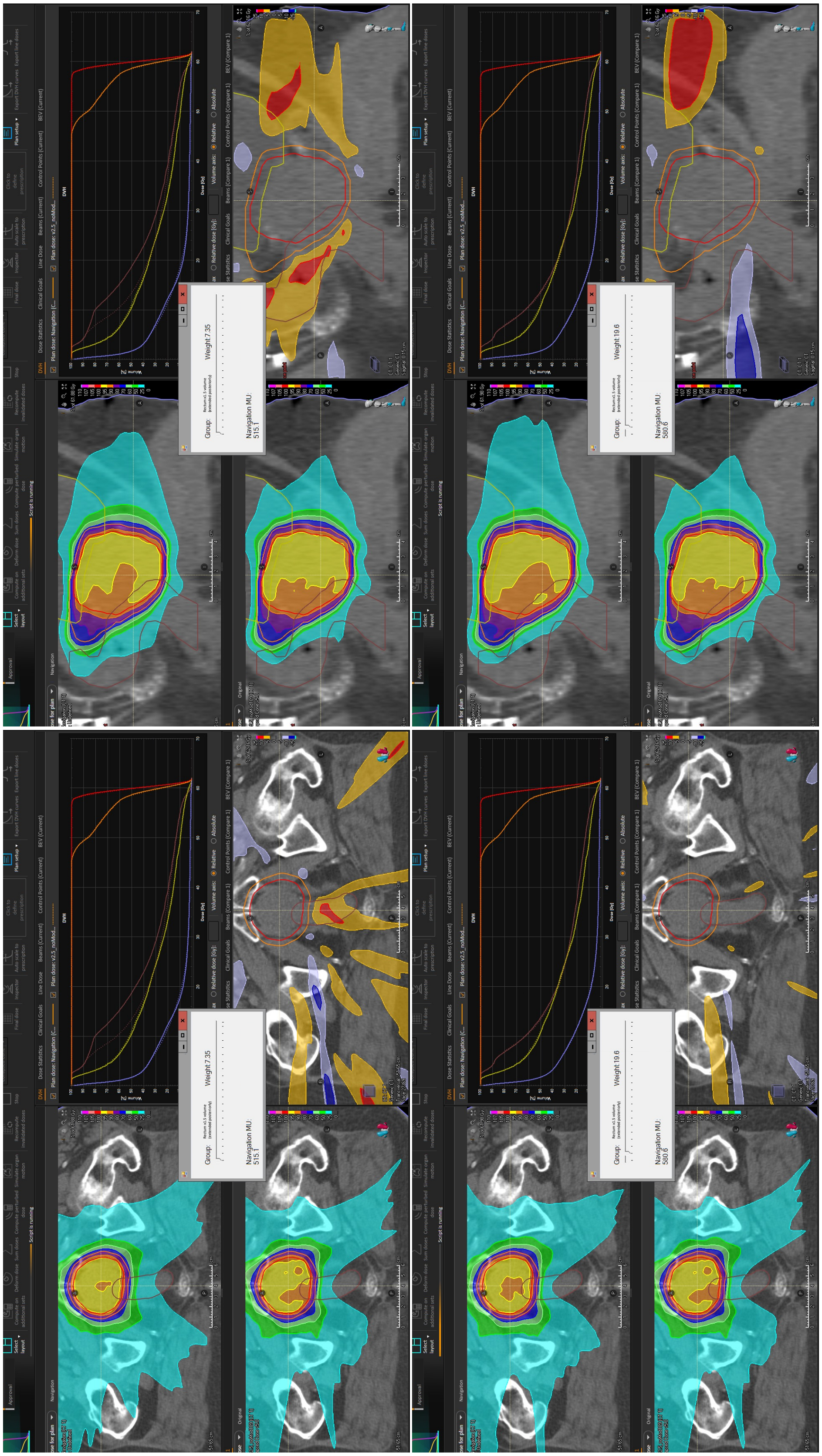
Figure 4.9: A sagittal slice of a prostate seminal vesicles patient. Delineated are the rectum (brown), bladder (yellow), low dose PTV (orange),/ high dose PTV (red), bowel 2cm superior of the low dose PTV (teal green) and external (lilac)

metric planning with the clinically relevant region of the Pareto front defined comparably by any expert.

4.7 Anatomy simulation

4.7.1 Introduction

As illustrated by the dosimetric outcomes of the previous two studies, dosimetry is determined following inverse optimisation and cannot be assessed apriori from the input parameters alone. Therefore, the direct relationships between input parameters, anatomy and dosimetry can be difficult to understand and can be a road block to efficient planning. Given planning is a strongly geometric problem, studying patient geometry and dose distribution in relation to gold standard planning parameters is valuable. However,



Group:
Weight 7.35
Navigation MU:
515.1

Group:
Weight 19.6
Navigation MU:
580.6

Figure 4.10: An example navigation session. Shown is the chosen weighing factor for a 1.5 times volume posterior expansion of the rectum that results in a similar dose distribution to that of the original rectum.

studying variations in anatomy can inherently comes with confounding variables related to individual patients given no two patients are necessarily comparable.

In order to directly research the relationships between anatomy, planning parameters and dosimetry, an alternative approach could be to start with a single patient. For example, one could determine the necessary change in planning parameters required to maintain a certain dose distribution given augmentation of a specific anatomy. First assume an applicable planning solution (i.e. set of weighting factors) is defined for said patient; that is an expert-driven clinically applicable gold standard plan. If one (and only one) element of said patients anatomy varies, so will the dose distribution given the original planning parameters. To maintain a comparable dose distribution for the augmented anatomy to that of the original non-augmented version, planning parameters must be tweaked. The question is, is it possible to determine the necessary change in related weighting factors to results in a comparable dose distribution to the original gold standard plan for the augmented anatomy? If this is possible, it may be possible to directly model the relationship between anatomy, weighting factors and dosimetry. This will be a valuable finding for plan calibrations and dose prediction in general with this automated planning system.

4.7.2 Method

The chosen patient was taken from the set of PSV patients defined in section 3.6 and corresponds to patient 04 in that set. Figure 4.9 shows a CT scan slice of the PSV patient of choice. This patient was chosen given their standard anatomical qualities including ROI sizes and PTV overlaps with OARs. This patient's base anatomy was used in all simulated cases. As found in the intra- and inter-planner studies, expert-driven PGAP can be used to calibrate a clinically applicable solution for patients and a single expert medical physicist navigated a solution for this patient using EdgeVcc's Pareto navigation sliding interface and an AutoPlan protocol defined based on the navigated weighting factors. As a base, the clinically approved PSV planning protocol presented in Tables 4.1a-4.1d was used. This was a preliminary exploration, and for simplicity only the three most high priority PGs were navigated: 1. rectum D_{mean}, 2. bladder D_{mean} and 3. PTV conformality. Remaining PGs were held constant at their predefined values as of the clinically approved AutoPlan protocol similarly to the inter-planner study above. All new volumes were created in Raystation using the built-in ROI algebra functions that enable generation of new and derived volumes.

The PG most closely associated with the augmented ROI was re-navigated in each case with all other values remaining held at the level defined by the expert-driven gold standard PGAP navigation and base AutoPlan protocol. Therefore, for each augmentation, a new weighing factor was determined for the PG most related to the augmented ROIs only. During re-navigation, dose distribution of augmented anatomy was compared to that of the original and similarity between dose distributions was determined using the 3D dose and DVH curves. The chosen plan for each augmentation would minimise DVH deviations from the original plan especially for the non-augmented ROIs and the dose distribution would map comparably well to other possible navigation choices. Figure 4.10 shows a rectum D_{mean} re-navigation and the chosen plan resulted in a comparable dose distribution to that containing the original rectum contour.

Given the three PGs rectum D_{mean}, bladder D_{mean} and PTV conformality, the three associated ROIs selected for augmentation were rectum, bladder and external volumes respectively. ROIs were augmented one at a time and such that expansions or contractions were approximately equal to some predefined scaler of the original volume. For example, contractions of non-external volumes (i.e. bladder and rectum) were always 0.5 or 0.75 times the original volume and expansions were up to 3 times the original volume. Augmentations of differing kinds were considered e.g. posterior only. This meant not only could the general relationship between parameters and volume be explored, but also specific relationships between parameters and certain kinds of anatomical variations.

The main requirement for augmentation was to avoid excessively encroaching on PTVs beyond that found in the original case. Figure 4.9 shows the rectum and bladder contours encroaching on the two PTVs in the original anatomy. Rectum expansions are therefore not expanded in the anterior direction and inferior bladder extensions avoid further PTV overlap. Figure 4.11 illustrate some examples of expansion types. Generally only one internal ROI contraction was applied to each internal ROI and expansions were limited to those that would not encroach on or go outside of the external volume. No expansions of the external volume were considered and external volume contractions were limited to those not encroaching on internal ROIs. For a summary of rectum ROI augmentations, see Table 4.7. Similar approaches were taken for bladder and the external volumes and navigated weighting factors obtained for all of the augmented anatomy generated.

Five augmentation types were considered in the rectum scenario with seventeen aug-

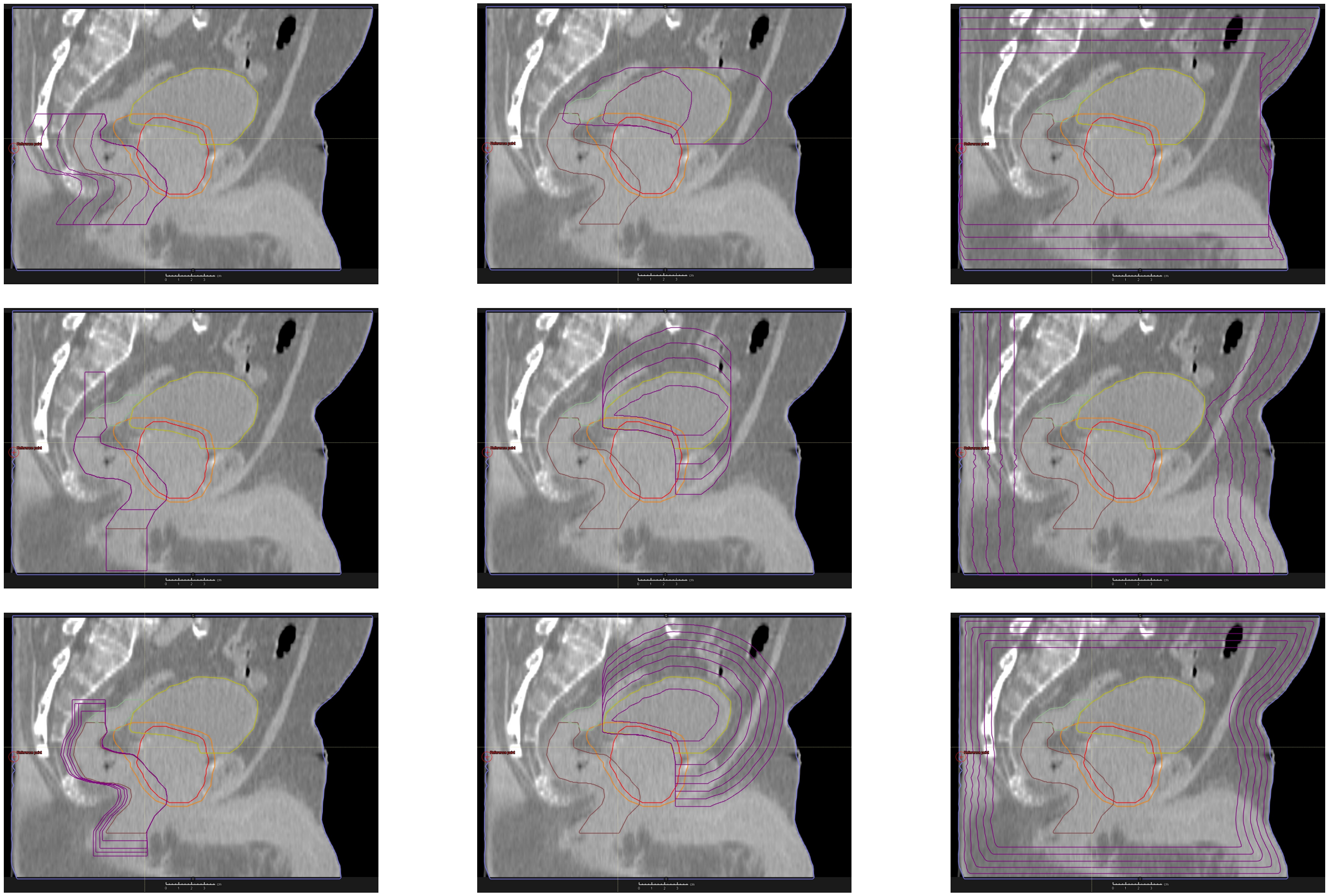


Figure 4.11: An illustration of some of the volume augmentation for rectum (left), bladder (middle) and external (right). Included are augmentations for rectum include: posterior only, superior-inferior and superior-inferior-posterior expansions. Expansions for the bladder volume include: posterior-anterior, superior-inferior and superior-inferior-anterior expansions. Contractions for the external include: superior-inferior, posterior-anterior and isotropic contractions.

Expansion Type	Relative volume to original OAR	Volume	Rectum D_{mean} navigated WF
Original	1	61.8	21.2
superior & inferior	0.75	47.3	29.4
	1.5	92.8	39.2
posterior only	0.5	30.6	44.1
	1.5	93.1	19.6
	2	124	24.5
	2.5	155	31.9
left & right	1.5	91.3	22.1
	2	124	34.3
	2.5	153	46.6
superior, inferior, left & right	0.75	45.5	19.6
	1.5	93.0	24.5
	2	122	31.9
	3	186	53.9
	4	248	85.8
superior, inferior, left, right & posterior	1.5	92.9	24.5
	2	120	29.4
	2.5	157	36.8

Table 4.7: Summary of re-navigated weighting factors for each of the augmented rectum volumes

mentations of the original rectum volume (Table 4.7). *Superior & inferior* expansions consisted of copies of the most superior and inferior transverse slices transposes such that the new volume extends directly up and down vertically with no change to the core shape of the original rectum. Contractions were created by removing superior and inferior transverse slices from the original volume. Posterior expansions were created using the Raystation ROI algebra function for new ROI geometry creation and volumes were grown and shrunk in the anterior-posterior directions only. This is similarly true for *left & right* expansions which were created using ROI algebra also.

When all volume had been re-navigated, the weighing factor absolute values were analysed. Relative weighting factor values were not considered in this study given weighting factors for one and only one PG were being changed at a time and any change in relative values of other PGs are due to this and only this change in every scenario. Three variables relating to ROI volume (volumetric features) were extracted for each that included 1. volume of ROI, 2. volume of ROI within the superior and inferior CT slices of the low dose PTV (volume in field or VIF) and 3. volume of ROI outside of the superior and inferior CT slices of the low dose PTV (volume out of field or VOF). Relationships were illustrated using scatter plots and assessed using lines-of-best fit and coefficients of determination (R^2 values) as goodness-of-fit metrics. These measures will be discussed in more detail in the following chapter.

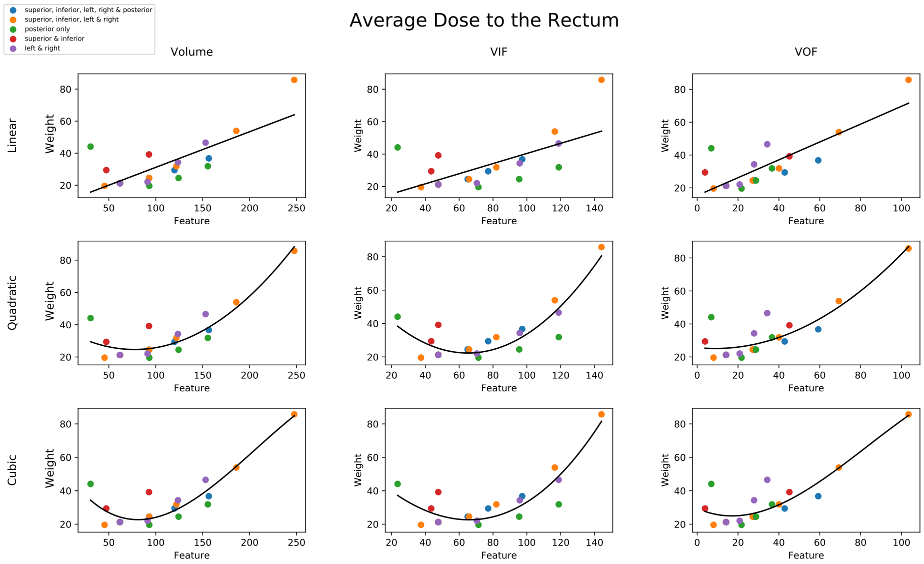


Figure 4.12: Plots showing the linear, quadratic and cubic line-of-best fit for rectum D_{mean} weighting factors plotted against three features: volume of the rectum contour (Volume), volume of the rectum contour in field of PTV48 (VIF) and volume of the rectum contour out of field of PTV48 (VOF).

4.7.3 Results

Seventeen, 20 and 21 different augmentations of the rectum, bladder and external were generated. See Figure 4.12-4.14 for scatter plots and representative linear, quadratic and cubic models. Navigation of rectum and bladder D_{mean} weighting factors was simple for each augmentation case with the most appropriate new weighing factor readily discernible. Navigation of a new weighing factor for PTV conformality was not always as readily discernible. This was due to DVHs for new anatomy not always being comparable to the original case and 3D dose was found to be quite different from that of the original anatomy.

Coefficients of determination (R^2 values) increase as the degree of the polynomial of the model increases. That is, as the model becomes increasingly complex and progresses from linear to quadratic to cubic, the models fits better and deviations of individual values decreases. Therefore, the best models (models with largest R^2 values) were all cubic with respect to R^2 metrics. For the rectum D_{mean} PG, the best model was produced using the Volume feature. For bladder D_{mean} PG VOF maximised R^2 and for PTV conformality VIF maximised R^2. R^2 values for the best models were 0.860, 0.837 and 0.328 for rectum D_{mean} bladder D_{mean} and PTV conformality respectively.

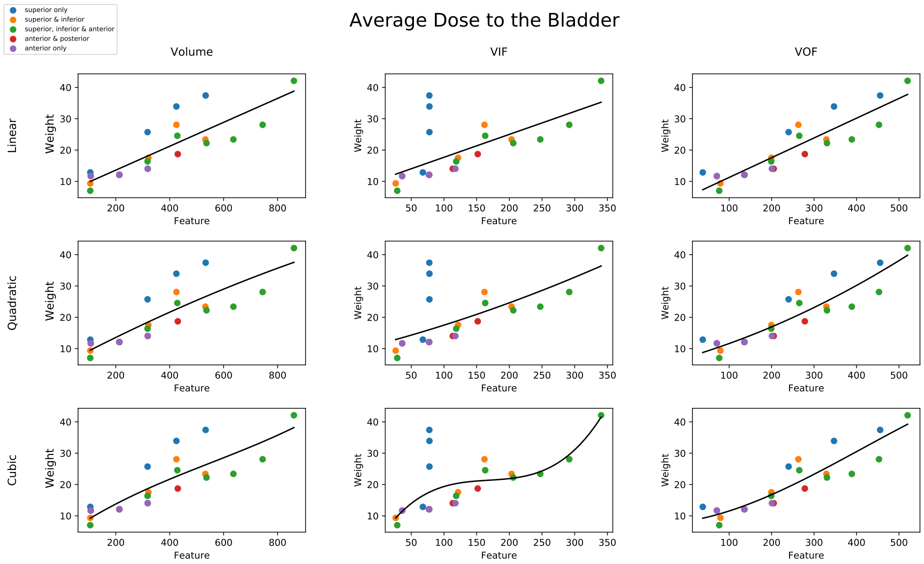


Figure 4.13: Plots showing the linear, quadratic and cubic line-of-best fit for bladder D_{mean} weighting factors plotted against three features: volume of the bladder contour (Volume), volume of the bladder contour in field of PTV48 (VIF) and volume of the bladder contour out of field of PTV48 (VOF).

4.7.4 Discussion

Bladder D_{mean} produced monotonic increasing models in all cases with an increase in volume resulting in an increase in weighing factor regardless of the nature of changes in volume or shape. Furthermore, the general relationship is strongly linear with coefficients of x^3 and x^2 close to zero in all cases. Rectum D_{mean} models were strongly quadratic given coefficients of x^3 were close to zero in all cases. Augmenting the external and re-navigating did not result in drastic or expected changes to the PTV conformality PG. In fact, choosing a new weighing factor was not always straight forward and could have been chosen differently based on the strategy taken. This may be due to the fact that this PG is not being wholly related to the external volume and suggests the relationship between PTV conformality weighting factors and anatomy is more complex than that seen for rectum and bladder D_{mean} PGs. This could require more advanced regression modelling such as multiple regression in which weighting factors are regressed against multiple variables at once or a different ML technique altogether. Nevertheless, although the relationship between PTV conformality and external volumetric features is more subtle than that of the other two PGs, models suggest a cubic relationship that may become more pronounced with advanced modelling such as multiple polynomial regression.

A key finding was the relationships between VIF and weighting factors of PGs re-

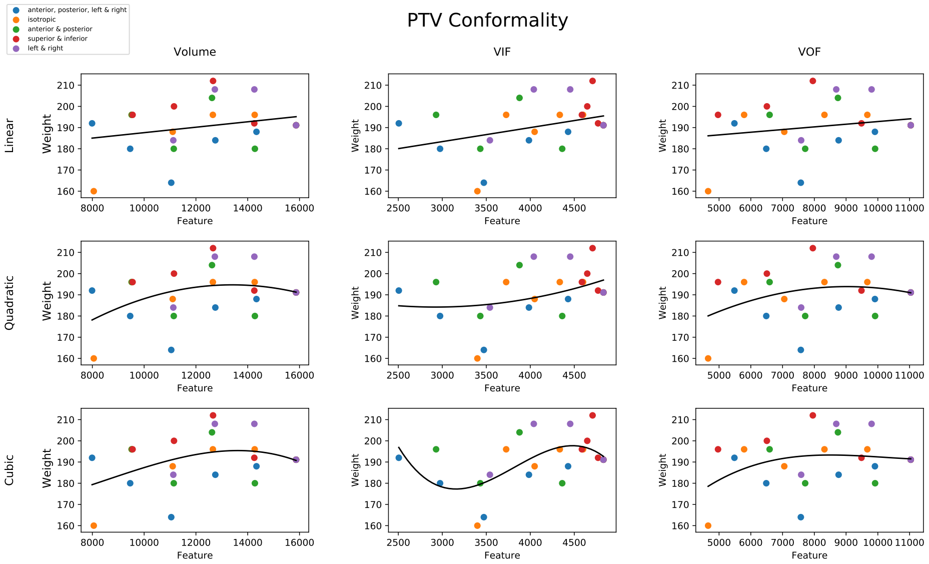


Figure 4.14: Plots showing the linear, quadratic and cubic line-of-best fit for PTV conformality weighting factors plotted against three features: volume of the external contour (Volume), volume of the external contour in field of PTV48 (VIF) and volume of the external contour out of field of PTV48 (VOF).

lated to volume increases outside of the PTV field. Rectum expansion showed increases in volume outside of the field resulted in an increase in the rectum D_{mean} weighing factor. Bladder D_{mean} did not show this relationship given increases in the bladder volume outside of the PTV field were shown to have no impact on the bladder D_{mean} weighing factor. This finding indicates there may be a relationship between the distance from the centre of an OAR from the centre of the target given the centre of the bladder volume was always outside of the PTV field. This is supported by research in automated planning that shows there is a relationship between OAR proximity to targets and dose[67,120,148,166].

4.7.5 Conclusion

This study suggests a relationship between volumetric anatomical features and weighting factors for the achievement of a given dose distribution and there is some evidence to suggest a spatial relationship between weighting factors and dose given expansions outside of the PTV field results in no weighing factor changes when the centre of the volume is located outside of the PTV field. PTV conformality weighting factors showed the weakest relationships potentially due to more complicated underlying relationships.

4.8 Chapter Summary

These three studies have together supplemented the knowledge found in the literature regarding the feasibility of defining gold standard planning and understanding relationships between anatomy and planning parameters. Findings of the intra-planner study indicate expert-driven PGAP can be expected to result in consistent planning with high degrees of similarity, few statistically significant differences and negligible impact on dose distribution. Planning choices were found to have high degrees of similarity and planner-specific gold standard weighing factor readily discernible. These findings suggest that modelling a single planners choices may be more desirable than attempting to define gold standard planning across multiple planning professionals. This study also highlighted the importance of considering planning parameters in their relative form given relative values are more closely related to objective functions.

The inter-planner study highlighted planning choice can vary between qualified planning professionals to notable degrees even amongst professions of the same training background and institution. However, this study highlighted that clinical preference can contain a broad range of choices especially after all clinical goals have already been met. Moreover, the automated planning system can be valuable at mitigating variances in individual planner choices as it led to dosimetrically comparable planning.

The volume expansion study highlighted there exist some simple relationships that are discernible for certain PGs based on volumetric features alone. Findings also suggested it may be necessary to explore more complicated models to uncover some of the underlying relationships. Spatial features are also thought to have an impact given the nature of relationships with features in proximity to targets. This study demonstrates there is a foundation for further exploration of ML techniques to model the relationships between weighting factors and anatomy.

However, there are limitations to these studies relating to the number of patient cases considered, the number of planning goals considered, the number of participants recruited and the number of treatment sites considered. The ability to make inferences and generalise findings is enhanced with more data. Obtaining more data will have been beneficial in all studies for avoidance of type I and type II errors. These limitations will be discussed in detail in section 7.1.1. In the next chapter, ML techniques are presented, explained and discussed. Included are the two chosen ML techniques including the reasons they were considered appropriate and ultimately chosen for further investigation.

Chapter 5

Modelling and cross validation

Given the modelling of weighting factors is a numerical problem, different mathematical approaches are examined for their efficacy. The field of ML has brought forward prediction techniques that may be useful in this regard and some of these established techniques are explored in this chapter and will be presented later. First, an introduction to ML itself.

5.1 About machine learning

The term "learning" refers to knowledge that is acquired from exposure to information, an established and well understood concept in a human context. Artificial intelligence is a term referring to any intelligence exhibited by a machine opposed to a human, animal or organism. It usually denotes the development of computer systems able to perform tasks previously requiring human intelligence. ML is a development in the broader field of artificial intelligence and posits it is feasible for machines to acquire knowledge[171,172]. ML can be considered a data-driven artificial intelligence given it requires a knowledge-base containing data in which to *train* on (i.e., learn from). The literature shows more than one ML approach has been developed, some of which will be discussed in this chapter, and all ML approaches can be defined in terms of two types of learning: supervised and unsupervised.

5.1.1 Supervised and unsupervised learning

The two types of learning refer to the nature of the outputs they produce. Supervised learning is categorised by the use of *labelled data* where the labels are determined in the training data[173]. That is, the models are built such that outputs have a predefined form.

Unsupervised learning does not contain such constraints and the outputs defined during modelling are based only on the relationships between data points with no prior assumptions made[173]. Some examples of supervised learning include classification, regression and neural networks, all of which define the desired output and use the data to establish the relationships between independent variables that best fit those labels. Some examples of unsupervised learning include automatic clustering, dimension reduction and association rule learning all of which are used to establish key relationships within the data that are later observed and labelled using human intelligence.

In this work, one of each learning type is explored for the purpose of weighing factor prediction. The supervised learning technique of choice was **regression** with the unsupervised choice being **automatic clustering**. These two methods will now be discussed in more detail including their comparison to other ML techniques and the reasons they were chosen over other methods.

5.2 Regression

Regression is a statistical technique used to determine a functional relationship between independent variables and dependent variables[174]. An *ordinary least squares* (OLS) method is defined such that the sum of squared differences between the points in the data and the model are minimised. All modelling in this work aims to estimate and ultimately predict the underlying relationships between predictive features and weighing factor with respect to dose. OLS regression produces statistics that are maximum likelihood estimators of population parameters (a quality that is not a given with all regression approaches)[175]. Therefore all regression models discussed in this thesis will be of the OLS type.

Linear regressions are the most common OLS regression models and are of the form[176]:

$$\hat{y_i} = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in}, \tag{5.1}$$

$$\beta_j = \frac{\sum_i (y_i - \bar{y})(x_{ij} - \bar{x}_j)}{\sum_i (x_{ij} - \bar{x}_j)^2} \tag{5.2}$$

where $\hat{y_i}$ is an m-dimensional model of the true dependent variable y_i, x_{ij} are independent variables (referred to as predictive features in this case) and $\bar{y}$ and $\bar{x}_j$ are the means of y_i and x_{ij} respectively. The β_j are known as the coefficients of the equation and chosen such that the sum of squared residuals of the model (SS_{RES}) are minimised i.e.

minimise

$$SS_{RES} = \sum_i (y_i - \hat{y}_i)^2. \tag{5.3}$$

Therefore linear models can take the general form:

$$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_n x_{in} + \epsilon_i, \tag{5.4}$$

where the ϵ_i are the *residuals* and represent the error between the model $\hat{y}_i$ and the true value y_i. In OLS regression, values of ϵ_i are normally distributed and have a statistical expected value of zero.

As seen, a goodness-of-fit metric commonly used with OLS regression is the coefficient of determination or R^2 and this can be thought of as the ratio between the sum of squares due to the regression (SS_{REG}) with the total sum of squares (SS_{TOT}). Equivalently, it can be thought of as one minus the ratio between the residual sum of squares with the total sum of squares. It is defined as:

$$\begin{aligned} \mathrm{R}^2 = \frac{SS_{REG}}{SS_{TOT}} &= 1 - \frac{SS_{RES}}{SS_{TOT}} && (5.5)\\ = \frac{\sum_i(\hat{y}_i - \bar{y})^2}{\sum_i(y_i - \bar{y})^2} &= 1 - \frac{\sum_i(y_i - \hat{y}_i)^2}{\sum_i(y_i - \bar{y})^2}. && (5.6) \end{aligned}$$

This coefficient takes a value between zero and 1 and helps to give an indication of the level of variability about the model. A model with a value close to 1 can be considered a better fit to the data than models with lower R^2 values given the former achieves smaller error terms and therefore describes more of the variability in the data. Nevertheless, there are other goodness-of-fit metrics that can be used to assess a regression model and these will be discussed in more detail in section 5.4.

Regression modelling need not be purely linear however[176]. Although the most well known form of regression is linear and will often contain a single variable, regression models can be built using more complex statistics. For example, in the same way linear regression models are expressed in a single degree polynomial (i.e., x_{ij} to the power of 1), quadratic, cubic and higher degree polynomials are also possible. Permitting higher degree terms in the formula such as x_{ij}^2 terms can increase the degrees of freedom and the complexity of the model allowing for a better fit. However, although this may be good for finding a better model that expresses the nature of the data at hand (the training data), it may not be wholly positive if models are to be used for other reasons such as predicting unseen data. This issue relates closely to the concept of overfitting and this will be discussed in section 5.4.

5.3 Clustering

Automatic clustering algorithms (also known as data clustering and cluster analysis) refers to any ML technique in which sets of data are grouped together such that members within a group are considered more similar to each other than the data outside of the group[177]. These methods will be referred to simply as *clustering* from here onward. When data are effectively clustered, the clusters aid in reducing the overall dimensions of the data and potentially help to filter out noise and establish relationships that were previously unknown. Compared to regression which produces a continuous model and defines an infinite range of possible outcomes, clustering is thought to be valuable for parameter prediction as the solution space will contain a finite number of elements hence reducing the overall optimisation problem.

Many clustering methods exist in the literature[178] and define the criteria used to determine inclusion and exclusion of each datum to a cluster. Certain clustering approaches make special assumptions about the nature of the clusters. For example, *distribution-based clustering* makes the assumption that all clusters follow a certain distribution such as a Gaussian distribution[177,179]. Similarly, *Density-based clustering* methods such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[177,180,181] assumes clusters to be high density with discernible boundaries between them. These are powerful techniques with strong applications. However, they require some knowledge about the nature of the data apriori in order to determine if they meet the assumption criteria. These models can also be difficult to visualise and conceptualise, difficult to use for modelling and prediction of unseen data, and can be computationally expensive to achieve. These approaches are not appropriate in this case given little is known about the nature of the data. The majority of clustering approaches make far fewer assumptions and these methods can be categorised into two groups[177]: hierarchical and partitional clustering.

5.3.1 Key types of clustering

5.3.1.1 Hierarchical clustering

Hierarchical clustering methods can be agglomerative (a bottom-up approach) or divisive (a top-down approach). During each iteration of an agglomerative hierarchical clustering, each data point begins as a member of its own cluster. Based on an aggregated cluster

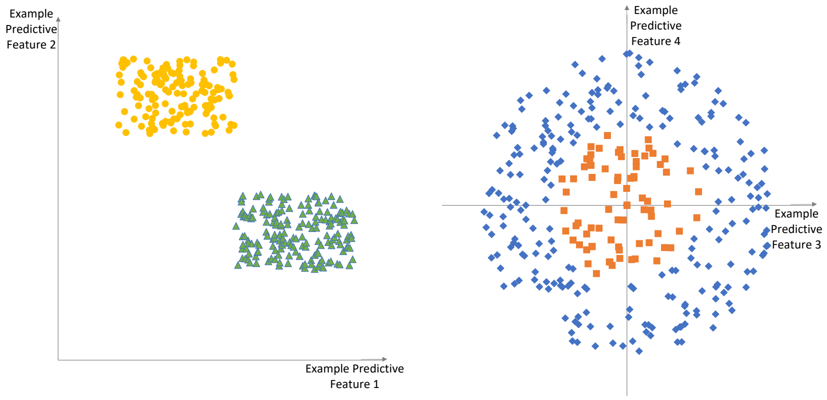


Figure 5.1: An illustration of clustering scenarios over two features. The plot on the left represents a clustering that is linearly separable. The solution on the right cannot be separated with a linear function therefore is not linearly separable problem. Adapted from Baranwal et al. (2018)[13]

value (e.g., the sum or the mean), the algorithm determines the threshold of similarity for the merge of existing clusters and clusters meeting this threshold are merged to form a new cluster in the next level of the hierarchy. In the final iteration, all remaining clusters are linked together into one master cluster. For example, in an agglomerative Ward Linkage clustering, clusters are merged if the increase in the sum of squared deviation for the mean due to the merge is minimal for all possible merges[182]. Divisive algorithms begin with all data a member of a single cluster and during each iteration, the algorithm will determine the point of greatest differentiation between data points within a cluster for separation. At the final iteration, each datum will be a member of its own cluster.

A hierarchical clustering has a few advantages over partitional clustering including simple and comprehensible visualisation and aposteriori selection of the most appropriate cluster formation. Hierarchical clustering is not dependent on an initial state but on explicitly defined values within the data and is therefore comparatively robust[183] and has been shown to maintain clustering structure well even when noise is presented into the data[184]. This form of clustering is also often well visualised using a dendrogram that shows the level at which the algorithm determined existing clusters be merged or separated to form new clusters and hence form the hierarchy. This can be very useful in choosing the most appropriate value of K especially given that for a partitional clustering it is a requirement that the number of clusters be defined upfront. This is an obstacle of partitional methods even when for the data have low dimensions (e.g., up to three variables) and can be visualised using scatter plots. When the data have higher dimensions that this, visualisation may be infeasible and the value of K can be difficult to choose.

5.3.1.2 Partitional clustering

Partitional clustering (also known as centroid- or center-based clustering) differs from hierarchical clustering in a few key ways. Partitional clustering is an optimisation algorithm with improved solutions saved after each pass. To instantiate a partitional clustering, data points are assigned to clusters arbitrarily. That is, they are clustered in some random state that can then be used as a starting point to leverage the algorithm to improve on the current state. Unlike hierarchical clustering where the appropriate number of clusters is chosen aposteriori, partitional clustering usually requires the number of clusters be defined prior to the algorithm being run.

An example, for the partitional clustering method K-medoid (also known as Partitioning Around Medoids or PAM)[185], a predefined number of clusters, K, is chosen prior to the algorithm being run and a random selection of K data points are selected as the initial cluster centroids. Using a distance or cost metric such as the Euclidean distance, the other data points are assigned to a centroid and thus a cluster. With K clusters defined, there is an attempt to establish new cluster centroids for each existing cluster by swapping the initial centroid for another data point and recalculating the distance. New centroids are defined given the distance from the new centroid is smaller than the initial one. With new centroids established, existing clusters are discarded and a new pass begins. In this work, the partitional method K-means was used. The benefits of this and other partitional methods will be outline later in section 5.3.2.3.

5.3.2 K-means and why it was chosen

5.3.2.1 About K-means

K-means is a concept that has been around since the 1950s[186,187] and is referred to as one of the most efficient and widely used clustering algorithms[180]. As with many other partitional methods, the value of K is predefined and initial centroids established based on randomly chosen real data points. Using Euclidean distance, the distance between each point and each centroid is calculated and each point assigned to a cluster based on that which minimises the distance between it and a cluster centroid. When all data points have been assigned to a cluster, new centroids are calculated by taking the mean value across each dimension of the data. For example, given three predictive features, the data have three dimensions (therefore $n = 3$) and the centroid of cluster K as a coordinate is calculated as:

$$\left(\frac{\sum\limits_{p=1}^{q} x_{1p}}{q}, \frac{\sum\limits_{p=1}^{q} x_{2p}}{q}, \frac{\sum\limits_{p=1}^{q} x_{3p}}{q}\right) \text{ or } \left(\bar{x}_1^K, \bar{x}_2^K, \bar{x}_3^K\right) \quad \text{s.t. } q \leq m. \tag{5.7}$$

where q is the number of data points in cluster K and m is the number of total data points in the dataset. Once new cluster centroids have been calculated for each of the K clusters, the algorithm begins again and continues until either it has been stable for two consecutive iterations or the predefined maximum number of iterations is reached (e.g. 300).

5.3.2.2 Alternatives to K-means

K-means is a similar clustering approach to the K-medoid approach described earlier. In a K-medoid approach, the centroid is chosen to be the medoid, a real data point within the cluster that minimises the distance metrics with all other points in the cluster[188] and this is similarly true for variants that include CLARA, CLARANS and FANNY[189]. In a K-means approach, the centroid is not necessarily a real data point, but a virtual data point. The centroid in K-means is the geometric centre or arithmetic mean of the points and is equivalent to the centre of gravity of the cluster if the cluster is thought of in terms of real matter.

Other similar approaches include K-medians that takes the median value opposed to the mean as new cluster centroids, and K-modes which can be useful for categorical data. These alternative approaches were developed to try to solve some of the issues associated with taking means such as the nature of means to be influenced more strongly by outliers than other estimators or the fact the real data points can be interpreted whereas the mean is a "virtual" value that in some cases may not carrying any intrinsic value.

There also exist alternative methods to the classic K-means approach. For example, classic K-means is an exclusive clustering approach given data points are assigned to one and only one cluster with final clusters being mutually *exclusive*. This is actually true for all classic partitional methods as well as hierarchical clustering. However, inclusive K-means approaches exist which allow data points to be assigned to more than one cluster[190]. Such approaches are less restrictive than a classic approach and may have some interesting applications for parameter prediction such as the ability to derive weighting factors by aggregating values of each associated cluster.

Another alternative method aims to mitigate the fact most standard partitional clustering approaches optimise for the clustering of linearly separable data. That is, given the clusters are separable using a linear function (such as a straight line), most clustering algorithms will converge quickly. However, this is not always necessarily the case as can be seen by the right-hand plot in Figure 5.1. Clustering approaches have been developed to try to tackle this phenomenon. Given the data are not linearly separable, a DBSCAN clustering can be used. Or, for an intuitive and potentially faster approach, kernel K-mean can be employed. This approach takes advantage of a function (known as a kernel function) that projects the data into additional dimensions and uses those new dimension to create a linearly separable data space.

5.3.2.3 Why K-means was chosen

There are a number of clustering approaches that could have been applied in this work but only K-means was chosen. The benefits and pitfalls of each method were considered and related to points discussed in the previous section. Here those points are addressed including why a partitional method was chosen over a hierarchical method, why K-medians was not chosen, why fuzzy clustering method were not considered and why kernel K-means was not used.

A benefit of partitional methods are how well suited they are to larger datasets[187] given they are generally less computationally expensive to generate. In this work it is also not valuable to determine a hierarchy of clusters, therefore it is more computationally efficient to cycling through all possible values of K to determine the optimal value. The optimal value of K can be chosen in an intuitive and efficient way, for example using metrics such as silhouette scores or the MSE from the benchmark. The MSE metrics was introduced in section 5.6.1 and silhoutte scores will be discussed in more detail in section 5.4. For this reason, hierarchical clustering was not used in this work and a partitional method was chosen.

As mentioned, the centroid defined via K-means can be criticised for being a virtual value with no intrinsic meaning. However in this work, a virtual centroid is not inappropriate given there is no restriction on obtaining a real valued points found within the training data. Also, in this work, all predictive features will be standardised (discussed in section 5.5.1) prior to clustering. Given all variables will be on a comparable scale and of a comparable distribution, K-means is considered as appropriate for use. Also, given the widespread use of K-means and the increased computational expense of other methods, with comparable performance expected from all methods following standardisation of the data, K-means was the most sensible choice.

Regarding alternatives to classic exclusive K-means clustering, fuzzy approaches may have merit especially in this work. However, application of these approaches are not widespread and remain an area of research and development[191] with many programming applications yet to incorporate validated modules and packages into their systems[192]. Also, application of these techniques require expert knowledge of the algorithms and some knowledge about the nature of the data apriori for appropriate use[193]. Research of the validity and usage of more established techniques (such as K-means) enables more confidence not only in the application but also interpretation of results. There are also partial clustering approaches that cluster based on density with certain outlying data points being assigned to no clusters. Such approaches are definitely not useful in this work, given it is necessary all data points obtain a predicted value following modelling.

Although K-means assumes data are linearly separable when in truth they many not be, alternative choices such as kernel K-means can be useful for solving this problem. However, the choice of kernel is not straight forward and can be complex[194]. Given the choice of kernel can require extensive research of its own and the nature of underlying relationships in the data are not known, the simple assumption of linear separability has been applied in this work and a classic approach taken with mutually exclusive and complete clusters considered in all cases. That is, defined clusters do not permit overlap with other clusters, all data point will be assigned to a cluster and the only assumption applied is that the data are expected to be linearly separable.

5.4 Goodness-of-fit: underfitting and overfitting

ML models can be used to discover if a relationship exists between the independent variables and the dependent variables and the nature of those relationship. This is useful given once models have been generated and relationships established, they can be used to predict the outcome of unobserved data and form the basis of a ML technique such as clustering and regression. When used for predictive means, independent variables can be referred to as *predictive features*

Once predictive ML models have been generated, goodness-of-fit metrics can be calculated and used to give an indication of how well the model defines the training data and may help to provide confidence in the model for it's predictive abilities especially when the model is simple. Examples of goodness-of-fit metrics are R^2 for regression and silhouette scores for clustering, the latter of which is defined as:

$$a(i) = \frac{1}{|q-1|} \sum_{i=1}^{q} \sum_{j=1}^{q} d(i,j), \tag{5.8}$$

where q is the number of data points in cluster K and $d(i,j)$ is the distance between points i and j. This metrics provides a value between -1 and 1 that indicates the level of coherence within clusters and the level of separation between clusters[182]. Values close to 1 indicate clusters that are desirably dense and separated with values less than zero meaning the clusters are overlapping i.e., all silhouette scores under a classic K-means clustering will be between 0 and 1.

However, robust model development should also involve an enhanced validation process. This is because goodness-of-fit metrics may give an indication of the fit of the known observations but they are not a definitive indication of the performance for novel data. This is especially true when the model is thought to be overfitting to the data. Overfitting occurs when a model fits not only the underlying relationship in the training database data but also attempts fit to the noise in the training data. In this way, goodness-of-fit metrics may appear excellent (i.e., values close to 1) but the predictive power of the model will likely be minimal when validated and tested on previously unseen data. Overfitting becomes increasingly likely as the complexity in the model increases. Therefore, as the number of predictive features and polynomial degree increases for regression, so does the likelihood of overfitting. Likewise is true for clustering when the number of clusters chosen increase. For this reason, when models are being built for predictive purposes, combinations of inputs should be considered and validated either using a separate database or using a re-sampling technique or both. Both were used in this work and the re-sampling technique will be outlined in section 5.6.

Therefore, goodness-of-fit metrics may be key statistics for analysing the quality of models, but have some limitations. For example, the more predictive features used for a regression model the higher the R^2 values by virtue of the way R^2 values are calculation. This is the case even when variables are known to have low (or no) variance and therefore little impact on overall regression outcomes. To overcome this situation for regression, the *adjusted R^2* can be employed to generate a comparable metric that is adjusted for the number of variables considered. The adjust R^2 is defined as:

$$R^2_{adj} = 1 - \frac{(1-R^2)(m-1)}{m-n-1}, \tag{5.9}$$

where m and n are the number of observations and variables respectively. Similarly, models built using higher degree polynomials will almost always yield R^2 values greater than that of a lower degree polynomials given the increased degrees of freedom available during modelling but there is no reason to mitigate this as it is an expected outcome of increasing the complexity of a model.

5.5 Predictive features and preprocessing

Dose distribution is geometry dependent and an aim of this work is to understand the relationships between geometric anatomy and the weighting factors that determine dose distribution in RBP AP. Therefore, only geometric anatomical features have been considered and all variables chosen were continuous and appropriate for use with the two ML approaches chosen (i.e., regression and clustering). Variables considered for predictive features included:

- volumetric features - variables associated with the volumes of ROIs e.g. those considered in the anatomical simulation study in the previous chapter
- spatial features such as the distance between ROIs
- other derived features such as the ratio between two volumetric features and the slope between two versions of the same type of volumetric feature.

Across the three treatment sites considered, the predictive features chosen were similar. The simple delineated volume of each ROI was taken including the external and any PTVs. In addition to whole volumes, OAR overlap volumes with PTVs were also considered. Studies have shown the dose distribution to be highly dependent on proximity of key OAR to PTVs[68] and OAR dose distribution is inversely correlated to its distance from the surface of PTVs[195]. For this reason, spatial features were also considered including maximum and average distances of OAR to PTVs. Moreover, in line with previous work in this area, distance-to-target style feature were also considered[109,148]. These features have been shown to have strong predictive qualities for dose distribution and factors related to planning. A summary of the types of variables considered is found in Table 5.1.

Type of Feature	Feature	Variant	Example
Volumetric	Volume	individual OARs and total OARs	volume of rectum (cm^3)
	Overlap of OAR with PTV	None	$OV_{bladder,PTV48}$: volume of bladder in PTV48 (cm^3)
	Volume-in-field of PTV: OAR volume within the superior-inferior slices of a PTV	None	bladder VIF_{PTV60}: volume of the bladder within the superior-inferior slices of PTV60 (cm^3)
	Volume-out-of-field of PTV: OAR volume above the superior slices and below the inferior slice of a PTV	None	rectum VOF_{PTV48}: volume of the rectum above superior slice and below the inferior slices of PTV48 (cm^3)
	Volume defined by nested PTVs (i.e., PTV annulus)	None	volume of PTV48 minus PTV60 (cm^3)
Spatial	Distance between ROIs	minimum, maximum and average surface-to-surface distance and distance between centres-of-mass	minimum distance between rectum and bladder (cm)
Derived	Overlap volume with expanded PTV	0.2cm increments of isotropic expansion up to 2.4cm	$OV_{rectum,PTV60_{1.4cm}}$: volume of rectum in $PTV60_{1.4cm}$ (cm^3)
	Rate of change (slope) between overlap volumes of adjacent expanded PTVs with OARs	None	slope between $OV_{rectum,PTV60_{1.4cm}}$ and $OV_{rectum,PTV60_{1.6cm}}$ (cm^3)
	Ratio of two ROI volumes	None	ratio of volume of rectum to volume of PTV48

Table 5.1: Summary of variables considered for FeatureDS1 and FeatureDS2. Features fall into three categories: volume related (volumetric), distance related (spatial) and derivations of volumetric and/or spatial (derived). Variants are denoted where multiple features of their kind are generated.

In terms of volumetric features, simple volumes were recorded for all ROIs including targets, OARs and the external, but also excluding certain delineated ROIs used in planning. Simple overlaps volumes of OARs with targets were recorded as well as derived overlap volume with isotropic expansions of target volumes which simulate distance-to-target histogram data. In terms of spatial features, three kinds were considered: average maximum distance, average distance and distance from ROI centre-to-centre. Average and maximum distances were calculated for internal ROIs (i.e. not for the external) and were calculated using Raystation's RoiSurfaceToSurfaceDistanceBasedOnDT() function within its Statetree. According to the function description it *"measures the distance between the surfaces of two ROI geometries using a distance transform based approach. Each point (voxel) on the surface of the target ROI will be assigned the minimum distance to a point (voxel) on the surface of the reference ROI."* ROI centre-to-centre differences are defined by the difference from the centre-of-mass of one ROI in DICOM coordinates to another.

In each case, data cleaning was performed. This was to ensure robustness during modelling[196], better modelling performance (reduction of type I and II errors)[197] and computational efficiency[198]. Variables will be chosen such that they are potentially applicable for prediction in each case. Therefore, incomplete data will not be used in any modeling. Low and no variance variables will also be removed given the lack of information they provide and the potential inefficiency of leaving them in for modelling.

The number of variables extracted varied for each treatment site given the numbers of ROIs vary. For PSV, 130 predictive features were identified. This site contained rectum, bladder, external and bowel as non-target ROIs. Given bowel contours consisted of delineations up to 2cm superior of PTVs, it was not considered a complete ROI and although is applicable for planning, was not considered reliable to be used to create predictive features in these cases. Therefore, no variables derived of the bowel contour were considered. A total of 142 variables were extracted in which bladder and rectum were treated as the main OARs. Simple volumes were recorded for all ROIs including targets and the external but excluding bowel as mentioned. A bladder-rectum combined volume was also recorded. Simple overlaps volumes of the rectum and bladder with PTVs were recorded as well as derived overlap volumes with isotropic PTV expansions. PTVs were isotropically expanded in 0.2cm increments (up to 2.4cm) and the overlap volumes determined. From these new volumes, the slope (rate of change).

For rectum, 340 variable were extracted of which 281 remained after cleaning. Of the 59 variables removed, 56 were zero variance and the remaining three were removed due to low variances. All zero variance variables were uniquely zero in all instances and low variance variables were zero in 85% of the instances or more. Of the 59, 52 were stoma related and due to zero variance, four were related to bowel bag/aux ant and due to zero variance and the three low variance were due to the genitals. Stoma related variable included $\mathrm{OV}_{\mathrm{PT45,stoma}}$ and related PTV expansions given stomas values were generally very far from the PTV or not a present contour for the patient.

For lung, 242 predictive features were identified from a total 313 extracted variables. Of the 313, 4 variable were removed for having no variance and 47 variable were removed for having low variance. Similarly to rectum, all zero variance variables were uniquely zero in all instances and low variance variables were zero in 85% of the instances or more. This site contained one target volume (PTV55), an external and five OARs including the heart, oesophagus, ipsilateral lung, contralateral lung and cord. Of these, not all possible variables relating to OARs were extracted due to missing data where the OAR had not been delineated for certain patients such as the liver or brachial plexus.

5.5.1 Standardisation

Before performing multivariate predictive analyses, the predictive features must be scaled. The purpose of this is to produce models bases on the variance within the features and not the variance between them. Min-max is one applicable data scaling method that involves identifying the smallest and largest value of each feature in the training database and setting them to some predefined values (e.g., zero and 1 or -1 and +1) and assigning all intermediate values a new value proportional to the the newly assigned min and max values. However, in this work, the transform of choice was a standard normal distribution scaling transform.

Normal distributions are defined as a continuous probability density functions and are used to illustrate the expected pattern followed by the majority of continuous data as they grow large. In this study, due to the nature of the predictive features chosen and size of the training databases, the underlying distribution of each is expected to be approximately normal and an unbiased sample. Under this assumption, predictive feature i can be expressed by a standard normal distribution given by

$$x_i \sim N(\mu_i, \sigma_i), \tag{5.10}$$

where μ_i and σ_i are the population parameters (mean and standard deviation respectively) of the underlying population of the normally distributed predictive feature. In this case, it is simple to recast them in terms of the standard normal distribution (or Z distribution) which is given by

$$Z \sim N(0,1), \quad Z = \frac{x_i - \mu_i}{\sigma_i}. \tag{5.11}$$

As the datasets used here contain sample data, instead of using population parameters μ and σ, their unbiased minimum variance estimators $\bar{x}_i$ and s_i are used respectively. Therefore

$$Z \approx \frac{x_i - \bar{x}_i}{s_i} \quad \text{where} \quad \bar{x}_i = \frac{\sum\limits_{j=1}^{m} x_{ij}}{m} \quad \text{and} \quad s_i = \sqrt{\frac{\sum\limits_{j=1}^{m} (x_{ij} - \bar{x}_i)^2}{m-1}}. \tag{5.12}$$

Following a Shapiro-Wilks test of normality, no raw features for PSV or Rectum were found to differ significantly from a normal distribution. This validates the appropriateness of a standard normal scaler for feature standardisation for these sites. Figure

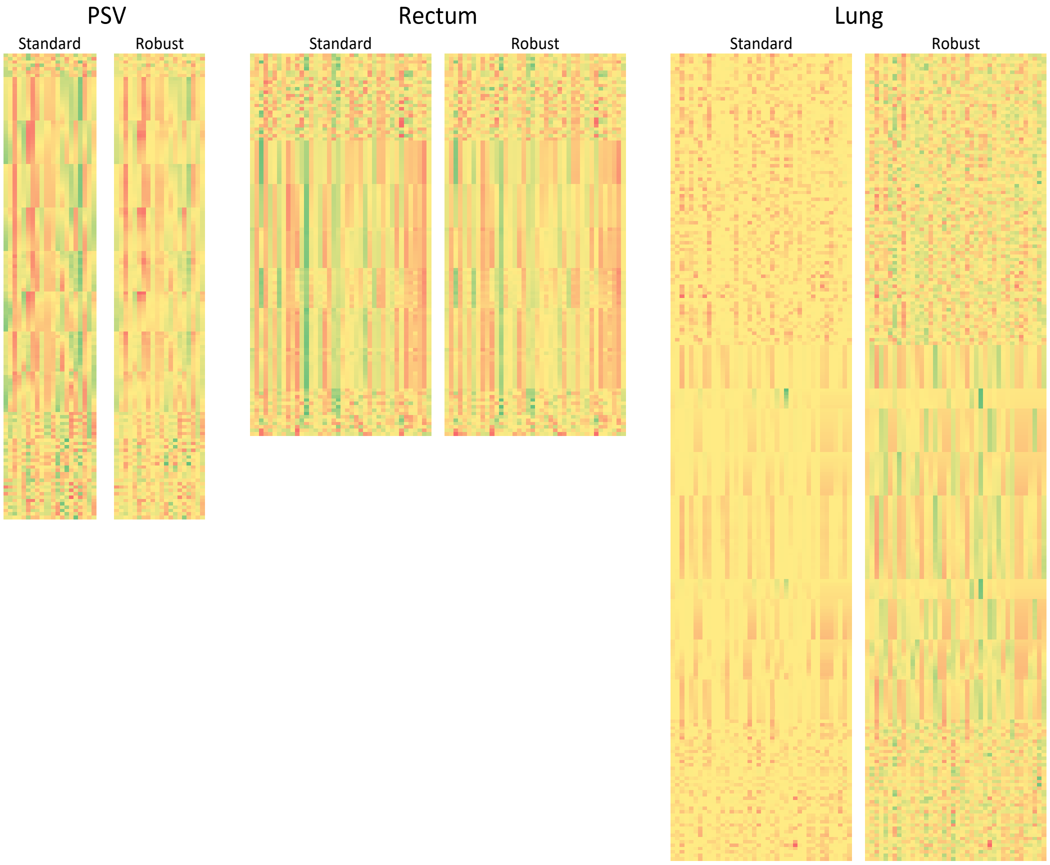


Figure 5.2: Distribution of features under a standard and robust scaling. All values are represented on a green-yellow-red scale with the highest values in green and lowest values in red. For legibility, feature databases have been transposed and represented a (features × patients) matrix. Similarities between scaling methods indicate that either method is appropriate. Also see appendices B.1-B.6

5.2 illustrates the distribution of scaled values under different scalers: a standard normal scaler that uses mean and standard deviation and a robust scaler that uses the median and inter-quartile range. Given all variables are approximately normally distributed, distributions are comparable regardless of the method. However, of the 241 cleaned features chosen for Lung, 26 were found to differ significantly from a normal distribution. This means they are highly skewed and a standard normal scaler may not be appropriate. In this case, both a standard and robust scaler are explored for use during modelling. Also see Appendix B for box plot representations of value distributions under each scaler.

5.5.2 Data cleaning and feature reduction

When independent variables are generated, to be considered for ML it is vital to assess their quality and usability. Variables with low or no variance provide little information for discrimination between data points and variables that are highly correlated to other variables also have minimum utility. To generate models using such variables can create some undesirable issues. For example, modelling over a large number of variables can be computational inefficient due to there being more data to manage and removing ineffectual variables is an easy way to improve the efficiency and may reduce modelling time too[198]. Research has also shown that failure to appropriately clean data can lead to poor modelling performance. This includes generating models with poorer performance than otherwise might be achieved or models that are misleading (i.e., causing type I and II errors)[197]. However, data cleaning and variable selection is a large area of data science especially for larger datasets and more than one approach can be taken.

Feature reduction (or selection) is a complex field and techniques are classified in a few ways. *Learner dependent* methods reduce features based on performance of a ML approach with respect to the dependent variable. *Learner independent* methods reduce features based on considerations among the predictive features only and with considerations with respect to the dependent variable. Aside from learner dependence, there is also the selection approach. One the most common learner dependent feature selection approaches is a wrapper method characterised by exploration of feature subsets after training a ML model. Common learner independent methods are filter approaches characterised by exploring correlations[199].

Wrapper approaches consist of searching for combinations of features with a view of identifying the optimal feature set for mapping to the dependent variable using the modelling technique of choice. Searching methods can be random, systematic or exhaustive. Random methods often involve selecting random sets of features and settling on that which yields the most desirable solution. An exhaustive search refers to a complete search of all possible combinations and the categorical choice of the best solution. Systematic searches involve using a strategic method such as forward selection where the best single feature is identified using an exhaustive search and new features are added based on that which yields the greatest improvement. Filter methods involve exploring relationships among the features and removing variables with minimal utility. This usually involves managing variables with missing values, those with low variance and those that are highly correlated with other variables. Filter methods can be that include filter methods such as removing one in a pair or correlated features.

By definition, wrapper methods are learner dependent methods and filter approach are learner independent. Both have benefits and given there is no restriction on choosing just one of them, the benefits of combining them can be exploited. Learner dependent approaches are considered the best given given they are exploratory and ensure the best possible selection but they can be computationally expensive especially in comparison to filter techniques especially for big data or when using a fully exhaustive search method. They also use the dependent variable potentially leading to overfitting. Filter methods are much more computational efficient and ensure the data do not contain redundant variables. However, they do not guarantee the resulting data are useful for prediction of the dependent variable and given the filter method chosen, may lead to usual variable being dropped from the database[200].

In this work, a combination of filter and wrapper techniques were used for data cleaning and feature selection. Data cleaning involved eliminating incomplete features (if they existed), removing zero-variance features (e.g., all zeros) and removing those with low variance. Low variance features were considered based on percentage of values that were constant. For example, the percentage of the variable that had the value 1. The threshold was chosen to be 85% of the variable i.e. if 85% of the variable had the same value, the variable was dropped. This threshold has been applied for other ML models and is considered an appropriate cutoff as these variables can be considered constant. An important area of data cleaning involves managing outliers and this can be important given outliers can skew the data and lead to misleading models or cannot be generalised or used for prediction. However, in this work, outliers have not been removed. The aim of this work is to establish the relationship between weighting factors and anatomy. Given all the data were collected and checked systematically and with no generation errors expected, outlying values are true and should be included. Outliers are also expected to aid in the illustration of the strength of relationship between weighting factors and anatomy where they exist.

5.6 Cross validation

A very popular re-sampling technique is cross validation and involves leaving out a portion of the training database, modelling on the rest, validating using the left-out data and then re-sampling again. This method produces a series of models built and validated on different samples of the database. Findings can then be aggregated and used to intuitively forecast the expected performance of a model built on the entire database and validated on separate database of unseen data. The most well known cross validation method is leave-one-out cross validation (LOOCV). As the name implies, the process involves leaving a single observation out of the training database, modelling on the remainder, producing a predicted value for the left-out observation, adding the observation back and running the process again for a new left-out patient.

In this work, a cross validation technique was employed for efficient data validation and model selection. In all cases a LOOCV was employed and used to determine the optimal model to take forward for final validation. Regression was thought to work well for this given the preliminary work and similar work found in the literature.

Assuming a simple linear regression model formation, a LOOCV is performed by leaving a single patient out of the training database and generating a linear regression model over the remaining patients using one of standardised features. The left-out patient can then be assigned a predicted value. When all patients have been left-out, the predicted values can be compared against the true values. This can be done in a number of ways and these will be discussed in the following section.

See Figure 5.3 for three examples of single feature regression models given patient 1 has been left out. The weighing factor in each case is for the PSV PG *rectum* D_{mean} and the predict feature in each case is the volumetric feature Volume of the External (cm^3). In these examples, a Pearson correlation coefficient was used and weak positive correlations are observed between the weighing factor and the feature. Nevertheless, each model is favourable for predicting the left-out patient's weighing factor. The error between the model and the true value under the linear, quadratic and cubic model are 0.631, 2.37 and 1.45 respectively. Therefore, although the correlations were weak, each of these models can be considered highly useful for estimating the true value of the patient that happened to be left-out in this case. However, this will not always necessarily be the case and this will now be discussed in the following section.

5.6.1 Assessing model performance after LOOCV

Once all of the cases have been considered within the LOOCV, performance of each is assessed. There are a few ways of doing this including:

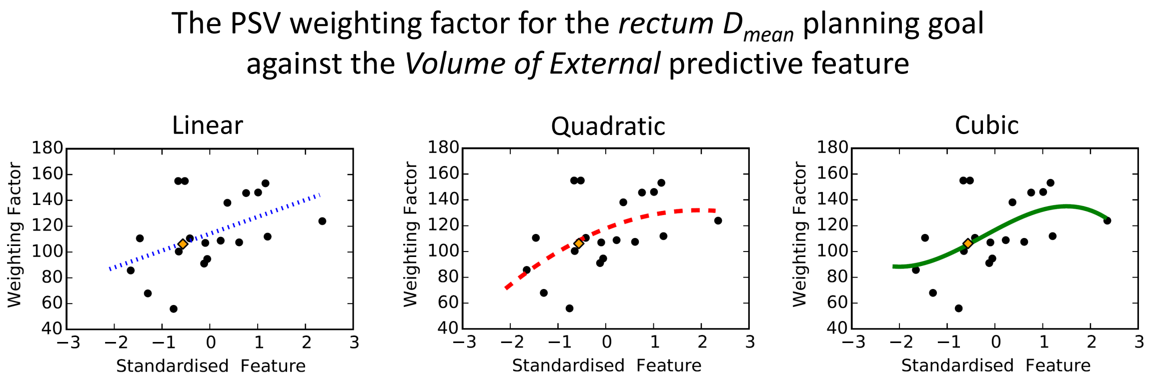


Figure 5.3: Regressions produced for PSV planning goal *rectum D_{mean}* and generated using *Volume of the External (cm^3)*. Left to right shows a linear, quadratic and cubic model with patient 1 left out of model generation but overlaid on the axes as a yellow point. Models have R^2 values 0.2145, 0.2419 and 0.2548 for the linear, quadratic and cubic model respectively.

- analysis of correlation coefficients (e.g, R) or coefficients of determination (R^2)
- analysis of the distribution of differences including aggregated differences such as the mean and median
- analysis of the mean squared error OR MSE

Analyses using R^2 for example can have some intuitive benefits. Following LOOCV, if the average R^2 (or adjusted R^2) were calculated over left-out patients, this provides not only an idea of how well the feature fits to the data in general but could also be an indication of how well such a model is likely to perform for unseen data. In this example adjusted R^2 is considered only but R^2 values are also considered in the main study as well as the adjusted values. A mean adjusted R^2 with a strong value (e.g., greater than 0.65) and low standard deviation indicates a stable model with strong predictive power and a model that is likely to perform well for unseen data.

Analysis of the distribution of the absolute differences found between modelled values and the true values is also highly values. Model with high levels of dispersion about the true values is an indication that the model performs undesirably overall and can help to filter poor models out of future consideration. Aggregated values such as the mean and median can further help to highlight how models perform overall at the databases population level.

MSE, however, is the standard metric of choice in such cases and is defined as the sum of the mean of the squared differences between the predicted value and the true value. It is also closely linked to the predicted residual error sum of square or PRESS statistic[201],

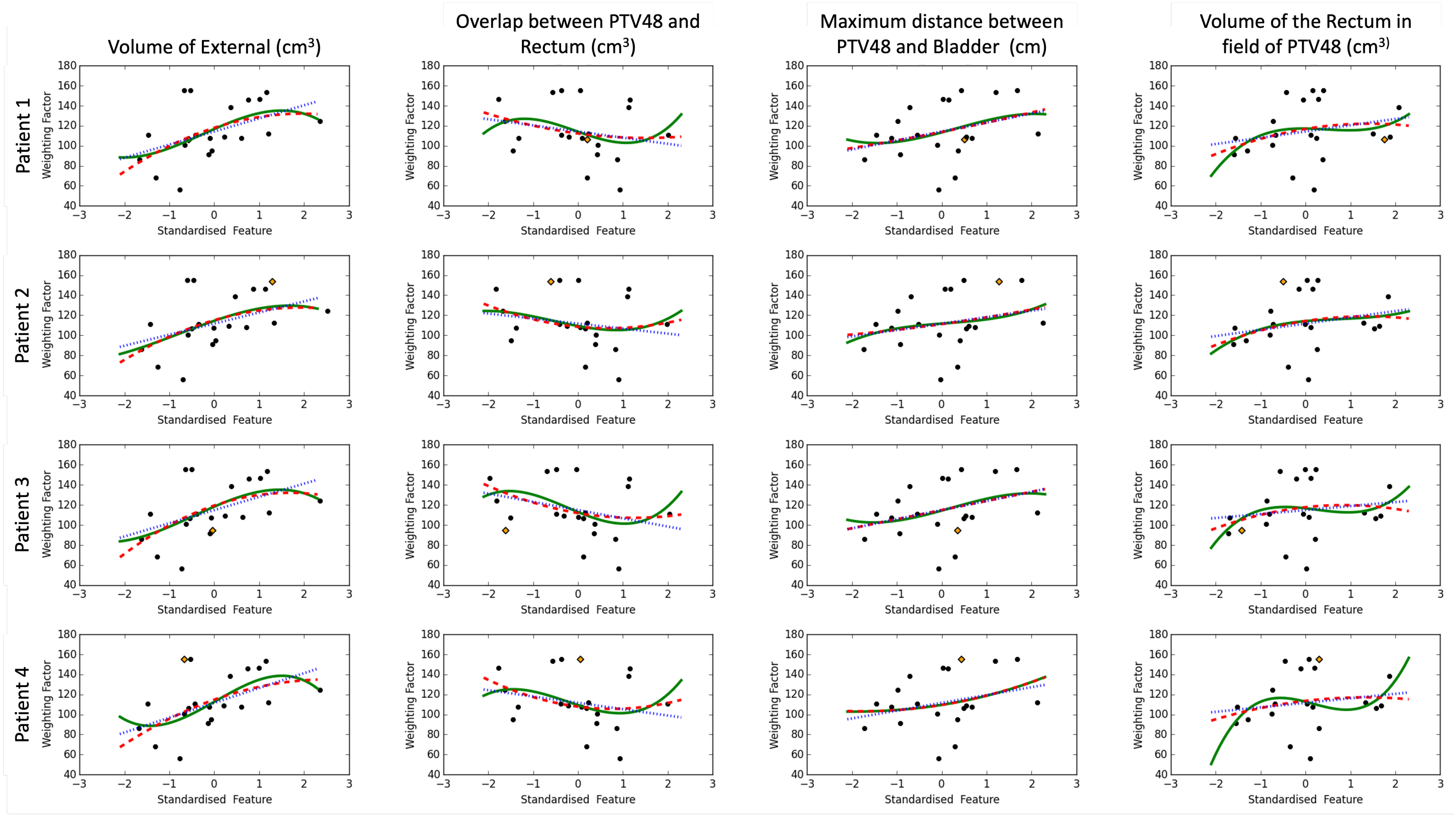


Figure 5.4: Scatter plots with overlaying regression models for four predictive features against weighing factor for patients 1-4.

which can be used to determine the optimal model from a selection of candidate models during cross validation. PRESS and MSE are defined as:

$$\text{PRESS} = \sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2, \tag{5.13}$$

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_{ij} - \hat{y}_{ij})^2 \tag{5.14}$$

and are particularly useful for a few reasons. First of all, like the other methods mentioned, MSE is a single metric that can determine the performance of a model formation in comparison to other formations but has the quality of penalising larger differences more than smaller differences hence making the differentiation between good and bad fitting models much greater than the other metrics. It is noteworthy that MSE is asymptotically equivalent to the Akaike Information Criterion, another well known and appropriate metric for assessing performance[202]

5.7 Regression LOOCV example

To illustrate the modelling and cross validation methodology, a reduced method is employed. Consider the models illustrated in the Figure 5.4. For the purpose of this illustration, these four patients (patient 1-4) are deemed representative of the entire training database of 20 patients therefore only these data will be used to demonstrate the LOOCV method. Also for illustration purposes only, only single feature models of the four features presented are considered for modelling in this example. However, in all latter portions of this thesis, all 20 patients will be considered during LOOCV as well a greater number of variables including those containing multiple features.

The aim is to establish the optimal feature set and in this example there are four feature sets each containing a single feature. Features include: volume of external or $\text{Volume}_{\text{External}}$ (cm^3), overlap volume between PTV48 and the rectum volume or $\text{OV}_{\text{PTV48,Rectum}}$ (cm^3), the maximum distance between PTV48 and the rectum or $\text{DistMax}_{\text{PTV48,Rectum}}$ (cm), the volume of the rectum in field of (i.e. within the most inferior and superior slice of PTV48) or $\text{VIF}_{\text{PTV48,Rectum}}$ (cm) and distance between the geometric center of PTV48 and the rectum or $\text{DistCentre}_{\text{PTV48,Rectum}}$ (cm). The MSE (or PRESS) is calculated between the true and modelled values (i.e., for the modelling error) and used to determine optimality. In addition, other metrics are used to assist in the analysis and include the average adjusted R^2 as well as metrics related to the difference such as mean and median

Model	Metric	Adj R^2				Error				%Error			
		P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4	P_1	P_2	P_3	P_4
$\text{Volume}_{\text{External}}$	Linear	0.215	0.123	0.160	0.229	-45.4	-26.3	22.3	-50.0	-29.9%	-17.2%	23.5%	-32.2%
	Quadratic	0.242	0.160	0.217	0.260	-43.7	-26.0	27.5	48.8	-28.7%	-17.0%	29.0%	31.5%
	Cubic	0.255	0.160	0.217	0.275	-40.4	-19.4	20.9	51.5	-26.6%	-12.7%	22.1%	33.2%
$\text{OV}_{\text{PTV48,Rectum}}$	Linear	0.047	0.027	0.075	0.044	-39.3	-36.3	35.4	-41.3	-25.8%	-23.7%	37.4%	-26.6%
	Quadratic	0.054	0.031	0.080	0.051	-41.1	-37.6	37.5	-43.4	-27.0%	-24.5%	39.6%	-28.0%
	Cubic	0.080	0.054	0.110	0.070	-40.3	-30.9	31.4	-42.0	-26.5%	-20.2%	33.2%	-27.1%
$\text{DistMax}_{\text{PTV48,Rectum}}$	Linear	0.101	0.076	0.126	0.103	-33.5	-29.2	26.3	-37.0	-22.0%	-19.0%	27.8%	-23.9%
	Quadratic	0.101	0.077	0.129	0.104	-33.9	-29.3	27.7	-37.5	-22.3%	-19.1%	29.3%	-24.2%
	Cubic	0.104	0.078	0.143	0.107	-32.6	-26.0	23.2	-36.2	-21.4%	-17.0%	24.5%	-23.4%
$\text{VIF}_{\text{PTV48,Rectum}}$	Linear	0.049	0.125	0.071	0.088	-27.1	-44.1	11.6	-38.4	-17.8%	-28.8%	12.3%	-24.8%
	Quadratic	0.062	0.126	0.074	0.088	-30.6	-43.7	9.94	-38.2	-20.1%	-28.5%	10.5%	-24.7%
	Cubic	0.073	0.126	0.080	0.111	-23.3	-36.9	9.27	-38.6	-15.3%	-24.1%	9.80%	-24.9%
$\text{DistCentre}_{\text{PTV48,Rectum}}$	Linear	0.104	0.059	0.026	0.035	-53.6	-43.7	18.9	-40.4	-35.2%	-28.5%	20.0%	-26.1%
	Quadratic	0.107	0.081	0.042	0.077	-57.8	-44.6	18.8	-45.7	-38.0%	-29.1%	19.9%	-29.5%
	Cubic	0.125	0.128	0.054	0.106	-27.8	-43.0	13.3	-36.5	-18.3%	-28.1%	14.1%	-23.5%

Table 5.2: A summary of the raw metrics for example data

absolute error and mean and median parentage difference. To determine MSE and these other metrics, some values are taken which include the absolute error in the model for each left-out patient. A summary of these values are found in Table 5.2.

To gain the insights needed to determine the optimal model, these data are aggregated and compared and a summary of this is found in Table 5.3. Ultimately, the choice of optimal model is chosen based on that which minimises the MSE or PRESS which in this example is a cubic model built using $\text{VIF}_{\text{PTV48,Rectum}}$ with a LOOCV MSE of 870. This model is closely followed by a cubic model built on $\text{DistMax}_{\text{PTV48,Rectum}}$. It is favourable that optimal models show desirable characteristics for the other metrics also. This helps to inspire confidence in the chosen model and establish it as the true optimum and not a spurious outcome based on the data sample. However, that is not the case in this example which shows desirable characteristics for quadratic and cubic models built using $\text{Volume}_{\text{External}}$, a very different feature altogether. This may be evidence that $\text{VIF}_{\text{PTV48,Rectum}}$ is not the true optimal model but is more likely a consequence of not basing this exploration on sufficient data.

The additional metrics are not only helpful for identifying the optimal model but

Model	Metric	MSE	Average Adjusted R^2	Mean Error	Median Error	Mean Percentage Error	Median Percentage Error
$\text{Volume}_{\text{External}}$	Linear	1438	0.184	-24.9	-35.9	-14.0%	-23.6%
	Quadratic	1431	0.220	**1.65**	**0.750**	**3.70%**	6.00%
	Cubic	1274	**0.227**	3.15	**0.750**	4.00%	**4.70%**
$\text{OV}_{\text{PTV48,Rectum}}$	Linear	1455	0.048	-20.4	-37.8	-9.68%	-24.8%
	Quadratic	1598	0.054	-21.2	-39.4	-10.0%	-25.8%
	Cubic	1332	0.079	-20.5	-35.6	-10.2%	-23.4%
$\text{DistMax}_{\text{PTV48,Rectum}}$	Linear	1009	0.102	-18.4	-31.4	-9.28%	-20.5%
	Quadratic	1045	0.103	-18.3	-31.6	-9.08%	-20.7%
	Cubic	897	0.108	-17.9	-29.3	-9.33%	-19.2%
$\text{VIF}_{\text{PTV48,Rectum}}$	Linear	1072	0.083	-24.5	-32.8	-14.8%	-21.3%
	Quadratic	1101	0.088	-25.6	-34.4	-15.7%	-22.4%
	Cubic	870	0.098	-22.4	-30.1	-13.6%	-19.7%
$\text{DistCentre}_{\text{PTV48,Rectum}}$	Linear	1693	0.056	-29.7	-42.1	-17.5%	-27.3%
	Quadratic	1943	0.077	-32.3	-45.2	-19.2%	-29.3%
	Cubic	1003	0.103	-23.5	-32.2	-14.0%	-20.9%

Table 5.3: A summary of key aggregated metrics for the example data.

also help establish the expected performance of that model. That is, the most desirable models (optimal or not) will have an R^2 values close to 1 and error metrics close to 0 and 0%. Under these criteria, all models in this example can be said to be performing poorly. This may be due to the data being ill-defined for the presiding methodology (regression), but is most likely related to the limited data used in the example and poor underlying relationships between the parameter being predicted (rectum D_{mean}) and the features sets considered. When the LOOCV is applied to the entire training databases and more feature sets are considered, more meaningful relationships may be established that are applicable for making predictions.

5.8 Chapter summary and next steps

The purpose of the work presented in this chapter was to consider different ML techniques and explore the approach to be used in the main study. This was achieved successfully with a simulated example illustrating the process with example output for the cross validation stage.

The next chapter builds on this work in the following ways. Firstly, no assumptions will be made about the representation of a sample of the training database as in this example, instead the entire database will be included in all stages of the LOOCV. This will provide a larger and more representative foundation that better indicates the performance each model will have in truth hence will aid the identification of the optimal model in each case as well as the identification of features that consistently perform well regarding prediction.

Secondly, a greater number of feature sets will be used including combinations of up to five sets of features. This will not only mean modelling on a single feature at a time as in this example, but will consider all possible combinations of the features up to sets of five. This will enable the identification of the performance of not only individual features but the relationships between interacting features. Not only will using a larger set of features allow for deeper exploration and increased likelihood of identifying the best models but modelling over all possible combinations will help gauge understanding of which features are interacting in a way that supports prediction. Similarly to this example, statistics considered in addition to MSE will be calculated to assist critique of the optimal model with the most desirable model achieving an R^2 close to 1 and a error and percentage error terms close to zero and zero percent.

Chapter 6

Regression and cluster modelling

Under the paradigms outlined in previous chapters, each of the two ML methods (i.e., regression and clustering) will be applied to try to predict gold standard planning across each of the three clinical sites. Research questions include:

1. Is there a relationship between gold standard weighting factors and numerical anatomical features such as volume of ROIs and distance between ROIs?

2. Can this relationship be determined using a ML technique such as multiple polynomial regression or K-means clustering method and used to generate weighting factors for automated plan calibration?

3. Do modelled weighting factors lead to dosimetrically comparable gold standard planning via PBAIO?

Rules-based AP methods including the Erasmus iCycle's lexicographic order method, Pinnacle's Auto-Planning software and Velindre Cancer centre's EdgeVcc have all been used to deliver clinically applicable planning and RBP planning has been implemented for treatment sites including PSV, lung, rectum, breast and oesophagus. Nevertheless, RBP planning requires some apriori calibration of which a "one size fits all" or universal standard prioritisation is usually defined for all patients (Std).

However, studies have shown that such calibration methods are not always appropriate. Vanderstraeten et al. (2018) found benefits of a Std approach with regard to planning time and dosimetric outcomes were found to be appropriate for the majority of plans with no need for further fine tuning (>75%). However, there was a significant proportion of cases that failed the clinical objectives[203]. Zhang et al. (2021) found that not only can a Std approach show some clinically undesirable characteristics in certain cases, it suggested there is a spatial (distance-related) relationship between the calibration solution and the dosimetric outcomes[204]. Janssen et al. (2019) noted the benefits of an RBP approach but stressed the need for independent planning QA for plan assessment given there is evidence to suggest a fully optimal plan is not always guaranteed[205]. Given the anatomical variance study presented in Chapter 4, there is strong evidence of weighting factor dependence of geometry and applying this knowledge of weighting factor generation may be the key to fully optimal and clinically applicable planning with these AP methods.

This chapter presents the full methodology of each of the two ML approaches chosen in light of the examples presented in the Chapter 5 with the results of each including a dosimetric comparison to the gold standard.

6.1 The RATING framework

This work was completed with reference to the RATINGS framework defined by Hansen (2020)[206]. The aim of the framework is to "improve the scientific quality of treatment planning studies and papers". The framework contains a list of 76 considerations for researchers to make when completing a study. A self-rating is applied for each and aggregated to an overall score used for reference and comparison against other studies. This framework encourages quality implementation of the methodology[207–209] hence improving the reproducible of the results. All studies in this work were completed under comparable conditions hence all obtained the same self-rated score: 173/203. Points were lost, for example, given there are some restriction to data and software that can be made available to a wide audience for replication of results. Given this 85% quality metric, studies in this work are considered reasonable quality against these criteria. See Appendix A for the full criteria of scoring.

6.2 Patients

Patient selection is defined in Section 3.6. For PSV, the full patient dataset consisted of 40 randomly selected prostate seminal vesicles (PSV) patients. Of those 40, 20 were randomly assigned to the training cohort (Patient 01-20) and 20 to the validation cohort (Patient 21-40). The number of patients selected for training reflected numbers found in previous work related to RBP[96,105,120] and planning parameter prediction for PSV[148].

For rectum and lung, the full patient dataset in each case consisted of 60 randomly selected patients of which 40 were assigned to the training cohort (Patient 01-40) and 20 to validation (Patient 41-60). Given rectum and lung have had far fewer AP cases reviewed with this RBP method than PSV, more patients were considered to increase the statistical power of the results in these cases.

6.3 Feature Set Databases

Collinearity between variables can leading to modelling bias[210] therefore to improve modelling efficiency and improve model performance, associations between extracted features were explored and a subset of features defined by removing one in every pair of highly correlated features. To determine the strength of associations between pairs of features in the training database, a Pearson correlation coefficient was calculated. For coefficients greater than 0.85, one of the two features was randomly removed. A value of 0.85 was considered a reasonable cut-off and is in-line with other ML studies in the general ML literature[211–213]. A second feature dataset is therefore defined for each treatment site: the full set of cleaned features (FeatureDS1) and a subset of FeatureDS1 containing uncorrelated features (FeatureDS2). The full list of features in FeatureDS1 datasets can be found in Appendix C.

Note that Pearson's correlation coefficient determines the strength of a linear relationship between features where one exists. In reality, there are a greater number of possible relationships that could exist between pairs of features but exploring all of them is impracticable. Removing pairs of linearly correlated features will nonetheless always have a positive outcome on modelling both in terms of bias and computational efficiency.

6.4 Modelling approach

Training and validation was performed using a "gold standard" dataset, where patient-specific weighting factors were obtained following navigation by an expert practitioner familiar with the system and qualified to create and validate plans for each site considered. The practitioner responsible for navigating the gold standard plans in all cases was a medical physicist with 15 years experience who was highly familiar with EdgeVcc and it's functionality and experienced in plan creation and validation for each site.

The weighting factors obtained during gold standard calibration were used to gen-

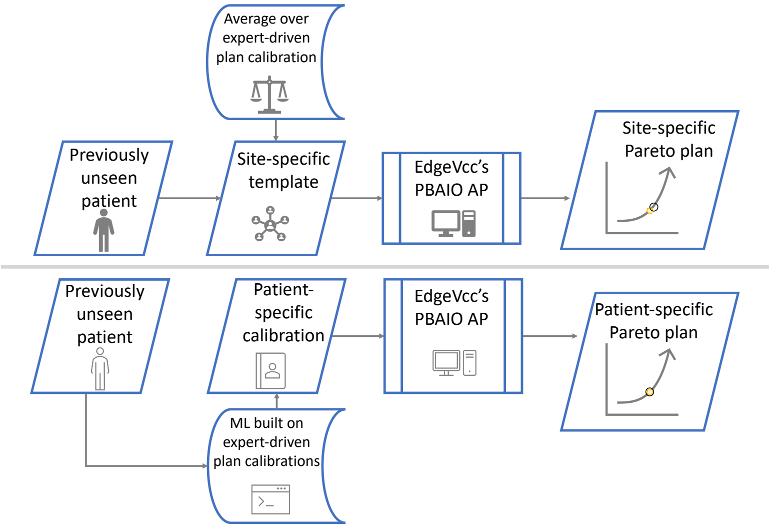

Average over expert-driven plan calibration
Previously unseen patient
Site-specific template
EdgeVcc's PBAIO AP
Site-specific Pareto plan
Previously unseen patient
Patient-specific calibration
EdgeVcc's PBAIO AP
Patient-specific Pareto plan
ML built on expert-driven plan calibrations

Figure 6.1: An outline of how the process defined in this work (bottom) differs from the more classic site-specific methods (top)

erate plans via PBAIO and are considered the modelling ground truth for comparison (MCO_{gs}). Predictive ML models were trained on this MCO_{gs} calibrated dataset with the aim of identifying the relationships between anatomical features and patient-specific weighting factors obtains via PGAP calibration. Once trained, predicted weighting factors can be generated for novel patients and used to form the inputs for the PBAIO system with the aim of generating plans of equivalent quality to MCO_{gs}. This method contrasts with classic approach to calibration (Std) where all patients are planned with the same site-specific calibration of RBP parameters. In this work, Std was defined by taking the mean gold standard weighting factor values for each patient in the training dataset. Std calibration, regression ML calibration (ML_{reg}) and cluster calibration (ML_{clus}) weighting factors were validated against MCO_{gs} using an independent set of patients.

All plans in these studies were generated within RayStation (Raysearch Laboratories, Stockholm, version 8B) using a single 360° VMAT arc. Patients were planned according to 20 fractions with a simultaneous integrated boost technique for PTVs. PGs for three all sites were derived from local clinical goals with PSV PGs based on the UK PIVOTAL trial[144]. Rectum cases contained a combination of PGs given genitals and stoma were not delineated for all patients.

For lung many originally chosen PGs were later set to static values resulting from limited clinical impact. These choices were made retrospectively given the expert operator determined their own choices to be redundant for all patients regardless of the weight chosen. PG4 (liver D_{mean}) was determined to have negligible impact. PG5 (oesophagus/brachial plexus D_{mean}) also had negligible impact except for one patient. PG7 (PTV conformality precision) was not considered clinically relevant in these cases and PG8 (high dose to lung) was not navigated given little impact shown to the dose distribution. Lung PG1-3 and 6 were navigated as normal and correspond to Lung D_{mean}, Heart D_{max}, PTV conformality and Lung D_{max}. PG 4,5 and 8 were set to the static values zero, 2445 and zero respectively. Optimal models were defined using the training data via a LOOCV with final models defined using all training patients and validated on unseen cases.

6.4.1 Regression

Two approaches were explored for regression modelling: (1) modelling using combinations of raw features within FeatureDS2 (reg-raw), (2) forward selection using Principal Components generated using FeatureDS1 (reg-PCA). In all cases the same method was followed and regressions built using the SKlearn Version 0.15.2 Linear Model and Preprocessing algorithms. Linear and polynomial regression models were explored in-line with the literature[117,148,214,215] and preliminary research. Modelling and prediction were performed for each PG individually.

As raw features are not ordinal, all possible combinations of features (feature sets) were considered in the reg-raw approach. To limit the search space, up to a maximum of 5 features were allowed within a feature set. A separate ‘feature set selection’ step was performed prior to model selection to identify the optimum feature set per model formation. The methodology involved identifying the feature set with the smallest MSE under each model formation.

Given PCA features are ordinal, in the reg-PCA approach FeatureDS1 is transformed to Principal Components and models generated using forward selection i.e. the first Principal Component (PC1) was used for all one feature models, PC1 and PC2 for all two feature models and so on up to the maximum features. For both approaches (reg-raw and reg-PCA), models explored were linear, quadratic and cubic. Therefore for each PSV PG, 15 model formations were defined using the reg-raw approach and 60 for reg-PCA. A single choice was made from among the 75 model formation given that which minimised MSE.

6.4.2 Clustering

K-means clustering was facilitated by the SKlearn Version 0.15.2 Cluster package. The two approaches considered were: (1) clustering over FeatureDS2 (clus-raw), (2) clustering over Principal Components of FeatureDS1 (clus-PCA). Training patients were clustered over all data available using a random initial state of 42 and 300 maximum iterations with all possible values of K considered.

The initial state defines a random state for the initial position of centroids and 42 was chosen arbitrarily and kept constant to ensure repeatability should the code be run again. The maximum number of iterations defines the number of K-means passes before should the model not stabilise prior to reaching this point. This threshold ensures the model does not run for an infinite number of passes should there be no absolute optimum and should be chosen such that it is large enough that the model is approximately stable once the threshold is reached. The value 300 was considered significantly high given the number of data points and computationally practicable.

Once clusters had been defined using the training cases, the mean average PG weight over the training cases in each cluster was calculated. These PG average weights are the machine learned PG weights. Validation patients were then assigned to clusters based on centroids that minimised Euclidean distance and a machine learned PG weight assigned. Hence, models are defined once using the training patients and machine learned weights defined. The model is then applied to all novel cases that had not been used to train it. In this way, the Std approach can be compared to a $\mathrm{ML}_{\mathrm{clus}}$ approach in which all training patients (and henceforth novel cases) are assigned to a single cluster for all PGs.

To aid the analysis of cluster performance, two metrics were calculated for each model formation: (1) the sum of the squared differences between each point and its cluster centroid (SSE), (2) a silhouette coefficient - a value between -1 and 1 that scores the goodness-of-fit of each formation based on average inter- and intra-cluster distances. SSE values close to zero and silhouette scores close to 1 indicate models that are well defined.

6.4.3 Validation and statistical analysis

All patients in the validation dataset were planned according to the four approaches: $\mathrm{MCO}_{\mathrm{gs}}$, Std, $\mathrm{ML}_{\mathrm{reg}}$ and $\mathrm{ML}_{\mathrm{clus}}$. Given weighting factors are essentially relative values but can theoretically range between zero and infinity, for the purposes of analysis all weighting factors will be converted to relative values and expressed as a percentage by dividing by the summed weight of all PGs.

For the validation cohort, the difference between the modelled relative PG weights and gold standard ($\mathrm{MCO_{gs}}$) relative PG weights are the primary metric used to assess model quality, with MSE additionally calculated to aid in the comparison with training results. Plans are compared against $\mathrm{MCO_{gs}}$ in terms of relative weighting factors as well as dosimetric features with statistical testing carried out.

For PSV, dose metrics of interest have been adapted from the UK PIVOTAL trials[144]. Rectum and lung have been defined internally based on current clinical practice at Velindre Cancer Center. PTV homogeneity index (HI) and Paddick's conformality index (CI) were also calculated for the analysis[216] and all outliers were defined as values outside of the range $[Q1 - (1.5 \times IQR), Q3 + (1.5 \times IQR)]$, where Q1, Q3 and IQR are quartile 1, quartile 3 and inter-quartile range (Q3-Q1) respectively.

Given a gold standard was explicitly defined in this work, all other plans are compared directly to this baseline. Therefore, no composite plan quality metrics have been considered to compare different plans produced for the same patient. Prediction of $\mathrm{MCO_{gs}}$ is the aim hence comparison against $\mathrm{MCO_{gs}}$ is done at all stages. Nevertheless, dosimetric improvements over $\mathrm{MCO_{gs}}$ for specific cases will be highlighted and discussed where they exist.

6.4.4 Definition of PCA features

Given there are a greater number of features in FeatureDS1 than patients in all cases, fewer Principal Components were generated than features in FeatureDS1. For any $m \times n$ matrix A

$$A = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{21} & \dots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \dots & a_{mn} \end{bmatrix},$$

when A is reduced to Principal Components, the number of Principal Components returned depends on $m > n$ or otherwise. Given $m > n$, an $m \times n$ matrix is returned i.e. the same shape as the original data. When $m \leq n$, an $n \times n$ matrix is returned where the nth Principal Component is a zero vector. This is due to the extra dimensions beyond $n - 1$ becoming redundant given the number of data points.

For example, given three data points $x_1, x_2, x_3 \in \mathbb{R}^3$, the relationship between the three points can be fully determined using only two-dimensions (i.e., axes or variables). Similarly, given four point $x_1, x_2, x_3, x_4 \in \mathbb{R}^4$, when transforming the axes such that the number of necessary axes is minimised (as in PCA), given all four points can be defined in $\mathbb{R}^3$ space, the extra dimension becomes redundant.

6.5 PSV Modelling Results

6.5.1 Predictive features

As mentioned in section 5.5, a total of 139 predictive features were generated for FeatureDS1. Of these, 27 were retained for FeatureDS2. A summary of features retained in FeatureDS2 can be found in Table 6.1 along with the number of excluded features that were correlated with it in FeatureDS1. Excluded features may have been correlated with more than one retained feature.

For reg-PCA PSV LOOCV, a 20×20 matrix is defined given FeatureDS1 represents a 130×20 matrix. The first PC accounts for 46.5% of the variance with the first eight PCs accounting over 95% of the variance in the entire database.

6.5.2 Leave-one-out summary

To define the optimal model, LOOCV of the training cohort is used. For reg-raw PSV LOOCV, a total of 36,656,880 individual regression models were generated:

$$\underset{\text{(PGs)}}{6} \quad \times \quad \underset{\text{(patients)}}{20} \quad \times \quad \underset{\text{(degrees)}}{3} \quad \times \quad \underset{\text{(combinations)}}{\sum_{k=1}^{5} \binom{27}{k}} \quad = \quad \underset{\text{(total)}}{36{,}656{,}880.}$$

Therefore, given the number of PG and degrees, 18 full LOOCV models were defined over the training patients for each combination type (1-5 feature combinations). One model was chosen for each of the six PGs given MSE was minimised as in the method outlined in section 5.7. Building on work from the previous chapter where only 1-feature models were built, more feature sets and patients have been considered here. The results of chosen models for each PG following LOOCV can be found in Table 6.2. The mean R^2 value was greater than 0.8 for all PGs indicating a strong positive relationship on average. The validation MSE indicates a poorer fit to the validation database than the training database. No models chosen were a higher degree than quadratic which suggests the relationships between weights and anatomical features are simple.

Type	Feature	Excluded Features
Derived	Slope between $\mathrm{OV}_{\mathrm{bladder,PTV48}_{0.2\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{bladder,PTV48}_{0.4\mathrm{cm}}}$	36
Derived	Slope between $\mathrm{OV}_{\mathrm{bladder,PTV48}_{1.2\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{bladder,PTV48}_{1.4\mathrm{cm}}}$	25
Volumetric	$\mathrm{OV}_{\mathrm{bladder,PTV60}}$ (cm^3)	24
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV48}_{1.4\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV48}_{1.6\mathrm{cm}}}$	14
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV60}_{0.8\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV60}_{1.0\mathrm{cm}}}$	14
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV48}_{0.6\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV48}_{0.8\mathrm{cm}}}$	13
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV48}_{0.2\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV48}_{0.4\mathrm{cm}}}$	11
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV60}_{0.2\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV60}_{0.4\mathrm{cm}}}$	11
Volumetric	Bladder (cm^3)	7
Volumetric	$\mathrm{OV}_{\mathrm{rectum,PTV48}}$ (cm^3)	5
Volumetric	$\mathrm{OV}_{\mathrm{rectum,PTV60}}$ (cm^3)	5
Spatial	Maximum distance between bladder and PTV48 (cm)	3
Spatial	Maximum distance between rectum and PTV48 (cm)	3
Derived	Slope between $\mathrm{OV}_{\mathrm{rectum,PTV48}_{2.2\mathrm{cm}}}$ and $\mathrm{OV}_{\mathrm{rectum,PTV48}_{2.4\mathrm{cm}}}$	3
Spatial	Distance between the center of bladder and rectum (cm)	4
Volumetric	Volume of the rectum (cm^3)	2
Spatial	Distance between the center of rectum and PTV48 (cm)	2
Volumetric	Volume of the PTV48 (cm^3)	1
Derived	Ratio of PTV48 to bladder	1
Derived	Ratio of PTV48 to rectum	1
Volumetric	Total OAR $\mathrm{VIF}_{\mathrm{PTV48}}$ (cm^3)	1
Volumetric	Total OAR $\mathrm{VIF}_{\mathrm{PTV60}}$ (cm^3)	1
Volumetric	Rectum $\mathrm{VIF}_{\mathrm{PTV60}}$ (cm^3)	1
Spatial	Distance between the center of PTV60 and PTV48 (cm)	0
Volumetric	Volume of the PTV48 minus PTV60 (cm^3)	0
Volumetric	Volume of the external (cm^3)	0
Derived	Ratio of bladder to rectum	0

Table 6.1: Summary of the features chosen for FeatureDS2 from FeatureDS1. The number of correlated features in FeatureDS1 that were excluded from FeatureDS2 are summarised in the columm Excluded Features.

For reg-PCA PSV LOOCV, 7,200 models were created:

$$\underset{\text{(PGs)}}{6} \times \underset{\text{(patients)}}{20} \times \underset{\text{(degrees)}}{3} \times \underset{\text{(PC combinations)}}{20} = \underset{\text{(total)}}{7{,}200.}$$

However, none of the optimal models used Principal Components. With only raw features chosen and high R^2 values, this indicates key relationships between the weighting factors and the specified variables.

To generate all 1-feature reg-raw model combinations and all reg-PCA model combinations, the LOOCV python scripts ran for approximately 0.05 minutes. Two-feature reg-raw models took approximately 0.35 minutes, 3-features took approximately 2.5 minutes, 4-features approximately 40 minutes and 5-features approximately 300 minutes. This illustrates the exponential increase in computation time given the exponential increase in the number of feature set combinations as the total number of predictive features increases.

For clus-raw and clus-PCA, a total of 400 models were generated for each:

$$\underset{\text{(patients)}}{20} \times \underset{\text{(no. clusters)}}{20} = \underset{\text{(total)}}{400.}$$

For the generation of clus-raw and clus-PCA solutions, python scripts took approximately 1 minute to produce all clusters. The maximum number of possible clusters is determined by the number of data points which is determine by the number of training patients in this case. A summary of the LOOCV for ML_{clus} can be found in Table 6.3. Of the six PGs, the optimal model for the PTV Dose Falloff PG was defined using the mean of all training patients and therefore the outcomes of the clustering methods (i.e., reg-raw or rerg-PCA) were equivalent. Of the remaining PGs, two were defined over clusters of raw features and three over PCA features. This indicates the advantage of reducing the dimensions of the predictive features prior to clustering for the PSV treatment site.

Comparing optimal ML_{reg} and ML_{clus} models following LOOCV, MSE values indicate regression performed better at modelling training PGs than clustering for four of the six PGs: rectum D_{mean}, PTV Conformality and rectum D_{max}. Nevertheless, MSE values for the validation database indicate ML_{clus} may be better at predicting MCO_{gs} than ML_{reg}. Validation MSE values indicate ML_{clus} is optimal for rectum D_{mean}, PTV Conformality, PTV dose falloff and bladder $\text{D}_{\text{max}}$and ML_{reg} is optimal for bladder D_{mean} and rectum D_{max}.

Although, the ML_{reg} model was optimal for the two PGs bladder D_{mean} and rectum D_{max}, the ML_{clus} model for them is comparable for each in terms of performance given validation MSE values are similar. All other PGs show stronger MSE values for under their optimal models.

6.5.3 Weight Summary

See Table 6.4 and Figure 6.2 for an overview of relative weight calibrations for the validation dataset and Figure 6.3 for the underlying patient level distributions for differences from MCO_{gs}. No weighing factor distributions met ANOVA assumptions and all were tested using a Friedman test. Significant differences relative weighing factor differences were observed between MCO_{gs} and at least one of Std, ML_{reg} and ML_{clus} for six of the

Planning	Regression	Features	Training		Validation
Goal	equation		Av adj R^2	MSE	MSE
Rectum	3 features	Volume of the external (cm^3)	0.835	368	7025
D_{mean}	quadratic	Rectum VIF_{PTV48} (cm^3)			
		Slope between $OV_{rectum,PTV48_{0.2cm}}$ and $OV_{rectum,PTV48_{0.4cm}}$			
Bladder	5 features	Volume of the rectum (cm^3)	0.858	24.5	271
D_{mean}	linear	$OV_{rectum,PTV48}$ (cm^3)			
		Total OAR VIF_{PTV60} (cm^3)			
		Distance from center of PTV48 to the center of rectum (cm)			
		Ratio between PTV48 and rectum volume			
PTV	5 features	Volume of the PTV48 (cm^3)	0.907	1441	19442
Conformality	linear	Distance from center of PTV48 to the center of rectum (cm)			
		Slope between $OV_{rectum,PTV48_{0.2cm}}$ and $OV_{rectum,PTV48_{0.4cm}}$			
		Ratio between bladder and rectum volume			
		Ratio between PTV48 and bladder volume			
Rectum	4 features	Volume of the rectum (cm^3)	0.997	0.125	5.82
D_{max}	quadratic	Distance from center of PTV48 to the center of rectum (cm)			
		Distance from center of PTV48 to the center of PTV60 (cm)			
		Ratio between bladder and rectum volume			
PTV Dose	4 features	Volume of the PTV48 (cm^3)	0.998	2.62	495
Falloff	quadratic	Rectum VIF_{PTV48} (cm^3)			
		Distance from center of bladder to the center of rectum (cm)			
		Distance from center of PTV48 to the center of rectum (cm)			
Bladder	4 features	$OV_{rectum,PTV60}$ (cm^3)	0.999	0.309	69.3
D_{max}	quadratic	Total OAR VIF_{PTV48} (cm^3)			
		Distance from center of bladder to the center of rectum (cm)			
		Slope between $OV_{bladder,PTV48_{1.2cm}}$ and $OV_{bladder,PTV48_{1.4cm}}$			

Table 6.2: Summary of ML_{reg} model formations determined automatically for PSV via leave-one-out.

Planning goal	**Number of clusters**	**Feature Type**	**Cluster SSE**	**Silhouette Score**	**Training WF MSE**	**Validation WF MSE**
Rectum D_{mean}	2	Raw	416	0.173	698	1273
Bladder D_{mean}	11	Raw	123	0.058	9.89	298
PTV Conformality	9	PCA	562	0.162	2320	4037
Rectum D_{max}	7	PCA	802	0.182	0.592	6.42
PTV Dose Falloff	1	n/a	540	n/a	10.8	133
Bladder D_{max}	12	PCA	416	0.128	0.791	37.9

Table 6.3: Summary of ML_{clus} model formations determined automatically for PSV via leave-one-out.

eight PG groups. These included rectum D_{mean} and D_{max} PTV conformality and dose falloff, bowel D_{medium} and PG_H.

For rectum D_{mean}, MCO_{gs} differed from Std only. Mean and median difference from MCO_{gs} for Std was greatest compared with other methods with a value of 0.079%. Importantly, Std weighting factors were higher than MCO_{gs} weighting factors for 17/20 patients leading to rank differences within statistical tests indicating Std weighting factors to be significantly greater than MCO_{gs} weighting factors. Significant differences were observed for rectum D_{max} between MCO_{gs} and all other methods. Relative weighting factors for Std, ML_{reg} and ML_{clus} were significantly lower than MCO_{gs} of which mean the largest median difference was observed for ML_{clus}.

PTV conformality difference against MCO_{gs} was observed for Std only. Similarly to the rectum D_{mean}case, Std showed the greatest mean and median difference from MCO_{gs} and weighing factor values were greater than MCO_{gs} values in 17/20 cases. PTV dose falloff showed differences from MCO_{gs} for Std and ML_{reg}. ML_{reg} weighing factor values are less than MCO_{gs} in 16/20 cases and is similarly true for Std. The largest deviations from MCO_{gs} were observed for Std with mean and median deviations of 0.231% 0.296% relative weight difference respectively.

Bowel D_{medium} and PG_H differences from MCO_{gs} were observed from Std only. The largest mean and median differences were observed for this method with values lower than MCO_{gs} in 15/20 cases for each these PGs. In addition, differences were observed for rectum D_{mean} between Std and $ML_{reg}$with Std greater than ML_{reg} in 14/20 cases. However, although statistically significant differences were observed, differences were considered to be clinically insignificant. The greatest percentile difference from MCO_{gs}

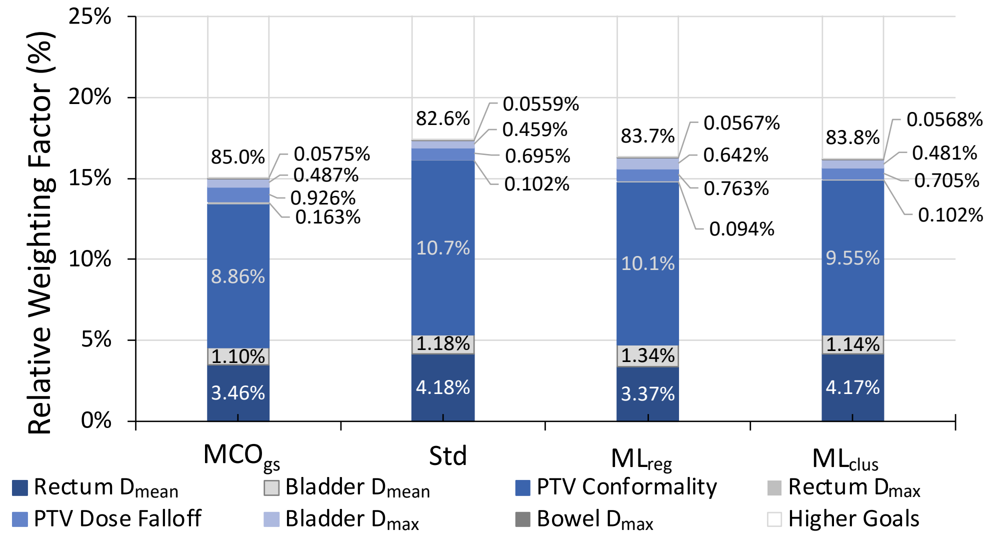


Figure 6.2: Plots showing relative weight distribution for all calibration methods for the PSV validation dataset.

were all within 10% and comparable on aggregate across patients for relative weights.

ML$_{\text{clus}}$ was considered the closest to MCO$_{\text{gs}}$.Mean differences from MCO$_{\text{gs}}$ of less than 1.17% were observed for all PGs and the median difference from MCO$_{\text{gs}}$ was closest to zero for three of the eight PGs. Although median weighing factor differences of ML$_{\text{reg}}$ from MCO$_{\text{gs}}$ was closest to zero for four PGs, ML$_{\text{clus}}$ was comparably close to zero in all four cases with notably larger distribution spreads observed within the data for ML$_{\text{reg}}$. For this reason, ML$_{\text{reg}}$ can be considered the poorest performer overall with deviations as great as 2.49% and 3.57% for PTV conformality and PG$_{\text{H}}$ respectively. Also, the most extreme ranges were observed for ML$_{\text{reg}}$ model for all PGs and included some of the most extreme outliers. This indicates ML$_{\text{reg}}$ models are the most volatile and prone to undesirable outliers. Std and ML$_{\text{clus}}$ performed comparably to each other given similar medians, ranges and inter-quartile ranges for many key dosimetric features. This is particularly true for rectum D$_{\text{mean}}$, rectum D$_{\text{max}}$ and PTV dose falloff.

Figure 6.2 illustrates relative weight deviations from MCO$_{\text{gs}}$ at patient level for all three methods. In general, patient-level deviations were moderately small overall. Maximum deviations of 7.39%, 17.1% and 5.58% were observed for Std, ML$_{\text{reg}}$ and ML$_{\text{clus}}$ respectively. ML$_{\text{clus}}$ was considered the optimal calibration method given median relative weighing factor difference from MCO$_{\text{gs}}$ close to zero in all cases and the small difference ranges. ML$_{\text{reg}}$ was considered the poorest performer of the three methods given the largest range and inter-quartile range differences from MCO$_{\text{gs}}$ in all cases.

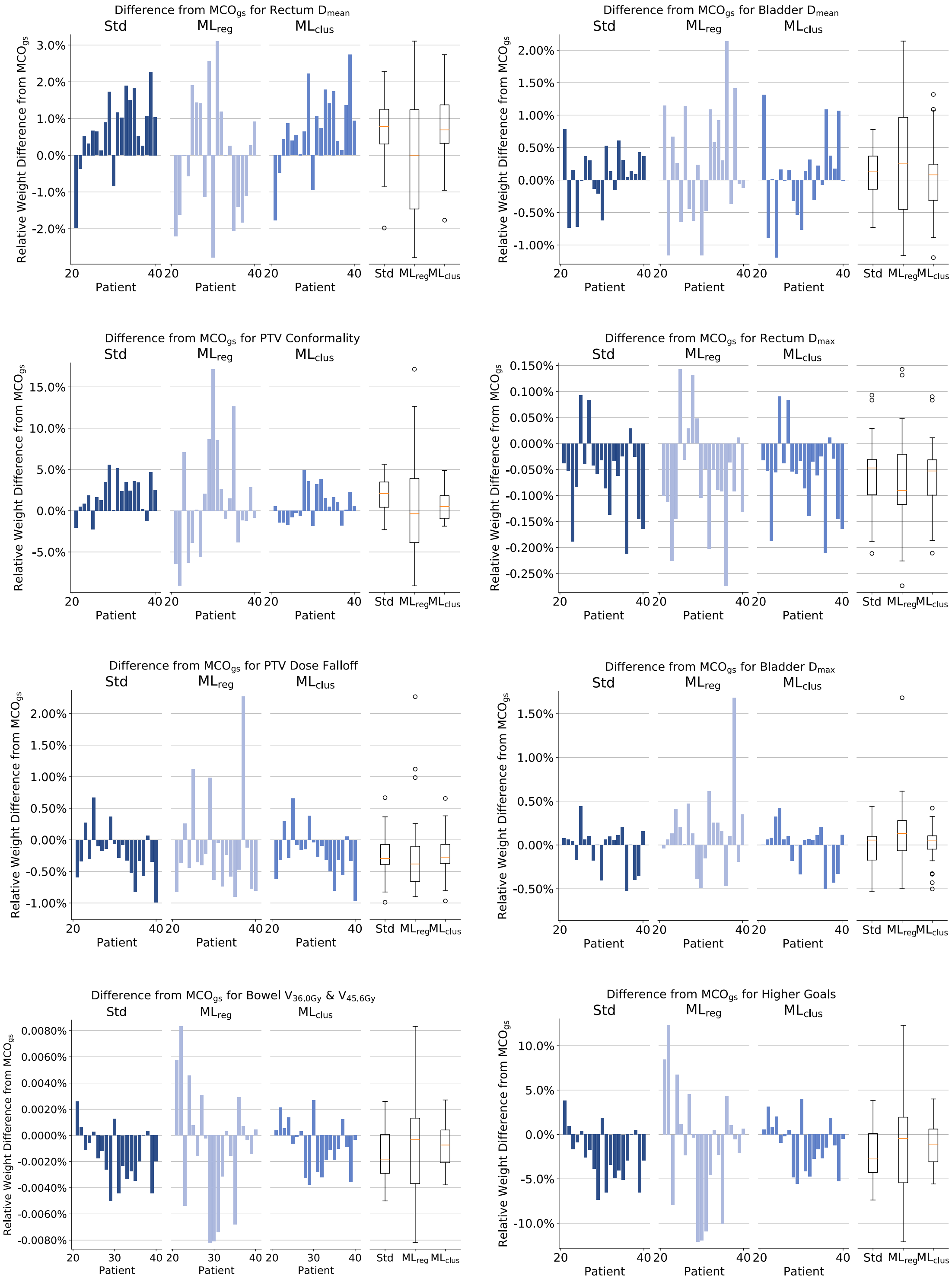


Figure 6.3: Plots showing relative weight difference from MCO_{gs} for the validation dataset. Bar chart are order patient 21-40 and box plot represent the overall distribution.

Weight metric	$\mathrm{MCO}_{\mathrm{gs}}$	Std	$\mathrm{ML}_{\mathrm{reg}}$	$\mathrm{ML}_{\mathrm{clus}}$
Rectum D_{mean}	3.46% ± 0.999%	**4.18%**	3.37% ± 1.51%	4.17% ± 0.257%
Bladder D_{mean}	1.10% ± 0.421%	1.18%	1.34% ± 0.820%	1.14% ± 0.404%
PTV Conformality	8.86% ± 2.252%	**10.7%**	10.1% ± 6.27%	9.55% ± 1.75%
Rectum D_{max}	0.163% ± 0.0800%	**0.102%**	**0.0943% ± 0.0771%**	**0.102% ± 0.00820%**
PTV Dose Falloff	0.926% ± 0.390%	**0.695%**	**0.763% ± 0.720%**	0.705% ± 0.0153%
Bladder D_{max}	0.487% ± 0.239%	0.459%	0.641% ± 0.598%	0.481% ± 0.116%
bowel D_{medium}	0.0575% ± 0.00204%	**0.0559%**	**0.0567% ± 0.00400%**	0.0568% ± 0.00123%
Higher Goals	85.0% ± 3.02%	**82.6%**	83.6% ± 5.87%	83.8% ± 1.82%

Table 6.4: Summary of PG relative weights for PSV. Values are mean averages across the the validation dataset ± one standard deviation. Boldface indicates statistically significant differences from $\mathrm{MCO}_{\mathrm{gs}}$ at the 95% level.

6.5.4 Dose Summary

See Table 6.5 for a dosimetric summary of the calibration methods. Also, see Figure 6.4 for an illustration of dosimetric differences from $\mathrm{MCO}_{\mathrm{gs}}$ for key dose-related metrics for each patient in the validation dataset. Of the 24 key metrics tested, ANOVA assumptions were met by seven. The rest were tested using a Friedman test. Of the 24 metrics, statistically significant difference were observed for five including PTV60 $D_{98\%}$ (Gy), rectum $V_{24.3\mathrm{Gy}}$ (%), rectum D_{mean}, homogeneity indices of PTV60 (CI60) and homogeneity indices of PTV48 (CI48).

Regression models are clearly the poorest performers in this scenario given substantial evidence they do not reflect $\mathrm{MCO}_{\mathrm{gs}}$weighting factors. Median differences are not minimised for any metrics and large variance show low concordance between predicted weighing factor values and those of $\mathrm{MCO}_{\mathrm{gs}}$. $\mathrm{ML}_{\mathrm{reg}}$was not anaylsed further and the remainder of this PSV dosimetric summary will refer to Std and $\mathrm{ML}_{\mathrm{clus}}$ only. PTV coverage and hotspot metrics (i.e., PTV60 and PTV48 $D_{98\%}$, $D_{50\%}$ and $D_{2\%}$) for Stdand $\mathrm{ML}_{\mathrm{clus}}$ were within 0.642Gy of $\mathrm{MCO}_{\mathrm{gs}}$ and OAR objectives were within 2.49% and 1.27Gy for volume and dose metrics respectively.

$\mathrm{ML}_{\mathrm{reg}}$and $\mathrm{ML}_{\mathrm{clus}}$ are considered equivalent given of the 24 metrics of interest, median dosimetric difference from $\mathrm{MCO}_{\mathrm{gs}}$ are minimised by Std for 12 metrics and $\mathrm{ML}_{\mathrm{reg}}$for 12 metrics. In addition, these two methods are highly dosimetrically comparable as in general dosimetric deviations from $\mathrm{MCO}_{\mathrm{gs}}$ across all patients were considered small, likely not of clinical significance and of a similar magnitude at per-patient.

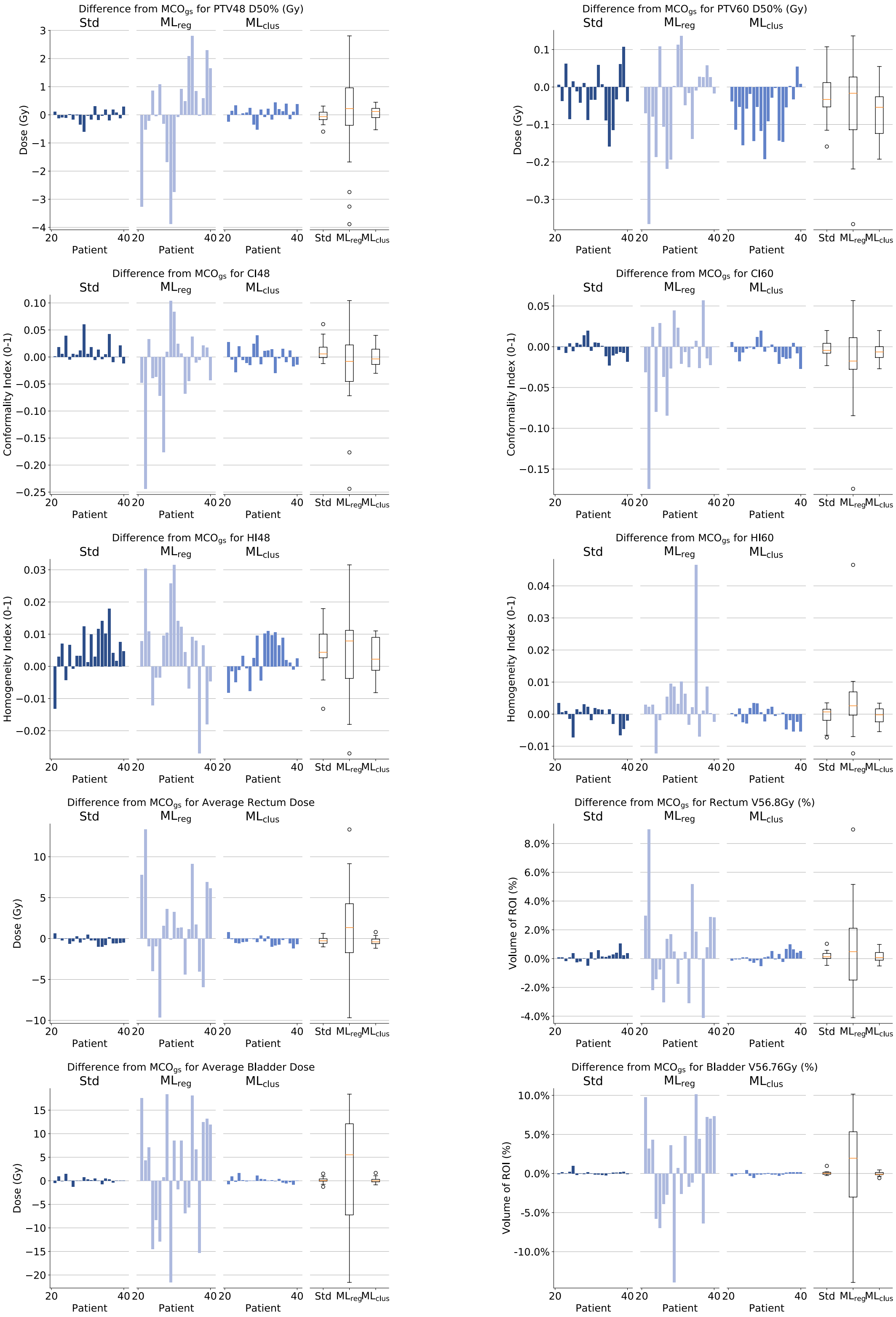


Figure 6.4: Plots for PSV showing absolute difference of Std, $\mathrm{ML_{reg}}$ and $\mathrm{ML_{clus}}$ from $\mathrm{MCO_{gs}}$. Distributions are across the validation dataset and show key dose related metrics for each of the three calibration techniques.

Metric		MCO_{gs}	Std	ML_{reg}	ML_{clus}
PTV60	$D_{98\%}$ (Gy)	57.5 ± 0.200	57.5 ± 0.171	**57.3 ± 0.570**	57.5 ± 0.134
	$D_{50\%}$ (Gy)	60.0 ± 0.0748	59.9 ± 0.0611	59.9 ± 0.122	59.9 ± 0.0395
	$D_{2\%}$ (Gy)	61.7 ± 0.0879	61.7 ± 0.0853	61.7 ± 0.104	61.7 ± 0.0794
	CI	0.853 ± 0.00910	0.851 ± 0.0108	**0.834 ± 0.0523**	0.848 ± 0.0112
	HI	0.0700 ± 0.00435	0.0696 ± 0.00358	0.0742 ± 0.0108	**0.0694 ± 0.00309**
PTV48	$D_{98\%}$ (Gy)	46.3 ± 0.532	46.1 ± 0.407	46.2 ± 0.592	46.2 ± 0.422
	$D_{50\%}$ (Gy)	53.3 ± 1.32	53.2 ± 1.20	53.3 ± 1.48	53.3 ± 1.22
	$D_{2\%}$ (Gy)	59.1 ± 0.277	59.2 ± 0.242	59.3 ± 0.501	59.1 ± 0.236
	CI	0.812 ± 0.0327	0.823 ± 0.0210	0.789 ± 0.0779	0.813 ± 0.0291
	HI	0.241 ± 0.0112	**0.246 ± 0.00892**	0.246 ± 0.0141	0.243 ± 0.0101
Rectum	$V_{24.3Gy}$ (%)	29.1% ± 8.47%	28.5% ± 7.94%	32.3% ± 11.4%	**28.4% ± 8.21%**
	$V_{32.4Gy}$ (%)	23.7% ± 7.44%	23.2% ± 7.14%	25.8% ± 9.07%	23.3% ± 7.29%
	$V_{40.5Gy}$ (%)	18.6% ± 6.17%	18.2% ± 6.01%	20.0% ± 7.26%	18.3% ± 6.11%
	$V_{48.6Gy}$ (%)	12.8% ± 4.41%	12.6% ± 4.38%	13.5% ± 5.09%	12.7% ± 4.46%
	$V_{52.7Gy}$ (%)	9.32% ± 3.28%	9.23% ± 3.26%	9.77% ± 3.69%	9.32% ± 3.37%
	$V_{56.8Gy}$ (%)	5.32% ± 2.12%	5.48% ± 2.20%	5.97% ± 2.61%	5.46% ± 2.31%
	V_{60Gy} (%)	0.299% ± 0.445%	0.271% ± 0.221%	0.596% ± 0.770%	0.180% ± 0.168%
	$V_{60.8Gy}$ (%)	0.0690% ± 0.129%	0.0430% ± 0.0419%	0.0220% ± 0.0357%	0.0223% ± 0.0351%
	D_{mean}(Gy)	18.7 ± 3.72	18.4± 3.50	20.1 ± 5.30	**18.3 ± 3.69**
Bladder	$V_{40.5Gy}$ (%)	18.0% ± 11.3%	18.0% ± 11.3%	20.9% ± 10.4%	18.1% ± 11.4%
	$V_{48.6Gy}$ (%)	12.2% ± 7.83%	12.0% ± 7.70%	14.3% ± 7.55%	12.3% ± 7.83%
	$V_{52.7Gy}$ (%)	9.46% ± 6.33%	9.37% ± 6.25%	11.1% ± 6.02%	9.47% ± 6.29%
	$V_{56.8Gy}$ (%)	6.49% ± 4.58%	6.54% ± 4.65%	7.35% ± 4.01%	6.44% ± 4.59%
	D_{mean}(Gy)	20.2 ± 8.72	20.3 ± 8.77	22.2 ± 8.32	20.3 ± 8.91

Table 6.5: Summary of key dose metrics for PSV. Values shown are Mean ± 1 Standard Deviation. Statistical difference from MCO_{gs} at the 95% level of significance is indicated by boldface.

All three methods have comparable dosimetric qualities to MCO_{gs} with deviations either not statistically significant at the 95% level, or of a small magnitude.

6.5.5 Conclusions

Following exhaustive searches for the optimal models in each case, ML_{clus} was found to predict MCO_{gs} more closely than any other method. Nevertheless, all methods performed favourably and achieved clinically acceptable planning. This was positive outcome especially given the small size of the training database.

6.6 Rectum Modelling Results

The rectum site is a peculiar case for the generation of calibration parameters for an AP system given the PGs are inhomogenous. There are cases that are missing genitals and/or stoma delineations. To manage this, regression and clustering were defined in a similar way to other sites but predicted values set to zero where the volume of the related ROI is zero. Therefore, models are defined using homogenous predictive features only and cases with missing delineations automatically set to zero.

FeatureDS1 contained 114 predictive features, of which 23 were retained for FeatureDS2. A summary of variables in FeatureDS2 can be found in Table 6.6. Of the 40 Principal Components defined using FeatureDS1, PC1 accounts for 61.2% of the variance with PC1-6 accounting to approximately 95% of the variance in all 40.

Type	**Feature**	**Excluded**
		Features
Volumetric	$\text{OV}_{\text{PTV45+0.2,reduced Bowel Bag}}$	44
Derived	Slope $\text{OV}_{\text{PTV45+2.0,External}}$ and $\text{OV}_{\text{PTV45+2.2,External}}$	24
Volumetric	Volume of the PTV45	23
Volumetric	$\text{OV}_{\text{PTV45,Full Bowel Bag}}$	12
Volumetric	Volume of the Full Bowel Bag	7
Volumetric	Volume of the External	5
Spatial	Av. distance between the Full Bowel Bag and External	4
Spatial	Min. distance between the Full Bowel Bag and External	2
Derived	Ratio of Full Bowel Bag and External	1
Spatial	Distance between the centers of the Full Bowel Bag and External	1
Spatial	Max. distance between PTV45 and reduced Bowel Bag	1
Derived	$\text{OV}_{\text{VIFPT+45.4500,reduced Bowel Bag}}$	1
Spatial	Av. distance between the Full Bowel Bag and reduced Bowel Bag	1
Spatial	Distance between the centers of the PTV45 and reduced Bowel Bag	1
Spatial	Max. distance between PTV45 and Full Bowel Bag	1
Spatial	Distance between the centers of the Full Bowel Bag and reduced Bowel Bag	0
Derived	Ratio of PTV45 and Full Bowel Bag	0
Derived	Ratio of PTV45 and reduced Bowel Bag	0
Derived	Ratio of Full Bowel Bag and reduced Bowel Bag	0
Spatial	Distance between the centers of the PTV45 and External	0
Spatial	Min. distance between PTV45 and External	0
Spatial	Max. distance between the Full Bowel Bag and reduced Bowe Bag	0

Derived	Ratio of PTV45 and External	0

Table 6.6: Summary of the rectum predictive features chosen for FeatureDS2 from FeatureDS1. The number of correlated features in FeatureDS1 that were excluded from FeatureDS2 are summarised in the column Excluded Features.

6.6.1 Leave-one-out summary

For reg-raw rectum LOOCV, a total of 21,384,480 individual regression models were generated:

$$\begin{array}{ccccccccc} 4 & \times & 40 & \times & 3 & \times & \sum\limits_{k=1}^{5}\binom{23}{k} & = & 21{,}384{,}480. \\ \text{(PGs)} & & \text{(patients)} & & \text{(degrees)} & & \text{(combinations)} & & \text{(total)} \end{array}$$

Therefore, given the number of PG and degrees, 12 full LOOCV models were defined over the training patient for each combination type. For each PG and degree, it took approximately 0.15 minutes to generate all 1-feature LOOCV models, approximately 0.5 minutes for 2-feature models, approximately 3.5 minutes for three, approximately 15 minutes for four and approximately 65 minutes for five. For reg-PCA, 19,200 LOOCV models were generated:

$$\begin{array}{ccccccccc} 4 & \times & 40 & \times & 3 & \times & 40 & = & 19{,}200, \\ \text{(PGs)} & & \text{(patients)} & & \text{(degrees)} & & \text{(PC combinations)} & & \text{(total)} \end{array}$$

and it took comparably long to generate each reg-PCA model for any one PG and degree as a 1-feature reg-raw model.

Of the four PGs, two of the optimal models were defined using raw features (i.e., PTV conformality and stoma D_{mean}) and two using PCA features (i.e., bowel bag $D_{\text{mean}}$and genitals D_{mean}). Models defined using raw features yielded higher average R^2 values than those defined using PCA features. Similarly to PSV, this indicates a strong linear relationships between raw predictive features and weighting factors. All regression models predicting genitals D_{mean} weighting factors were very low given the LOOCV model that minimised MSE obtained an average R^2 of less than 0.1. This suggests there are no strong relationships between the weighting factors and any of the predictive features considered.

Of the 23 raw predictive features in FeatureDS2, seven were used across the two optimal reg-raw models. Of the 40 PCA features generated from FeatureDS1, three were used in the two optimal reg-PCA models (PC1-3). PC1-2 and PC1-3 accounted

Planning Goal	**Regression Model Type**	**Features**	**Training**		**Validation**
			Av adj R^2	**MSE**	**MSE**
PTV Conformalty	3 features cubic	Volume of PTV45 Minimum distance between PTV45 and the External Ratio between PTV45 and Bowel Bag	0.841	269,664	1,050,892
Bowel Bag D_{mean}	3 features linear	PC1 PC2 PC3	0.339	30,645	72,799
Genitals D_{mean}	2 features linear	PC1 PC2	0.09	25,782	35,469
Stoma D_{mean}	4 features cubic	Volume of the Bowel Bag Centre of the Bowel Bag to centre of Full Bowel Bag Volume of the External in field of PTV45 Volume of Full Bowel Bag in field of PTV45	0.974	3,123	96,396

Table 6.7: Summary of ML_{reg} model formations determined automatically for Rectum via leave-one-out.

for 76% and 86% of the variance in FeatureDS1 respectively. So much of the variance being described with a small number of Principal Components is likely the reason PCA features were found to be optimal during LOOCV.

Of the four PGs, all optimal cluster models contained PCA features except stoma D_{mean}. With a similar outcome noted for PSV, this suggests the prepossessing of the variance to Principal Components and modelling over fewer dimensions is better suited to clustering than raw features are. In comparison to optimal regression models, MSE indicates cluster models are a better fit to MCO_{gs} than regression models given lower validation MSE values for all PGs with validation values even lower than training MSE values for two Ps: PTV conformality and bowel bag. Nevertheless, this data are not considered typically well suited to clustering in general given high SSE values (very far from zero) and lower silhouette scores (close to zero) in addition to large MSE values. However, given the results observed for PSV weighting factors and dosimetry, this does not necessarily mean poor performance.

Comparing regression models to the cluster models, clustering showed clear modelling benefits over regression. MSE value were smaller for all PGs for both training and validation. This is particularly true for PTV conformality and Bowel Bag D_{mean} where significant reductions are noted in validation MSE values.

Planning goal	Number of clusters	Feature Type	Cluster SSE	Silhouette Score	Training WF MSE	Validation WF MSE
PTV Conformalty	7	PCA	1331	0.18369	366,939	296,567
Bowel Bag D_{mean}	10	PCA	989	0.15633	39,339	5,985
Stoma D_{mean}	5	Raw	471	0.1503	10,727	31,573
Genitals D_{mean}	4	PCA	1895	0.18526	27,086	75,603

Table 6.8: Summary of ML_{clus} model formations determined automatically for Rectum via leave-one-out.

6.6.2 Weight Summary

Weight summaries across the validation database can be found in Table 6.9 and Figure 6.5. Patient level distributions can be found in Figure 6.6. Overall, distributions were found were found to be highly comparable across PGs regardless of the method chosen. Mean deviations from MCO_{gs} did not differ by more than 2.4% and following a Friedman test (a non-parametric analysis of variance test for repeated measures), no significant differences were found even at very lenient levels (e.g., 20% level of significance).

However, although there was no statistical evidence to suggest differences between models, standard deviation values indicate smaller weighing factor variance for models than MCO_{gs}. Therefore, much of the variation seen in the difference plots in Figure 6.6

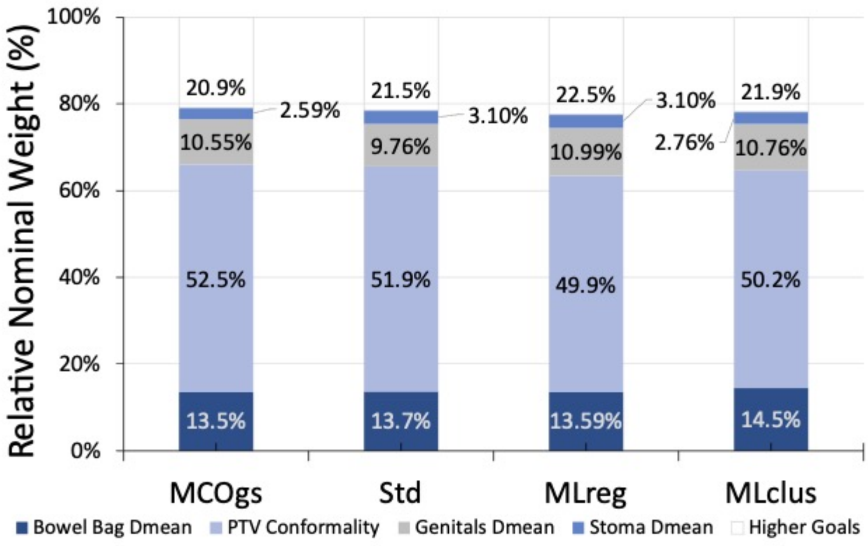


Figure 6.5: Plots showing relative weighting factor distribution for all calibration methods for the Rectum validation dataset.

is actually due to the variance in $\mathrm{MCO}_{\mathrm{gs}}$. Given the median difference from $\mathrm{MCO}_{\mathrm{gs}}$ illustrated in Figure 6.6, Std is the optimal method as the median is closest to zero in three of the five cases: PTV conformality, stoma $\mathrm{D}_{\mathrm{mean}}$ and PG_{H}. Nevertheless, given comparable ranges, similar medians and few outlier patients, $\mathrm{ML}_{\mathrm{clus}}$ can be considered equivalent to Std. $\mathrm{ML}_{\mathrm{reg}}$ was considered the poorest performed given some extreme outliers especially for conformality, stoma $\mathrm{D}_{\mathrm{mean}}$ and PG_{H}. Regardless, all differences were considered to be small overall given small ranges and comparable averages.

At patient-level, the most extreme differences were observed for patient 45 under Std and $\mathrm{ML}_{\mathrm{reg}}$ and patient 60 under $\mathrm{ML}_{\mathrm{reg}}$. For example, Patient 45's $\mathrm{ML}_{\mathrm{reg}}$ weighing factor for the PTV conformality PG was 38.7% higher than $\mathrm{MCO}_{\mathrm{gs}}$, the most extreme outlier seen among the predicted weighting factors. Patient 60's weighing factor for this PG and method was also extreme at 30.9% lower than $\mathrm{MCO}_{\mathrm{gs}}$. Patient 45 and 60 under $\mathrm{ML}_{\mathrm{reg}}$ were also notably extreme for PG_{H} at -14.9% and +17.3% respectively. This indicates the model prioritised navigated PGs more highly than $\mathrm{MCO}_{\mathrm{gs}}$ for this patient and this may be due to the fact patient 45's $\mathrm{MCO}_{\mathrm{gs}}$ weights were notable lower than the average $\mathrm{MCO}_{\mathrm{gs}}$ weight.

Given similar weighting factors, similar dose related metrics are expected. This would mean methods are comparable with few clinical benefits of gold standard planning or complex modelling. This would imply that even basic calibration methods such as Std can be used to achieve acceptable planning in general. Dose-related summaries are discussed in the following section.

Planning Goal	$\mathrm{MCO}_{\mathrm{gs}}$	Std	$\mathrm{ML}_{\mathrm{reg}}$	$\mathrm{ML}_{\mathrm{clus}}$
Bowel Bag $\mathrm{D}_{\mathrm{mean}}$	13.5% ± 6.37%	13.7.0% ± 2.88%	13.6% ± 4.18%	14.5% ± 5.78%
PTV Conformality	52.5% ± 10.5%	51.9% ± 6.32%	49.9% ± 8.51%	50.2% ± 7.07%
Genitals $\mathrm{D}_{\mathrm{mean}}$	10.6% ± 5.95%	9.76% ± 2.69%	11.0% ± 3.99%	10.8% ± 3.29%
Stoma $\mathrm{D}_{\mathrm{mean}}$	2.59% ± 4.83%	3.10% ± 5.50%	3.10% ± 6.83%	2.76% ± 4.97%
Higher Goals	20.9% ± 3.17%	21.5% ± 1.03%	22.5% ± 5.39%	21.9% ± 2.23%

Table 6.9: Summary of PG relative weights for Rectum. Values are mean averages across the the validation dataset ± one standard deviation. Boldface indicates statistically significant differences from $\mathrm{MCO}_{\mathrm{gs}}$ at the 95% level.

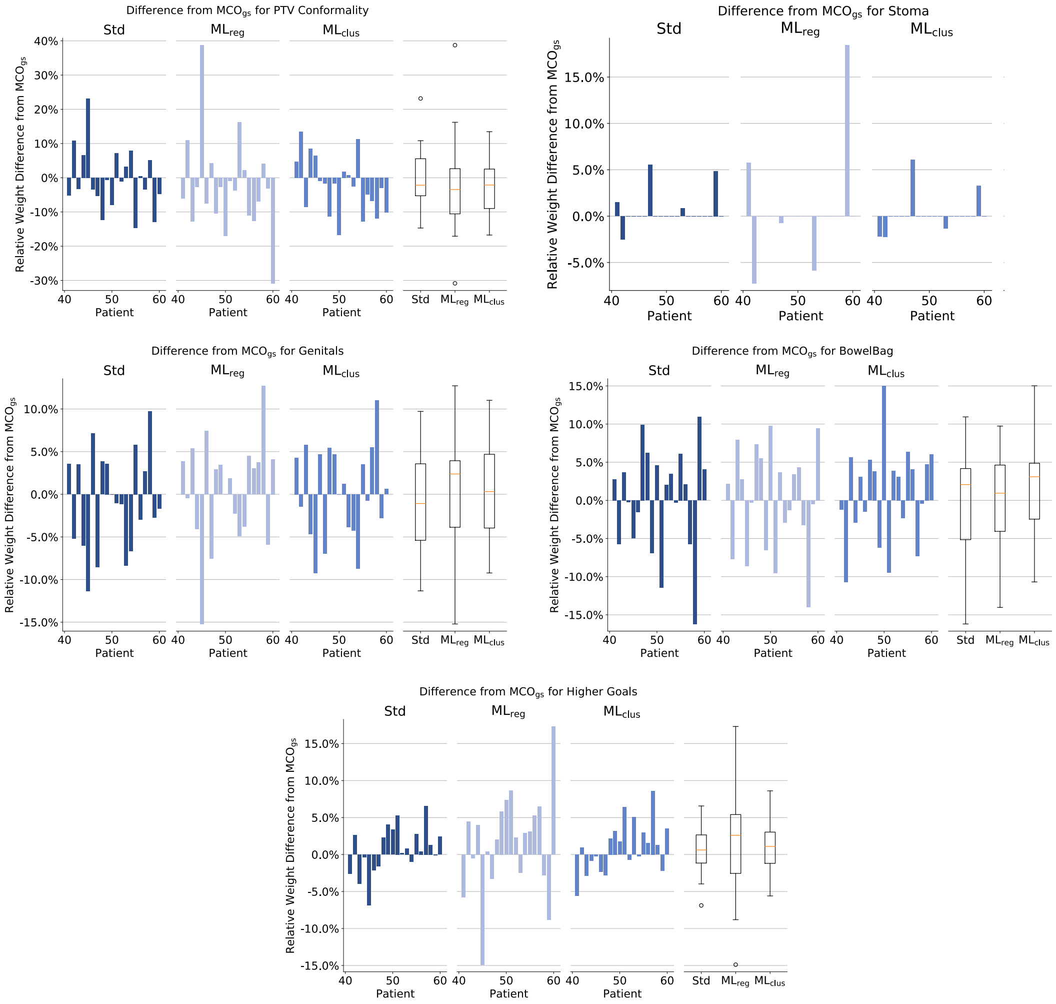


Figure 6.6: Plots showing relative weight difference from MCO_{gs} for the Rectum validation dataset. Bar chart are order patient 41-60 and box plot represent the overall distribution.

6.6.3 Dose Summary

A summary of dosimetric features related to plans can be found in Table 6.10 with patient-level distributions illustrated in Figure 6.7. No statistically significant dosimetric differences were observed between the generated plans under each model. therefore, no boldface is used in Table 6.10. Not only were mean values comparable, standard deviations were very similar also. This is reflected at patient-level given small ranges of the individual differences and comparable medians for all key dose-related metrics. At population level, the relative weighting factors between methods different by less than 2.07% for all PGs and suggests all methods will perform similarly to MCO_{gs}.

There is also no clear optimal method. Regarding median differences from MCO_{gs} across the metrics, both Std and ML_{clus} perform favourably with Std medians closest

ROI Name	DVH Statistic Type	MCO$_{gs}$	Std	ML$_{reg}$	ML$_{clus}$
PTV45	$D_{98\%}$ (Gy)	43.4 ± 0.107	43.5 ± 0.0751	43.5 ± 0.107	43.4 ± 0.112
	$D_{50\%}$ (Gy)	44.9 ± 0.0536	44.9 ± 0.0576	44.9 ± 0.0599	44.9 ± 0.0508
	$D_{2\%}$ (Gy)	46.4 ± 0.142	46.3 ± 0.121	46.3 ± 0.127	46.3 ± 0.124
	CI	0.879 ± 0.0213	0.875 ± 0.0190	0.872 ± 0.0269	0.873 ± 0.0191
	HI	0.0651 ± 0.00532	0.0637 ± 0.00410	0.0637 ± 0.00491	0.0643 ± 0.00509
Stoma Region	D_{mean}	0.473 ± 0.945	0.436 ± 0.909	0.488 ± 0.994	0.482 ± 0.995
Genital Region	D_{mean}	2.59 ± 2.00	2.48 ± 1.81	2.47 ± 1.82	2.42 ± 1.79
Bowel Bag	D_{mean}	26.8 ± 3.28	26.5 ± 3.13	26.5 ± 3.29	26.4 ± 3.37

Table 6.10: Summary of key dose metrics for Rectum. Values shown are Mean ± 1 Standard Deviation. Statistical difference at the 95% level of significance is indicated by boldface.

to zero for four metrics (i.e., PTV45 $D_{98\%}$, PTV45 $D_{2\%}$, HI45 and genital region) and ML$_{clus}$ closest to zero for the other four (i.e., PTV45 $D_{50\%}$, CI45, bowel bag D_{mean} and Stoma D_{mean}). These finding shows so evidence to suggest the PTV and other PG$_H$ related dose metrics are prioritise more highly under Std than ML$_{clus}$ and navigated PG related metrics are prioritised more highly under ML$_{clus}$. There is not evidence suggesting this to a significant difference however.

Patient 60 is a notable outlier for CI45 under ML$_{reg}$ and patient 57 is a notable outlier in a number of cases: PTV45 $D_{98\%}$ under Std, PTV45 $D_{2\%}$ under ML$_{clus}$, HI45 under ML$_{clus}$. Patient 60 is likely an outlier for CI45 under ML$_{reg}$ given a notably high PTV conformality EF for this patient that deviated patterns seen in training database. Patient 57 is an anatomical outlier with a notably large PTV, a case not well defined given cases considered in the training database.

6.6.4 Conclusion

Even given the homogeneity of the site, all calibration methods were comparable to MCO$_{gs}$. Even the simplest method, Std, yielded clinically applicable planning with few outliers. This is though to be due planning being simple for this site given a small number of PGs and few OAR to balance. Many anatomical features were also found to be highly correlated and followed a standard normal distribution making them very well suited to scaling and machine learning.

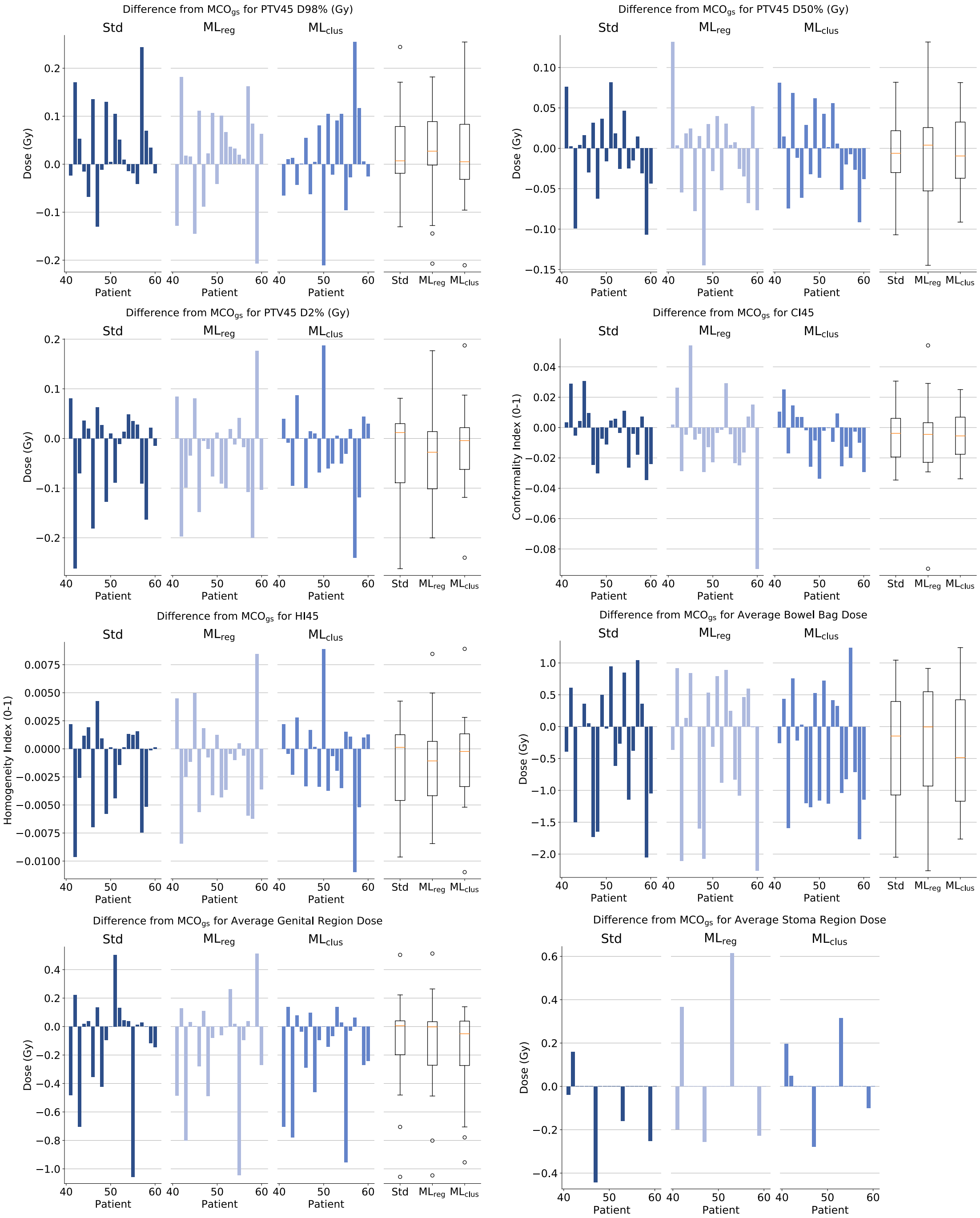


Figure 6.7: Plots for Rectum showing absolute difference of Std, ML_{reg} and ML_{clus} from MCO_{gs}. Distributions are across the validation dataset and show key dose related metrics for each of the three calibration techniques.

6.7 Lung Modelling Results

A total of 241 predictive features were defined for FeatureDS1 of which 95 were retained for FeatureDS2. Following a Shapiro-Wilks test for normality, 23 of the 241 predictive features were found to differ significantly from a normal distribution hence a standard normal distribution scaler with mean centralisation and standard deviation scaling could not be employed as in the PSV and rectum cases. For this reason, a robust scaler based on median centralisation and inter-quartile range scaling was considered in addition to a standard scaler. See Figure 5.2 for a comparsion of each scaler and Appendix B for box plot representations under each scaler. This was done for comparison but also to ensured PCA features were generated appropriately and K-means clustering was valid. Regarding regression, scaling features has no mathematical implications for the predicted feature, only the ability to interpret the coefficients of the equation. Interpretation of regression coefficients is not valuable in this context hence regression is considered unaffected by scaling.

However, PCA transformation is highly dependent on the scale of each variable. If the variance within variables is more valuable than the variance between them (as in this work), applying an appropriate scaling technique to the raw variables is vital especially when the resulting Principal Components are to be used for prediction. Not doing so can result in misleading results as variables with inordinately large scales are essentially being given more weight in the PCA transform than those on smaller scales. Similarly, centroid-based clustering such as K-means uses an isotropic search and is highly dependent on the spatial relationships between data points. That is, data points are assigned to clusters based on minimum distance. Therefore, allowing any one variable to have an inordinately large scale is comparable to assigning a lower priority to that variable than other variables with smaller scales. Given no one predictive feature in this work is considered more valuable than another, all must be appropriately scaled.

6.7.1 Leave-one-out summary

Given the number of patients and number of features in FeatureDS2, 4- and 5-feature reg-raw models were not generated using an exhaustive search but instead stepwise forward selection was employed using the optimal 3-feature model in each case. In total, 68,719,200 reg-raw models were defined using LOOCV:

Planning	Regression	Features	Training		MSE
Goal	Equation		Av adj R2	training	validation
Lung D_{mean}	5 features quadratic	Volume of the Oesophagus $OV_{\text{Oespahagus,PTV55}}$ $VIF_{\text{Ips Lung,PTV55}}$ Maximum distance between the Cont Lung and PTV55 Distance from the center of of the Spinal Cord and Heart	0.859	9,600	40,177
Heart D_{max}	5 features linear	$VIF_{\text{Spinal Cord,PTV55}}$ Distance from the center of the Spinal Cord to the External Slope of $OV_{\text{PTV}_{1.6\text{cm}},\text{Spinal Cord}}$ and $OV_{\text{PTV}_{1.8\text{cm}},\text{Spinal Cord}}$ $OV_{\text{Cont Lung,PTV55}}$ Ratio of External and the Spinal Cord	0.703	481	842
PTV Conformality	3 features quadratic	Volume of the Cont Lung in field of the PTV55 Distance from the center of Ips Lung and Combined Lung Maximum distance between the Cont Lung and the Heart	0.593	11,878	34,824
Lung D_{max}	3 features quadratic	Distance from the center of External and Cont Lung Distance from the center of Ips Lung and Combined Lung Ratio between PTV55 and Ipsi Lung	0.596	931	18,194

Table 6.11: Summary of ML_{reg} model formations determined automatically for Lung via leave-one-out.

$$\begin{array}{ccccccccc} 4 & \times & 40 & \times & 3 & \times & \left(3\binom{95}{1}+\binom{95}{2}+\binom{95}{3}\right) & = & 68{,}719{,}200. \\ \text{(PGs)} & & \text{(patients)} & & \text{(degrees)} & & \text{(combinations)} & & \text{(total)} \end{array}$$

Therefore, given the number of PG and degrees, 12 full LOOCV models were defined over the training patients for each combination type. Following an exhaustive search, 1-feature LOOCV models for each lung PG and degree took approximately 0.75 minutes, 2-features took approximately 15 minutes each and 3-feature models took approximately 450 minutes each. Running models in parallel did not reduced the time taken for the code to run. All 4- and 5-feature models each took approximately 0.75 minutes; comparatively long as the 1-feature cases. When optimal 3-feature models had been defined, 4-features models were chosen by exploring all combinations of four features given the first three were fixed as features defined in by optimal 3-feature model. Similarly, once the optimal 4-feature model had been established, the optimal 5-feature model was determined similarly: by fixing the first four features given the optimal 4-feature model and exploring all combinations of five features varying only feature number 5. For reg-PCA models, as in

Scaler	Planning	Number	Feature	SSE	Silhouette	Training	Validation
	goal	of clusters	Type		Score	MSE	MSE
Standard	Lung D_{mean}	15	Raw	1427	0.0406	17481	33443
	Heart D_{max}	11	PCA	3376	0.124	580	869
	PTV Conformality	1	Raw	3800		22107	19042
	Lung D_{max}	15	PCA	1427	0.0406	664	915
Robust	Lung D_{mean}	1	Raw	8662		26499	16950
	Heart D_{max}	8	PCA	3398	0.103	861	975
	PTV Conformality	16	Raw	1040	0.0400	21041	45664
	Lung D_{max}	10	Raw	1521	0.0390	1262	1408

Table 6.12: Summary of ML_{clus} model formations determined automatically for Lung via leave-one-out using a standard and a robust scaler.

the rectum reg-PCA case, 19,200 LOOCV models were generated in total.

A summary of optimal models following reg-raw LOOCV is found in Table 6.11. Optimal model for all PGs contained raw features only. Of the features within feature sets, 15 unique features were used of which only one occurred in more than one feature set: *distance from the center of Ipsilateral Lung and Combined Lung*. This suggests each PG is influenced by very different anatomical factors. Average R^2 values were high for PG1 (lung D_{mean}) and PG2 (heart D_{max}) indicating these PGs have a strong curvilinear relationship that is well defined for regression in general. Poor average R^2 values for PG3 and PG4 suggests a poor curvilinear relationship and may be best modelled using a clustering technique other than K-means.

A summary of cluster LOOCV optimal models can be found in Table 6.12. Given 23 of the features in FeatureDS1 did not follow a normal distribution, both a standard and robust scaler were considered for this treatment site. Training MSE values were lower using a standard scaler than a robust scaler with the exception of PTV conformality which was relatively comparably given a value 4.82% lower than the robust scaler. Therefore, only the standard scaler was used.

Silhouette scores and SSE values indicate the solutions were not well defined for K-means clustering in general. Nevertheless, regarding validation MSE values, clustering indicates solutions that optimised for MCO_{gs} for all PG given validation MSE values were minimised using a clustering solutions (standard scaler).

6.7.2 Weight Summary

Weighing factor summaries across the validation database can be found in Table 6.13 and Figure 6.8. Patient level distributions can be found in Figure 6.9. Similarly to PSV and rectum, overall distributions were found to be highly comparable across PGs regardless of the method chosen. Mean deviations from MCO_{gs} did not differ by more than 1.72% and following a Friedman χ^2 test, there is very little evidence to suggest a difference given a p-value of 0.978.

Even with very little evidence to suggest differences, Std may be considered the most optimal given median deviations from MCO_{gs} across the validation database differ least for three of the six PG groups: PG4, PG5 and PG_{H}. Std also had the smallest ranges ML_{clus} minimised the median the difference from MCO_{gs} for two PGs (lung D_{mean} and PTV conformality) and ML_{reg} for one PG (Heart D_{max}).

The poorest performing method was Std. The median difference from MCO_{gs} was largest for ML_{reg} for five of the six PG groups. Therefore, although the median difference was minimised for Heart D_{max}, ML_{reg} showed the largest deviations from MCO_{gs} on average for all other PGs. This method also led to the most extreme outliers for all PG groups. The two most notable outliers were observed for patient 55 lung D_{mean} and patient 53 lung D_{max} under ML_{reg} showing relative weight differences of 11.4% and 8.93% respectively.

Planning Goal	MCOgs	Std	MLreg	MLclus
Lung D_{mean}	8.94% ± 1.99%	7.94%	7.61% ± 2.87%	8.25% ± 1.85%
Heart D_{max}	0.698% ± 0.711%	0.84%	0.677% ± 0.45%	0.608% ± 0.549%
PTV Conformality	7.50% ± 2.23%	6.69%	5.78% ± 2.49%	6.70% ± 0.140%
Oesophagus + Brachial Plexus D_{mean}	44.8% ± 1.25%	45.70%	46.0% ± 2.04%	45.8% ± 0.955%
Lung D_{max}	1.42% ± 0.556%	1.51%	2.25% ± 1.97%	1.25% ± 0.284%
Higher Goals	36.6% ± 1.03%	37.40%	37.7% ± 1.67%	37.4% ± 0.781%

Table 6.13: Summary of relative weights for Lung. Values are mean averages across the the validation dataset ± one standard deviation. Boldface indicates statistically significant differences from MCO_{gs} at the 95% level.

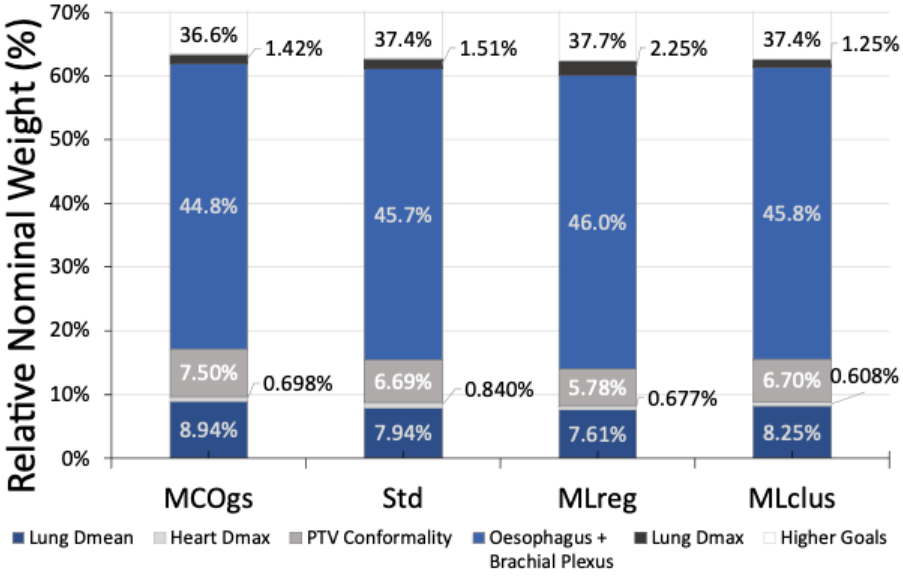


Figure 6.8: Plots for showing relative weight distribution for all calibration methods for the Lung validation dataset.

6.7.3 Dose Summary

A summary of key dosimetric outcomes can be found in Table 6.14 and illustrated in Figure 6.10. There were no statistically significant differences observed and there was considered to be no clinically significant difference between the calibration methods. On average across the validation database, PTV differences between all methods were within ± 0.043 Gy.

When compared to MCO$_{\text{gs}}$, the largest patient-level deviations observed were for average dose to the heart with a maximum deviation of 3.52 Gy observed for patient 48 under ML$_{\text{reg}}$. Patient 50 is a notable outlier patient under Std given comparatively extreme values observed for PTV55 D$_{98\%}$, PTV55 D$_{2\%}$ and PTV55 homogeneity. Std and ML$_{\text{clus}}$ are comparatively optimal given median differences from MCO$_{\text{gs}}$ were minimised by each method for five of the key metrics. Std minimised median differences from MCO$_{\text{gs}}$ for CI55, HI55, average dose to the heart, heart D$_{\text{48ccGy}}$, ipsilateral lung D$_{\text{mean}}$, ipsilateral lung D$_{\text{19ccGy}}$, Contralateral Lung D$_{\text{mean}}$. contralateral lung D$_{\text{19ccGy}}$ and Oesophagus D$_{\text{mean}}$. ML$_{\text{reg}}$ was comparatively the poorest performer given median difference from MCO$_{\text{gs}}$ were largest with this method for eight of the key metrics. A zero median difference was observed for all methods for the heart D$_{\text{cc48Gy}}$ metric.

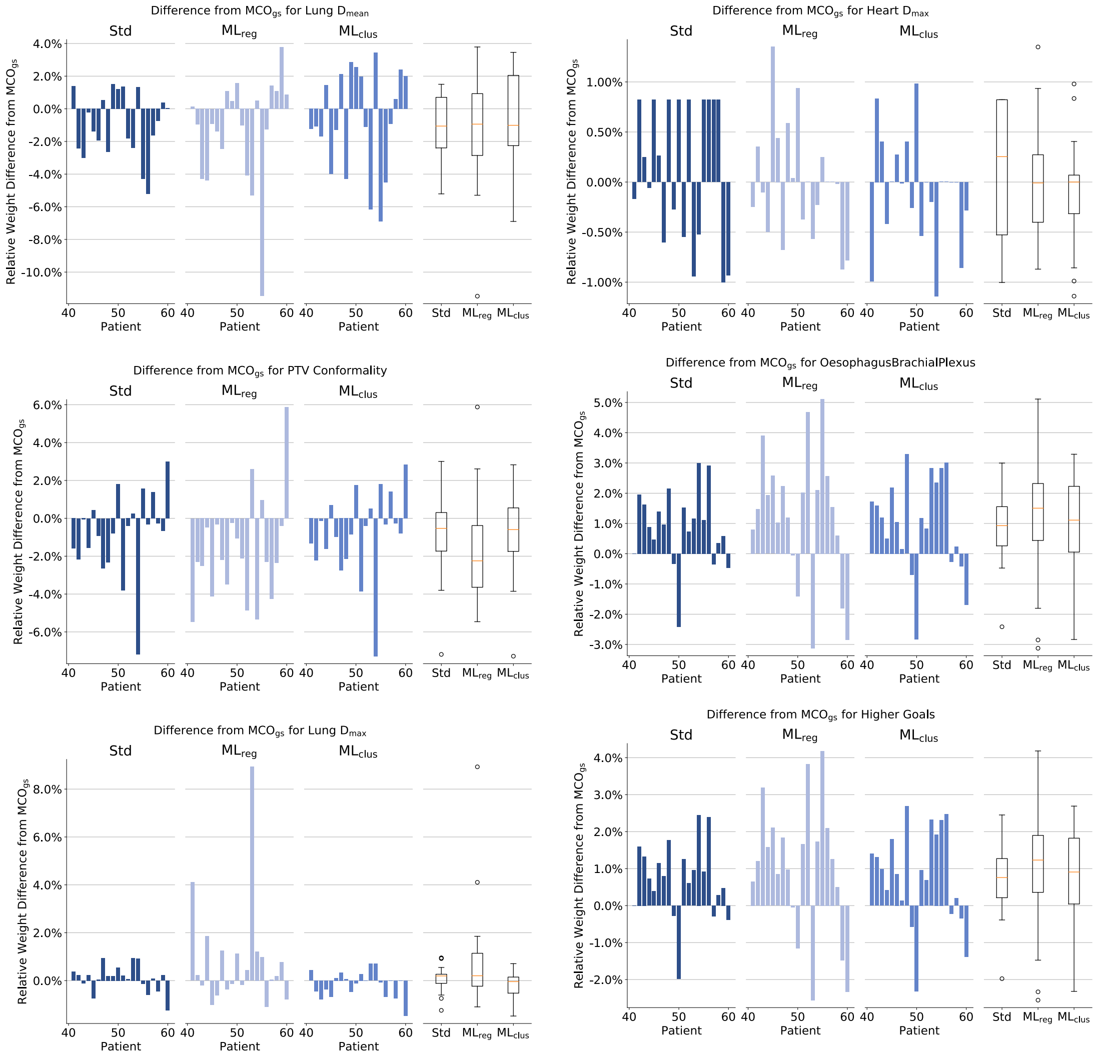


Figure 6.9: Plots showing relative weight difference for Std, ML_{reg} and ML_{clus} from MCO_{gs} for the Lung validation dataset. Bar chart are ordered patient 41-60 and box plot represent the overall distribution.

6.7.4 Conclusion

Similarly to PSV and rectum, differing calibration approaches were not found to have have clinically significant differences for weighting factors with all approaches highly comparable.

6.8 Chapter Summary

These studies have shown evidence expert-driven RBP calibration parameters can be effectively modelled using ML. Clustering techniques in particular show notable promise given median differences from MCO_{gs} consistently tended towards zero more often than

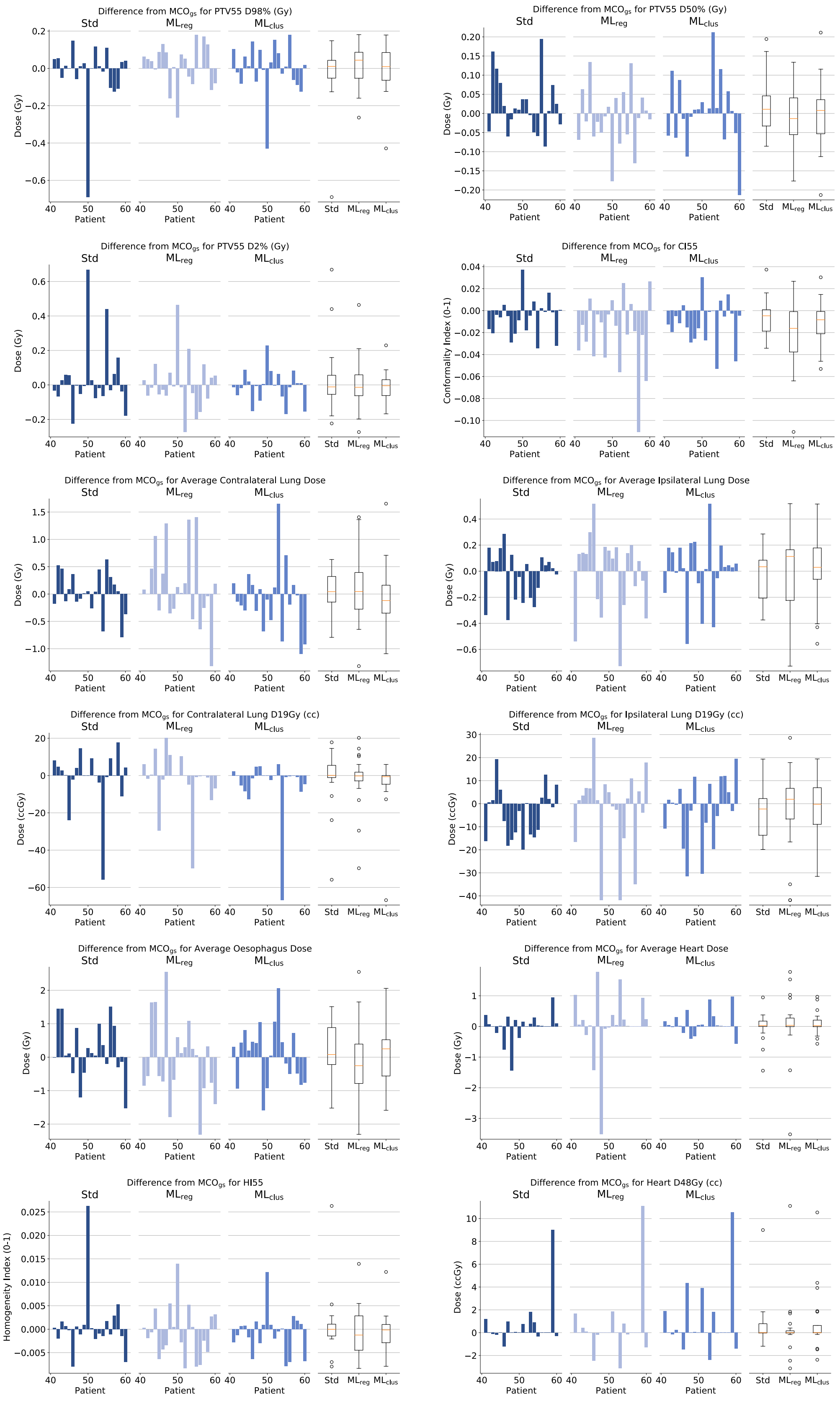


Figure 6.10: Plots for Lung showing absolute difference of Std, ML_{reg} and ML_{clus} from MCO_{gs}. Distributions are across the validation dataset and show key dose related metrics for each of the three calibration techniques.

ROI Name	DVH Statistic Type	MCO_{gs}	Std	ML_{reg}	ML_{clus}
PTV55	$D_{98\%}$ (Gy)	53.4 ± 0.179	53.4 ± 0.276	53.4 ± 0.197	53.4 ± 0.202
	$D_{50\%}$ (Gy)	55.3 ± 0.274	55.4 ± 0.280	55.3 ± 0.241	55.4 ± 0.266
	$D_{2\%}$ (Gy)	56.8 ± 0.242	56.9 ± 0.372	56.9 ± 0.304	56.8 ± 0.268
	CI	0.809 ± 0.0281	0.802 ± 0.0307	0.789 ± 0.043	0.798 ± 0.0339
	HI	0.0703 ± 0.00829	0.0711 ± 0.0121	0.0696 ± 0.00931	0.0695 ± 0.00925
Heart	D_{mean}(Gy)	8.12 ± 8.00	8.12 ± 8.04	8.18 ± 8.31	8.22 ± 8.15
	$D_{48\text{Gy}}$ (ccGy)	8.32 ± 13.9	8.95 ± 15.1	8.76 ± 14.9	9.2 ± 15.2
Ipsilateral Lung	D_{mean}(Gy)	17.4 ± 6.29	17.4 ± 6.19	17.4 ± 6.19	17.4 ± 6.25
	$D_{19\text{Gy}}$ (ccGy)	477 ± 157	473 ± 157	474 ± 163	474 ± 153
Contralateral Lung	D_{mean}(Gy)	3.91 ± 2.18	3.94 ± 1.98	4.06 ± 2.23	3.82 ± 1.93
	$D_{19\text{Gy}}$ (ccGy)	30.2 ± 77.3	29.1 ± 64.9	27.8 ± 67.6	25.5 ± 62.5
Oesophagus	D_{mean}(Gy)	16.7 ± 10.9	16.9 ± 10.8	16.6 ± 11.2	16.8 ± 11.1

Table 6.14: Summary of key dose metrics for Lung. Values shown are Mean ± 1 Standard Deviation. Statistical difference at the 95% level of significance is indicated by boldface.

alternative methods. Not only did clustering present a feasible solution in general, patient level analysis showed deviations from the MCO_{gs} to be consistently small given tight distribution in comparison to alternative methods especially the regression technique. Clustering and standard approach in this work were found to have high degrees of congruence in performance for both weighs and dosimetry. This is likely linked to the methodology for ML_{clus} weighting factors selection. The method involved taking means across the training database of clustered patient weighting factors and can be thought of as an enhanced version of the standard approach.

Regression methods consistently showed poorer performance against other methods. The average deviations of ML_{reg} from MCO_{gs} were larger than other methods in majority of cases the distribution of values about MCO_{gs} indicated poor congruence with MCO_{gs} weighting factors in general. The reason for the performance of each metho this will be more thoroughly assessed in the following chapter.

Chapter 7

Discussion, future work and conclusions

7.1 Discussion

7.1.1 Hypothesis generation

The overall aim the work presented in this thesis was to develop a proof-of-concept that a fully automated planning system delivering patient-tailored plans is feasible. Ensuring clinical preference is achieved is a key challenge of automated planning today and with RBP systems such as the in-house built PBAIO system used in work, the most significant manual stage is achieving clinically acceptable calibration of the PGs. In Chapter 4 an intra- and inter-planner study were conducted using a novel and intuitive PGAP approach within the planning system. It had been hypothesised that incorporation of MCO techniques within the planning process is expected to have a positive association with clinical preference[67,105] hence MCO techniques were explored for efficacy and for the feasibility of defining gold standard planning.

7.1.1.1 Intra-planner study

Regarding calibration of the AP system using PGAP to generate patient-tailored plans, findings of the intra-planning study suggest experts perform comparably between sessions. This was a positive finding for a few reasons. Firstly, given experts use their experience and knowledge to make judgements during calibration, it implies the desirable consistency observed during AP calibration is applied by experts in other methods of planning such as standard manual planning. More notably, this is a positive finding in this work because it implies gold standards planning (as defined by an individual planner) is a strong benchmark comparison to ML performance. In addition, given the robust procedure and rich data capture within the intra-planner study, there is reason to believe the findings can be generalised.

However, one of the main limitations of this study is it contains only a single participant. Evidence of intra-planner choices illustrating this degree of consistency across multiple participants has not been collected. Therefore, defining gold standard using calibration choices of a single participant is only acceptable for the participant in question and not others. This limitation has been mitigated by doing just that. All gold standard planning considered in the main studies presented in Chapter 6 were defined using the calibration choices of a single qualified expert, the participant of this intra-planner study.

7.1.1.2 Inter-planner study

The inter-planner study was novel due to the calibration method of the AP system and served to fill a gap in the literature regarding rules-based AP. Nevertheless, the procedure, number of participants and data captured is comparable to that of other inter-planner work in this field[168]. Therefore, a strength was its ability to be compared to other studies as it has scientific validity. Moreover, it was also important for helping identify differences in planning choices between different expert planning professionals as although different professionals are qualified to create and validate plans in the clinic, differences in calibration choices for this AP method were previously unknown. The study not only indicates difference in choice but also similarities and some of the reasoning behind certain choices. For example, one of the oncologists expressed a preference for sparing the rectum hence compromising on other PGs and leading to a detriment for these PGs in comparison. These findings aid in building an understanding of what constitutes ‘clinical preference’ and can be built up in future work. This will be discussed in more detail in section 7.2.

However, the degree of difference in planning choices between participants indicate clinical preference is nuanced and the study did not aid in the definition of “gold standard” in thesis. This work could be improved in a number of ways. For example, a small number of participants were considered and two of them were oncologists. A larger group of participants from a wider range of backgrounds will have enriched the findings such as participants from a wider range of tenures and institutions. A small number of patient cases were also considered and this may have implications regarding the representation of cases typically seen in the clinic. Also, given participants performed the task only once per patient case, their intra-planner variability was not captured. This would have enabled inference about inter- and intra-planner variability in one study. It would have lead to more insights about planning choices among qualified professionals, insights around definition of clinical preference among different groups of professionals and the exploration of interacting factors such as background and tenure on planning choices. These are limiting factors to the data obtained when it comes to making statistical generalisation from the findings and all of these limitations relate to the finite capacity of such professionals to take part in research studies. Nevertheless, the studies were still scientifically robust in methodology and the data collected. The study also showed the AP approach was robust enough to mitigate differences observed during calibration as it generated plans that were dosimetrically comparable.

7.1.1.3 Anatomy simulation study

The objective of the anatomy simulation study was to explore planning parameters and the changes necessary to achieve comparable dose distributions even when anatomy varies greatly between cases. This was done using a single patient's original anatomy, augmenting it, and establishing the change in the planning parameter necessary to achieve a dose distribution comparable to that of the original gold standard plan. There are many benefits to such an approach. Most notably, because it is a computationally cheap in silico method, it can be used to generated a range of data points for analysis very quickly. Generation of the data enabled a rudimentary view of what constitutes consistent planning between patients of varying anatomy. Another strength of this study is the ability to augment anatomy beyond physical constraints. In reality there are limitations to the size a human rectum can be for example. Being able to produce ROI volumes exceeding these limitations has mathematical benefits for prediction that should hold true even when the variance observed between cases is less exaggerated.

However, this study has a key weakness in not being clinically relevant. That is, anatomy in each case was contrived and not representative of true cases. Although there are mathematical benefits of this, results must be validated with true cases. To overcome this limitation, the findings of this study were used for generation of hypotheses and heuristics only. This study considered only a small number of predictive features but these features and those found in the literature were used to identify which kinds of spatial, volumetric and derived variables would be considered in the main study. The full list of variables considered in the main study (FeatureDS1 variables) can be found in Appendix C.

Furthermore, there were limitations due to the generation of the data. ROI volumes were augmented in a structured way such that the new volume was a predefined ratio of the original. This was done to ensure there was a variance of volumes to considered when modelling. However, some augmentation types resulted in a greater number of data points being captured than for others. This was due to certain physiological constraints being applied during augmentation to ensure the augmented scenario did not contravene that of a true scenario. For example, in Figure 4.12 where rectum augmentations are seen, models appear highly influenced by a single volume. The largest superior-inferior-left-right expansions in this case is the only volume of that size in the dataset and given the constraints on the creation of new volumes (e.g. cannot overlap the External volume), no other volumes of this size could be created for comparison. This resulted in a single data point seemingly having a large influence on the final model and potentially biases the true strength of the relationship that is reported in the study.

Related to this, a further issue with the approach was the study was only carried out for one ROI at a time when in reality, patients differ based on many anatomical attributes. Not augmenting more than one ROI at a time may have lead to misleading models or results that do not take into consideration the interaction of features related to more than one change occurring at once. These issues are confounded by the fact only a single patients anatomy was considered. Considering more than a single patients anatomy will have enriched the data and produced more holistic models with a better view of which variables are the most predictive of parameter changes. These developments to this work will be discussed in more detail in section 7.2.

7.1.2 Model definition

ML was a natural choice given a broad scope of possible approaches could be considered. ML models were built using numerical anatomical features hence there was a reliance on a dataset of geometric information derived from delineated patient anatomy. The chosen dataset came from a single planner in each case. This follows on from the insights gained from Chapter 4 where a single qualified professional was found to produce consistent planning both in terms of weighting factors and dose metrics. A criticism of this approach is models are planner dependent and reflective of only a single persons judgement. However, a justification for this approach is that this planner was qualified and familiar with these treatment sites and consistent planning was evident in previous planning. From a ML perspective, this stability in the gold standard data is most suitable to modelling.

More than one ML method was considered for comparison in this work. This is a benefit to the work given different models are better suited to some data and not others and the underlying nature of this data had not previously been explored in this regard. Nevertheless, exploring a third ML model would have been advantageous. It would have not only provided another comparison, but may also have allowed for the nature of the data to be better understood. For example, the clustering methods chosen, K-means, is well defined for data that are linearly separable. However, the linear separability of this data is not known. Exploration of a fuzzy clustering approach has merit in this regard as there is no assumption of linear separability. For this reason it would allow definition of planning parameters based on an aggregation of cluster values calculated on the degree of proximity each has to cluster centroids.

Introducing ML upstream of the PBAIO AP algorithm served in the development of a hybrid KBP-RBP planning approach, a notable development in the area of AP. However, a particular advantage of this work is that unlike traditional planning studies, generated models were compared to gold standard PGAP planning opposed to trial-and-error AP or manual planning[204,217–219].

ML techniques used were not new to radiotherapy planning. PCA[214], regression[214,215] and clustering[121] have all been used in KBP to make predictions based on anatomical features with notable success. This work builds upon this knowledge with a method to evaluate the performance of different model formations using a LOOCV decision framework, such that the optimal model for a given site is selected. This allows for an automatic and unbiased choice from among candidate model and removes the requirement for a homogeneous ML approach, which may not always be appropriate.

7.1.3 Modelling

Regarding multiple polynomial regressions, the maximum number of features was limited to five. In relation to Chapter 4, exhaustive search using more than one model fulfilled the need to explore true patient cases in more detail than the contrived case. However, this approach applies a restriction on the exhaustive search. Although the computing power and time to go beyond 5 features was available, the justification for not was to limit the probability of overfitting the model to the training data as well as to limit the complexity of resulting models and prevent it becoming a black box.

With regard to results, the largest variances in difference from MCO_{gs} for all studies was observed for ML_{reg} and this was in terms of both weighting factors and dosimetry. Std and ML_{clus} were more robust models especially for PSV, with small deviations from MCO_{gs} observed even for outlier patients. Given regression allows predictions to be extrapolated beyond the bounds defined by the training dataset, robustness of Std and ML_{clus} compared to ML_{reg} is thought to be due to Std and ML_{clus} prediction weights being bounded by the training data. For outlier patients ML_{reg} could therefore lead to predictions less consistent with the training data and lead to extreme outliers. Therefore, despite some of the promising results obtained during hypothesis testing, results of the larger study indicate regression methods may not be the best suited for routine clinical application.

Of the three methods, ML_{clus} was considered to have the best congruence with MCO_{gs} based on overall distributions, number of outliers and number of statistically significant differences observed. However, the superiority of ML_{clus} over the standard method was marginal and this is due to the relation between the methods. Results indicate that marginal improvements may be gained over a traditional “one size fits all” approach for patients who are anatomical outliers. The logic being, clustering isolates major deviations and assigns a weighing factor. Anatomical variance between patients may be due in general to patients that have large anatomical outliers such as those with a large bladder for example. This work therefore suggests ML may help to provide improvements over standard methods where larger anatomical variations would have otherwise caused the patient’s plan to be subpar under a standard “one size fits all” approach.

7.2 Future work

The studies presented in the thesis served the purpose of fulfilling the goals and objectives of the overall project. Nevertheless, they could each be developed to build on the work presented and serve as a basis for future work. First of all, repeating each of the studies will be beneficial for determination of replicability and hence how reliable the results of each study are. Secondly, the exclusion criteria could be slackened to include a wider range of patient cases such as those with non-standard areas of treatment avoidance.

But there are specifically a number of developments that could be made to preliminary studies. All of the preliminary studies were carried out for a single treatment site: prostate and seminal vesicles. An assumption has been made that findings can be generalised to other sites but this has not been explored in this work. There is also scope in all studies to consider a larger range of patient cases. This will ensure greater statistical power when making inferences and will lead to more reliable models particularly for the anatomy simulation study. Specifically regarding the intra-planner study, given a limitation of the study relates to the recruitment of a single participant, recruiting a range of planning professionals could be a development. The results of the study in this work suggest intra-planner variability is low but results may be heavily biased by the participant recruited and variability may be higher for planning professionals in general. Studying this will aid in the definition of the gold standard and illustrate how justifiable it is to use a single participant.

Regarding the inter-planner study, this could be developed in a number of ways. Firstly, the inclusion of more participants from a range of backgrounds and more patient cases. Secondly, repeat expose of participants to the same cases. This study could be designed to gauge a much larger and more holistic view of planning behaviour across planning professionals. This will not only enable a view of similarities and difference between individuals, but also within and between professions, tenures and institutions. Moreover, assuming each participant is required to perform the task more than once, intra-planner variability will be captured in the same study building on the richness of the data. Such a study would then present evidence of the range of planning practices across professions and provide a foundation for the definitions of a universal gold standard in planning.

The anatomy simulation study could also be developed in a number of ways. As mentioned, the augmentation type was a limitation due to the number of data points for certain volumes. Also, only one ROI was altered at a time. A new study could be designed to include a greater number of augmentation types in which more than on ROI is altered in a scenario as well as smaller incremental changes. This will provide more evidence of relationships between anatomy and planning parameters using this in silico method. Replicating this for more than one original patient case will determine whether this approach can be generalised across patients and the degree of difference in planning parameters that can be expected among patients.

To build on the modelling work, there could be a consideration of other features that may help improve versatility and modelling accuracy. For example, utilisation of neural network generated features may be promising and has been explored by other researchers[122,220]. Neural networks could be utilised to directly generate patient-specific AP protocols or used in a two step approach to generate dosimetric features (rather than anatomical features) from which PG weights are derived[221].

But also, in this study, PG weight predictions were considered individually with their own optimal model defined. This made performing regression and clustering straight forward and helped to identify anatomical features that are important when optimising a given trade-off. An alternative approach could be to use multi-output ML technique such as multi-output regression or deep learning to predict not only PG weights but relative PG weights. There is the potential that such an approach can be generalised because PG weights are strongly relative in plan optimisation.

And given favourable outcomes have been observed for clustering, deep diving into and exploring difference approaches may have value. Despite the ultimate dosimetric strength of the clustering technique used here, there is some evidence to suggest the method may be improvable. For example, silhouette scores were not always necessarily favourable given the the K-means approach and this may be because the data are better suited to techniques that do not rely on the assumption of linear separability. A study that explores only clustering techniques may yield some interesting findings, desirable results and lead to a greater understanding of the underlying nature of the data. A class of clustering techniques that may be of particular valuable are fuzzy clustering approaches. With a hard clustering technique like K-means, novel cases are assigned one of a finite number of values for each PG. With fuzzy clustering technique where data points may be assigned to more than one cluster, relative distance from centroids may enable assignation of a unique and more appropriate machine learned value leading to true patient-tailored planning.

One of the key challenges related to this work was actually generation of the Pareto plans to enable PGAP. The process is time consuming and resource intensive taking up bandwidth to generate and computing memory to store. A pivotal development to the this work would therefore not only relate to the development of a machine learned calibration solution, but also machine learned Pareto sets to calibrate over. This will save time, computing power and potentially computing memory making the process more efficient.

7.3 Final conclusions

The goal of this work was to develop a fully automated planning system for optimal patient-tailored planning. Using an existing rules-based AP system, the objectives were to train models to predict gold standard parameters and to determine dosimetric differences between plans generated via different methods. The two objectives were achieved successfully. Regarding the first objective, this thesis presents the successful incorporation of ML into an AP planning procedure to generate a fully automated hybrid RBP-KBP AP method.

Regarding the second objective, heuristics were gained regarding expert-driven planning parameters, dosimetry and anatomy to determine what constitutes gold standard planning. Although there is strong evidence the PBAIO method used is a robust and clinically applicable planning methodology with comparable dose distributions achieved via different calibration methods. Additionally, this work serves to supplement the body of knowledge regarding intra- and inter-planner behaviours.

Therefore the overall goal of developing a fully automated planning system for optimal patient-tailored planning has also been achieved. However, there is evidence the underlying AP method is fit for purpose as is. The hypothesis was that patient-tailored planning requires bespoke calibration. Although some dosimetric differences were observed between different calibration methods, differences were often small and statistically or clinically insignificant. Therefore, this work illustrates it may be possible to achieve incremental improvements using advanced calibration methods but the dosimetric benefits are arguably negligible.

Bibliography

[1] "Scientists end 13 year debate proving non-ionizing rf microwave effect causes cell phone radiation dna damage." https://www.rfsafe.com/scientists-end-13-year-debate-proving-non-ionizing-rf-microwave-effect-
Accessed: 2022-12-01.

[2] "The cell cycle." https://teachmephysiology.com/biochemistry/cell-growth-death/cell-cycle/. Accessed: 2023-08-01.

[3] I. J. Chetty, B. Curran, J. E. Cygler, J. J. DeMarco, G. Ezzell, B. A. Faddegon, I. Kawrakow, P. J. Keall, H. Liu, C.-M. C. Ma, *et al.*, "Report of the aapm task group no. 105: Issues associated with clinical implementation of monte carlo-based photon and electron external beam treatment planning," *Medical Physics*, vol. 34, no. 12, pp. 4818–4853, 2007.

[4] R. K. Parajuli, M. Sakai, R. Parajuli, and M. Tashiro, "Development and applications of compton camera—a review," *Sensors*, vol. 22, no. 19, p. 7374, 2022.

[5] M. Reda, A. F. Bagley, H. Y. Zaidan, and W. Yantasee, "Augmenting the therapeutic window of radiotherapy: A perspective on molecularly targeted therapies and nanomaterials," *Radiotherapy and Oncology*, vol. 150, pp. 225–235, 2020.

[6] F. Carlsson, *Utilizing problem structure in optimization of radiation therapy*. PhD thesis, KTH, 2008.

[7] H. Rocha, J. Dias, B. Ferreira, and M. Lopes, "From fluence map optimization to fluence map delivery: the role of combinatorial optimization," tech. rep., Inescc Research Report 05/2011, ISSN: 1645–2631. Available at: www. inescc . . . , 2011.

[8] S. Webb, "The physical basis of imrt and inverse planning," *The British journal of radiology*, vol. 76, no. 910, pp. 678–689, 2003.

[9] "Volume modulated arc therapy (vmat)." https://www.mgcancerhospital.com/volume-modulated-arc-therapy-vmat/. Accessed: 2022-12-01.

[10] S. Breedveld, D. Craft, R. Van Haveren, and B. Heijmen, "Multi-criteria optimization and decision-making in radiotherapy," *European Journal of Operational Research*, vol. 277, no. 1, pp. 1–19, 2019.

[11] C. M. Rebello, M. A. Martins, D. D. Santana, A. E. Rodrigues, J. M. Loureiro, A. M. Ribeiro, and I. B. Nogueira, "From a pareto front to pareto regions: A novel standpoint for multiobjective optimization," *Mathematics*, vol. 9, no. 24, p. 3152, 2021.

[12] "Rapidplan knowledge-based planning: Harness the power of machine learning to improve treatment planning in eclipse™." https://www.varian.com/en-gb/products/radiotherapy/treatment-planning/rapidplan-knowledge-based-planning. Accessed: 2022-12-01.

[13] M. Baranwal and S. M. Salapaka, "Weighted kernel deterministic annealing: A maximum-entropy principle approach for shape clustering," in *2018 Indian Control Conference (ICC)*, pp. 1–6, IEEE, 2018.

[14] "Cancer statistics for the uk," *Cancer Resarch UK*, 2023.

[15] "Cancer treatment statistics," *Cancer Resarch UK*, 2023.

[16] J. Fraser, E. Hope, J. Anderson, W. Verstraete, R. Sandhu, and S. McPhai, "Chemotherapy, radiotherapy and surgical tumour resections in england," *Public Health England Official Statistics*, 2020.

[17] "Facts about radiotherapy," *Royal Free London NHS Foundation Trust*, 2023.

[18] P. Mayles, A. Nahum, and J.-C. Rosenwald, *Handbook of radiotherapy physics: theory and practice*. CRC Press, 2007.

[19] G. Borrego-Soto, R. Ortiz-López, and A. Rojas-Martínez, "Ionizing radiation-induced dna injury and damage detection in patients with breast cancer," *Genetics and Molecular Biology*, vol. 38, no. 4, pp. 420–32, 2015.

[20] M. Beyzadeoglu, G. Ozyigit, and C. Ebruli, *Basic radiation oncology*, vol. 71. Springer, 2010.

[21] G. G. Steel, T. J. McMillan, and J. Peacock, "The 5rs of radiobiology," *International journal of radiation biology*, vol. 56, no. 6, pp. 1045–1048, 1989.

[22] C. M. Yashar, "Basic principles in gynecologic radiotherapy," in *Clinical Gynecologic Oncology*, pp. 586–605, Elsevier, 2018.

[23] N. Chatterjee and G. C. Walker, "Mechanisms of dna damage, repair, and mutagenesis," *Environmental and molecular mutagenesis*, vol. 58, no. 5, pp. 235–263, 2017.

[24] R. D. Kennedy and A. D. D'Andrea, "Dna repair pathways in clinical practice: lessons from pediatric cancer susceptibility syndromes," *Journal of Clinical Oncology*, vol. 24, no. 23, pp. 3799–3808, 2006.

[25] R. Tarnawski, J. Fowler, K. Skladowski, A. Świerniak, R. Suwiński, B. Maciejewski, and A. Wygoda, "How fast is repopulation of tumor cells during the treatment gap?," *International Journal of Radiation Oncology* Biology* Physics*, vol. 54, no. 1, pp. 229–236, 2002.

[26] J. Yang, J.-B. Yue, J. Liu, and J.-M. Yu, "Repopulation of tumor cells during fractionated radiotherapy and detection methods," *Oncology letters*, vol. 7, no. 6, pp. 1755–1760, 2014.

[27] T. Aruga, K. Ando, M. Iizuka, S. Koike, K. Fukutsu, H. Itsukaichi, and N. Arimizu, "Radiosensitivity and cell cycle redistribution of cultured human tumour cells during fractionated daily 2-gy irradiations," *International journal of radiation biology*, vol. 67, no. 1, pp. 65–70, 1995.

[28] B. S. Bhutta, F. Alghoula, and I. Berim, "Hypoxia," in *StatPearls [Internet]*, StatPearls Publishing, 2022.

[29] W. Boulefour, E. Rowinski, S. Louati, S. Sotton, A.-S. Wozny, P. Moreno-Acosta, B. Mery, C. Rodriguez-Lafrasse, and N. Magne, "A review of the role of hypoxia in radioresistance in cancer therapy," *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, vol. 27, pp. e934116–1, 2021.

[30] S. Laskar, S. Kakoti, N. Khanna, J. J. Manjali, A. Mangaj, A. Puri, A. Gulia, P. Nayak, P. Pai, D. Nair, *et al.*, "Outcomes of osteosarcoma, chondrosarcoma and chordoma treated with image guided-intensity modulated radiation therapy," *Radiotherapy and Oncology*, vol. 164, pp. 216–222, 2021.

[31] M. Levis, A. R. Filippi, C. Fiandra, V. De Luca, S. Bartoncini, D. Vella, R. Ragona, and U. Ricardi, "Inclusion of heart substructures in the optimization process of volumetric modulated arc therapy techniques may reduce the risk of heart disease in hodgkin's lymphoma patients," *Radiotherapy and Oncology*, vol. 138, pp. 52–58, 2019.

[32] D. Rades, A. J. Conde-Moreno, J. Cacicedo, B. Segedin, V. Rudat, and S. E. Schild, "Excellent outcomes after radiotherapy alone for malignant spinal cord compression from myeloma," *Radiology and oncology*, vol. 50, no. 3, p. 337, 2016.

[33] A. Brahme, *Comprehensive Biomedical Physics*. Newnes, 2014.

[34] J. G. Bazan, Q.-T. Le, and D. Zips, "Radiobiology of lung cancer," in *IASLC Thoracic Oncology*, pp. 330–336, Elsevier, 2018.

[35] T. M. Pawlik and K. Keyomarsi, "Role of cell cycle in mediating sensitivity to radiotherapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 59, no. 4, pp. 928–942, 2004.

[36] C. de Groot, J. C. Beukema, J. A. Langendijk, H. P. van der Laan, P. van Luijk, J. P. van Melle, C. T. Muijs, and N. H. Prakken, "Radiation-induced myocardial fibrosis in long-term esophageal cancer survivors," *International Journal of Radiation Oncology, Biology, Physics*, 2021.

[37] H. P. van der Laan, L. Van den Bosch, E. Schuit, R. J. Steenbakkers, A. van der Schaaf, and J. A. Langendijk, "Impact of radiation-induced toxicities on quality of life of patients treated for head and neck cancer," *Radiotherapy and Oncology*, vol. 160, pp. 47–53, 2021.

[38] A. Yaney, A. S. Ayan, X. Pan, S. Jhawar, E. Healy, S. Beyer, K. Lindsey, K. Kuhn, K. Tedrick, J. R. White, *et al.*, "Dosimetric parameters associated with radiation-induced esophagitis in breast cancer patients undergoing regional nodal irradiation," *Radiotherapy and Oncology*, vol. 155, pp. 167–173, 2021.

[39] K. S. Gleisner, E. Spezi, P. Solny, P. M. Gabina, F. Cicone, C. Stokke, C. Chiesa, M. Paphiti, B. Brans, M. Sandström, *et al.*, "Variations in the practice of molecular radiotherapy and implementation of dosimetry: results from a european survey," *EJNMMI physics*, vol. 4, no. 1, p. 28, 2017.

[40] J. Skowronek, "Current status of brachytherapy in cancer treatment–short overview," *Journal of contemporary brachytherapy*, vol. 9, no. 6, p. 581, 2017.

[41] M. Gaze and G. Flux, "Molecular radiotherapy—the radionuclide raffle?," *The British journal of radiology*, vol. 83, no. 996, pp. 995–7, 2010.

[42] E. B. P. Hoskin, *External beam therapy (radiotherapy in practice)*. Oxford University Press, 2012.

[43] E. Bakiu, E. Telhaj, E. Kozma, F. Ruçi, and P. Malkaj, "Comparison of 3d crt and imrt tratment plans," *Acta Informatica Medica*, vol. 21, no. 3, p. 211, 2013.

[44] T. de la Fuente Herman, E. Schnell, J. Young, K. Hildebrand, Ö. Algan, E. Syzek, T. Herman, and S. Ahmad, "Dosimetric comparison between imrt delivery modes: Step-and-shoot, sliding window, and volumetric modulated arc therapy—for whole pelvis radiation therapy of intermediate-to-high risk prostate adenocarcinoma," *Journal of Medical Physics/Association of Medical Physicists of India*, vol. 38, no. 4, p. 165, 2013.

[45] E. Schnell, T. de la Fuente Herman, J. Young, K. Hildebrand, O. Algan, E. Syzek, T. Herman, and S. Ahmad, "Dosimetric comparison of volumetric modulated arc therapy, step-and-shoot, and sliding window imrt for prostate cancer," in *AIP Conference Proceedings*, vol. 1494, pp. 23–26, American Institute of Physics, 2012.

[46] E. M. Quan, X. Li, Y. Li, X. Wang, R. J. Kudchadker, J. L. Johnson, D. A. Kuban, A. K. Lee, and X. Zhang, "A comprehensive comparison of imrt and vmat plan quality for prostate cancer treatment," *International Journal of Radiation Oncology, Biology, Physics*, vol. 83, no. 4, pp. 1169–78, 2012.

[47] K. Gomarteli, J. Fleckenstein, M. Meyer, T. Henzler, S. Kirschner, B. Kraenzlin, M. Brockmann, G. Welzel, G. Glatting, F. Wenz, *et al.*, "Focus on the low-dose bath: no increased cancer risk after mediastinal vmat versus ap/pa irradiation in a tumor-prone rat model," *International journal of radiation oncology, biology, physics*, vol. 99, no. 2, pp. S76–S77, 2017.

[48] K. Gomarteli, J. Fleckenstein, S. Kirschner, V. Bobu, M. A. Brockmann, T. Henzler, M. Meyer, P. Riffel, S. O. Schönberg, M. R. Veldwijk, *et al.*, "Radiation-induced malignancies after intensity-modulated versus conventional mediastinal radiotherapy in a small animal model," *Scientific reports*, vol. 9, no. 1, pp. 1–10, 2019.

[49] M. Treutwein, F. Steger, R. Loeschel, O. Koelbl, and B. Dobler, "The influence of radiotherapy techniques on the plan quality and on the risk of secondary tumors in patients with pituitary adenoma," *BMC cancer*, vol. 20, no. 1, p. 88, 2020.

[50] G. Wang, H. Wang, H. Zhuang, and R. Yang, "An investigation of non-coplanar volumetric modulated radiation therapy for locally advanced unresectable pancreatic cancer using a trajectory optimization method," *Frontiers in Oncology*, p. 3653, 2021.

[51] W. Segars, J. Bond, J. Frush, S. Hon, C. Eckersley, C. H. Williams, J. Feng, D. J. Tward, J. Ratnanather, M. Miller, *et al.*, "Population of anatomically variable 4d xcat adult phantoms for imaging research and optimization," *Medical Physics*, vol. 40, no. 4, p. 043701, 2013.

[52] T. Landberg, J. Chavaudra, J. Dobbs, J.-P. Gerard, G. Hanks, J.-C. Horiot, K.-A. Johansson, T. Möller, J. Purdy, N. Suntharalingam, *et al.*, "2. volumes," *Reports of the International Commission on Radiation Units and Measurements*, no. 1, pp. 3–20, 1999.

[53] F. M. Khan, *The physics of radiation therapy*. Lippincott Williams & Wilkins, 2010.

[54] J. H. Kim, "Three principles for radiation safety: time, distance, and shielding," *The Korean journal of pain*, vol. 31, no. 3, pp. 145–146, 2018.

[55] T. Speer, C. Knowlton, M. Mackay, C. Ma, L. Wang, L. Daugherty, B. Fisher, J. Wong, B. Hasson, D. Michalski, *et al.*, "Dose calculation algorithms," *Encyclopedia of Radiation Oncology; Springer: Berlin/Heidelberg, Germany*, pp. 158–166, 2013.

[56] D. Craft, "Multi-criteria optimization methods in radiation therapy planning: a review of technologies and directions," *arXiv preprint arXiv:1305.1546*, 2013.

[57] J. K. Selvaraj, *Modelling the effect of geometric uncertainties, clonogen distribution and IMRT interplay effect on tumour control probability*. PhD thesis, University of Liverpool, 2013.

[58] F. Tommasino, A. Nahum, and L. Cella, "Increasing the power of tumour control and normal tissue complication probability modelling in radiotherapy: recent trends and current issues," *Transl Cancer Res*, vol. 6, no. S5, pp. S807–21, 2017.

[59] R. Nuraini and R. Widita, "Tumor control probability (tcp) and normal tissue complication probability (ntcp) with consideration of cell biological effect," in *Journal of Physics: Conference Series*, vol. 1245, p. 012092, IOP Publishing, 2019.

[60] O. Hamming-Vrieze, N. Depauw, D. Craft, A. Chan, C. Rasch, M. Verheij, J. Sonke, and H. Kooy, "Impact of setup and range uncertainties on tcp and ntcp following vmat or impt of oropharyngeal cancer patients," *Physics in Medicine & Biology*, vol. 64, no. 9, p. 095001, 2019.

[61] W. D. Bidgood Jr, S. C. Horii, F. W. Prior, and D. E. Van Syckle, "Understanding and using dicom, the data interchange standard for biomedical imaging," *Journal of the American Medical Informatics Association*, vol. 4, no. 3, pp. 199–212, 1997.

[62] D. M. Aleman, "Fluence map optimization in intensity-modulated radiation therapy treatment planning," *Decision Analytics and Optimization in Disease Prevention and Treatment*, pp. p–285, 2018.

[63] R. Laboratories, "Vmat optimisation in raystation," Tech. Rep. 2017-04-20, RaySearch Laboratories AB, P.O. Box 3297, SE-103 65 Stockholm, Sweden, Apr 2017.

[64] S. Webb, *Handbook of radiotherapy physics: theory and practice*, ch. 43 Conformal and Intensity-Modulated Radiotherapy, pp. 943–971. CRC Press, 2007.

[65] R. Bokrantz, *Multicriteria optimization for managing tradeoffs in radiation therapy treatment planning*. PhD thesis, KTH Royal Institute of Technology, 2013.

[66] S. Breedveld, D. Craft, R. Van Haveren, and B. Heijmen, "Multi-criteria optimisation and decision-making in radiotherapy," *European Journal of Operational Research*, vol. 277, pp. 1–19, 08 2019.

[67] D. L. Craft, T. S. Hong, H. A. Shih, and T. R. Bortfeld, "Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 82, no. 1, pp. e83–e90, 2012.

[68] M. A. Hunt, C.-Y. Hsiung, S. V. Spirou, C.-S. Chui, H. I. Amols, and C. C. Ling, "Evaluation of concave dose distributions created using an inverse planning system," *International Journal of Radiation Oncology, Biology, Physics*, vol. 54, no. 3, pp. 953–962, 2002.

[69] Y. Ge and Q. J. Wu, "Knowledge-based planning for intensity-modulated radiation therapy: a review of data-driven approaches," *Medical Physics*, vol. 46, no. 6, pp. 2760–2775, 2019.

[70] S. Petrovic, G. Khussainova, and R. Jagannathan, "Knowledge-light adaptation approaches in case-based reasoning for radiotherapy treatment planning," *Artificial intelligence in medicine*, vol. 68, pp. 17–28, 2016.

[71] C. McIntosh and T. G. Purdie, "Contextual atlas regression forests: multiple-atlas-based automated dose prediction in radiation therapy," *IEEE transactions on medical imaging*, vol. 35, no. 4, pp. 1000–1012, 2015.

[72] Y. Sheng, T. Li, Y. Zhang, W. R. Lee, F.-F. Yin, Y. Ge, and Q. J. Wu, "Atlas-guided prostate intensity modulated radiation therapy (imrt) planning," *Physics in Medicine & Biology*, vol. 60, no. 18, p. 7277, 2015.

[73] V. Chanyavanich, S. K. Das, W. R. Lee, and J. Y. Lo, "Knowledge-based imrt treatment planning for prostate cancer," *Medical Physics*, vol. 38, no. 5, pp. 2515–2522, 2011.

[74] C. McIntosh and T. G. Purdie, "Voxel-based dose prediction with multi-patient atlas selection for automated radiotherapy treatment planning," *Physics in Medicine & Biology*, vol. 62, no. 2, p. 415, 2016.

[75] C. McIntosh, M. Welch, A. McNiven, D. A. Jaffray, and T. G. Purdie, "Fully automated treatment planning for head and neck radiotherapy using a voxel-based dose prediction and dose mimicking method," *Physics in Medicine & Biology*, vol. 62, no. 15, p. 5926, 2017.

[76] X. Bai, B. Wang, S. Wang, Z. Wu, C. Gou, and Q. Hou, "Radiotherapy dose distribution prediction for breast cancer using deformable image registration," *BioMedical Engineering OnLine*, vol. 19, no. 1, pp. 1–20, 2020.

[77] D. Franceschini, L. Cozzi, A. Fogliata, B. Marini, L. Di Cristina, L. Dominici, R. Spoto, C. Franzese, P. Navarria, T. Comito, *et al.*, "Training and validation of a knowledge-based dose-volume histogram predictive model in the optimisation of intensity-modulated proton and volumetric modulated arc photon plans for pleural mesothelioma patients," *Radiation Oncology*, vol. 17, no. 1, pp. 1–10, 2022.

[78] A. Tudda, R. Castriconi, G. Benecchi, E. Cagni, A. Cicchetti, F. Dusi, P. G. Esposito, M. Guernieri, A. Ianiro, V. Landoni, *et al.*, "Knowledge-based multi-institution plan prediction of whole breast irradiation with tangential fields," *Radiotherapy and Oncology*, vol. 175, pp. 10–16, 2022.

[79] H. Liu, R. Clark, A. Magliari, R. Foster, F. Reynoso, M. Schmidt, V. Gondi, C. Abraham, H. Curry, P. Kupelian, *et al.*, "Rapidplan hippocampal sparing whole brain model version 2—how far can we reduce the dose?," *Medical Dosimetry*, 2022.

[80] R. J. Douglas, A. Olanrewaju, L. Zhang, B. M. Beadle, and L. E. Court, "Assessing the practicality of using a single knowledge-based planning model for multiple linac vendors," *Journal of Applied Clinical Medical Physics*, vol. 23, no. 8, p. e13704, 2022.

[81] V. A. Dumane, T.-C. Tseng, R.-D. Sheu, Y.-C. Lo, V. Gupta, A. Saitta, K. E. Rosenzweig, and S. Green, "Training and evaluation of a knowledge-based model for automated treatment planning of multiple brain metastases," *Journal of Cancer Metastasis and Treatment*, vol. 5, 2019.

[82] A. Fogliata, F. Belosi, A. Clivio, P. Navarria, G. Nicolini, M. Scorsetti, E. Vanetti, and L. Cozzi, "On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer," *Radiotherapy and Oncology*, vol. 113, no. 3, pp. 385–391, 2014.

[83] J. S. Munter and J. Sjölund, "Dose-volume histogram prediction using density estimation," *Physics in Medicine & Biology*, vol. 60, no. 17, p. 6923, 2015.

[84] S. F. Petit and W. van Elmpt, “Accurate prediction of target dose-escalation and organ-at-risk dose levels for non-small cell lung cancer patients,” *Radiotherapy and Oncology*, vol. 117, no. 3, pp. 453–458, 2015.

[85] S. Shiraishi and K. L. Moore, “Knowledge-based prediction of three-dimensional dose distributions for external beam radiotherapy,” *Medical Physics*, vol. 43, no. 1, pp. 378–387, 2016.

[86] A. M. Barragán-Montero, D. Nguyen, W. Lu, M.-H. Lin, R. Norouzi-Kandalan, X. Geets, E. Sterpin, and S. Jiang, “Three-dimensional dose prediction for lung imrt patients with deep neural networks: robust learning from heterogeneous beam configurations,” *Medical Physics*, vol. 46, no. 8, pp. 3679–3691, 2019.

[87] W. T. Hrinivich and J. Lee, “Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning,” *Medical Physics*, vol. 47, no. 12, pp. 6140–6150, 2020.

[88] K.-W. Jee, D. L. McShan, and B. A. Fraass, “Lexicographic ordering: intuitive multicriteria optimization for imrt,” *Physics in Medicine & Biology*, vol. 52, no. 7, p. 1845, 2007.

[89] A. C. Spalding, K.-W. Jee, K. Vineberg, M. Jablonowski, B. A. Fraass, C. C. Pan, T. S. Lawrence, R. K. Ten Haken, and E. Ben-Josef, “Potential for dose-escalation and reduction of risk in pancreatic cancer using imrt optimization with lexicographic ordering and geud-based cost functions,” *Medical Physics*, vol. 34, no. 2, pp. 521–529, 2007.

[90] S. Breedveld, P. R. Storchi, P. W. Voet, and B. J. Heijmen, “icycle: Integrated, multicriterial beam angle, and profile optimization for generation of coplanar and noncoplanar imrt plans,” *Medical Physics*, vol. 39, no. 2, pp. 951–63, 2012.

[91] H. Wang and L. Xing, “Application programming in c# environment with recorded user software interactions and its application in autopilot of vmat/imrt treatment planning,” *Journal of Applied Clinical Medical Physics*, vol. 17, no. 6, pp. 189–203, 2016.

[92] R.-P. Li and F.-F. Yin, “Optimization of inverse treatment planning using a fuzzy weight function,” *Medical Physics*, vol. 27, no. 4, pp. 691–700, 2000.

[93] I. Xhaferllari, E. Wong, K. Bzdusek, M. Lock, and J. Z. Chen, "Automated imrt planning with regional optimization using planning scripts," *Journal of Applied Clinical Medical Physics*, vol. 14, no. 1, pp. 176–191, 2013.

[94] J. P. Tol, M. Dahele, J. Peltola, J. Nord, B. J. Slotman, and W. F. Verbakel, "Automatic interactive optimization for volumetric modulated arc therapy planning," *Radiation Oncology*, vol. 10, no. 1, pp. 1–12, 2015.

[95] D. Gintz, K. Latifi, J. Caudell, B. Nelms, G. Zhang, E. Moros, and V. Feygelman, "Initial evaluation of automated treatment planning software," *Journal of Applied Clinical Medical Physics*, vol. 17, no. 3, pp. 331–346, 2016.

[96] P. A. Wheeler, M. Chu, R. Holmes, M. Smyth, R. Maggs, E. Spezi, J. Staffurth, D. G. Lewis, and A. E. Millin, "Utilisation of pareto navigation techniques to calibrate a fully automated radiotherapy treatment planning solution," *Physics and Imaging in Radiation Oncology*, vol. 10, pp. 41–8, 2019.

[97] C. Cotrutz and L. Xing, "Imrt dose shaping with regionally variable penalty scheme," *Medical Physics*, vol. 30, no. 4, pp. 544–551, 2003.

[98] P. Wheeler, M. Chu, A. Mazurek, R. Maggs, R. Jadon, J. Staffurth, C. Hanna, T. Perrett, D. Lewis, and A. Millin, "Oc-0384: Development of fully optimised single iteration vmat class solutions and their clinical application," *Radiotherapy and Oncology*, no. 111, p. S149, 2014.

[99] P. W. Voet, M. L. Dirkx, S. Breedveld, A. Al-Mamgani, L. Incrocci, and B. J. Heijmen, "Fully automated volumetric modulated arc therapy plan generation for prostate cancer patients," *International Journal of Radiation Oncology, Biology, Physics*, vol. 88, no. 5, pp. 1175–1179, 2014.

[100] L. Marrazzo, I. Meattini, C. Arilli, S. Calusi, M. Casati, C. Talamonti, L. Livi, and S. Pallotta, "Auto-planning for vmat accelerated partial breast irradiation," *Radiotherapy and Oncology*, vol. 132, pp. 85–92, 2019.

[101] B. Wu, M. Kusters, M. Kunze-busch, T. Dijkema, T. McNutt, G. Sanguineti, and D. Pang, "Mo-g-201-01: A multi-institutional study investigating the performance of a knowledge-based planning system against pinnacle auto-planning engine in sib-imrt for the head-and-neck cancer," *Medical Physics*, vol. 43, no. 6Part32, pp. 3723–3724, 2016.

[102] R. Bijman, L. Rossi, A. W. Sharfo, W. Heemsbergen, L. Incrocci, S. Breedveld, and B. Heijmen, "Automated radiotherapy planning for patient-specific exploration of the trade-off between tumor dose coverage and predicted radiation-induced toxicity—a proof of principle study for prostate cancer," *Frontiers in Oncology*, vol. 10, p. 943, 2020.

[103] C. R. Hansen, A. Bertelsen, I. Hazell, R. Zukauskaite, N. Gyldenkerne, J. Johansen, J. G. Eriksen, and C. Brink, "Automatic treatment planning improves the clinical quality of head and neck cancer treatment plans," *Clinical and translational radiation oncology*, vol. 1, pp. 2–8, 2016.

[104] M. Cokelek, E. Holt, F. Kelly, A. Rolfo, M. Ng, B. Foley, S. Ryan, H. Ho, A. Brown, J. McAlpine, *et al.*, "Automation: The future of radiotherapy," *International Journal of Radiation Oncology, Biology, Physics*, vol. 108, no. 3, p. e314, 2020.

[105] P. A. Wheeler, M. Chu, R. Holmes, O. W. Woodley, C. S. Jones, R. Maggs, J. Staffurth, N. Palaniappan, E. Spezi, D. G. Lewis, *et al.*, "Evaluating the application of pareto navigation guided automated radiotherapy treatment planning to prostate cancer," *Radiotherapy and Oncology*, vol. 141, pp. 220–26, 2019.

[106] J. Krayenbuehl, M. Zamburlini, S. Ghandour, M. Pachoud, S. Tanadini-Lang, J. Tol, M. Guckenberger, and W. Verbakel, "Planning comparison of five automated treatment planning solutions for locally advanced head and neck cancer," *Radiation Oncology*, vol. 13, no. 1, p. 170, 2018.

[107] L. Lu, Y. Sheng, J. Donaghue, Z. Liu Shen, M. Kolar, Q. J. Wu, and P. Xia, "Three imrt advanced planning tools: A multi-institutional side-by-side comparison," *Journal of Applied Clinical Medical Physics*, vol. 20, no. 8, pp. 65–77, 2019.

[108] M. Hussein, B. J. Heijmen, D. Verellen, and A. Nisbet, "Automation in intensity modulated radiotherapy treatment planning—a review of recent innovations," *The British journal of radiology*, vol. 91, no. 1092, p. 20180270, 2018.

[109] L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu, "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in imrt plans," *Medical Physics*, vol. 39, no. 11, pp. 6868–78, 2012.

[110] E. Cagni, A. Botti, Y. Wang, M. Iori, S. F. Petit, and B. J. Heijmen, "Pareto-optimal plans as ground truth for validation of a commercial system for knowledge-based dvh-prediction," *Physica Medica*, vol. 55, pp. 98–106, 2018.

[111] A. Smith, A. Granatowicz, C. Stoltenberg, S. Wang, X. Liang, C. A. Enke, A. O. Wahl, S. Zhou, and D. Zheng, "Can the student outperform the master? a plan comparison between pinnacle auto-planning and eclipse knowledge-based rapidplan following a prostate-bed plan competition," *Technology in cancer research & treatment*, vol. 18, p. 1533033819851763, 2019.

[112] C. R. Hansen, W. Crijns, M. Hussein, L. Rossi, P. Gallego, W. Verbakel, J. Unkelbach, D. Thwaites, and B. Heijmen, "Radiotherapy treatment planning study guidelines (rating): A framework for setting up and reporting on scientific treatment planning studies," *Radiotherapy and Oncology*, vol. 153, pp. 67–78, 2020.

[113] L. Vandewinckele, M. Claessens, A. Dinkla, C. Brouwer, W. Crijns, D. Verellen, and W. van Elmpt, "Overview of artificial intelligence-based applications in radiotherapy: recommendations for implementation and quality assurance," *Radiotherapy and Oncology*, vol. 153, pp. 55–66, 2020.

[114] I. El Naqa, J. M. Boone, S. H. Benedict, M. M. Goodsitt, H.-P. Chan, K. Drukker, L. Hadjiiski, D. Ruan, and B. Sahiner, "Ai in medical physics: Guidelines for publication," 2021.

[115] B. Liang, H. Yan, Y. Tian, X. Chen, L. Yan, T. Zhang, Z. Zhou, L. Wang, and J. Dai, "Dosiomics: extracting 3d spatial features from dose distribution to predict incidence of radiation pneumonitis," *Frontiers in oncology*, vol. 9, p. 269, 2019.

[116] M. Ma, N. Kovalchuk, M. K. Buyyounouski, L. Xing, and Y. Yang, "Dosimetric features-driven machine learning model for dvh prediction in vmat treatment planning," *Medical Physics*, vol. 46, no. 2, pp. 857–867, 2019.

[117] J. Swamidas, S. Pradhan, S. Chopra, S. Panda, Y. Gupta, S. Sood, S. Mohanty, J. Jain, K. Joshi, R. Ph, L. Gurram, U. Mahantshetty, and J. P. Agarwal, "Development and clinical validation of knowledge-based planning for volumetric modulated arc therapy of cervical cancer including pelvic and para aortic fields," *Physics and Imaging in Radiation Oncology*, vol. 18, pp. 61–67, 2021.

[118] A. Babier, J. J. Boutilier, M. B. Sharpe, A. L. McNiven, and T. C. Chan, "Inverse optimization of objective function weights for treatment planning using clinical dose-volume histograms," *Physics in Medicine & Biology*, vol. 63, no. 10, p. 105004, 2018.

[119] A. Babier, B. Zhang, R. Mahmood, K. L. Moore, T. G. Purdie, A. L. McNiven, and T. C. Chan, "Openkbp: The open-access knowledge-based planning grand challenge and dataset," *Medical Physics*, vol. 48, no. 9, pp. 5549–5561, 2021.

[120] J. J. Boutilier, T. Lee, T. Craig, M. B. Sharpe, and T. C. Chan, "Models for predicting objective function weights in prostate cancer imrt," *Medical Physics*, vol. 42, no. 4, pp. 1586–95, 2015.

[121] A. Goli, J. J. Boutilier, T. Craig, M. B. Sharpe, and T. C. Chan, "A small number of objective function weight vectors is sufficient for automated treatment planning in prostate cancer," *Physics in Medicine & Biology*, vol. 63, no. 19, p. 195004, 2018.

[122] T. Zhang, R. Bokrantz, and J. Olsson, "Probabilistic feature extraction, dose statistic prediction and dose mimicking for automated radiation therapy treatment planning," *arXiv preprint arXiv:2102.12569*, 2021.

[123] Z. Liu, X. Chen, K. Men, J. Yi, and J. Dai, "A deep learning model to predict dose–volume histograms of organs at risk in radiotherapy treatment plans," *Medical Physics*, vol. 47, no. 11, pp. 5467–5481, 2020.

[124] D. Nguyen, R. McBeth, A. Sadeghnejad Barkousaraie, G. Bohara, C. Shen, X. Jia, and S. Jiang, "Incorporating human and learned domain knowledge into training deep neural networks: A differentiable dose-volume histogram and adversarial inspired framework for generating pareto optimal dose distributions in radiation therapy," *Medical Physics*, vol. 47, no. 3, pp. 837–49, 2020.

[125] H.-H. Tseng, Y. Luo, R. K. Ten Haken, and I. El Naqa, "The role of machine learning in knowledge-based response-adapted radiotherapy," *Frontiers in oncology*, vol. 8, p. 266, 2018.

[126] C. Ling, X. Han, P. Zhai, H. Xu, J. Chen, J. Wang, and W. Hu, "A hybrid automated treatment planning solution for esophageal cancer," *Radiation Oncology*, vol. 14, no. 1, pp. 1–7, 2019.

[127] C. Shen, D. Nguyen, L. Chen, Y. Gonzalez, R. McBeth, N. Qin, S. B. Jiang, and X. Jia, "Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning," *Medical Physics*, vol. 47, no. 6, pp. 2329–2336, 2020.

[128] J. Fan, J. Wang, Z. Zhang, and W. Hu, "Iterative dataset optimization in automated planning: Implementation for breast and rectal cancer radiotherapy," *Medical Physics*, vol. 44, no. 6, pp. 2515–31, 2017.

[129] M. Monz, K.-H. Küfer, T. R. Bortfeld, and C. Thieke, "Pareto navigation—algorithmic foundation of interactive multi-criteria imrt planning," *Physics in Medicine & Biology*, vol. 53, no. 4, p. 985, 2008.

[130] R. Bokrantz and K. Miettinen, "Projections onto the pareto surface in multicriteria radiation therapy optimization," *Medical Physics*, vol. 42, no. 10, pp. 5862–5870, 2015.

[131] R. G. Kierkels, R. Visser, H. P. Bijl, J. A. Langendijk, A. A. van't Veld, R. J. Steenbakkers, and E. W. Korevaar, "Multicriteria optimization enables less experienced planners to efficiently produce high quality treatment plans in head and neck cancer radiotherapy," *Radiation oncology*, vol. 10, no. 1, pp. 1–9, 2015.

[132] RaySearch Laboratories AB, "White paper: Multi-criteria optimization in raystation," tech. rep., RaySearch Laboratories, AB, P.O. Box 3297, SE-103 65 Stockholm, Sweden, April 2017.

[133] Varian Medical Systems, "Multi-criteria optimization: Creating high-quality treatment plans in a fraction of the time," tech. rep., Varian Medical Systems, Inc., 3100 Hansen Way, Palo Alto, CA 94304, United States, February 2019.

[134] A. Lindström, "Comparison of multi-criteria optimization in two different treatment planning systems mastere thesis," *University of Gothenburg, Institute of Clinical Science*, 2019.

[135] E. Miguel-Chumacero, G. Currie, A. Johnston, and S. Currie, "Effectiveness of multi-criteria optimization-based trade-off exploration in combination with rapidplan for head & neck radiotherapy planning," *Radiation Oncology*, vol. 13, no. 1, p. 229, 2018.

[136] K. Teichert, G. Currie, K.-H. Küfer, E. Miguel-Chumacero, P. Süss, M. Walczak, and S. Currie, “Targeted multi-criteria optimisation in imrt planning supplemented by knowledge based model creation,” *Operations Research for Health Care*, vol. 23, p. 100185, 2019.

[137] C. Huang, Y. Yang, N. Panjwani, S. Boyd, and L. Xing, “Pareto optimal projection search (pops): Automated radiation therapy treatment planning by direct search of the pareto surface,” *IEEE Transactions on Biomedical Engineering*, 2021.

[138] R. Van Haveren, B. J. Heijmen, and S. Breedveld, “Automatically configuring the reference point method for automated multi-objective treatment planning,” *Physics in Medicine & Biology*, vol. 64, no. 3, p. 035002, 2019.

[139] R. Bijman, A. W. Sharfo, L. Rossi, S. Breedveld, and B. Heijmen, “Pre-clinical validation of a novel system for fully-automated treatment planning,” *Radiotherapy and Oncology*, vol. 158, pp. 253–261, 2021.

[140] M. Uto, T. Mizowaki, K. Ogura, and M. Hiraoka, “Non-coplanar volumetric-modulated arc therapy (vmat) for craniopharyngiomas reduces radiation doses to the bilateral hippocampus: a planning study comparing dynamic conformal arc therapy, coplanar vmat, and non-coplanar vmat,” *Radiation Oncology*, vol. 11, no. 1, pp. 1–8, 2016.

[141] S.-T. Kim, H. J. An, J.-i. Kim, J.-R. Yoo, H. J. Kim, and J. M. Park, “Non-coplanar vmat plans for lung sabr to reduce dose to the heart: a planning study,” *The British Journal of Radiology*, vol. 93, no. 1105, p. 20190596, 2020.

[142] H. Kamal Sayed, M. Herman, and C. Beltran, “A pareto-based beam orientation optimization method for spot scanning intensity-modulated proton therapy,” *Medical Physics*, vol. 47, no. 5, pp. 2049–2060, 2020.

[143] “Cancer incidence for common cancers,” Dec 2020.

[144] D. Dearnaley, C. L. Griffin, R. Lewis, P. Mayles, H. Mayles, O. F. Naismith, V. Harris, C. D. Scrase, J. Staffurth, I. Syndikus, *et al.*, “Toxicity and patient-reported outcomes of a phase 2 randomized trial of prostate and pelvic lymph node versus prostate only radiotherapy in advanced localised prostate cancer (pivotal),” *International Journal of Radiation Oncology, Biology, Physics*, vol. 103, no. 3, pp. 605–617, 2019.

[145] P. Ziegenhein, C. P. Kamerling, and U. Oelfke, "Interactive dose shaping part 1: a new paradigm for imrt treatment planning," *Physics in Medicine & Biology*, vol. 61, no. 6, p. 2457, 2016.

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

[147] B. E. Huitema, "Single-participant research designs," *The Corsini Encyclopedia of Psychology*, pp. 1–3, 2010.

[148] T. Lee, M. Hammad, T. C. Chan, T. Craig, and M. B. Sharpe, "Predicting objective function weights from patient anatomy in prostate imrt treatment planning," *Medical Physics*, vol. 40, no. 12, p. 121706, 2013.

[149] C. Harrer, W. Ullrich, and J. J. Wilkens, "Prediction of multi-criteria optimization (mco) parameter efficiency in volumetric modulated arc therapy (vmat) treatment planning using machine learning (ml)," *Physica Medica*, vol. 81, pp. 102–113, 2021.

[150] D. Craft, T. Halabi, H. A. Shih, and T. Bortfeld, "An approach for practical multi-objective imrt treatment planning," *International Journal of Radiation Oncology, Biology, Physics*, vol. 69, no. 5, pp. 1600–1607, 2007.

[151] C. Collicott, E. Bonacker, I. Lammel, K. Teichert, M. Walzcak, and P. Süss, "Interactive navigation of multiple convex patches," *Journal of Multi-Criteria Decision Analysis*, vol. 28, no. 5-6, pp. 311–321, 2021.

[152] K. Chircop and D. Zammit-Mangion, "On-constraint based methods for the generation of pareto frontiers," *Journal of Mechanics Engineering and Automation*, vol. 3, no. 5, pp. 279–289, 2013.

[153] P. Mishra, C. M. Pandey, U. Singh, A. Gupta, C. Sahu, and A. Keshri, "Descriptive statistics and normality tests for statistical data," *Annals of cardiac anaesthesia*, vol. 22, no. 1, p. 67, 2019.

[154] J. W. Mauchly, "Significance test for sphericity of a normal n-variate distribution," *The Annals of Mathematical Statistics*, vol. 11, no. 2, pp. 204–209, 1940.

[155] S. W. Greenhouse and S. Geisser, "On methods in the analysis of profile data," *Psychometrika*, vol. 24, no. 2, pp. 95–112, 1959.

[156] O. J. Dunn, "Multiple comparisons among means," *Journal of the American statistical association*, vol. 56, no. 293, pp. 52–64, 1961.

[157] C. Tian, X. Manfei, T. Justin, W. Hongyue, and N. Xiaohui, "Relationship between omnibus and post-hoc tests: An investigation of performance of the f test in anova," *Shanghai archives of psychiatry*, vol. 30, no. 1, p. 60, 2018.

[158] S. Lewis, M. Chan, J. Weiss, H. Raziee, B. Driscoll, A. Bezjak, A. Sun, B. Lok, D. Vines, J. Cho, *et al.*, "3'-deoxy-3'-(18f) fluorothymidine positron emission tomography/computed tomography in non-small cell lung cancer treated with stereotactic body radiation therapy: A pilot study," *Advances in Radiation Oncology*, vol. 7, no. 6, p. 101037, 2022.

[159] Y. H. Zhang, E. Cha, K. Lynch, R. Gennarelli, J. Brower, M. V. Sherer, D. W. Golden, S. Chimonas, D. Korenstein, and E. F. Gillespie, "Attitudes and access to resources and strategies to improve quality of radiotherapy among us radiation oncologists: A mixed methods study," *Journal of Medical Imaging and Radiation Oncology*, 2022.

[160] B. Nelson, M. Lamba, S. Ewart, N. Ike, L. Lewis, and L. Pater, "Normal tissue exposure and second malignancy risk in vertebral-body-sparing craniospinal irradiation," *Medical Dosimetry*, vol. 47, no. 2, pp. 142–145, 2022.

[161] M. Friedman, "The use of ranks to avoid the assumption of normality implicit in the analysis of variance," *Journal of the American statistical association*, vol. 32, no. 200, pp. 675–701, 1937.

[162] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *The Journal of Machine learning research*, vol. 7, pp. 1–30, 2006.

[163] S. Irmak, L. Zimmermann, D. Georg, P. Kuess, and W. Lechner, "Cone beam ct based validation of neural network generated synthetic cts for radiotherapy in the head region," *Medical Physics*, vol. 48, no. 8, pp. 4560–4571, 2021.

[164] J. Litoborska, T. Piotrowski, and J. Malicki, "Evaluation of three vmat-tmi plan-

ning methods to find an appropriate balance between plan complexity and the resulting dose distribution," *Physica Medica*, vol. 75, pp. 26–32, 2020.

[165] I. Das, K. Cashon, K. Chopra, K. Khadivi, H. Malhotra, and C. Mayo, "We-e-t-617-06: Intra-and inter-planner dosimetric variations in inverse planning of imrt," *Medical Physics*, vol. 32, no. 6Part20, pp. 2147–2147, 2005.

[166] X. Zhang, X. Li, E. M. Quan, X. Pan, and Y. Li, "A methodology for automatic intensity-modulated radiation treatment planning for lung cancer," *Physics in Medicine & Biology*, vol. 56, no. 13, p. 3873, 2011.

[167] S. van Beek, A. Betgen, M. Buijs, J. Stam, L. Hartgring, B. van Triest, and P. Remeijer, "Pre-clinical experience of an adaptive plan library strategy in radiotherapy of rectal cancer: An inter-observer study," *Physics and imaging in radiation oncology*, vol. 6, pp. 89–93, 2018.

[168] E. Y. Erkal, A. Karabey, M. Sahin, A. Karayel, B. Tirpanci, and H. Erkal, "Intensity-modulated radiotherapy planning for prostate cancer: The evaluation of inter-observer variability and treatment delivery efficiency," *International Journal of Radiation Research*, vol. 20, no. 1, pp. 49–53, 2022.

[169] S. Srivastava, O. Nohadani, C. Medawar, C. Cheng, and I. Das, "Su-e-t-604: Inter planner dosimetric variations in imrt," *Medical Physics*, vol. 39, no. 6Part19, pp. 3845–3845, 2012.

[170] M. Esposito, G. Maggi, C. Marino, L. Bottalico, E. Cagni, C. Carbonini, M. Casale, S. Clemente, V. D'Alesio, D. Fedele, *et al.*, "Multicentre treatment planning inter-comparison in a national context: the liver stereotactic ablative radiotherapy case," *Physica Medica*, vol. 32, no. 1, pp. 277–283, 2016.

[171] J. Hurwitz and D. Kirsch, "Machine learning for dummies," *IBM Limited Edition*, vol. 75, 2018.

[172] M. P. Deisenroth, A. A. Faisal, and C. S. Ong, *Mathematics for machine learning*. Cambridge University Press, 2020.

[173] K. K. Hiran, R. K. Jain, K. Lakhwani, and R. Doshi, *Machine Learning: Master Supervised and Unsupervised Learning Algorithms with Real Examples (English Edition)*. BPB Publications, 2021.

[174] A. Chaturvedi, "Handbook of regression analysis with applications in r," 2022.

[175] T. Dwivedi and V. Srivastava, "Optimality of least squares in the seemingly unrelated regression equation model," *Journal of Econometrics*, vol. 7, no. 3, pp. 391–395, 1978.

[176] N. R. Draper and H. Smith, *Applied regression analysis*, vol. 326. John Wiley & Sons, 1998.

[177] G. Gan, C. Ma, and J. Wu, *Data clustering: theory, algorithms, and applications*. SIAM, 2020.

[178] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms," *Annals of Data Science*, vol. 2, pp. 165–193, 2015.

[179] C. Bouveyron, G. Celeux, T. B. Murphy, and A. E. Raftery, *Model-based clustering and classification for data science: with applications in R*, vol. 50. Cambridge University Press, 2019.

[180] C. C. Aggarwal and C. K. Reddy, "Data clustering," *Algorithms and applications. Chapman&Hall/CRC Data mining and Knowledge Discovery series, Londra*, 2014.

[181] J. Sander, *Generalized density based clustering for spatial data mining*. Herbert Utz Verlag, 1999.

[182] G. Bonaccorso, *Hands-On Unsupervised Learning with Python: Implement machine learning and deep learning models using Scikit-Learn, TensorFlow, and more*. Packt Publishing Ltd, 2019.

[183] Y. Lu, C. A. Phillips, and M. A. Langston, "A robustness metric for biological data clustering algorithms," *BMC bioinformatics*, vol. 20, no. 15, pp. 1–8, 2019.

[184] A. Ben-Hur, A. Elisseeff, and I. Guyon, "A stability based method for discovering structure in clustered data," in *Biocomputing 2002*, pp. 6–17, World Scientific, 2001.

[185] C. Lesmeister, *Mastering machine learning with R*. Packt Publishing Ltd, 2017.

[186] H. Steinhaus, "Sur la division des corps matériels en parties," *Bulletin L'Académie Polonaise des Science*, vol. 4, no. 12, pp. 801–804, 1956.

[187] J. MacQueen, “Classification and analysis of multivariate observations,” in *5th Berkeley Symp. Math. Statist. Probability*, pp. 281–297, 1967.

[188] J. Wu, “Cluster analysis and k-means clustering: an introduction,” in *Advances in K-means Clustering*, pp. 1–16, Springer, 2012.

[189] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009.

[190] S. Khanmohammadi, N. Adibeig, and S. Shanehbandy, “An improved overlapping k-means clustering method for medical applications,” *Expert Systems with Applications*, vol. 67, pp. 12–18, 2017.

[191] B. Beltrán and D. Vilariño, “Survey of overlapping clustering algorithms,” *Computación y Sistemas*, vol. 24, no. 2, pp. 575–581, 2020.

[192] Z. Huo and G. Tseng, “Integrative sparse k-means with overlapping group lasso in genomic applications for disease subtype discovery,” *The annals of applied statistics*, vol. 11, no. 2, p. 1011, 2017.

[193] S. Baadel, F. Thabtah, and J. Lu, “Overlapping clustering: A review,” in *2016 SAI Computing Conference (SAI)*, pp. 233–237, IEEE, 2016.

[194] L. Chen, S. Zhou, J. Ma, and M. Xu, “Fast kernel k-means clustering using incomplete cholesky factorization,” *Applied Mathematics and Computation*, vol. 402, p. 126037, 2021.

[195] X. Zhu, Y. Ge, T. Li, D. Thongphiew, F.-F. Yin, and Q. J. Wu, “A planning quality evaluation tool for prostate adaptive imrt based on machine learning,” *Medical Physics*, vol. 38, no. 2, pp. 719–726, 2011.

[196] I. F. Ilyas and X. Chu, *Data cleaning*. Morgan & Claypool, 2019.

[197] J. W. Osborne, *Best practices in data cleaning: A complete guide to everything you need to do before and after collecting your data*. Sage, 2013.

[198] M. Van der Loo and E. De Jonge, *Statistical data cleaning with applications in R*. John Wiley & Sons, 2018.

[199] M. Kuhn and K. Johnson, *Feature engineering and selection: A practical approach for predictive models*. Chapman and Hall/CRC, 2019.

[200] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Computational Statistics & Data Analysis*, vol. 143, p. 106839, 2020.

[201] D. M. Allen, "The relationship between variable selection and data agumentation and a method for prediction," *technometrics*, vol. 16, no. 1, pp. 125–127, 1974.

[202] M. Stone, "An asymptotic equivalence of choice of model by cross-validation and akaike's criterion," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 39, no. 1, pp. 44–47, 1977.

[203] B. Vanderstraeten, B. Goddeeris, K. Vandecasteele, M. Van Eijkeren, C. De Wagter, and Y. Lievens, "Automated instead of manual treatment planning? a plan comparison based on dose-volume statistics and clinical preference," *International Journal of Radiation Oncology, Biology, Physics*, vol. 102, no. 2, pp. 443–450, 2018.

[204] Q. Zhang, L. Ou, Y. Peng, H. Yu, L. Wang, and S. Zhang, "Evaluation of automatic vmat plans in locally advanced nasopharyngeal carcinoma," *Strahlentherapie und Onkologie*, vol. 197, pp. 177–187, 2021.

[205] T. M. Janssen, M. Kusters, Y. Wang, G. Wortel, R. Monshouwer, E. Damen, and S. F. Petit, "Independent knowledge-based treatment planning qa to audit pinnacle autoplanning," *Radiotherapy and Oncology*, vol. 133, pp. 198–204, 2019.

[206] C. R. Hansen, W. Crijns, M. Hussein, L. Rossi, P. Gallego, W. Verbakel, J. Unkelbach, D. Thwaites, and B. Heijmen, "Radiotherapy treatment planning study guidelines (rating): A framework for setting up and reporting on scientific treatment planning studies," *Radiotherapy and Oncology*, vol. 153, pp. 67–78, 2020.

[207] P. Mancosu, V. Hernandez, M. Esposito, C. Moustakis, S. Russo, and O. Blanck, "Application of the rating score: in regards to hansen et al.," *Radiotherapy and Oncology: Journal of the European Society for Therapeutic Radiology and Oncology*, 2021.

[208] J. Leitão, R. Bijman, A. W. Sharfo, Y. Brus, L. Rossi, S. Breedveld, and B. Heijmen, "Automated multi-criterial planning with beam angle optimization to establish non-coplanar vmat class solutions for nasopharyngeal carcinoma," *Physica Medica*, vol. 101, pp. 20–27, 2022.

[209] L. Devlin, L. Grocutt, B. Hunter, H. Chemu, A. Duffton, A. McDonald, N. Macleod, P. McLoone, and S. M. O'Cathail, "The in-silico feasibility of dose escalated, hypofractionated radiotherapy for rectal cancer," *Clinical and translational radiation oncology*, vol. 36, pp. 24–30, 2022.

[210] T. Dasu and T. Johnson, *Exploratory data mining and data cleaning*, vol. 479. John Wiley & Sons, 2003.

[211] C. Capinha and P. Anastácio, "Assessing the environmental requirements of invaders using ensembles of distribution models," *Diversity and Distributions*, vol. 17, no. 1, pp. 13–24, 2011.

[212] J. Elith, C. H. Graham, R. P. Anderson, M. Dudík, S. Ferrier, A. Guisan, R. J. Hijmans, F. Huettmann, J. R. Leathwick, A. Lehmann, J. Li, L. G. Lohmann, B. A. Loiselle, G. Manion, C. Moritz, M. Nakamura, Y. Nakazawa, J. M. M. Overton, A. T. Peterson, S. J. Phillip, K. Richardson, R. cachetti Pereira, R. E. Schapire, J. Soberón, S. Williams, M. S. Wisz, and N. E. Zimmermann, "Novel methods improve prediction of species' distributions from occurrence data," *Ecography*, vol. 29, no. 2, pp. 129–151, 2006.

[213] M. M. Syfert, M. J. Smith, and D. A. Coomes, "The effects of sampling bias and model complexity on the predictive performance of maxent species distribution models," *PloS one*, vol. 8, no. 2, p. e55158, 2013.

[214] L. Yuan, Y. Ge, W. R. Lee, F. F. Yin, J. P. Kirkpatrick, and Q. J. Wu, "Quantitative analysis of the factors which affect the interpatient organ-at-risk dose sparing variation in imrt plans," *Medical Physics*, vol. 39, no. 11, pp. 6868–6878, 2012.

[215] E. van der Bijl, Y. Wang, T. Janssen, and S. Petit, "Predicting patient specific pareto fronts from patient anatomy only," *Radiotherapy and Oncology*, vol. 150, pp. 46–50, 2020.

[216] I. Paddick, "A simple scoring ratio to index the conformity of radiosurgical treatment plans," *Journal of neurosurgery*, vol. 93, no. supplement_3, pp. 219–222, 2000.

[217] K. L. Moore, R. Schmidt, V. Moiseenko, L. A. Olsen, J. Tan, Y. Xiao, J. Galvin, S. Pugh, M. J. Seider, A. P. Dicker, *et al.*, "Quantifying unnecessary normal tissue

complication risks due to suboptimal planning: A secondary study of rtog 0126," *International Journal of Radiation Oncology, Biology, Physics*, vol. 92, no. 2, pp. 228–235, 2015.

[218] S. Breedveld, A. B. Bennan, S. Aluwini, D. R. Schaart, I.-K. K. Kolkman-Deurloo, and B. J. Heijmen, "Fast automated multi-criteria planning for hdr brachytherapy explored for prostate cancer," *Physics in Medicine & Biology*, vol. 64, no. 20, p. 205002, 2019.

[219] Y. Li, H. Bai, D. Huang, F. Chen, and Y. Xia, "Evaluation of auto-planning for left-side breast cancer after breast-conserving surgery based on geometrical relationship," *Technology in Cancer Research & Treatment*, vol. 20, p. 15330338211033050, 2021.

[220] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu, "Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique," *Medical Physics*, vol. 46, no. 1, pp. 370–81, 2019.

[221] A. Babier, R. Mahmood, A. L. McNiven, A. Diamant, and T. C. Chan, "Knowledge-based automated planning with three-dimensional generative adversarial networks," *Medical Physics*, vol. 47, no. 2, pp. 297–306, 2020.

Appendices

Appendix A

RATINGS framework criteria and self rated score

	RATING score sheet	Points
Questions for the Introduction		
	The study aim formulated by research questions	
1	Does the study have a concise and precise study aim, defined with a restricted number of interconnected questions?	10
	The motivation for the research questions	
2	Has relevant up to date literature been included to support the need for the current study?	5
3	Does the study address an existing knowledge gap?	10
Questions for Materials and Methods		
4	Is the global study design adequate for answering the posed research questions?	10
5	Is the global study design described in sufficient detail for others to interpret and reproduce the results?	5
	Patient cohort	
6	Are the inclusion and exclusion criteria of the patient cohort described?	1
7	Is the clinical patient information of the cohort presented, including disease type, site(s) and clinical staging?	1
8	Is the included number of patients stated, explained and justified?	1
9	Has there been consideration of the need for ethical and/or legal approval for the study and if needed, is there a statement about this?	5
	Imaging procedures	
10	Have the scanning parameters been reported in sufficient detail (image modalities, equipment model, slice thickness, voxel size, patient position (e.g. head first, supine, etc.) etc.)?	1
11	Has the applied immobilisation equipment been described, (e.g. vendor and type, standard settings, etc.) where relevant?	1
	Treatment machine and settings	
12	Have the treatment machine and relevant parameters been described with sufficient detail (model, beam energy, MLC, etc.)?	1
13	Have the monitor unit reference conditions been defined, where relevant?	1
	Definition of targets and OARs	

14	Has GTV definition been described in sufficient detail, with references if possible?	1
15	Has CTV definition been described in sufficient detail, with references if possible?	1
16	Has the establishment of PTVs (or alternatively robustness settings) been described in sufficient detail?	1
17	Have PTV sizes in the patient cohort been described?	1
18	Have OAR definitions been described in sufficient detail, with references if possible?	1
19	Have PRV margins been described in sufficient detail, with references if available?	1
	Treatment planning system and dose calculation	
20	Have all applied dose calculation algorithms been described in sufficient detail?	1
21	For any commercial software used, have the manufacturer, algorithms and specific versions been stated?	1
22	Have all relevant user parameters and settings in the TPS been reported, e.g. beams, dose grid, control point spacing?	1
23	Have all volumes been evaluated with the same software/methodology?	1
	Planning aims and optimisation	
24	Are clear planning aims defined, including imposed hard constraints and planning objectives (with or without soft constraints)?	5
25	Has the ranking of planning objectives (priorities) been described?	5
26	Is the dose prescription clearly defined?	10
27	Is there a narrative description of the applied optimisation process, including the handling of all objectives with their ranking?	5
28	If manual intervention during or after optimisation is allowed, has this been described?	1
	Bias mitigation	
29	Have enough study details been provided such that bias issues could be noted?	5
30	Has bias been sufficiently mitigated to reliably answer the posed research question?	10
	Plan acceptability – minor and major protocol deviations	
31	Was the procedure for assessment of plan acceptability well-described?	1
32	Was the procedure for assessment of minor and major protocol deviations well described?	1
	Plan (re-)normalisation for plan comparisons	
33	Has plan (re-)normalisation been described sufficiently?	1
	Dose-volume parameters for plan evaluation and comparison	
34	Have sufficiently comprehensive dose-volume parameters been used for plan evaluations and comparisons?	5
	Population-mean DVHs	
35	Has the algorithm for creating population-mean/median DVHs been reported?	1
36	Have the definitions of confidence intervals been included?	1
	Plan evaluations by clinicians	
37	Have clinicians scored plans to assess quality?	1
38	Were plan comparisons by clinicians blinded?	1
	Predicted tumour control probability and normal tissue complication probabilities for plan evaluation and comparison	
39	Have any applied TCP models been described and referenced?	1
40	Have any applied NTCP models been described and referenced?	1
	Plan deliverability and complexity	
41	Have methods used to assess plan deliverability and complexity been described in sufficient detail?	1

	Composite plan quality metrics	
42	Is there a sufficient basis (e.g. in the literature) for any selected composite plan quality metrics?	1
43	Is there an adequate description of the calculation of the composite plan quality metrics?	1
	Planning and delivery times	
44	Has measurement of planning times been described in sufficient detail?	1
45	Has the establishment of delivery times been described in sufficient detail?	1
=	*Statistical analysis*	
46	Have proper statistical methods been used and described in sufficient detail?	5
47	In case of multiple testing for research questions, has this been handled appropriately?	1
Questions for Results		
48	Does the provided data contribute to (at least partly) answering all aspects of the research questions, e.g. plan acceptability, dosimetric quality, deliverability and planning and delivery times?	10
	Dose distribution reporting	
49	Are complete summaries of the dose distributions in the patient cohort provided (low doses, high doses, OARs, PTV, patient, etc.)?	5
50	Are tables and figures optimised to clearly present the results obtained?	1
51	Have the answers to the research questions been illustrated for an example patient by providing dose distributions, DVHs, etc.?	1
	Plan acceptability reporting – minor and major protocol deviations	
52	In case of treatment technique or planning technique comparisons, was plan acceptability reported separately for each technique?	1
53	Has plan acceptability been reported in sufficient detail: how many plans were acceptable, how many were not and for what reasons (e.g. violation of hard constraints, violation of soft constraints, other reasons)?	1
54	Was there adequate reporting of minor and major protocol deviations?	1
	Deliverability and complexity reporting	
55	Has the deliverability of the plans been adequately reported?	1
56	Have plan deliverability and complexity been investigated in sufficient detail in relation to the posed research questions?	1
	Planning and delivery times reporting	
57	Have planning and delivery times been adequately evaluated and reported?	1
	Patient-specific analyses reporting	
58	Is there sufficient description of inter-patient variations in the results presented?	1
59	Have outlier patients been reported and has any exclusion from population analyses been sufficiently motivated and explained?	1
	Statistical reporting	
60	Are the p-values reported appropriately?	1
61	Are there confidence intervals for the appropriate parameters?	1
Questions for discussions		
62	Is there an overall interpretation of the data presented in the Results section as to how the posed research questions are answered?	10
	Comparison with literature	
63	Has the study been sufficiently discussed in the context of existing literature?	5
	Clinical and statistical significance	
64	Does the discussion focus on statistically significant results?	1

65	Is the potential clinical significance of the results clearly discussed (assuming practical application would be feasible)?	5
	Clinical applicability of the study	
66	Is future the clinical applicability sufficiently discussed?	1
	Study limitations	
67	Has the impact of the study limitations on the provided answers to the research questions been sufficiently discussed?	10
	Future work	
68	Has the potential future work arising from the study been discussed?	1
Questions for conclusions		
69	Do the presented conclusions represent answers to the posed research questions?	5
70	Are the conclusions fully supported by the results?	5
71	Are the conclusions a fair summary of all results?	5
Questions for supplementary		
	Supplementary materials	
72	Is the information presented in the supplementary material of sufficient relevance?	1
73	Is the presentation of the included information of sufficient quality, including readability?	1
74	Has sufficient underlying data been made available or a willingness to share data been indicated, within local data sharing restrictions?	5
RATING remarks		
75	Is the RATING score added to the manuscript?	5
76	Is the accompanying question table added to the cover letter or the supplementary material?	1

Appendix B

Distribution of features under various scalers

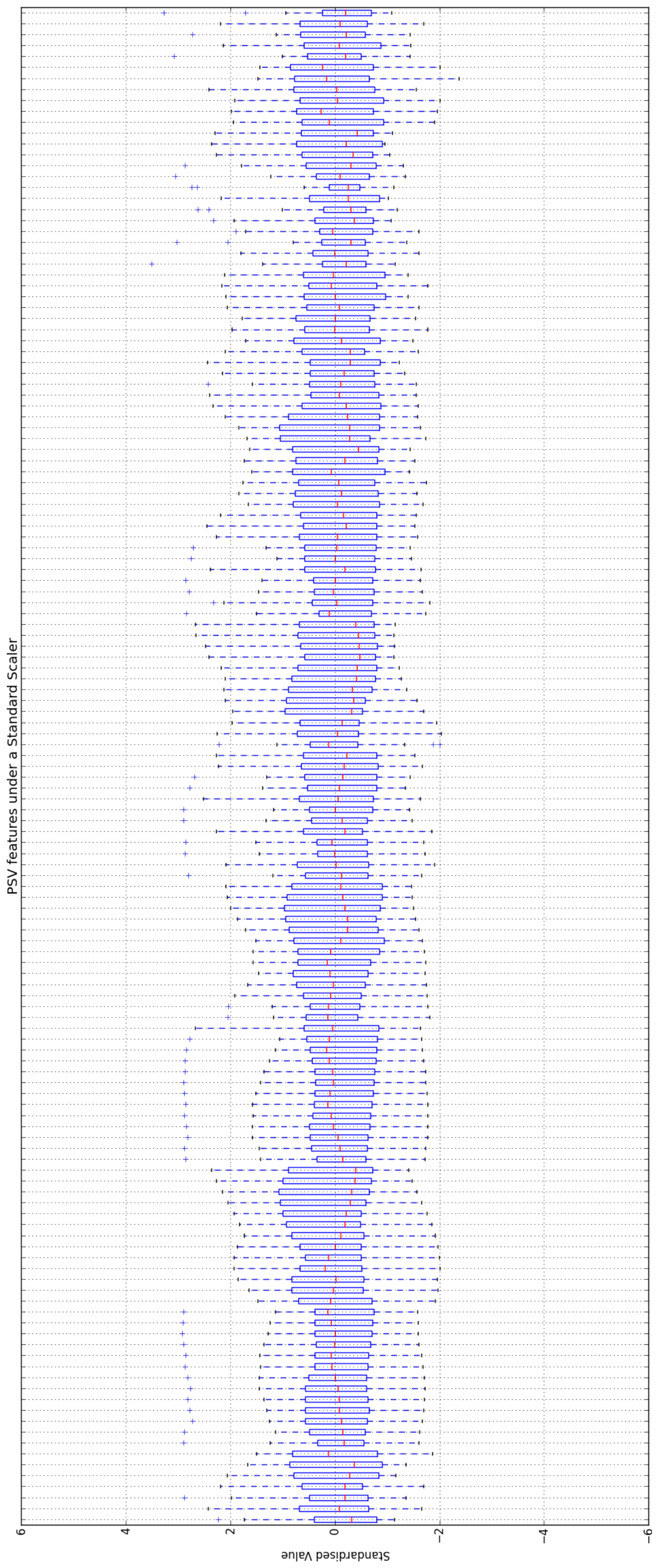


Figure B.1: PSV features under a standard scaler

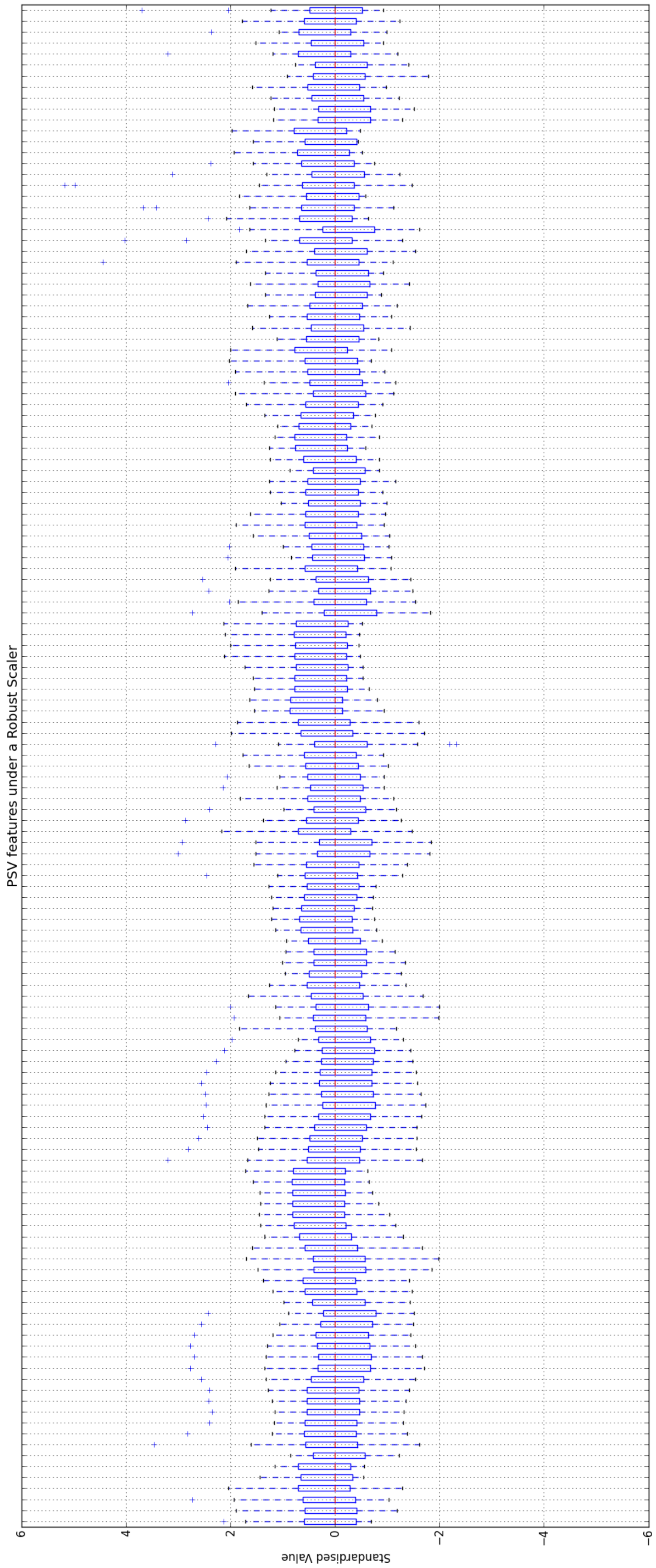


Figure B.2: PSV features under a robust scaler

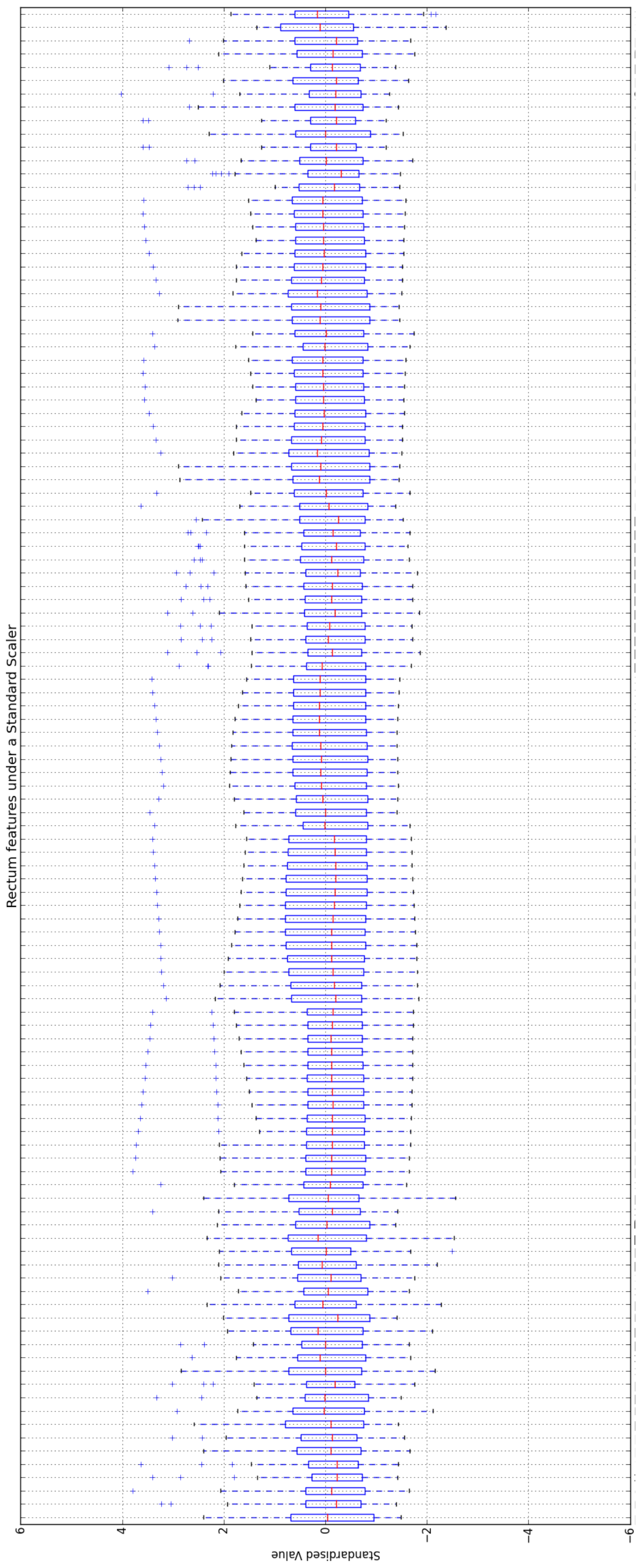


Figure B.3: Rectum features under a standard scaler

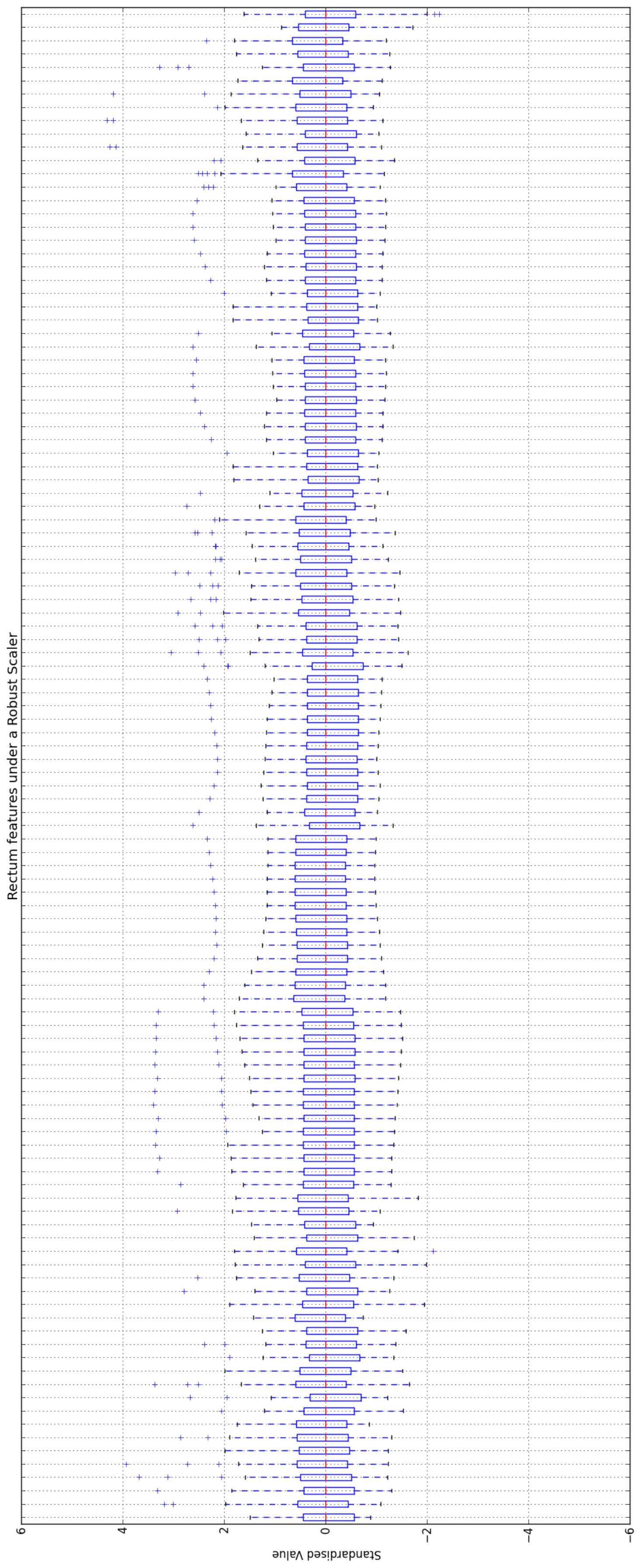


Figure B.4: Rectum features under a robust scaler

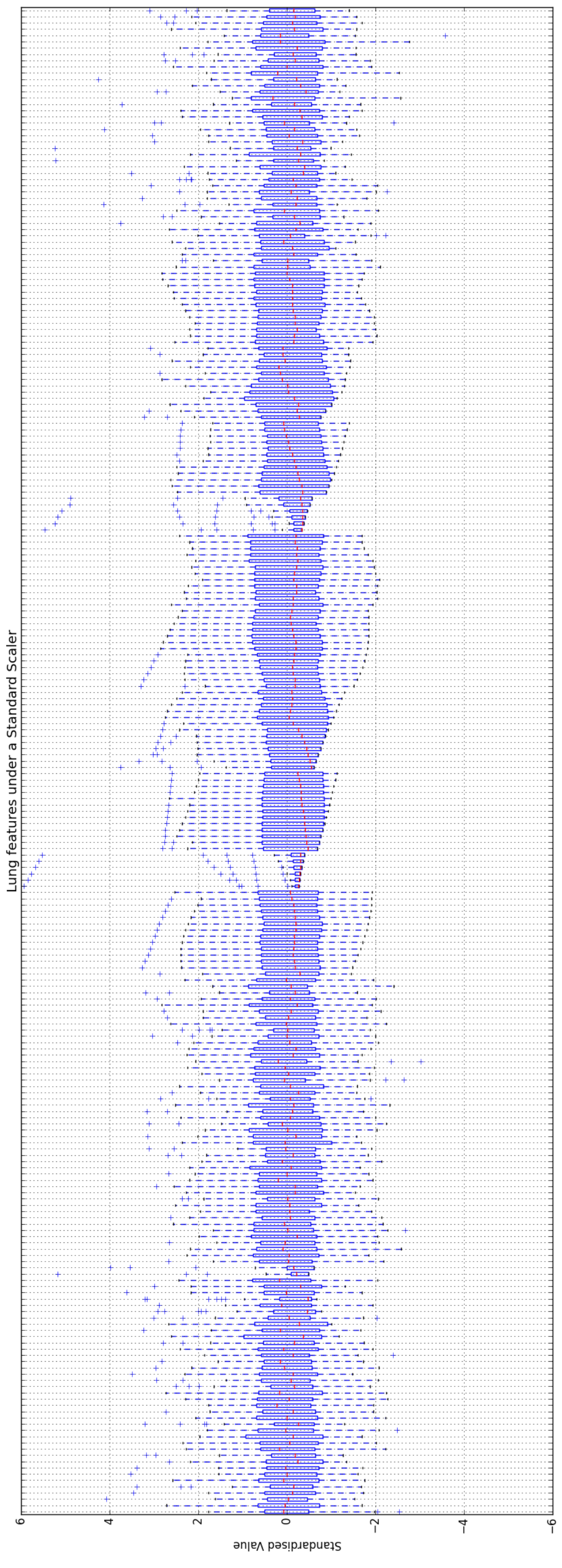


Figure B.5: Lung features under a standard scaler

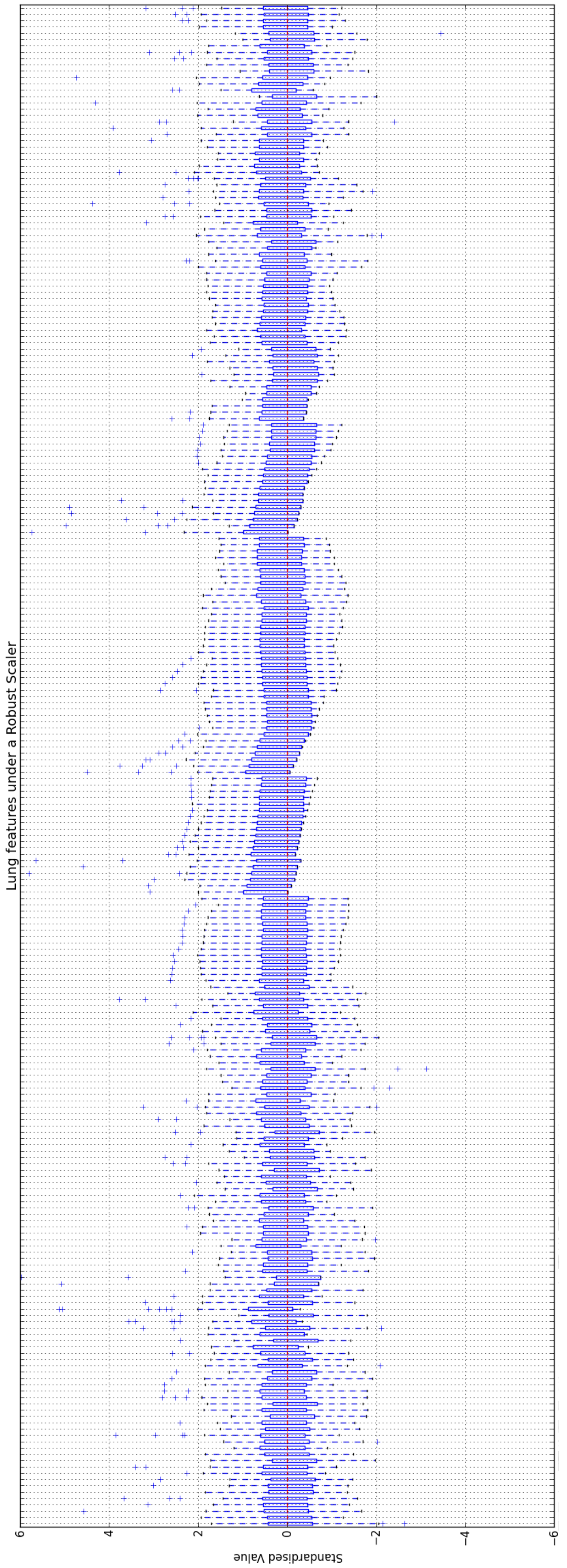


Figure B.6: Lung features under a robust scaler

Appendix C

Predictive Features in FeatureDS1

C.1 PSV features

Variable Alias	Variable
Bladder	Volume of the bladder
External	Volume of the external
Rectum	Volume of the rectum
dm_ptv48p00_sum	Volume of PTV48 minus PTV60
total_OAR	Sum of rectum and bladder volume
PTV60	Volume of PTV60
PTV48	Volume of PTV48
PTV60_0p0_Bladder	Volume of PTV60 expanded 0cm isoptropically overlapping the bladder volume
PTV60_0p2_Bladder	Volume of PTV60 expanded 0.2cm isoptropically overlapping the bladder volume
PTV60_0p4_Bladder	Volume of PTV60 expanded 0.4cm isoptropically overlapping the bladder volume
PTV60_0p6_Bladder	Volume of PTV60 expanded 0.6cm isoptropically overlapping the bladder volume
PTV60_0p8_Bladder	Volume of PTV60 expanded 0.8cm isoptropically overlapping the bladder volume
PTV60_1p0_Bladder	Volume of PTV60 expanded 1cm isoptropically overlapping the bladder volume
PTV60_1p2_Bladder	Volume of PTV60 expanded 1.2cm isoptropically overlapping the bladder volume
PTV60_1p4_Bladder	Volume of PTV60 expanded 1.4cm isoptropically overlapping the bladder volume

PTV60_1p6_Bladder Volume of PTV60 expanded 1.6cm isoptropically overlapping the bladder volume

PTV60_1p8_Bladder Volume of PTV60 expanded 1.8cm isoptropically overlapping the bladder volume

PTV60_2p0_Bladder Volume of PTV60 expanded 2cm isoptropically overlapping the bladder volume

PTV60_2p2_Bladder Volume of PTV60 expanded 2.2cm isoptropically overlapping the bladder volume

PTV60_2p4_Bladder Volume of PTV60 expanded 2.4cm isoptropically overlapping the bladder volume

PTV60_0p0_Rectum Volume of PTV60 expanded 0cm isoptropically overlapping the rectum volume

PTV60_0p2_Rectum Volume of PTV60 expanded 0.2cm isoptropically overlapping the rectum volume

PTV60_0p4_Rectum Volume of PTV60 expanded 0.4cm isoptropically overlapping the rectum volume

PTV60_0p6_Rectum Volume of PTV60 expanded 0.6cm isoptropically overlapping the rectum volume

PTV60_0p8_Rectum Volume of PTV60 expanded 0.8cm isoptropically overlapping the rectum volume

PTV60_1p0_Rectum Volume of PTV60 expanded 1cm isoptropically overlapping the rectum volume

PTV60_1p2_Rectum Volume of PTV60 expanded 1.2cm isoptropically overlapping the rectum volume

PTV60_1p4_Rectum Volume of PTV60 expanded 1.4cm isoptropically overlapping the rectum volume

PTV60_1p6_Rectum Volume of PTV60 expanded 1.6cm isoptropically overlapping the rectum volume

PTV60_1p8_Rectum Volume of PTV60 expanded 1.8cm isoptropically overlapping the rectum volume

PTV60_2p0_Rectum Volume of PTV60 expanded 2cm isoptropically overlapping the rectum volume

PTV60_2p2_Rectum Volume of PTV60 expanded 2.2cm isoptropically overlapping the rectum volume

PTV60_2p4_Rectum Volume of PTV60 expanded 2.4cm isoptropically overlapping the rectum volume

PTV48_0p0_Bladder Volume of PTV48 expanded 0cm isoptropically overlapping the bladder volume

PTV48_0p2_Bladder Volume of PTV48 expanded 0.2cm isoptropically overlapping the bladder volume

PTV48_0p4_Bladder	Volume of PTV48 expanded 0.4cm isoptropically overlapping the bladder volume
PTV48_0p6_Bladder	Volume of PTV48 expanded 0.6cm isoptropically overlapping the bladder volume
PTV48_0p8_Bladder	Volume of PTV48 expanded 0.8cm isoptropically overlapping the bladder volume
PTV48_1p0_Bladder	Volume of PTV48 expanded 1cm isoptropically overlapping the bladder volume
PTV48_1p2_Bladder	Volume of PTV48 expanded 1.2cm isoptropically overlapping the bladder volume
PTV48_1p4_Bladder	Volume of PTV48 expanded 1.4cm isoptropically overlapping the bladder volume
PTV48_1p6_Bladder	Volume of PTV48 expanded 1.6cm isoptropically overlapping the bladder volume
PTV48_1p8_Bladder	Volume of PTV48 expanded 1.8cm isoptropically overlapping the bladder volume
PTV48_2p0_Bladder	Volume of PTV48 expanded 2cm isoptropically overlapping the bladder volume
PTV48_2p2_Bladder	Volume of PTV48 expanded 2.2cm isoptropically overlapping the bladder volume
PTV48_2p4_Bladder	Volume of PTV48 expanded 2.4cm isoptropically overlapping the bladder volume
PTV48_0p0_Rectum	Volume of PTV48 expanded 0cm isoptropically overlapping the rectum volume
PTV48_0p2_Rectum	Volume of PTV48 expanded 0.2cm isoptropically overlapping the rectum volume
PTV48_0p4_Rectum	Volume of PTV48 expanded 0.4cm isoptropically overlapping the rectum volume
PTV48_0p6_Rectum	Volume of PTV48 expanded 0.6cm isoptropically overlapping the rectum volume
PTV48_0p8_Rectum	Volume of PTV48 expanded 0.8cm isoptropically overlapping the rectum volume
PTV48_1p0_Rectum	Volume of PTV48 expanded 1cm isoptropically overlapping the rectum volume
PTV48_1p2_Rectum	Volume of PTV48 expanded 1.2cm isoptropically overlapping the rectum volume
PTV48_1p4_Rectum	Volume of PTV48 expanded 1.4cm isoptropically overlapping the rectum volume
PTV48_1p6_Rectum	Volume of PTV48 expanded 1.6cm isoptropically overlapping the rectum volume

PTV48_1p8_Rectum	Volume of PTV48 expanded 1.8cm isoptropically overlapping the rectum volume
PTV48_2p0_Rectum	Volume of PTV48 expanded 2cm isoptropically overlapping the rectum volume
PTV48_2p2_Rectum	Volume of PTV48 expanded 2.2cm isoptropically overlapping the rectum volume
PTV48_2p4_Rectum	Volume of PTV48 expanded 2.4cm isoptropically overlapping the rectum volume
PTV60_0p0_0p2_slope_Bladder	Slope between PTV60 expanded 0cm overlapping the bladder and PTV60 expanded 0.2cm overlapping the bladder
PTV60_0p2_0p4_slope_Bladder	Slope between PTV60 expanded 0.2cm overlapping the bladder and PTV60 expanded 0.4cm overlapping the bladder
PTV60_0p4_0p6_slope_Bladder	Slope between PTV60 expanded 0.4cm overlapping the bladder and PTV60 expanded 0.6cm overlapping the bladder
PTV60_0p6_0p8_slope_Bladder	Slope between PTV60 expanded 0.6cm overlapping the bladder and PTV60 expanded 0.8cm overlapping the bladder
PTV60_0p8_1p0_slope_Bladder	Slope between PTV60 expanded 0.8cm overlapping the bladder and PTV60 expanded 1cm overlapping the bladder
PTV60_1p0_1p2_slope_Bladder	Slope between PTV60 expanded 1cm overlapping the bladder and PTV60 expanded 1.2cm overlapping the bladder
PTV60_1p2_1p4_slope_Bladder	Slope between PTV60 expanded 1.2cm overlapping the bladder and PTV60 expanded 1.4cm overlapping the bladder
PTV60_1p4_1p6_slope_Bladder	Slope between PTV60 expanded 1.4cm overlapping the bladder and PTV60 expanded 1.6cm overlapping the bladder
PTV60_1p6_1p8_slope_Bladder	Slope between PTV60 expanded 1.6cm overlapping the bladder and PTV60 expanded 1.8cm overlapping the bladder
PTV60_1p8_2p0_slope_Bladder	Slope between PTV60 expanded 1.8cm overlapping the bladder and PTV60 expanded 2cm overlapping the bladder
PTV60_2p0_2p2_slope_Bladder	Slope between PTV60 expanded 2cm overlapping the bladder and PTV60 expanded 2.2cm overlapping the bladder
PTV60_2p2_2p4_slope_Bladder	Slope between PTV60 expanded 2.2cm overlapping the bladder and PTV60 expanded 2.4cm overlapping the bladder
PTV60_0p0_0p2_slope_Rectum	Slope between PTV60 expanded 0cm overlapping the bladder and PTV60 expanded 0.2cm overlapping the bladder
PTV60_0p2_0p4_slope_Rectum	Slope between PTV60 expanded 0.2cm overlapping the bladder and PTV60 expanded 0.4cm overlapping the bladder
PTV60_0p4_0p6_slope_Rectum	Slope between PTV60 expanded 0.4cm overlapping the bladder and PTV60 expanded 0.6cm overlapping the bladder
PTV60_0p6_0p8_slope_Rectum	Slope between PTV60 expanded 0.6cm overlapping the bladder and PTV60 expanded 0.8cm overlapping the bladder

PTV60_0p8_1p0_slope_Rectum	Slope between PTV60 expanded 0.8cm overlapping the bladder and PTV60 expanded 1cm overlapping the bladder
PTV60_1p0_1p2_slope_Rectum	Slope between PTV60 expanded 1cm overlapping the bladder and PTV60 expanded 1.2cm overlapping the bladder
PTV60_1p2_1p4_slope_Rectum	Slope between PTV60 expanded 1.2cm overlapping the bladder and PTV60 expanded 1.4cm overlapping the bladder
PTV60_1p4_1p6_slope_Rectum	Slope between PTV60 expanded 1.4cm overlapping the bladder and PTV60 expanded 1.6cm overlapping the bladder
PTV60_1p6_1p8_slope_Rectum	Slope between PTV60 expanded 1.6cm overlapping the bladder and PTV60 expanded 1.8cm overlapping the bladder
PTV60_1p8_2p0_slope_Rectum	Slope between PTV60 expanded 1.8cm overlapping the bladder and PTV60 expanded 2cm overlapping the bladder
PTV60_2p0_2p2_slope_Rectum	Slope between PTV60 expanded 2cm overlapping the bladder and PTV60 expanded 2.2cm overlapping the bladder
PTV60_2p2_2p4_slope_Rectum	Slope between PTV60 expanded 2.2cm overlapping the bladder and PTV60 expanded 2.4cm overlapping the bladder
PTV48_0p0_0p2_slope_Bladder	Slope between PTV48 expanded 0cm overlapping the bladder and PTV48 expanded 0.2cm overlapping the bladder
PTV48_0p2_0p4_slope_Bladder	Slope between PTV48 expanded 0.2cm overlapping the bladder and PTV48 expanded 0.4cm overlapping the bladder
PTV48_0p4_0p6_slope_Bladder	Slope between PTV48 expanded 0.4cm overlapping the bladder and PTV48 expanded 0.6cm overlapping the bladder
PTV48_0p6_0p8_slope_Bladder	Slope between PTV48 expanded 0.6cm overlapping the bladder and PTV48 expanded 0.8cm overlapping the bladder
PTV48_0p8_1p0_slope_Bladder	Slope between PTV48 expanded 0.8cm overlapping the bladder and PTV48 expanded 1cm overlapping the bladder
PTV48_1p0_1p2_slope_Bladder	Slope between PTV48 expanded 1cm overlapping the bladder and PTV48 expanded 1.2cm overlapping the bladder
PTV48_1p2_1p4_slope_Bladder	Slope between PTV48 expanded 1.2cm overlapping the bladder and PTV48 expanded 1.4cm overlapping the bladder
PTV48_1p4_1p6_slope_Bladder	Slope between PTV48 expanded 1.4cm overlapping the bladder and PTV48 expanded 1.6cm overlapping the bladder
PTV48_1p6_1p8_slope_Bladder	Slope between PTV48 expanded 1.6cm overlapping the bladder and PTV48 expanded 1.8cm overlapping the bladder
PTV48_1p8_2p0_slope_Bladder	Slope between PTV48 expanded 1.8cm overlapping the bladder and PTV48 expanded 2cm overlapping the bladder
PTV48_2p0_2p2_slope_Bladder	Slope between PTV48 expanded 2cm overlapping the bladder and PTV48 expanded 2.2cm overlapping the bladder
PTV48_2p2_2p4_slope_Bladder	Slope between PTV48 expanded 2.2cm overlapping the bladder and PTV48 expanded 2.4cm overlapping the bladder

PTV48_0p0_0p2_slope_Rectum	Slope between PTV48 expanded 0cm overlapping the bladder and PTV48 expanded 0.2cm overlapping the bladder
PTV48_0p2_0p4_slope_Rectum	Slope between PTV48 expanded 0.2cm overlapping the bladder and PTV48 expanded 0.4cm overlapping the bladder
PTV48_0p4_0p6_slope_Rectum	Slope between PTV48 expanded 0.4cm overlapping the bladder and PTV48 expanded 0.6cm overlapping the bladder
PTV48_0p6_0p8_slope_Rectum	Slope between PTV48 expanded 0.6cm overlapping the bladder and PTV48 expanded 0.8cm overlapping the bladder
PTV48_0p8_1p0_slope_Rectum	Slope between PTV48 expanded 0.8cm overlapping the bladder and PTV48 expanded 1cm overlapping the bladder
PTV48_1p0_1p2_slope_Rectum	Slope between PTV48 expanded 1cm overlapping the bladder and PTV48 expanded 1.2cm overlapping the bladder
PTV48_1p2_1p4_slope_Rectum	Slope between PTV48 expanded 1.2cm overlapping the bladder and PTV48 expanded 1.4cm overlapping the bladder
PTV48_1p4_1p6_slope_Rectum	Slope between PTV48 expanded 1.4cm overlapping the bladder and PTV48 expanded 1.6cm overlapping the bladder
PTV48_1p6_1p8_slope_Rectum	Slope between PTV48 expanded 1.6cm overlapping the bladder and PTV48 expanded 1.8cm overlapping the bladder
PTV48_1p8_2p0_slope_Rectum	Slope between PTV48 expanded 1.8cm overlapping the bladder and PTV48 expanded 2cm overlapping the bladder
PTV48_2p0_2p2_slope_Rectum	Slope between PTV48 expanded 2cm overlapping the bladder and PTV48 expanded 2.2cm overlapping the bladder
PTV48_2p2_2p4_slope_Rectum	Slope between PTV48 expanded 2.2cm overlapping the bladder and PTV48 expanded 2.4cm overlapping the bladder
av_dist_PTV60_Bladder	Average distance between PTV60 and the bladder
av_dist_PTV60_Rectum	Average distance between PTV60 and the rectum
av_dist_PTV48_Bladder	Average distance between PTV48 and the bladder
av_dist_PTV48_Rectum	Average distance between PTV48 and the rectum
max_dist_PTV60_Bladder	The largest distance between PTV60 and the bladder
max_dist_PTV60_Rectum	The largest distance between PTV60 and the rectum
max_dist_PTV48_Bladder	The largest distance between PTV48 and the bladder
max_dist_PTV48_Rectum	The largest distance between PTV48 and the rectum
VIF_PTV60_Bladder	Volume of the bladder between the most superior and inferior transverse slices of PTV60
VIF_PTV60_Rectum	Volume of the rectum between the most superior and inferior transverse slices of PTV60
VIF_PTV48_Bladder	Volume of the bladder between the most superior and inferior transverse slices of PTV48
VIF_PTV48_Rectum	Volume of the rectum between the most superior and inferior transverse slices of PTV48

VOF_PTV60_Bladder	Volume of the bladder above the most superior and below the most inferior transverse slices of PTV60
VOF_PTV60_Rectum	Volume of the rectum above the most superior and below the most inferior transverse slices of PTV60
VOF_PTV48_Bladder	Volume of the bladder above the most superior and below the most inferior transverse slices of PTV48
VOF_PTV48_Rectum	Volume of the rectum above the most superior and below the most inferior transverse slices of PTV48
Total_VIF_PTV60	Rectum and bladder between the most superior and inferior transverse slices of PTV60
Total_VIF_PTV48	Rectum and bladder between the most superior and inferior transverse slices of PTV48
Total_VOF_PTV60	Rectum and bladder above the most superior and below the most inferior transverse slices of PTV60
Total_VOF_PTV48	Rectum and bladder above the most superior and below the most inferior transverse slices of PTV48
centre_dist_ptv60p00_ptv48p00	Distance between the centre of PTV60 and the centre of PTV48
centre_dist_ptv60p00_Bladder	Distance between the centre of PTV60 and the centre of the bladder
centre_dist_ptv60p00_Rectum	Distance between the centre of PTV60 and the centre of the rectum
centre_dist_ptv48p00_Bladder	Distance between the centre of PTV48 and the centre of the bladder
centre_dist_ptv48p00_Rectum	Distance between the centre of PTV48 and the centre of the rectum
centre_dist_Bladder_Rectum	Distance between the centre of the bladder and the centre of the rectum
ratio_ptv60p00_ptv48p00	The voume of PTV60 divided by the volume of PTV48
ratio_ptv60p00_Bladder	The volume of PTV60 divided by the volume of the bladder
ratio_ptv60p00_Rectum	The volume of PTV60 divided by the volume of the rectum
ratio_ptv48p00_Bladder	The volume of PTV48 divided by the volume of the bladder
ratio_ptv48p00_Rectum	The volume of PTV48 divided by the volume of the rectum
ratio_Bladder_Rectum	The volume of the bladder divided by the volume of the rectum

C.2 Rectum features

Variable alias	**Variable**
External	Volume of the external
BowelBag	Volume of the Bowel Bag
PTV45	Volume of PTV45
aux_ant	Volume of aux ant (bowel bag minus PTV45)
total_OAR	Volume of bowel bag, genitals and stoma
min_dist_BowelBag_External	The smallest distance between the bowel bag and the external
min_dist_aux_ant_External	The smallest distance between the aux ant and the external
min_dist_PTV_4500_External	The smallest distance between PTV45 and the external
centre_dist_PTV_4500_BowelBag	The distance between the centre of PTV45 and the centre of the bowel bag
centre_dist_BowelBag_External	The distance between the centre of the bowel bag and the centre of the external
centre_dist_PTV_4500_External	The distance between the centre of PTV45 and the centre of the external
centre_dist_PTV_4500_aux_ant	The distance between the centre of PTV45 and aux ant
centre_dist_BowelBag_aux_ant	The distance between the centre of the bowl bag and aux ant
centre_dist_aux_ant_External	The distance between the centre of aux ant and the external
max_dist_BowelBag_External	The largest distance between the bowel bag and the external
max_dist_PTV_4500_BowelBag	The largest distance between PTV45 and the bowel bag
max_dist_PTV_4500_External	The largest distance between PTV45 and the external
max_dist_PTV_4500_aux_ant	The largest distance between PTV45 and aux ant
max_dist_BowelBag_aux_ant	The largest distance between the bowel bag and aux ant
max_dist_aux_ant_External	The largest distance between aux ant and the external
av_dist_BowelBag_External	The average distance between bowel bag and the external
av_dist_PTV_4500_External	The average distance between PTV45 and the external
av_dist_PTV_4500_BowelBag	The average distance between PTV45 and the bowel bag
av_dist_BowelBag_aux_ant	The average distance between bowel bag and the aux ant
av_dist_aux_ant_External	The average distance between aux ant and the external
av_dist_PTV_4500_aux_ant	The average distance between PTV45 and aux ant
PTV_4500_0p0_External	The volume of PTV45 expanded 0cm isotropically overlapping the external
PTV_4500_0p2_External	The volume of PTV45 expanded 0.2cm isotropically overlapping the external
PTV_4500_0p4_External	The volume of PTV45 expanded 0.4cm isotropically overlapping the external
PTV_4500_0p6_External	The volume of PTV45 expanded 0.6cm isotropically overlapping the external

PTV_4500_0p8_External	The volume of PTV45 expanded 0.8cm isotropically overlapping the external
PTV_4500_1p0_External	The volume of PTV45 expanded 1cm isotropically overlapping the external
PTV_4500_1p2_External	The volume of PTV45 expanded 1.2cm isotropically overlapping the external
PTV_4500_1p4_External	The volume of PTV45 expanded 1.4cm isotropically overlapping the external
PTV_4500_1p6_External	The volume of PTV45 expanded 1.6cm isotropically overlapping the external
PTV_4500_1p8_External	The volume of PTV45 expanded 1.8cm isotropically overlapping the external
PTV_4500_2p0_External	The volume of PTV45 expanded 2cm isotropically overlapping the external
PTV_4500_2p2_External	The volume of PTV45 expanded 2.2cm isotropically overlapping the external
PTV_4500_2p4_External	The volume of PTV45 expanded 2.4cm isotropically overlapping the external
PTV_4500_0p0_BowelBag	The volume of PTV45 expanded 0cm isotropically overlapping the bowel bag
PTV_4500_0p2_BowelBag	The volume of PTV45 expanded 0.2cm isotropically overlapping the bowel bag
PTV_4500_0p4_BowelBag	The volume of PTV45 expanded 0.4cm isotropically overlapping the bowel bag
PTV_4500_0p6_BowelBag	The volume of PTV45 expanded 0.6cm isotropically overlapping the bowel bag
PTV_4500_0p8_BowelBag	The volume of PTV45 expanded 0.8cm isotropically overlapping the bowel bag
PTV_4500_1p0_BowelBag	The volume of PTV45 expanded 1cm isotropically overlapping the bowel bag
PTV_4500_1p2_BowelBag	The volume of PTV45 expanded 1.2cm isotropically overlapping the bowel bag
PTV_4500_1p4_BowelBag	The volume of PTV45 expanded 1.4cm isotropically overlapping the bowel bag
PTV_4500_1p6_BowelBag	The volume of PTV45 expanded 1.6cm isotropically overlapping the bowel bag
PTV_4500_1p8_BowelBag	The volume of PTV45 expanded 1.8cm isotropically overlapping the bowel bag
PTV_4500_2p0_BowelBag	The volume of PTV45 expanded 2cm isotropically overlapping the bowel bag

PTV_4500_2p2_BowelBag	The volume of PTV45 expanded 2.2cm isotropically overlapping the bowel bag
PTV_4500_2p4_BowelBag	The volume of PTV45 expanded 2.4cm isotropically overlapping the bowel bag
PTV_4500_0p2_aux_ant	The volume of PTV45 expanded 0cm isotropically overlapping the aux ant
PTV_4500_0p4_aux_ant	The volume of PTV45 expanded 0.2cm isotropically overlapping the aux ant
PTV_4500_0p6_aux_ant	The volume of PTV45 expanded 0.4cm isotropically overlapping the aux ant
PTV_4500_0p8_aux_ant	The volume of PTV45 expanded 0.6cm isotropically overlapping the aux ant
PTV_4500_1p0_aux_ant	The volume of PTV45 expanded 0.8cm isotropically overlapping the aux ant
PTV_4500_1p2_aux_ant	The volume of PTV45 expanded 1cm isotropically overlapping the aux ant
PTV_4500_1p4_aux_ant	The volume of PTV45 expanded 1.2cm isotropically overlapping the aux ant
PTV_4500_1p6_aux_ant	The volume of PTV45 expanded 1.4cm isotropically overlapping the aux ant
PTV_4500_1p8_aux_ant	The volume of PTV45 expanded 1.6cm isotropically overlapping the aux ant
PTV_4500_2p0_aux_ant	The volume of PTV45 expanded 1.8cm isotropically overlapping the aux ant
PTV_4500_2p2_aux_ant	The volume of PTV45 expanded 2cm isotropically overlapping the aux ant
PTV_4500_2p4_aux_ant	The volume of PTV45 expanded 2.2cm isotropically overlapping the aux ant
PTV_4500_0p0_0p2_slope_External	Slope between PTV45 expanded 0cm overlapping the external and PTV45 expanded 0.2cm overlapping the external
PTV_4500_0p2_0p4_slope_External	Slope between PTV45 expanded 0.2cm overlapping the external and PTV45 expanded 0.4cm overlapping the external
PTV_4500_0p4_0p6_slope_External	Slope between PTV45 expanded 0.4cm overlapping the external and PTV45 expanded 0.6cm overlapping the external
PTV_4500_0p6_0p8_slope_External	Slope between PTV45 expanded 0.6cm overlapping the external and PTV45 expanded 0.8cm overlapping the external
PTV_4500_0p8_1p0_slope_External	Slope between PTV45 expanded 0.8cm overlapping the external and PTV45 expanded 1cm overlapping the external
PTV_4500_1p0_1p2_slope_External	Slope between PTV45 expanded 1cm overlapping the external and PTV45 expanded 1.2cm overlapping the external

PTV_4500_1p2_1p4_slope_External	Slope between PTV45 expanded 1.2cm overlapping the external and PTV45 expanded 1.4cm overlapping the external
PTV_4500_1p4_1p6_slope_External	Slope between PTV45 expanded 1.4cm overlapping the external and PTV45 expanded 1.6cm overlapping the external
PTV_4500_1p6_1p8_slope_External	Slope between PTV45 expanded 1.6cm overlapping the external and PTV45 expanded 1.8cm overlapping the external
PTV_4500_1p8_2p0_slope_External	Slope between PTV45 expanded 1.8cm overlapping the external and PTV45 expanded 2cm overlapping the external
PTV_4500_2p0_2p2_slope_External	Slope between PTV45 expanded 2cm overlapping the external and PTV45 expanded 2.2cm overlapping the external
PTV_4500_2p2_2p4_slope_External	Slope between PTV45 expanded 2.2cm overlapping the external and PTV45 expanded 2.4cm overlapping the external
PTV_4500_0p0_0p2_slope_BowelBag	Slope between PTV45 expanded 0cm overlapping the bowel bag and PTV45 expanded 0.2cm overlapping the bowel bag
PTV_4500_0p2_0p4_slope_BowelBag	Slope between PTV45 expanded 0.2cm overlapping the bowel bag and PTV45 expanded 0.4cm overlapping the bowel bag
PTV_4500_0p4_0p6_slope_BowelBag	Slope between PTV45 expanded 0.4cm overlapping the bowel bag and PTV45 expanded 0.6cm overlapping the bowel bag
PTV_4500_0p6_0p8_slope_BowelBag	Slope between PTV45 expanded 0.6cm overlapping the bowel bag and PTV45 expanded 0.8cm overlapping the bowel bag
PTV_4500_0p8_1p0_slope_BowelBag	Slope between PTV45 expanded 0.8cm overlapping the bowel bag and PTV45 expanded 1cm overlapping the bowel bag
PTV_4500_1p0_1p2_slope_BowelBag	Slope between PTV45 expanded 1cm overlapping the bowel bag and PTV45 expanded 1.2cm overlapping the bowel bag
PTV_4500_1p2_1p4_slope_BowelBag	Slope between PTV45 expanded 1.2cm overlapping the bowel bag and PTV45 expanded 1.4cm overlapping the bowel bag
PTV_4500_1p4_1p6_slope_BowelBag	Slope between PTV45 expanded 1.4cm overlapping the bowel bag and PTV45 expanded 1.6cm overlapping the bowel bag
PTV_4500_1p6_1p8_slope_BowelBag	Slope between PTV45 expanded 1.6cm overlapping the bowel bag and PTV45 expanded 1.8cm overlapping the bowel bag
PTV_4500_1p8_2p0_slope_BowelBag	Slope between PTV45 expanded 1.8cm overlapping the bowel bag and PTV45 expanded 2cm overlapping the bowel bag
PTV_4500_2p0_2p2_slope_BowelBag	Slope between PTV45 expanded 2cm overlapping the bowel bag and PTV45 expanded 2.2cm overlapping the bowel bag
PTV_4500_2p2_2p4_slope_BowelBag	Slope between PTV45 expanded 2.2cm overlapping the bowel bag and PTV45 expanded 2.4cm overlapping the bowel bag
PTV_4500_0p0_0p2_slope_aux_ant	Slope between PTV45 expanded 0cm overlapping aux ant and PTV45 expanded 0.2cm overlapping aux ant
PTV_4500_0p2_0p4_slope_aux_ant	Slope between PTV45 expanded 0.2cm overlapping the aux ant and PTV45 expanded 0.4cm overlapping the aux ant

PTV_4500_0p4_0p6_slope_aux_ant	Slope between PTV45 expanded 0.4cm overlapping the aux ant and PTV45 expanded 0.6cm overlapping the aux ant
PTV_4500_0p6_0p8_slope_aux_ant	Slope between PTV45 expanded 0.6cm overlapping the aux ant and PTV45 expanded 0.8cm overlapping the aux ant
PTV_4500_0p8_1p0_slope_aux_ant	Slope between PTV45 expanded 0.8cm overlapping the aux ant and PTV45 expanded 1cm overlapping the aux ant
PTV_4500_1p0_1p2_slope_aux_ant	Slope between PTV45 expanded 1cm overlapping the aux ant and PTV45 expanded 1.2cm overlapping the aux ant
PTV_4500_1p2_1p4_slope_aux_ant	Slope between PTV45 expanded 1.2cm overlapping the aux ant and PTV45 expanded 1.4cm overlapping the aux ant
PTV_4500_1p4_1p6_slope_aux_ant	Slope between PTV45 expanded 1.4cm overlapping the aux ant and PTV45 expanded 1.6cm overlapping the aux ant
PTV_4500_1p6_1p8_slope_aux_ant	Slope between PTV45 expanded 1.6cm overlapping the aux ant and PTV45 expanded 1.8cm overlapping the aux ant
PTV_4500_1p8_2p0_slope_aux_ant	Slope between PTV45 expanded 1.8cm overlapping the aux ant and PTV45 expanded 2cm overlapping the aux ant
PTV_4500_2p0_2p2_slope_aux_ant	Slope between PTV45 expanded 2cm overlapping the aux ant and PTV45 expanded 2.2cm overlapping the aux ant
PTV_4500_2p2_2p4_slope_aux_ant	Slope between PTV45 expanded 2.2cm overlapping the aux ant and PTV45 expanded 2.4cm overlapping the aux ant
VIF_PTV_4500_BowelBag	Volume of the bowel bag between the most superior and most inferior transverse slice of PTV45
VIF_PTV_4500_External	Volume of the external between the most superior and most inferior transverse slice of PTV45
VIF_PTV_4500_aux_ant	Volume of the aux ant between the most superior and most inferior transverse slice of PTV45
VOF_PTV_4500_BowelBag	Volume of the bowel bag above the most superior below the most inferior transverse slice of PTV45
VOF_PTV_4500_External	Volume of the external above the most superior below the most inferior transverse slice of PTV45
VOF_PTV_4500_aux_ant	Volume of the aux ant above the most superior below the most inferior transverse slice of PTV45
Total_VIF_PTV45	Volume of bowel bag, genital region and stoma region between the most superior and most inferior transverse slice of PTV45
Total_VOF_PTV45	Volume of thebowel bag, genital region and stoma region above the most superior below the most inferior transverse slice of PTV45
ratio_BowelBag_External	Volume of the bowel bag divided by the volume of the external
ratio_PTV_4500_External	Volume of PTV45 divided by the volume of the external
ratio_PTV_4500_BowelBag	Volume of PTV45 divided by the volume of the bowel bag
ratio_aux_ant_External	Volume of aux ant divided by the volume the external
ratio_PTV_4500_aux_ant	Volume of PTV45 divided by the volume of aux ant

ratio_BowelBag_aux_ant	Volume of the bowel bag divided by the volume of aux ant

C.3 Lung features

Variable alias	**Variable**
External	Volume of the external
IpsLung	Volume of the ipsilateral lung minus GTV volume
ConLung	Volume of the contralateral lung
Cord	Volume of the spinal cord
Heart	Volume of the heart
Oesophagus	Volume of the oesophagus
CombinedLungs - GTV	Volume of the contralateral lung and ipsilateral lung minus the GTV
total_OAR	Volume of the heart, combined lungs (minus GTV volume) and
PTV55	Volume of PTV55
av_dist_ptv55p00_IpsLung	The average distance between PTV55 and the ipsilateral lung
av_dist_ptv55p00_ConLung	The average distance between PTV55 and the ipsilateral lung
av_dist_ptv55p00_Cord	The average distance between PTV55 and the ipsilateral lung
av_dist_ptv55p00_Heart	The average distance between PTV55 and the ipsilateral lung
av_dist_ptv55p00_Oesophagus	The average distance between PTV55 and the ipsilateral lung
av_dist_ptv55p00_CombinedLungs - GTV	The average distance between PTV55 and the ipsilateral lung
av_dist_IpsLung_ConLung	The average distance between the ipsilateral lung and the contralateral lung
av_dist_IpsLung_Cord	The average distance between the ipsilateral lung and the spinal cord
av_dist_IpsLung_Heart	The average distance between the ipsilateral lung and the heart
av_dist_IpsLung_Oesophagus	The average distance between the ipsilateral lung and the oesophagus
av_dist_ConLung_Cord	The average distance between contralateral lung and the spinal cord
av_dist_ConLung_Heart	The average distance between contralateral lung and the heart
av_dist_ConLung_Oesophagus	The average distance between contralateral lung and the oesophagus
av_dist_Cord_Heart	The average distance between the spinal cord and the heart
av_dist_Cord_Oesophagus	The average distance between the spinal cord and the oesophagus
av_dist_Cord_CombinedLungs - GTV	The average distance between the spinal cord and the spinal cord
av_dist_Heart_Oesophagus	The average distance between the heart and the oesophagus
av_dist_Heart_CombinedLungs - GTV	The average distance between the heart and the combine lungs
av_dist_Oesophagus_CombinedLungs - GTV	The average distance between the oesophagus and the combine lungs
min_dist_ptv55p00_ConLung	The smallest distance between PTV55 and the contralateral lung
min_dist_ptv55p00_Cord	The smallest distance between PTV55 and spinal cord
min_dist_IpsLung_ConLung	The smallest distance between the ipsilateral lung and the contralateral lung
min_dist_IpsLung_Cord	The smallest distance between the ipsilateral lung and the spinal cord
min_dist_IpsLung_Oesophagus	The smallest distance between the ipsilateral lung and the oesophagus

min_dist_ConLung_Cord	The smallest distance between the contralateral lung and the spinal cord
min_dist_ConLung_Oesophagus	The smallest distance between the contralateral lung and the oesophagus
min_dist_Cord_Heart	The smallest distance between the contralateral lung and the heart
min_dist_Cord_Oesophagus	The smallest distance between the spinal cord and the oesophagus
min_dist_Cord_CombinedLungs - GTV	The smallest distance between the spinal cord and the combined lungs
min_dist_Heart_Oesophagus	The smallest distance between the heart and the oesophagus
min_dist_Oesophagus_CombinedLungs - GTV	The smallest distance between the oesophagus and the combined lungs
centre_dist_ptv55p00_External	The distance from the centre of the PTV55 volume to the centre of the external volume
centre_dist_ptv55p00_IpsLung	The distance from the centre of the PTV55 volume to the centre of the ipsilateral lung volume
centre_dist_ptv55p00_ConLung	The distance from the centre of the PTV55 volume to the centre of the contralateral lung volume
centre_dist_ptv55p00_Cord	The distance from the centre of the PTV55 volume to the centre of the spinal cord volume
centre_dist_ptv55p00_Heart	The distance from the centre of the PTV55 volume to the centre of the heart volume
centre_dist_ptv55p00_Oesophagus	The distance from the centre of the PTV55 volume to the centre of the oesophagus volume
centre_dist_ptv55p00_CombinedLungs - GTV	The distance from the centre of the PTV55 volume to the centre of the combined lungs volume
centre_dist_External_IpsLung	The distance from the centre of the external volume to the centre of the ipsilateral lung volume
centre_dist_External_ConLung	The distance from the centre of the external volume to the centre of the contralateral lung volume
centre_dist_External_Cord	The distance from the centre of the external volume to the centre of the spinal cord volume
centre_dist_External_Heart	The distance from the centre of the external volume to the centre of the heart volume
centre_dist_External_Oesophagus	The distance from the centre of the external volume to the centre of the oesophagus volume
centre_dist_External_CombinedLungs - GTV	The distance from the centre of the external volume to the centre of the combined lungs volume
centre_dist_IpsLung_ConLung	The distance from the centre of the ipsilateral lung volume to the centre of the contralateral lung volume
centre_dist_IpsLung_Cord	The distance from the centre of the ipsilateral lung volume to the centre of the spinal cord volume
centre_dist_IpsLung_Heart	The distance from the centre of the ipsilateral lung volume to the centre of the heart volume

centre_dist_IpsLung_Oesophagus	The distance from the centre of the ipsilateral lung volume to the centre of the oesophagus volume
centre_dist_IpsLung_CombinedLungs - GTV	The distance from the centre of the ipsilateral lung volume to the centre of the combined lungs volume
centre_dist_ConLung_Cord	The distance from the centre of the contralateral lung volume to the centre of the spinal cord volume
centre_dist_ConLung_Heart	The distance from the centre of the contralateral lung volume to the centre of the heart volume
centre_dist_ConLung_Oesophagus	The distance from the centre of the contralateral lung volume to the centre of the oesophagus volume
centre_dist_ConLung_CombinedLungs - GTV	The distance from the centre of the contralateral lung volume to the centre of the combined lungs volume
centre_dist_Cord_Heart	The distance from the centre of the spinal cord volume to the centre of the heart volume
centre_dist_Cord_Oesophagus	The distance from the centre of the spinal cord volume to the centre of the oesophagus volume
centre_dist_Cord_CombinedLungs - GTV	The distance from the centre of the spinal cord volume to the centre of the combined lungs volume
centre_dist_Heart_Oesophagus	The distance from the centre of the heart volume to the centre of the oesophagus volume
centre_dist_Heart_CombinedLungs - GTV	The distance from the centre of the heart volume to the centre of the combined lungs volume
centre_dist_Oesophagus_CombinedLungs - GTV	The distance from the centre of the oesophagus volume to the centre of the combined lungs volume
max_dist_ptv55p00_IpsLung	The largest distance beween PTV55 to the the ipsilateral lung volume
max_dist_ptv55p00_ConLung	The largest distance beween PTV55 to the the contralateral lung volume
max_dist_ptv55p00_Cord	The largest distance beween PTV55 to the the spinal cord volume
max_dist_ptv55p00_Heart	The largest distance beween PTV55 to the the heart volume
max_dist_ptv55p00_Oesophagus	The largest distance beween PTV55 to the the oesophagus volume
max_dist_ptv55p00_CombinedLungs - GTV	The largest distance beween PTV55 to the the combined lungs volume
max_dist_IpsLung_ConLung	The largest distance beween the ipsilateral lung to the the contralateral lung volume
max_dist_IpsLung_Cord	The largest distance beween the ipsilateral lung volume to the the spinal cord volume
max_dist_IpsLung_Heart	The largest distance beween the ipsilateral lung volume to the the heart volume
max_dist_IpsLung_Oesophagus	The largest distance beween the ipsilateral lung volume to the the oesophagus volume
max_dist_ConLung_Cord	The largest distance beween the contralateral lung volume to the the spinal cord volume

max_dist_ConLung_Heart	The largest distance beween the contralateral lung volume to the the heart volume
max_dist_ConLung_Oesophagus	The largest distance beween the contralateral lung volume to the the oesophagus volume
max_dist_Cord_Heart	The largest distance beween the spinal cord volume to the the heart volume
max_dist_Cord_Oesophagus	The largest distance beween the spinal cord volume to the the oesophagus volume
max_dist_Cord_CombinedLungs - GTV	The largest distance beween the spinal cord volume to the the combined lungs volume
max_dist_Heart_Oesophagus	The largest distance beween the heart volume to the the oesophagus volume
max_dist_Heart_CombinedLungs - GTV	The largest distance beween the heart volume to the the combined lungs volume
max_dist_Oesophagus_CombinedLungs - GTV	The largest distance beween the oesophagus volume to the the combined lungs volume
ptv55p00_0p0_IpsLung	The volume of PTV55 expanded 0cm isoptropically overlapping the ipsilateral lung
ptv55p00_0p2_IpsLung	The volume of PTV55 expanded 0.2cm isoptropically overlapping the ipsilateral lung
ptv55p00_0p4_IpsLung	The volume of PTV55 expanded 0.4cm isoptropically overlapping the ipsilateral lung
ptv55p00_0p6_IpsLung	The volume of PTV55 expanded 0.6cm isoptropically overlapping the ipsilateral lung
ptv55p00_0p8_IpsLung	The volume of PTV55 expanded 0.8cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p0_IpsLung	The volume of PTV55 expanded 1cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p2_IpsLung	The volume of PTV55 expanded 1.2cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p4_IpsLung	The volume of PTV55 expanded 1.4cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p6_IpsLung	The volume of PTV55 expanded 1.6cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p8_IpsLung	The volume of PTV55 expanded 1.8cm isoptropically overlapping the ipsilateral lung
ptv55p00_2p0_IpsLung	The volume of PTV55 expanded 2cm isoptropically overlapping the ipsilateral lung
ptv55p00_2p2_IpsLung	The volume of PTV55 expanded 2.2cm isoptropically overlapping the ipsilateral lung

ptv55p00_2p4_IpsLung	The volume of PTV55 expanded 2.4cm isoptropically overlapping the ipsilateral lung
ptv55p00_1p4_ConLung	The volume of PTV55 expanded 1.4cm isoptropically overlapping the contralateral lung
ptv55p00_1p6_ConLung	The volume of PTV55 expanded 1.6cm isoptropically overlapping the contralateral lung
ptv55p00_1p8_ConLung	The volume of PTV55 expanded 1.8cm isoptropically overlapping the contralateral lung
ptv55p00_2p0_ConLung	The volume of PTV55 expanded 2cm isoptropically overlapping the contralateral lung
ptv55p00_2p2_ConLung	The volume of PTV55 expanded 2.2cm isoptropically overlapping the contralateral lung
ptv55p00_2p4_ConLung	The volume of PTV55 expanded 2.4cm isoptropically overlapping the contralateral lung
ptv55p00_0p0_Heart	The volume of PTV55 expanded 0cm isoptropically overlapping the heart
ptv55p00_0p2_Heart	The volume of PTV55 expanded 0.2cm isoptropically overlapping the heart
ptv55p00_0p4_Heart	The volume of PTV55 expanded 0.4cm isoptropically overlapping the heart
ptv55p00_0p6_Heart	The volume of PTV55 expanded 0.6cm isoptropically overlapping the heart
ptv55p00_0p8_Heart	The volume of PTV55 expanded 0.8cm isoptropically overlapping the heart
ptv55p00_1p0_Heart	The volume of PTV55 expanded 1cm isoptropically overlapping the heart
ptv55p00_1p2_Heart	The volume of PTV55 expanded 1.2cm isoptropically overlapping the heart
ptv55p00_1p4_Heart	The volume of PTV55 expanded 1.4cm isoptropically overlapping the heart
ptv55p00_1p6_Heart	The volume of PTV55 expanded 1.6cm isoptropically overlapping the heart
ptv55p00_1p8_Heart	The volume of PTV55 expanded 1.8cm isoptropically overlapping the heart
ptv55p00_2p0_Heart	The volume of PTV55 expanded 2cm isoptropically overlapping the heart
ptv55p00_2p2_Heart	The volume of PTV55 expanded 2.2cm isoptropically overlapping the heart
ptv55p00_2p4_Heart	The volume of PTV55 expanded 2.4cm isoptropically overlapping the heart

ptv55p00_0p0_Oesophagus	The volume of PTV55 expanded 0cm isoptropically overlapping the oesophagus
ptv55p00_0p2_Oesophagus	The volume of PTV55 expanded 0.2cm isoptropically overlapping the oesophagus
ptv55p00_0p4_Oesophagus	The volume of PTV55 expanded 0.4cm isoptropically overlapping the oesophagus
ptv55p00_0p6_Oesophagus	The volume of PTV55 expanded 0.6cm isoptropically overlapping the oesophagus
ptv55p00_0p8_Oesophagus	The volume of PTV55 expanded 0.8cm isoptropically overlapping the oesophagus
ptv55p00_1p0_Oesophagus	The volume of PTV55 expanded 1cm isoptropically overlapping the oesophagus
ptv55p00_1p2_Oesophagus	The volume of PTV55 expanded 1.2cm isoptropically overlapping the oesophagus
ptv55p00_1p4_Oesophagus	The volume of PTV55 expanded 1.4cm isoptropically overlapping the oesophagus
ptv55p00_1p6_Oesophagus	The volume of PTV55 expanded 1.6cm isoptropically overlapping the oesophagus
ptv55p00_1p8_Oesophagus	The volume of PTV55 expanded 1.8cm isoptropically overlapping the oesophagus
ptv55p00_2p0_Oesophagus	The volume of PTV55 expanded 2cm isoptropically overlapping the oesophagus
ptv55p00_2p2_Oesophagus	The volume of PTV55 expanded 2.2cm isoptropically overlapping the oesophagus
ptv55p00_2p4_Oesophagus	The volume of PTV55 expanded 2.4cm isoptropically overlapping the oesophagus
ptv55p00_0p0_CombinedLungs - GTV	The volume of PTV55 expanded 0cm isoptropically overlapping the combined lungs
ptv55p00_0p2_CombinedLungs - GTV	The volume of PTV55 expanded 0.2cm isoptropically overlapping the combined lungs
ptv55p00_0p4_CombinedLungs - GTV	The volume of PTV55 expanded 0.4cm isoptropically overlapping the combined lungs
ptv55p00_0p6_CombinedLungs - GTV	The volume of PTV55 expanded 0.6cm isoptropically overlapping the combined lungs
ptv55p00_0p8_CombinedLungs - GTV	The volume of PTV55 expanded 0.8cm isoptropically overlapping the combined lungs
ptv55p00_1p0_CombinedLungs - GTV	The volume of PTV55 expanded 1cm isoptropically overlapping the combined lungs
ptv55p00_1p2_CombinedLungs - GTV	The volume of PTV55 expanded 1.2cm isoptropically overlapping the combined lungs

ptv55p00_1p4_CombinedLungs - GTV	The volume of PTV55 expanded 1.4cm isoptropically overlapping the combined lungs
ptv55p00_1p6_CombinedLungs - GTV	The volume of PTV55 expanded 1.6cm isoptropically overlapping the combined lungs
ptv55p00_1p8_CombinedLungs - GTV	The volume of PTV55 expanded 1.8cm isoptropically overlapping the combined lungs
ptv55p00_2p0_CombinedLungs - GTV	The volume of PTV55 expanded 2cm isoptropically overlapping the combined lungs
ptv55p00_2p2_CombinedLungs - GTV	The volume of PTV55 expanded 2.2cm isoptropically overlapping the combined lungs
ptv55p00_2p4_CombinedLungs - GTV	The volume of PTV55 expanded 2.4cm isoptropically overlapping the combined lungs
ptv55p00_0p0_0p2_slope_IpsLung	Slope between PTV55 expanded 0cm overlapping the ipsilateral lung and PTV55 expanded 0.2cm overlapping the ipsilateral lung
ptv55p00_0p2_0p4_slope_IpsLung	Slope between PTV55 expanded 0.2cm overlapping the ipsilateral lung and PTV55 expanded 0.4cm overlapping the ipsilateral lung
ptv55p00_0p4_0p6_slope_IpsLung	Slope between PTV55 expanded 0.4cm overlapping the ipsilateral lung and PTV55 expanded 0.6cm overlapping the ipsilateral lung
ptv55p00_0p6_0p8_slope_IpsLung	Slope between PTV55 expanded 0.6cm overlapping the ipsilateral lung and PTV55 expanded 0.8cm overlapping the ipsilateral lung
ptv55p00_0p8_1p0_slope_IpsLung	Slope between PTV55 expanded 0.8cm overlapping the ipsilateral lung and PTV55 expanded 1cm overlapping the ipsilateral lung
ptv55p00_1p0_1p2_slope_IpsLung	Slope between PTV55 expanded 1cm overlapping the ipsilateral lung and PTV55 expanded 1.2cm overlapping the ipsilateral lung
ptv55p00_1p2_1p4_slope_IpsLung	Slope between PTV55 expanded 1.2cm overlapping the ipsilateral lung and PTV55 expanded 1.4cm overlapping the ipsilateral lung
ptv55p00_1p4_1p6_slope_IpsLung	Slope between PTV55 expanded 1.4cm overlapping the ipsilateral lung and PTV55 expanded 1.6cm overlapping the ipsilateral lung
ptv55p00_1p6_1p8_slope_IpsLung	Slope between PTV55 expanded 1.6cm overlapping the ipsilateral lung and PTV55 expanded 1.8cm overlapping the ipsilateral lung
ptv55p00_1p8_2p0_slope_IpsLung	Slope between PTV55 expanded 1.8cm overlapping the ipsilateral lung and PTV55 expanded 2cm overlapping the ipsilateral lung
ptv55p00_2p0_2p2_slope_IpsLung	Slope between PTV55 expanded 2cm overlapping the ipsilateral lung and PTV55 expanded 2.2cm overlapping the ipsilateral lung
ptv55p00_2p2_2p4_slope_IpsLung	Slope between PTV55 expanded 2.2cm overlapping the ipsilateral lung and PTV55 expanded 2.4cm overlapping the ipsilateral lung
ptv55p00_1p2_1p4_slope_ConLung	Slope between PTV55 expanded 1.2cm overlapping the contralateral lung and PTV55 expanded 1.4cm overlapping the contralateral lung
ptv55p00_1p4_1p6_slope_ConLung	Slope between PTV55 expanded 1.4cm overlapping the contralateral lung and PTV55 expanded 1.6cm overlapping the contralateral lung

ptv55p00_1p6_1p8_slope_ConLung Slope between PTV55 expanded 1.6cm overlapping the contralateral lung and PTV55 expanded 1.8cm overlapping the contralateral lung

ptv55p00_1p8_2p0_slope_ConLung Slope between PTV55 expanded 1.8cm overlapping the contralateral lung and PTV55 expanded 2cm overlapping the contralateral lung

ptv55p00_2p0_2p2_slope_ConLung Slope between PTV55 expanded 2cm overlapping the contralateral lung and PTV55 expanded 2.2cm overlapping the contralateral lung

ptv55p00_2p2_2p4_slope_ConLung Slope between PTV55 expanded 2.2cm overlapping the contralateral lung and PTV55 expanded 2.4cm overlapping the contralateral lung

ptv55p00_0p0_0p2_slope_Heart Slope between PTV55 expanded 0cm overlapping the heart and PTV55 expanded 0.2cm overlapping the heart

ptv55p00_0p2_0p4_slope_Heart Slope between PTV55 expanded 0.2cm overlapping the heart and PTV55 expanded 0.4cm overlapping the heart

ptv55p00_0p4_0p6_slope_Heart Slope between PTV55 expanded 0.4cm overlapping the heart and PTV55 expanded 0.6cm overlapping the heart

ptv55p00_0p6_0p8_slope_Heart Slope between PTV55 expanded 0.6cm overlapping the heart and PTV55 expanded 0.8cm overlapping the heart

ptv55p00_0p8_1p0_slope_Heart Slope between PTV55 expanded 0.8cm overlapping the heart and PTV55 expanded 1cm overlapping the heart

ptv55p00_1p0_1p2_slope_Heart Slope between PTV55 expanded 1cm overlapping the heart and PTV55 expanded 1.2cm overlapping the heart

ptv55p00_1p2_1p4_slope_Heart Slope between PTV55 expanded 1.2cm overlapping the heart and PTV55 expanded 1.4cm overlapping the heart

ptv55p00_1p4_1p6_slope_Heart Slope between PTV55 expanded 1.4cm overlapping the heart and PTV55 expanded 1.6cm overlapping the heart

ptv55p00_1p6_1p8_slope_Heart Slope between PTV55 expanded 1.6cm overlapping the heart and PTV55 expanded 1.8cm overlapping the heart

ptv55p00_1p8_2p0_slope_Heart Slope between PTV55 expanded 1.8cm overlapping the heart and PTV55 expanded 2cm overlapping the heart

ptv55p00_2p0_2p2_slope_Heart Slope between PTV55 expanded 2cm overlapping the heart and PTV55 expanded 2.2cm overlapping the heart

ptv55p00_2p2_2p4_slope_Heart Slope between PTV55 expanded 2.2cm overlapping the heart and PTV55 expanded 2.4cm overlapping the heart

ptv55p00_0p0_0p2_slope_Oesophagus Slope between PTV55 expanded 0cm overlapping the oesophagus and PTV55 expanded 0.2cm overlapping the oesophagus

ptv55p00_0p2_0p4_slope_Oesophagus Slope between PTV55 expanded 0.2cm overlapping the oesophagus and PTV55 expanded 0.4cm overlapping the oesophagus

ptv55p00_0p4_0p6_slope_Oesophagus Slope between PTV55 expanded 0.4cm overlapping the oesophagus and PTV55 expanded 0.6cm overlapping the oesophagus

ptv55p00_0p6_0p8_slope_Oesophagus Slope between PTV55 expanded 0.6cm overlapping the oesophagus and PTV55 expanded 0.8cm overlapping the oesophagus

ptv55p00_0p8_1p0_slope_Oesophagus	Slope between PTV55 expanded 0.8cm overlapping the oesophagus and PTV55 expanded 1cm overlapping the oesophagus
ptv55p00_1p0_1p2_slope_Oesophagus	Slope between PTV55 expanded 1cm overlapping the oesophagus and PTV55 expanded 1.2cm overlapping the oesophagus
ptv55p00_1p2_1p4_slope_Oesophagus	Slope between PTV55 expanded 1.2cm overlapping the oesophagus and PTV55 expanded 1.4cm overlapping the oesophagus
ptv55p00_1p4_1p6_slope_Oesophagus	Slope between PTV55 expanded 1.4cm overlapping the oesophagus and PTV55 expanded 1.6cm overlapping the oesophagus
ptv55p00_1p6_1p8_slope_Oesophagus	Slope between PTV55 expanded 1.6cm overlapping the oesophagus and PTV55 expanded 1.8cm overlapping the oesophagus
ptv55p00_1p8_2p0_slope_Oesophagus	Slope between PTV55 expanded 1.8cm overlapping the oesophagus and PTV55 expanded 2cm overlapping the oesophagus
ptv55p00_2p0_2p2_slope_Oesophagus	Slope between PTV55 expanded 2cm overlapping the oesophagus and PTV55 expanded 2.2cm overlapping the oesophagus
ptv55p00_2p2_2p4_slope_Oesophagus	Slope between PTV55 expanded 2.2cm overlapping the oesophagus and PTV55 expanded 2.4cm overlapping the oesophagus
ptv55p00_0p0_0p2_slope_CombinedLungs - GTV	Slope between PTV55 expanded 0cm overlapping the combined lungs and PTV55 expanded 0.2cm overlapping the combined lungs
ptv55p00_0p2_0p4_slope_CombinedLungs - GTV	Slope between PTV55 expanded 0.2cm overlapping the combined lungs and PTV55 expanded 0.4cm overlapping the combined lungs
ptv55p00_0p4_0p6_slope_CombinedLungs - GTV	Slope between PTV55 expanded 0.4cm overlapping the combined lungs and PTV55 expanded 0.6cm overlapping the combined lungs
ptv55p00_0p6_0p8_slope_CombinedLungs - GTV	Slope between PTV55 expanded 0.6cm overlapping the combined lungs and PTV55 expanded 0.8cm overlapping the combined lungs
ptv55p00_0p8_1p0_slope_CombinedLungs - GTV	Slope between PTV55 expanded 0.8cm overlapping the combined lungs and PTV55 expanded 1cm overlapping the combined lungs
ptv55p00_1p0_1p2_slope_CombinedLungs - GTV	Slope between PTV55 expanded 1cm overlapping the combined lungs and PTV55 expanded 1.2cm overlapping the combined lungs
ptv55p00_1p2_1p4_slope_CombinedLungs - GTV	Slope between PTV55 expanded 1.2cm overlapping the combined lungs and PTV55 expanded 1.4cm overlapping the combined lungs
ptv55p00_1p4_1p6_slope_CombinedLungs - GTV	Slope between PTV55 expanded 1.4cm overlapping the combined lungs and PTV55 expanded 1.6cm overlapping the combined lungs
ptv55p00_1p6_1p8_slope_CombinedLungs - GTV	Slope between PTV55 expanded 1.6cm overlapping the combined lungs and PTV55 expanded 1.8cm overlapping the combined lungs
ptv55p00_1p8_2p0_slope_CombinedLungs - GTV	Slope between PTV55 expanded 1.8cm overlapping the combined lungs and PTV55 expanded 2cm overlapping the combined lungs
ptv55p00_2p0_2p2_slope_CombinedLungs - GTV	Slope between PTV55 expanded 2cm overlapping the combined lungs and PTV55 expanded 2.2cm overlapping the combined lungs
ptv55p00_2p2_2p4_slope_CombinedLungs - GTV	Slope between PTV55 expanded 2.2cm overlapping the combined lungs and PTV55 expanded 2.4cm overlapping the combined lungs

VIF_ptv55p00_IpsLung	Volume of the ipsilateral lung between the most superior and most inferior transverse slice of PTV55
VIF_ptv55p00_ConLung	Volume of the contralateral lung between the most superior and most inferior transverse slice of PTV55
VIF_ptv55p00_Cord	Volume of the spinal cord between the most superior and most inferior transverse slice of PTV55
VIF_ptv55p00_Heart	Volume of the heart between the most superior and most inferior transverse slice of PTV55
VIF_ptv55p00_Oesophagus	Volume of the oesophagus between the most superior and most inferior transverse slice of PTV55
VIF_ptv55p00_CombinedLungs - GTV	Volume of the combined lungs between the most superior and most inferior transverse slice of PTV55
VOF_ptv55p00_IpsLung	Volume of the ipsilateral lung above the most superior and below the most inferior transverse slice of PTV55
VOF_ptv55p00_ConLung	Volume of the contralateral lung above the most superior and below the most inferior transverse slice of PTV55
VOF_ptv55p00_Cord	Volume of the spinal cord above the most superior and below the most inferior transverse slice of PTV55
VOF_ptv55p00_Heart	Volume of the heart above the most superior and below the most inferior transverse slice of PTV55
VOF_ptv55p00_Oesophagus	Volume of the oesophagus above the most superior and below the most inferior transverse slice of PTV55
VOF_ptv55p00_CombinedLungs - GTV	Volume of the combined lungs above the most superior and below the most inferior transverse slice of PTV55
Total_VIF_PTV55	Volume of the combined lungs, heart and oesophagus between the most superior and most inferior transverse slice of PTV55
Total_VOF_PTV55	Volume of the combined lungs, heart and oesophagus above the most superior and below the most inferior transverse slice of PTV55
ratio_ptv55p00_External	Volume of PTV55 divided by the external
ratio_ptv55p00_IpsLung	Volume of PTV55 divided by the ipsilateral lung
ratio_ptv55p00_ConLung	Volume of PTV55 divided by the contralateral lung
ratio_ptv55p00_Cord	Volume of PTV55 divided by the spinal cord
ratio_ptv55p00_Heart	Volume of PTV55 divided by the heart
ratio_ptv55p00_Oesophagus	Volume of PTV55 divided by the oesophagus
ratio_ptv55p00_CombinedLungs - GTV	Volume of PTV55 divided by the combined lungs
ratio_External_IpsLung	Volume of the external divided by the ipsilateral lung
ratio_External_ConLung	Volume of the external divided by the contralateral lung
ratio_External_Cord	Volume of the external divided by the spinal cord
ratio_External_Heart	Volume of the external divided by the heart
ratio_External_Oesophagus	Volume of the external divided by the oesophagus
ratio_External_CombinedLungs - GTV	Volume of the external divided by the combined lungs

ratio_IpsLung_ConLung	Volume of the ipsilalteral lung divided by the contralateral lung
ratio_IpsLung_Cord	Volume of the ipsilalteral lung divided by the spinal cord
ratio_IpsLung_Heart	Volume of the ipsilalteral lung divided by the heart
ratio_IpsLung_Oesophagus	Volume of the ipsilalteral lung divided by the oesophagus
ratio_IpsLung_CombinedLungs - GTV	Volume of the ipsilalteral lung divided by the combined lungs
ratio_ConLung_Cord	Volume of the contralateral lung divided by the spinal cord
ratio_ConLung_Heart	Volume of the contralateral lung divided by the heart
ratio_ConLung_Oesophagus	Volume of the contralateral lung divided by the oesophagus
ratio_ConLung_CombinedLungs - GTV	Volume of the contralateral lung divided by the combined lungs
ratio_Cord_Heart	Volume of the spinal cord divided by the heart
ratio_Cord_Oesophagus	Volume of the spinal cord divided by the oesophagus
ratio_Cord_CombinedLungs - GTV	Volume of the spinal cord divided by the combined lungs
ratio_Heart_Oesophagus	Volume of the spinal cord divided by the oesophagus
ratio_Heart_CombinedLungs - GTV	Volume of the heart divided by the combined lungs
ratio_Oesophagus_CombinedLungs - GTV	Volume of the oesophagus divided by the combined lungs

Appendix D

Regression python script

```
import time
import pandas as pd
import numpy as np
import ast
from sklearn.cross_validation import LeaveOneOut
import sklearn.linear_model as sm
from sklearn.metrics import mean_squared_error, r2_score

from sklearn.preprocessing import PolynomialFeatures
from sklearn.pipeline import make_pipeline

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
from sklearn.decomposition import PCA
pca = PCA()




##preparing the combinations of variable to be input into
   ↪ the MLR
def combinations(iterable, r):
   pool = tuple(iterable)
```

```
    n = len(pool)
    if r > n:
        return
    indices = range(r)
    yield list(pool[i] for i in indices)
    while True:
        for i in reversed(range(r)):
            if indices[i] != i + n - r:
                break
        else:
            return
        indices[i] += 1
        for j in range(i+1, r):
            indices[j] = indices[j-1] + 1
        yield list(pool[i] for i in indices)









def multiregRaw(site='PSV',metric='Weight',to_predict='
    ↪ Bladder',n_features=2,n_degrees=1,leaveout=1,same=1,
    ↪ select_features=[]):

    polyreg=make_pipeline(PolynomialFeatures(n_degrees),sm.
        ↪ LinearRegression())

    ##import data from excel
    File= "H:\Users\IonaF\DataWorkbook"+site+".xlsx"
    X_train = pd.read_excel(File,sheetname='
        ↪ FeaturesTrainingForRaw',index_col='Patient_ID')
```

```
y_train = pd.read_excel(File,sheetname=’WeightsTraining
   ↪ ’,index_col=’Patient_ID’)
y_train = y_train[y_train[’Metric’]==metric].loc[:,
   ↪ y_train.columns != ’Metric’]
y_train = y_train.loc[y_train[to_predict] != 0]
y_train.dropna(axis=1, how=’any’, inplace=True) #if
   ↪ there still remain predictive features with null
   ↪ values, delete.
X_train = X_train.loc[y_train.index]

if len(select_features) == 0:
   combis = combinations(X_train.columns, n_features)
else:
   combis = [select_features]

##initialising variables that will populate the "final"
   ↪  dictionary that will be converted to the "
   ↪ final_df" dataframe
for i in range(n_features):
   exec ’feat%s=[]’ % str(i+1)
r2 = []
leftout = []
novel = []
existing = []
predicting = []
feat_count=[]
poly_degree=[]
actuals=[]
predictions=[]
featsALL=[]
final = {}
SSE = []
```

```
if leaveout == 0:
    scaler.fit(X_train)
    if same == 0: #predicting for novel cases. Never a
        ↪ L1O with such data as all data is unseen
        X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)
        X_test = pd.read_excel(File,sheetname='
            ↪ FeaturesTestingForRaw',index_col='Patient_ID
            ↪ ')
        X_test.dropna(axis=1, how='any', inplace=True) #
            ↪ if there still remain predictive features
            ↪ with null values, delete.
        X_test = pd.DataFrame(scaler.transform(X_test),
            ↪ columns=X_test.columns,index=X_test.index)

        y_test = pd.read_excel(File,sheetname='
            ↪ WeightsTesting',index_col='Patient_ID')
        y_test = y_test[y_test['Metric']==metric].loc[:,
            ↪ y_test.columns != 'Metric']
        y_test = y_test[y_test[to_predict].notnull()] #
            ↪ clean for missing data
        y_test = y_test.loc[y_test[to_predict] != 0]
        X_test = X_test.loc[y_test.index]

        training_patients=y_train.index.tolist()
        testing_patients =y_test.index.tolist()
        novel_patients=[x for x in testing_patients if x
            ↪ not in training_patients] #shouldn't be any
            ↪ overlap but just a precaution

        for feature_set in combis:
            for novel_patient in novel_patients:
```

```
            for feat_count_minus1, current_feat in
                ↪ enumerate(feature_set):
                exec 'feat%s+=[current_feat]' % str(
                    ↪ feat_count_minus1+1) #splitting the
                    ↪  features
            novel += [novel_patient]
            predicting += [to_predict]
            feat_count += [n_features]
            poly_degree += [n_degrees]

            polyreg.fit(X_train[feature_set],y_train[
                ↪ to_predict])
            r2 += [polyreg.score(X_train[feature_set],
                ↪ y_train[to_predict])]

            X_test_ = X_test[feature_set].loc[
                ↪ novel_patient]#include only the novel
                ↪ patients
            y_test_ = y_test[to_predict].loc[
                ↪ novel_patient]

            y_pred = polyreg.predict(X_test_).tolist()
            actuals+=[y_test_]
            predictions+=y_pred
            featsALL+=[','.join(map(str, feature_set))]

        for feat_count_minus1, feat in enumerate(
            ↪ feature_set):
            exec "final['feat%s']=feat%s" % (str(
                ↪ feat_count_minus1+1), str(
                ↪ feat_count_minus1+1))
        final['r2'] = r2
        final['predicting'] = predicting
```

```
         final['number_of_feats']=feat_count
         final['degree'] = poly_degree
         final['novel␣patient']=novel
         final['actual␣y'] = actuals
         final['predicted␣y'] = predictions
         final['featsALL'] = featsALL
         final_df = pd.DataFrame(final)
         final_df['diff']=final_df['actual␣y']-final_df
            ↪ ['predicted␣y']
         final_df['perc␣diff']=(final_df['actual␣y']-
            ↪ final_df['predicted␣y'])/final_df['actual
            ↪ ␣y']
         final_df['squared␣error']=final_df['diff']**2
         col_list_order = ['predicting','
            ↪ number_of_feats','degree','novel␣patient'
            ↪ ,'feat1','feat2','feat3','feat4','feat5',
            ↪ 'r2','actual␣y','predicted␣y','squared␣
            ↪ error','diff','perc␣diff','featsALL']
         col_order=[]
         for col in col_list_order:
            if col in final_df.columns:
               col_order+=[col]
         final_df = final_df[col_order]


   elif same == 1: #no L1O but using training data.
      ↪ Used to view the final model created by the
      ↪ training data
      training_patients=X_train.index.tolist()
      X_train = pd.DataFrame(scaler.transform(X_train),
         ↪ columns=X_train.columns,index=X_train.index)
      for feature_set in combis:
         for training_patient in training_patients:
```

```
      for feat_count_minus1, current_feat in
         ↪ enumerate(feature_set):
         exec 'feat%s+=[current_feat]' % str(
            ↪ feat_count_minus1+1) #splitting the
            ↪  features
      existing += [training_patient]
      predicting += [to_predict]
      feat_count += [n_features]
      poly_degree += [n_degrees]

      polyreg.fit(X_train[feature_set],y_train[
         ↪ to_predict])

      r2 += [polyreg.score(X_train[feature_set],
         ↪ y_train[to_predict])]

      X_train_ = X_train[feature_set].loc[
         ↪ training_patient]
      y_train_ = y_train[to_predict].loc[
         ↪ training_patient]

      y_pred = polyreg.predict(X_train_).tolist()
      actuals+=[y_train_]
      predictions+=y_pred
      featsALL+=[','.join(map(str, feature_set))]

for feat_count_minus1, feat in enumerate(
   ↪ feature_set):
   exec "final['feat%s']=feat%s" % (str(
      ↪ feat_count_minus1+1), str(
      ↪ feat_count_minus1+1))
final['r2'] = r2
final['predicting'] = predicting
```

```
            final[’number_of_feats’]=feat_count
            final[’degree’] = poly_degree
            final[’existing␣patient’]=existing
            final[’actual␣y’] = actuals
            final[’predicted␣y’] = predictions
            final[’featsALL’] = featsALL
            final_df = pd.DataFrame(final)
            final_df[’diff’]=final_df[’actual␣y’]-final_df[’
               ↪ predicted␣y’]
            final_df[’perc␣diff’]=(final_df[’actual␣y’]-
               ↪ final_df[’predicted␣y’])/final_df[’actual␣y’
               ↪ ]
            final_df[’squared␣error’]=final_df[’diff’]**2
            col_list_order = [’predicting’,’number_of_feats’,
               ↪ ’degree’,’existing␣patient’,’feat1’,’feat2’,
               ↪ ’feat3’,’feat4’,’feat5’,’r2’,’actual␣y’,’
               ↪ predicted␣y’,’squared␣error’,’diff’,’perc␣
               ↪ diff’,’featsALL’]
            col_order=[]
            for col in col_list_order:
                if col in final_df.columns:
                    col_order+=[col]
            final_df = final_df[col_order]



    elif leaveout == 1: #same is redundant here. Only
        ↪ trainin data will ever be used n this step
        loo = LeaveOneOut(n=len(y_train))
        for feature_set in combis:
            leftout_ = []
            predicting_ = []
            feat_count_ = []
            poly_degree_ = []
```

```
r2_ = []
y_pred_ = []
actuals_ = []
predictions_ = []
featsALL_ = []
for included_patients,left_out_patient in loo:
   for feat_count_minus1, current_feat in
      ↪ enumerate(feature_set):
      exec 'feat%s+=[current_feat]' % str(
         ↪ feat_count_minus1+1)

   leftout_ += X_train.iloc[left_out_patient].
      ↪ index.tolist()
   predicting_ += [to_predict]
   feat_count_ += [n_features]
   poly_degree_ += [n_degrees]

   X_train_left_in = X_train[feature_set].iloc[
      ↪ included_patients]
   scaler.fit(X_train_left_in)
   X_train_left_in = scaler.transform(
      ↪ X_train_left_in)

   X_train_left_out = X_train[feature_set].iloc[
      ↪ left_out_patient]
   X_train_left_out = scaler.transform(
      ↪ X_train_left_out)

   y_train_left_in = y_train[to_predict].iloc[
      ↪ included_patients]
   y_train_left_out = y_train[to_predict].iloc[
      ↪ left_out_patient]
```

```
            polyreg.fit(X_train_left_in,y_train_left_in)

            r2_ += [polyreg.score(X_train_left_in,
                ↪ y_train_left_in)]

            y_pred_ += polyreg.predict(X_train_left_out).
                ↪ tolist()
            actuals_ += y_train_left_out.tolist()
            predictions_ = y_pred_
            featsALL_ += [','.join(map(str, feature_set))]
            #for feat_count, feat in enumerate(feature_set
                ↪ ):
            # exec "final['feat%s']=feat%s" % (str(
                ↪ feat_count+1), str(feat_count+1))
            SSE_ = sum([i**2 for i in np.subtract(y_pred_,
                ↪ actuals_).tolist()])

        if len(SSE)==0:
            leftout = leftout_
            predicting = predicting_
            feat_count = feat_count_
            poly_degree = poly_degree_
            r2 = r2_
            actuals = actuals_
            predictions = predictions_
            featsALL = featsALL_
        elif len(SSE)>0 and SSE_ < min(SSE):
            leftout = leftout_
            predicting = predicting_
            feat_count = feat_count_
            poly_degree = poly_degree_
            r2 = r2_
            actuals = actuals_
```

```
                predictions = predictions_
                featsALL = featsALL_
            SSE += [SSE_]



        final[’r2’] = r2
        final[’predicting’] = predicting
        final[’number_of_feats’]=feat_count
        final[’degree’] = poly_degree
        final[’left␣out␣patient’]=leftout
        final[’actual␣y’] = actuals
        final[’predicted␣y’] = predictions
        final[’featsALL’] = featsALL
        final_df = pd.DataFrame(final)
        final_df[’diff’]=final_df[’actual␣y’]-final_df[’
            ↪ predicted␣y’]
        final_df[’perc␣diff’]=(final_df[’actual␣y’]-final_df
            ↪ [’predicted␣y’])/final_df[’actual␣y’]
        final_df[’squared␣error’]=final_df[’diff’]**2
        final_df[’SSE’] = final_df.groupby(’featsALL’)[’
            ↪ squared␣error’].transform(’sum’)
        col_list_order = [’predicting’,’number_of_feats’,’
            ↪ degree’,’left␣out␣patient’,’feat1’,’feat2’,’
            ↪ feat3’,’feat4’,’feat5’,’r2’,’actual␣y’,’
            ↪ predicted␣y’,’squared␣error’,’diff’,’perc␣diff’
            ↪ ,’featsALL’]
        col_order=[]
        for col in col_list_order:
            if col in final_df.columns:
                col_order+=[col]
        final_df = final_df[col_order]

    return final_df
```

```
def multiregPCA(site='PSV',metric='Weight',to_predict='
    ↪ Bladder',n_features=2,n_degrees=1,leaveout=1,same=1):

    polyreg=make_pipeline(PolynomialFeatures(n_degrees),sm.
        ↪ LinearRegression())

    ##import data from excel
    File= "H:\Users\IonaF\DataWorkbook"+site+".xlsx"
    X_train = pd.read_excel(File,sheetname='
        ↪ FeaturesTrainingForPCA',index_col='Patient_ID')
    y_train = pd.read_excel(File,sheetname='WeightsTraining
        ↪ ',index_col='Patient_ID')
    y_train = y_train[y_train['Metric']==metric].loc[:,
        ↪ y_train.columns != 'Metric']
    y_train = y_train.loc[y_train[to_predict] != 0]
    y_train.dropna(axis=1, how='any', inplace=True)
    X_train = X_train.loc[y_train.index]


    ##initialising variables that will populate the "final"
        ↪  dictionary that will be converted to the "
        ↪ final_df" dataframe
    for i in range(n_features):
        exec 'feat%s=[]' % str(i+1)
    r2 = []
```

```
leftout = []
novel = []
existing = []
predicting = []
feat_count=[]
poly_degree=[]
y_pred_name=[]
actuals=[]
predictions=[]
featsALL=[]
predict_time=[]
fit_time=[]
final = {}

if leaveout == 0:
    scaler.fit(X_train)
    training_patients=y_train.index.tolist()
    if same == 0: #predicting for novel cases. Never a
        ↪ L1O with such data as all data is unseen
        y_test = pd.read_excel(File,sheetname=’
            ↪ WeightsTesting’,index_col=’Patient_ID’)
        y_test = y_test[y_test[’Metric’]==metric].loc[:,
            ↪ y_test.columns != ’Metric’]
        y_test = y_test[y_test[to_predict].notnull()] #
            ↪ clean for missing data
        y_test = y_test.loc[y_test[to_predict] != 0]

        X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)
        X_test = pd.read_excel(File,sheetname=’
            ↪ FeaturesTestingForPCA’,index_col=’Patient_ID
            ↪ ’)
```

```
X_test.dropna(axis=1, how='any', inplace=True) #
    ↪ if there still remain predictive features
    ↪ with null values, delete.
X_test = X_test.loc[y_test.index]
X_test = pd.DataFrame(scaler.transform(X_test),
    ↪ columns=X_test.columns,index=X_test.index)

pca.fit(X_train)
PC_cols=["PC"+str.zfill(str(x+1),2) for x in
    ↪ range(min(X_train.shape))]
X_train = pd.DataFrame(pca.transform(X_train),
    ↪ columns=PC_cols)
X_test = pd.DataFrame(pca.transform(X_test),
    ↪ columns=PC_cols)

testing_patients =y_test.index.tolist()
novel_patients=[x for x in testing_patients if x
    ↪ not in training_patients] #shouldn't be any
    ↪ overlap but just a precaution
feature_set=PC_cols[:n_features]
for novel_i,novel_patient in enumerate(
    ↪ novel_patients):
    novel += [novel_patient]
    predicting += [to_predict]
    feat_count += [n_features]
    poly_degree += [n_degrees]

    polyreg.fit(X_train[feature_set],y_train[
        ↪ to_predict])
    r2 += [polyreg.score(X_train[feature_set],
        ↪ y_train[to_predict])]
```

```
            X_test_ = X_test[feature_set].iloc[novel_i]#
               ↪ include only the novel patients
            y_test_ = y_test[to_predict].iloc[novel_i]

            y_pred = polyreg.predict(X_test_).tolist()
            actuals+=[y_test_]
            predictions+=y_pred
            featsALL+=[',’.join(map(str, feature_set))]

        final['r2'] = r2
        final['predicting'] = predicting
        final['number_of_PCs']=feat_count
        final['degree'] = poly_degree
        final['novel␣patient']=novel
        final['actual␣y'] = actuals
        final['predicted␣y'] = predictions
        final['featsALL'] = featsALL
        final_df = pd.DataFrame(final)
        final_df['diff']=final_df['actual␣y']-final_df['
           ↪ predicted␣y']
        final_df['perc␣diff']=(final_df['actual␣y']-
           ↪ final_df['predicted␣y'])/final_df['actual␣y'
           ↪ ]
        final_df['SSE'] = final_df.groupby('featsALL')['
           ↪ squared␣error'].transform('sum')
        final_df=final_df[final_df['SSE']==min(final_df['
           ↪ SSE'])]
        #final_df['squared error']=final_df['diff']**2
        col_list_order = ['predicting','number_of_PCs','
           ↪ degree','novel␣patient','r2','actual␣y','
           ↪ predicted␣y','squared␣error','diff','perc␣
           ↪ diff','featsALL']
        col_order=[]
        for col in col_list_order:
            if col in final_df.columns:
                col_order+=[col]
        final_df = final_df[col_order]


    elif same == 1: #no L1O but using training data.
        ↪ Used to view the final model created by the
        ↪ training data
        training_patients=y_train.index.tolist()
        X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)
        pca.fit(X_train)
        PC_cols=["PC"+str.zfill(str(x+1),2) for x in
            ↪ range(min(X_train.shape))]
        X_train = pd.DataFrame(pca.transform(X_train),
            ↪ columns=PC_cols)
        feature_set=PC_cols[:n_features]
        for training_i,training_patient in enumerate(
            ↪ training_patients):
            existing += [training_patient]
            predicting += [to_predict]
            feat_count += [n_features]
            poly_degree += [n_degrees]


            polyreg.fit(X_train[feature_set],y_train[
                ↪ to_predict])


            r2 += [polyreg.score(X_train[feature_set],
                ↪ y_train[to_predict])]


            X_train_ = X_train[feature_set].iloc[
                ↪ training_i]
            y_train_ = y_train[to_predict].iloc[training_i
                ↪ ]

            y_pred = polyreg.predict(X_train_).tolist()
            actuals+=[y_train_]
            predictions+=y_pred
            featsALL+=[','.join(map(str, feature_set))]

        final['r2'] = r2
        final['predicting'] = predicting
        final['number_of_PCs']=feat_count
        final['degree'] = poly_degree
        final['existing␣patient']=existing
        final['actual␣y'] = actuals
        final['predicted␣y'] = predictions
        final['featsALL'] = featsALL
        final_df = pd.DataFrame(final)
        final_df['diff']=final_df['actual␣y']-final_df['
            ↪ predicted␣y']
        final_df['perc␣diff']=(final_df['actual␣y']-
            ↪ final_df['predicted␣y'])/final_df['actual␣y'
            ↪ ]
        final_df['squared␣error']=final_df['diff']**2
        final_df['SSE'] = final_df.groupby('featsALL')['
            ↪ squared␣error'].transform('sum')
        final_df=final_df[final_df['SSE']==min(final_df['
            ↪ SSE'])]
        #final_df=final_df[final_df['r2']==max(final_df['
            ↪ r2'])]
        col_list_order = ['predicting','number_of_PCs','
            ↪ degree','existing␣patient','r2','actual␣y','
            ↪ predicted␣y','squared␣error','diff','perc␣
            ↪ diff','featsALL']
```

```
      col_order=[]
      for col in col_list_order:
         if col in final_df.columns:
            col_order+=[col]
      final_df = final_df[col_order]



elif leaveout == 1: #same is redundant here.
   loo = LeaveOneOut(n=len(y_train))
   for included_patients,left_out_patient in loo:

      leftout += X_train.iloc[left_out_patient].index.
         ↪ tolist()
      predicting += [to_predict]
      feat_count += [n_features]
      poly_degree += [n_degrees]

      X_train_left_in = X_train.iloc[included_patients]
      scaler.fit(X_train_left_in)
      X_train_left_in = scaler.transform(
         ↪ X_train_left_in)
      pca.fit(X_train_left_in)
      X_train_left_in = pca.transform(X_train_left_in)
      PC_cols=["PC"+str.zfill(str(x+1),2) for x in
         ↪ range(min(X_train_left_in.shape))]
      X_train_left_in = pd.DataFrame(X_train_left_in,
         ↪ columns=PC_cols)

      X_train_left_out = X_train.iloc[left_out_patient]
      X_train_left_out = pd.DataFrame(scaler.transform(
         ↪ X_train_left_out),columns=X_train_left_out.
         ↪ columns,index=X_train_left_out.index)
```

```
        X_train_left_out = pd.DataFrame(pca.transform(
           ↪ X_train_left_out),columns=PC_cols)
        feature_set=PC_cols[:n_features]

        y_train_left_in = y_train[to_predict].iloc[
           ↪ included_patients]
        y_train_left_out = y_train[to_predict].iloc[
           ↪ left_out_patient]

        polyreg.fit(X_train_left_in[feature_set],
           ↪ y_train_left_in)
        #r2 += [pd.to_numeric(polyreg.score(
           ↪ X_train_left_in[feature_set],y_train_left_in
           ↪ [feature_set]), errors=’coerce’).isnull()]
        r2 += [polyreg.score(X_train_left_in[feature_set
           ↪ ],y_train_left_in)]

        y_pred = polyreg.predict(X_train_left_out[
           ↪ feature_set])
        #y_pred_name+=X_train.index[left_out_patient]
        actuals+=y_train_left_out.tolist()
        predictions+=y_pred.tolist()
        featsALL+=[’,’.join(map(str, feature_set))]

    final[’r2’] = r2
    final[’predicting’] = predicting
    final[’number_of_PCs’]=feat_count
    final[’degree’] = poly_degree
    final[’left␣out␣patient’]=leftout
    final[’actual␣y’] = actuals
    final[’predicted␣y’] = predictions
    final[’featsALL’] = featsALL
    final_df = pd.DataFrame(final)
```

```
        final_df['diff']=final_df['actual y']-final_df['predicted y']
        final_df['perc diff']=(final_df['actual y']-final_df['predicted y'])/final_df['actual y']
        final_df['squared error']=final_df['diff']**2
        final_df['SSE'] = final_df.groupby('featsALL')['squared error'].transform('sum')
        col_list_order = ['predicting','number_of_PCs','degree','left out patient','r2','actual y','predicted y','squared error','diff','perc diff','featsALL']
        col_order=[]
        for col in col_list_order:
            if col in final_df.columns:
                col_order+=[col]
        final_df = final_df[col_order]

    return final_df
```

Appendix E

Clustering python script

```
import pandas as pd
import numpy as np
import sklearn.cluster as clus
from sklearn.metrics import mean_squared_error, r2_score,
   ↪ silhouette_samples, silhouette_score
from sklearn.model_selection import LeaveOneOut
loo = LeaveOneOut()

from sklearn.preprocessing import StandardScaler
scaler = StandardScaler()
from sklearn.decomposition import PCA
pca = PCA()

Stem = ’/Users/ionafoster/Desktop/ModellingCodes/’

def clusteringmethodRaw(site=’PSV’,metric=’Weight’,
   ↪ n_clusts=1,leaveout=0,same=0):

   kmeans = clus.KMeans(
            init="random",
            n_clusters=n_clusts,
            n_init=10,
            max_iter=300,
            random_state=42
            )

   ##import data from excel
   File= Stem+"DataWorkbook"+site+".xlsx"
   X_train = pd.read_excel(File,sheet_name=’
      ↪ FeaturesTrainingForRaw’,index_col=’Patient_ID’).
      ↪ fillna(0)
   y_train = pd.read_excel(File,sheet_name=’
      ↪ WeightsTraining’,index_col=’Patient_ID’)
   y_train = y_train[y_train[’Metric’]==metric].loc[:,
      ↪ y_train.columns != ’Metric’]
   y_train.dropna(axis=1, how=’any’, inplace=True) #if
      ↪ there still remain predictive features with null
      ↪ values, delete.
   y_train = y_train.replace(0, np.NaN)

   ##initialising variables that will populate the "final"
      ↪  dictionary that will be converted to the "
      ↪ final_df" dataframe
   actuals=[]
   predictions=[]
   sse=[]
   final = {}



   if leaveout==0:
      scaler.fit(X_train)
      if same==0:
         X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)
            ↪ .fillna(0)
```

```
X_test = pd.read_excel(File,sheet_name=’
    ↪ FeaturesTestingForRaw’,index_col=’Patient_ID
    ↪ ’).fillna(0)
X_test.dropna(axis=1, how=’any’, inplace=True) #
    ↪ if there still remain predictive features
    ↪ with null values, delete.
X_test = pd.DataFrame(scaler.transform(X_test),
    ↪ columns=X_test.columns,index=X_test.index)

y_test = pd.read_excel(File,sheet_name=’
    ↪ WeightsTesting’,index_col=’Patient_ID’)
y_test = y_test[y_test[’Metric’]==metric].loc[:,
    ↪ y_test.columns != ’Metric’]
y_test.dropna(axis=1, how=’any’, inplace=True)
y_test = y_test.replace(0, np.NaN)

kmeans.fit(X_train)
train_clus_groups = kmeans.predict(X_train)+1
y_train[’Cluster_Group’]=train_clus_groups
means = y_train.replace(0, np.NaN).groupby(’
    ↪ Cluster_Group’).mean()

test_clus_groups = kmeans.predict(X_test)+1
y_test[’Cluster_Group’]=test_clus_groups
y_pred = y_test[[’Cluster_Group’]].merge(means,
    ↪ left_on=’Cluster_Group’,right_index=True)

final[’actual␣y’] = y_test
final[’predicted␣y’] = y_pred
final[’sse’] = kmeans.inertia_
if n_clusts>1 and n_clusts<len(X_train):
    final[’silhouette_avg’] = silhouette_score(
        ↪ X_train, train_clus_groups)
```

```
            final['sample_silhouette_values'] = pd.
                ↪ DataFrame(silhouette_samples(X_train,
                ↪ train_clus_groups),columns=['
                ↪ silhouette_values'],index=X_train.index)
        diff_df=y_test.drop('Cluster_Group',axis=1)-
            ↪ y_pred.drop('Cluster_Group',axis=1)
        percent_diff_df=(y_test.drop('Cluster_Group',axis
            ↪ =1)-y_pred.drop('Cluster_Group',axis=1))/
            ↪ y_test.drop('Cluster_Group',axis=1)
        summary={}
        summary['MSE']=np.square(diff_df).replace(0, np.
            ↪ NaN).mean()
        summary['mean_diff']=diff_df.replace(0, np.NaN).
            ↪ mean()
        summary['median_diff']=diff_df.replace(0, np.NaN)
            ↪ .median()
        summary['mean_perc_diff']=percent_diff_df.replace
            ↪ (0, np.NaN).mean()
        summary['median_perc_diff']=percent_diff_df.
            ↪ replace(0, np.NaN).median()
        final['summary']=pd.DataFrame(summary)

    elif same==1:
        X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)

        training_patients=X_train.index.tolist()

        kmeans.fit(X_train)
        train_clus_groups = kmeans.predict(X_train)+1
        y_train['Cluster_Group']=train_clus_groups
        means = y_train.replace(0, np.NaN).groupby('
            ↪ Cluster_Group').mean()
        y_pred = y_train[['Cluster_Group']].merge(means,
            ↪ left_on='Cluster_Group',right_index=True)

        final['actual␣y'] = y_train
        final['predicted␣y'] = y_pred
        final['sse'] = kmeans.inertia_
        if n_clusts>1 and n_clusts<len(X_train):
            final['silhouette_avg'] = silhouette_score(
                ↪ X_train, train_clus_groups)
            final['sample_silhouette_values'] = pd.
                ↪ DataFrame(silhouette_samples(X_train,
                ↪ train_clus_groups),columns=['
                ↪ silhouette_values'],index=X_train.index)
        diff_df=y_train.drop('Cluster_Group',axis=1)-
            ↪ y_pred.drop('Cluster_Group',axis=1)
        percent_diff_df=(y_train.drop('Cluster_Group',
            ↪ axis=1)-y_pred.drop('Cluster_Group',axis=1))
            ↪ /y_train.drop('Cluster_Group',axis=1)
        summary={}
        summary['MSE']=np.square(diff_df).replace(0, np.
            ↪ NaN).mean()
        summary['mean_diff']=diff_df.replace(0, np.NaN).
            ↪ mean()
        summary['median_diff']=diff_df.replace(0, np.NaN)
            ↪ .median()
        summary['mean_perc_diff']=percent_diff_df.replace
            ↪ (0, np.NaN).mean()
        summary['median_perc_diff']=percent_diff_df.
            ↪ replace(0, np.NaN).median()
        final['summary']=pd.DataFrame(summary)

elif leaveout==1:
    loo.get_n_splits(X_train)
```

```
actuals=pd.DataFrame()
predictions=pd.DataFrame()
sse=[]
silhouette=[]
for included_patients,left_out_patient in loo.split(
    ↪ y_train):
    X_train_left_in = X_train.iloc[included_patients]
    scaler.fit(X_train_left_in)
    X_train_left_in = pd.DataFrame(scaler.transform(
        ↪ X_train_left_in),columns=X_train_left_in.
        ↪ columns,index=X_train_left_in.index)

    X_train_left_out = X_train.iloc[left_out_patient]
    X_train_left_out = pd.DataFrame(scaler.transform(
        ↪ X_train_left_out),columns=X_train_left_out.
        ↪ columns,index=X_train_left_out.index)

    y_train_left_in = y_train.iloc[included_patients]
    y_train_left_out = y_train.iloc[left_out_patient]

    kmeans.fit(X_train_left_in)
    left_in_clus_groups = kmeans.predict(
        ↪ X_train_left_in)+1
    y_train_left_in['Cluster_Group']=
        ↪ left_in_clus_groups
    means = y_train_left_in.replace(0, np.NaN).
        ↪ groupby('Cluster_Group').mean()

    left_out_clus_groups = kmeans.predict(
        ↪ X_train_left_out)+1
    y_train_left_out['Cluster_Group']=
        ↪ left_out_clus_groups
```

```
            y_pred = y_train_left_out[['Cluster_Group']].
                ↪ merge(means,left_on='Cluster_Group',
                ↪ right_index=True)

            sse+=[kmeans.inertia_]
            if n_clusts>1 and n_clusts<len(X_train)-1:
                silhouette+=[silhouette_score(X_train_left_in,
                    ↪  left_in_clus_groups)]

            actuals=pd.concat([actuals,y_train_left_out])
            predictions=pd.concat([predictions,y_pred])



        final['actual␣y'] = actuals
        final['predicted␣y'] = predictions
        final['sse'] = pd.DataFrame(sse,columns=['sse'],
            ↪ index=X_train.index)
        if n_clusts>1 and n_clusts<len(X_train):
            final['silhouette_avg'] = pd.DataFrame(silhouette
                ↪ ,columns=['silhouette_averages'],index=
                ↪ X_train.index)
        diff_df=actuals.drop('Cluster_Group',axis=1)-
            ↪ predictions.drop('Cluster_Group',axis=1)
        percent_diff_df=(actuals.drop('Cluster_Group',axis
            ↪ =1)-predictions.drop('Cluster_Group',axis=1))/
            ↪ actuals.drop('Cluster_Group',axis=1)
        summary={}
        summary['MSE']=np.square(diff_df).replace(0, np.NaN)
            ↪ .mean()
        summary['mean_diff']=diff_df.replace(0, np.NaN).mean
            ↪ ()
        summary['median_diff']=diff_df.replace(0, np.NaN).
            ↪ median()
        summary['mean_perc_diff']=percent_diff_df.replace(0,
           ↪  np.NaN).mean()
        summary['median_perc_diff']=percent_diff_df.replace
           ↪ (0, np.NaN).median()
        final['summary']=pd.DataFrame(summary)
    return final









def clusteringmethodPCA(site='PSV',metric='Weight',
    ↪ n_clusts=1,leaveout=0,same=0):

    kmeans = clus.KMeans(
             init="random",
             n_clusters=n_clusts,
             n_init=10,
             max_iter=300,
             random_state=42
             )

    ##import data from excel
    File= Stem+"DataWorkbook"+site+".xlsx"
    X_train = pd.read_excel(File,sheet_name='
        ↪ FeaturesTrainingForPCA',index_col='Patient_ID').
        ↪ fillna(0)
    y_train = pd.read_excel(File,sheet_name='
        ↪ WeightsTraining',index_col='Patient_ID')
```

```
y_train = y_train[y_train['Metric']==metric].loc[:,
   ↪ y_train.columns != 'Metric']
y_train.dropna(axis=1, how='any', inplace=True) #if
   ↪ there still remain predictive features with null
   ↪ values, delete.
y_train = y_train.replace(0, np.NaN)

##initialising variables that will populate the "final"
   ↪  dictionary that will be converted to the "
   ↪ final_df" dataframe
actuals=[]
predictions=[]
sse=[]
final = {}



if leaveout==0:
   scaler.fit(X_train)
   if same==0:
      X_train = pd.DataFrame(scaler.transform(X_train),
         ↪ columns=X_train.columns,index=X_train.index)
         ↪ .fillna(0)
      X_test = pd.read_excel(File,sheet_name='
         ↪ FeaturesTestingForPCA',index_col='Patient_ID
         ↪ ').fillna(0)
      X_test.dropna(axis=1, how='any', inplace=True) #
         ↪ if there still remain predictive features
         ↪ with null values, delete.
      X_test = pd.DataFrame(scaler.transform(X_test),
         ↪ columns=X_test.columns,index=X_test.index)

      y_test = pd.read_excel(File,sheet_name='
         ↪ WeightsTesting',index_col='Patient_ID')
```

```
y_test = y_test[y_test['Metric']==metric].loc[:,
    ↪ y_test.columns != 'Metric']
y_test.dropna(axis=1, how='any', inplace=True)
y_test = y_test.replace(0, np.NaN)

pca.fit(X_train)
PC_cols=["PC"+str.zfill(str(x+1),2) for x in
    ↪ range(min(X_train.shape))]
X_train = pd.DataFrame(pca.transform(X_train),
    ↪ columns=PC_cols)
X_test = pd.DataFrame(pca.transform(X_test),
    ↪ columns=PC_cols)

kmeans.fit(X_train)
train_clus_groups = kmeans.predict(X_train)+1
y_train['Cluster_Group']=train_clus_groups
means = y_train.replace(0, np.NaN).groupby('
    ↪ Cluster_Group').mean()

test_clus_groups = kmeans.predict(X_test)+1
y_test['Cluster_Group']=test_clus_groups
y_pred = y_test[['Cluster_Group']].merge(means,
    ↪ left_on='Cluster_Group',right_index=True)

final['actual␣y'] = y_test
final['predicted␣y'] = y_pred
final['sse'] = kmeans.inertia_
if n_clusts>1 and n_clusts<len(X_train):
    final['silhouette_avg'] = silhouette_score(
        ↪ X_train, train_clus_groups)
    final['sample_silhouette_values'] = pd.
        ↪ DataFrame(silhouette_samples(X_train,
        ↪ train_clus_groups),columns=['
```

```
                ↪ silhouette_values'],index=X_train.index)
        diff_df=y_test.drop('Cluster_Group',axis=1)-
            ↪ y_pred.drop('Cluster_Group',axis=1)
        percent_diff_df=(y_test.drop('Cluster_Group',axis
            ↪ =1)-y_pred.drop('Cluster_Group',axis=1))/
            ↪ y_test.drop('Cluster_Group',axis=1)
        summary={}
        summary['MSE']=np.square(diff_df).replace(0, np.
            ↪ NaN).mean()
        summary['mean_diff']=diff_df.replace(0, np.NaN).
            ↪ mean()
        summary['median_diff']=diff_df.replace(0, np.NaN)
            ↪ .median()
        summary['mean_perc_diff']=percent_diff_df.replace
            ↪ (0, np.NaN).mean()
        summary['median_perc_diff']=percent_diff_df.
            ↪ replace(0, np.NaN).median()
        final['summary']=pd.DataFrame(summary)

    elif same==1:
        training_patients=X_train.index.tolist()
        X_train = pd.DataFrame(scaler.transform(X_train),
            ↪ columns=X_train.columns,index=X_train.index)
        pca.fit(X_train)
        PC_cols=["PC"+str.zfill(str(x+1),2) for x in
            ↪ range(min(X_train.shape))]
        X_train = pd.DataFrame(pca.transform(X_train),
            ↪ columns=PC_cols)

        kmeans.fit(X_train)
        train_clus_groups = kmeans.predict(X_train)+1
        y_train['Cluster_Group']=train_clus_groups
```

```
means = y_train.replace(0, np.NaN).groupby(’
    ↪ Cluster_Group’).mean()
y_pred = y_train[[’Cluster_Group’]].merge(means,
    ↪ left_on=’Cluster_Group’,right_index=True)

final[’actual␣y’] = y_train
final[’predicted␣y’] = y_pred
final[’sse’] = kmeans.inertia_
if n_clusts>1 and n_clusts<len(X_train):
    final[’silhouette_avg’] = silhouette_score(
        ↪ X_train, train_clus_groups)
    final[’sample_silhouette_values’] = pd.
        ↪ DataFrame(silhouette_samples(X_train,
        ↪ train_clus_groups),columns=[’
        ↪ silhouette_values’],index=X_train.index)
diff_df=y_train.drop(’Cluster_Group’,axis=1)-
    ↪ y_pred.drop(’Cluster_Group’,axis=1)
percent_diff_df=(y_train.drop(’Cluster_Group’,
    ↪ axis=1)-y_pred.drop(’Cluster_Group’,axis=1))
    ↪ /y_train.drop(’Cluster_Group’,axis=1)
summary={}
summary[’MSE’]=np.square(diff_df).replace(0, np.
    ↪ NaN).mean()
summary[’mean_diff’]=diff_df.replace(0, np.NaN).
    ↪ mean()
summary[’median_diff’]=diff_df.replace(0, np.NaN)
    ↪ .median()
summary[’mean_perc_diff’]=percent_diff_df.replace
    ↪ (0, np.NaN).mean()
summary[’median_perc_diff’]=percent_diff_df.
    ↪ replace(0, np.NaN).median()
final[’summary’]=pd.DataFrame(summary)
```

```
    elif leaveout==1:
        loo.get_n_splits(X_train)
        actuals=pd.DataFrame()
        predictions=pd.DataFrame()
        sse=[]
        silhouette=[]
        for included_patients,left_out_patient in loo.split(
            ↪ y_train):
            X_train_left_in = X_train.iloc[included_patients]
            scaler.fit(X_train_left_in)
            X_train_left_in = pd.DataFrame(scaler.transform(
                ↪ X_train_left_in),columns=X_train_left_in.
                ↪ columns,index=X_train_left_in.index)

            pca.fit(X_train_left_in)
            PC_cols=["PC"+str.zfill(str(x+1),2) for x in
                ↪ range(min(X_train_left_in.shape))]
            X_train_left_in = pca.transform(X_train_left_in)
            X_train_left_in = pd.DataFrame(X_train_left_in,
                ↪ columns=PC_cols)

            X_train_left_out = X_train.iloc[left_out_patient]
            X_train_left_out = pd.DataFrame(scaler.transform(
                ↪ X_train_left_out),columns=X_train_left_out.
                ↪ columns,index=X_train_left_out.index)

            X_train_left_out = pd.DataFrame(pca.transform(
                ↪ X_train_left_out),columns=PC_cols)

            y_train_left_in = y_train.iloc[included_patients]
            y_train_left_out = y_train.iloc[left_out_patient]

            kmeans.fit(X_train_left_in)
```

```
      left_in_clus_groups = kmeans.predict(
         ↪ X_train_left_in)+1
      y_train_left_in['Cluster_Group']=
         ↪ left_in_clus_groups
      means = y_train_left_in.replace(0, np.NaN).
         ↪ groupby('Cluster_Group').mean()

      left_out_clus_groups = kmeans.predict(
         ↪ X_train_left_out)+1
      y_train_left_out['Cluster_Group']=
         ↪ left_out_clus_groups
      y_pred = y_train_left_out[['Cluster_Group']].
         ↪ merge(means,left_on='Cluster_Group',
         ↪ right_index=True)

      sse+=[kmeans.inertia_]
      if n_clusts>1 and n_clusts<len(X_train)-1:
         silhouette+=[silhouette_score(X_train_left_in,
            ↪  left_in_clus_groups)]

      actuals=pd.concat([actuals,y_train_left_out])
      predictions=pd.concat([predictions,y_pred])



   final['actual␣y'] = actuals
   final['predicted␣y'] = predictions
   final['sse'] = pd.DataFrame(sse,columns=['sse'],
      ↪ index=X_train.index)
   if n_clusts>1 and n_clusts<len(X_train):
      final['silhouette_avg'] = pd.DataFrame(silhouette
         ↪ ,columns=['silhouette_averages'],index=
         ↪ X_train.index)
```

```
        diff_df=actuals.drop(’Cluster_Group’,axis=1)-
            ↪ predictions.drop(’Cluster_Group’,axis=1)
        percent_diff_df=(actuals.drop(’Cluster_Group’,axis
            ↪ =1)-predictions.drop(’Cluster_Group’,axis=1))/
            ↪ actuals.drop(’Cluster_Group’,axis=1)
        summary={}
        summary[’MSE’]=np.square(diff_df).replace(0, np.NaN)
            ↪ .mean()
        summary[’mean_diff’]=diff_df.replace(0, np.NaN).mean
            ↪ ()
        summary[’median_diff’]=diff_df.replace(0, np.NaN).
            ↪ median()
        summary[’mean_perc_diff’]=percent_diff_df.replace(0,
            ↪  np.NaN).mean()
        summary[’median_perc_diff’]=percent_diff_df.replace
            ↪ (0, np.NaN).median()
        final[’summary’]=pd.DataFrame(summary)
    return final
```