Gezamenlijke opleiding UHasselt en KU Leuven

Universiteit Hasselt | Campus Diepenbeek | Faculteit Industriële Ingenieurswetenschappen | Agoralaan Gebouw H - Gebouw B | BE 3590 Diepenbeek

Universiteit Hasselt | Campus Diepenbeek | Agoralaan Gebouw D | BE 3590 Diepenbeek
Universiteit Hasselt | Campus Hasselt | Martelarenlaan 42 | BE 3500 Hasselt

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

Development and Clinical Validation of One-Click Batch VMAT Prostate Planning Using Dynamic Adjustment of Optimization Parameters

Robin Van Grunderbeeck

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

PROMOTOR :
Prof. dr. Brigitte RENIERS

PROMOTOR :
Dhr. Hasan CAVUS

COPROMOTOR :
Ms. Sc. Alexandra JANKELEVITCH

2024
2025

Faculteit Industriële Ingenieurswetenschappen

master in de industriële wetenschappen: nucleaire technologie

Masterthesis

Development and Clinical Validation of One-Click Batch VMAT Prostate Planning Using Dynamic Adjustment of Optimization Parameters

Robin Van Grunderbeeck

Scriptie ingediend tot het behalen van de graad van master in de industriële wetenschappen: nucleaire technologie, afstudeerrichting nucleair en medisch

PROMOTOR :

Prof. dr. Brigitte RENIERS

PROMOTOR :

Dhr. Hasan CAVUS

COPROMOTOR :

Ms. Sc. Alexandra JANKELEVITCH

UHASSELT

KU LEUVEN

Acknowledgments

I would like to express my sincere appreciation to all those who contributed to the completion of this master’s thesis.

First, I would like to thank my internal supervisor, Prof. Dr. Brigitte Reniers, for introducing me to the field of nuclear medicine. Her courses on radiotherapy and radiology sparked my initial interest and provided a strong foundation for this work.

I am also deeply grateful to Hasan Cavus, my external supervisor at Jessa Hospital, for offering the thesis topic, his guidance, valuable feedback, and support throughout the research process.

In addition, I would like to thank my co-promotor at Jessa Hospital, Alexandra Jankelevitch, for her insightful input, ongoing guidance, and constructive feedback during every phase of this thesis.

I would also like to thank the members of the medical physics team at Jessa Hospital for creating a welcoming and collaborative environment, which greatly enriched my research experience.

Finally, I want to thank Jessa Hospital for providing me with the opportunity, resources, and support to carry out this thesis. It has been a truly valuable and enriching experience.

Table of Contents

List of Tables

List of Figures

Glossary of Terms

Abbreviation	**Full term**
α	Significance level
$(1-\beta)$	Power
ARE	Asymptotic relative efficiency
CBCT	Cone-beam computed tomography
CM	Complexity metric
CT	Computed tomography
CTV	Clinical target volume
D95%, D99%, etc.	Dose received by 95%, 99%, etc., of the volume
DVH	Dose-volume histogram
FT	Fine-tuning
GTV	Gross target volume
H_0	Null hypothesis
H_a	Alternative hypothesis
ICRU	International Commission on Radiation Units and Measurements
IMRT	Intensity-modulated radiotherapy
KBP	Knowledge-based planning
MCO	Multi-criteria optimization
MLC	Multileaf collimator
MU	Monitor units
MU/Gy	Monitor units per Gray
OAR	Organ-at-risk
PTV	Planning target volume
Q-Q plot	Quantile-quantile plot
QA	Quality assurance
RP	RapidPlan
TPS	Treatment Planning System
VMAT	Volumetric modulated arc therapy
VxGy	Volume receiving at least x Gy

Abstract in English

Automated treatment planning methods have the potential to improve plan quality, standardization, and efficiency in radiotherapy. However, manual optimization is often still required. This thesis evaluates a fully automated fine-tuning process for prostate volumetric modulated arc therapy planning that dynamically adjusts optimization parameters to improve knowledge-based generated plans.

Treatment plans were generated for 200 previously treated prostate cases using a knowledge-based planning model and applying the fine-tuning process. Dose-volume histogram parameters and plan complexity were used to evaluate differences between the initial knowledge-based generated plan and the fine-tuned plan. Statistical analysis of the results was performed using the Wilcoxon signed-rank test.

In 165 of the cases, the knowledge-based generated plan required fine-tuning. Initially, 12 cases failed to complete the fine-tuning process. This problem was addressed by adding an extra constraint to the script. With this additional constraint, the fine-tuning algorithm resolved all unmet constraints in 104 of the cases. The results demonstrate that the fine-tuning algorithm effectively addressed unmet constraints of the knowledge-based generated plans in a significant number of cases. Additionally, the fine-tuning process reduced the plan complexity in most cases.

Abstract (in Dutch)

Automatisatie van het planningsproces in radiotherapie heeft het potentieel om de kwaliteit, standaardisatie en efficiëntie te verbeteren. Toch is handmatige optimalisatie van de behandelplannen vaak nog noodzakelijk. Deze masterproef evalueert een volledig geautomatiseerd finetuningproces voor prostaatbehandeling met volumetrisch gemoduleerde arctherapie, waarbij de optimalisatieparameters dynamisch worden aangepast om *knowledge-based* gegenereerde plannen te optimaliseren.

Voor 200 eerder behandelde prostaatpatiënten werden nieuwe behandelplannen gegenereerd met een knowledge-based-planningmodel, waarop het finetuningproces werd toegepast. Verschillen tussen het initiële knowledge-based-plan en het geoptimaliseerd plan werden geëvalueerd op basis van dosis-volumehistogramparameters en plancomplexiteit. De resultaten werden statistisch geanalyseerd met de Wilcoxon signed-rank-test.

In 165 gevallen was finetuning noodzakelijk. Initieel waren er 12 plannen die er niet in slaagden het finetuningproces volledig te doorlopen. Dit probleem werd opgelost door een extra dosisrestrictie aan het script toe te voegen. Met deze aanpassing kon het algoritme in 104 gevallen alle overschreden dosisrestricties oplossen. De resultaten tonen aan dat het algoritme in een aanzienlijk aantal gevallen de overschreden dosisrestricties van de knowledge-based-plannen oplost. Bovendien zorgde het finetuningalgoritme in de meeste gevallen ook voor een vermindering van de plancomplexiteit.

1 Introduction

Radiotherapy plays a central role in cancer treatment, with approximately 50 to 70 percent of all cancer patients requiring radiotherapy during treatment [1]. It is used for both curative and palliative treatment and can be combined with other treatment methods such as surgery, chemotherapy, immunotherapy, and hormone therapy depending on the cancer type and stage [2].

The radiotherapy workflow is a multi-step process that includes simulation imaging, target and organ at risk (OAR) contouring, treatment planning, quality assurance (QA) tests, treatment delivery, and post-treatment follow-up. This process requires the cooperation of a team of physicians, physicists, dosimetrists, and therapists [3].

The main objective of radiotherapy is to deliver the prescribed dose to the tumor while minimizing exposure to the surrounding healthy tissue. To optimize this objective, radiotherapy has evolved significantly in recent years. This evolution, driven by technological advances, has allowed for more precise treatment planning and dose delivery. However, it has also increased the complexity of the workflow and made it more time-consuming. As a result, different automation methods for each step in the workflow have been proposed to increase the quality, standardization, and efficiency of the workflow [4].

In recent years, several methods have been developed for the automation of segmentation, treatment planning, and quality assurance. This paper focuses on the automation of treatment planning, specifically for volumetric modulated arc therapy (VMAT) prostate treatment. It builds upon a feasibility study, performed at Jessa Hospital, that was published in 2024. In this preliminary study, a novel approach was introduced to automate the planning process of VMAT prostate treatment. An in-house script was developed that combined an existing knowledge-based planning (KBP) model with an automated fine-tuning (FT) process [5]. The fine-tuning process automatically adjusts optimization parameters based on unmet constraints. The script automates the different steps of the treatment planning and optimization process.

The goal of this study is to evaluate the software that was introduced in the previous study. The fine-tuning software is evaluated by comparing the plans generated by the KBP model with the plans optimized by the fine-tuning software. For each selected patient, a new treatment plan is generated using the automated script. The generated plans are collected in two datasets. The first dataset contains the initial plan generated by the KBP model and the second dataset contains the last plan that was optimized by the fine-tuning process. The two datasets are compared based on key parameters, such as target coverage, organs-at-risk sparing, and plan complexity. To evaluate plan complexity, monitor units per gray (MU/Gy) and a complexity metric (CM) are used. A statistical analysis is performed on these key parameters to assess the significance of the observed differences and the overall impact of the fine-tuning process.

This master’s thesis is structured into six main chapters that each focus on a specific aspect of the study. Following this introduction, the second chapter presents a literary review that covers relevant background information. The evolution of radiotherapy, the workflow used in current practice, different automation methods, and the preliminary study are discussed in greater detail. Chapter 3 provides an overview of the materials and methods used for this study. It discusses the patient selection criteria, plan generation and evaluation, and the process used for statistical analysis. The observed results and the results of the statistical analysis are presented in chapter 4 and the relevance and implications of the findings are discussed in chapter 5. Chapter 6 summarizes the most important conclusions of the study.

2 Literature Review

2.1 Radiotherapy

The evolution of radiotherapy treatment began with three important discoveries. In 1895, the discovery of the X-ray was announced by German scientist Wilhelm Röntgen [6]. Around the same time, Henri Becquerel discovered natural radioactivity, and in 1898, Marie and Pierre Curie discovered radium [7]. While the harmful effects of radiation were not initially understood, they quickly became apparent when both operators and patients started showing side effects [6]. These effects were taken into account to optimize the therapeutic ratio and improve radioprotection.

In the early years of radiotherapy, patients were treated using low-energy X-rays and long exposure times. The X-ray tube was placed close to the skin, or even in contact with it, and the treatment was delivered in a single fraction [6]. The equipment used for treatment was unreliable, and the dose distribution was imprecise. The introduction of more reliable equipment and the development of imaging methods marked the beginning of treatment planning in radiotherapy. However, the first treatment planning methods were still limited by image quality, and treatment was planned and delivered in 2D [8]. The development of megavoltage linear accelerator X-ray machines and the invention of computed tomography (CT) scans allowed for more precise treatment planning and delivery [7].

The increased use of radiotherapy in clinical practice led to the need for standard protocols and a common language. The International Commission on Radiation Units and Measurements (ICRU) introduced the concept of three main target volumes to promote uniform terminology. The ICRU 50 report defined the gross target volume (GTV), the clinical target volume (CTV), and the planning target volume (PTV) [9]. The GTV is the macroscopic tumor volume that is visible on imaging. Although it is the easiest to define, it can be difficult to delineate in practice and is highly dependent on image quality [10]. The CTV is defined as the GTV plus a margin to account for the microscopic extension of the tumor that is not visible on imaging. The margins used for the CTV depend on the imaging techniques, tumor site, and prior knowledge from historical cases [10]. The PTV is the volume that receives the prescribed dose during treatment. It includes the CTV and additional margins to account for motion and setup uncertainties, ensuring full coverage of the CTV [11]. In current practice, the GTV, CTV, and PTV are still used to delineate the tumor.

The development of 3D imaging methods and the introduction of computerized treatment planning systems laid the foundation for the transition from 2D to 3D planning [7]. This evolution continued with the development of intensity-modulated radiotherapy (IMRT) and volumetric modulated arc therapy. IMRT allows for highly conformal dose distribution by using variable intensity across multiple radiation beams. While IMRT minimizes the dose delivered to healthy tissue, it also increases treatment time. VMAT, which uses one or more arcs and allows simultaneous variation of the gantry rotation speed, dose rate, and multileaf collimator (MLC) leaf positions, enables faster dose delivery compared to IMRT [12]. Both IMRT and VMAT are currently used in clinical practice.

2.2 Current Workflow

The current radiotherapy workflow consists of five main steps. The first step is the acquisition of simulation images [13]. The patient is placed in the treatment position, and the user origin and skin markers are placed. These markers are used during treatment delivery to ensure consistent patient positioning. The simulation CT is then acquired and reviewed by the radiation oncologist. This scan is exported to the treatment planning system (TPS) and used during treatment planning.

The goal of the treatment planning step is to ensure that the PTV receives the prescribed dose while minimizing the dose to the OARs. This step starts by defining the specific location of the primary tumor and the extent of the spread around the tumor. The radiation oncologist uses the simulation CT to contour the target volumes and the organs at risk [3]. This contouring process relies on the image quality of the simulation CT and in some cases additional images like MRI and PET are used [14]. The second part of treatment planning is to determine the dose delivery plan. In current practice, most treatment plans are generated using inverse treatment planning methods. These methods start by setting initial planning goals and constraints, based on standard protocols. The first fluence map is then calculated using these initial parameters and the dose distribution is evaluated using institute constraints. The treatment plan is optimized by iteratively adjusting the planning goals and recalculating the fluence map until a clinically desirable plan is achieved [3].

Once the treatment plan is approved, quality assurance procedures are carried out to ensure that the delivered dose is as close to the planned dose as possible. The two types of QA are machine-specific QA and patient-specific QA. Machine-specific QA ensures that the equipment used to deliver treatment functions safely and correctly. Patient-specific QA ensures that the treatment plan can be correctly delivered by the equipment [15].

The final step is treatment delivery. The patient is secured to the treatment position and new images are acquired. These images are used to match the simulation CT and verify the patient's position. Skin markers or tattoos are often used as reference points to ensure the correct alignment of the patient. Set-up verification can be done with 2D images, either kV or MV portal images that are matched to the simulation CT based on anatomy landmarks. The setup can also be verified with 3D images such as cone-beam CT (CBCT), which offer better spatial resolution [16]. Treatment is delivered over one or multiple sessions and new images are regularly taken to verify the patient's position [13]. After treatment is completed, follow-up scans are performed to evaluate the outcome.

While the radiotherapy workflow is highly structured and standardized, it remains a complex and time-consuming process. Each step requires significant input from experienced professionals, and the planning process involves multiple iterations to achieve a clinically optimal plan.

To address the challenges and improve efficiency, consistency, and plan quality, multiple automation methods have been proposed [4]. In recent years, several automated planning techniques have been developed and implemented in clinical practice.

2.3 Automation in Treatment Planning

The evolution of treatment planning and delivery has been driven by the introduction of improved imaging and dose delivery techniques. Initially, treatment planning and delivery relied on low-quality 2D images, and the dose was delivered heterogeneously to large treatment fields [8]. The introduction of CT scans, the use of computerized algorithms, and treatment planning systems enabled a transition from 2D to 3D planning and execution [7]. The advanced delivery techniques and equipment allowed for more precise treatment planning and execution.

Currently, treatment planning is performed by medical physicists using simulation CT images and TPS [5]. Inverse planning methods, such as VMAT, are considered the gold standard in modern radiotherapy treatment planning. Inverse planning begins with the specification of the dose constraints for the OARs and the dose prescription to the PTV. The TPS generates an initial fluence map based on these constraints, which are derived from standard protocols. The medical physicist then adjusts the planning goals and constraints to optimize the dose distribution, and the TPS recalculates the new fluence map accordingly [3]. This trial-and-error process is repeated until a clinically acceptable plan is achieved. Depending on the plan's complexity, the optimization process can be very time-consuming.

Automating the treatment planning process aims to accelerate the workflow, improve plan quality, and reduce inter-operator variability. In recent years, several automation methods have been developed to enhance consistency and efficiency in treatment planning.

One of the first methods developed for automation in treatment planning is atlas-based KBP. This approach uses prior knowledge from high-quality plans from previously treated patients. The model identifies geometric similarities to match the new patient to a similar prior case (or set of prior cases) [17]. That prior case can then be used as a starting point for inverse optimization [3]. The effectiveness of this method depends heavily on patient similarity as well as the quality and number of atlases in the database. Atlas-based KBP does not fully automate the planning process, but it can accelerate the planning by providing a starting point.

Another method that has been developed is model-based KBP. This approach uses machine learning and high-quality plans from previously treated patients to train mathematical or statistical models [5]. The key difference from atlas-based planning KBP is that once the KBP model is trained, the original training data is no longer used. The model predicts the dose distribution for a new patient by applying prior knowledge and statistical inference [18]. Model-based KBP has been implemented in clinical practice through commercially available applications like Varian RapidPlan$^{\text{TM}}$ [3]. Although KBP reduces inter-operator variability and increases efficiency, manual refinement of the treatment plans is often still required [5].

A third auto-planning method is a posteriori multi-criteria optimization (MCO), which generates multiple treatment plans based on the Pareto optimal principle. A Pareto optimal plan is a treatment plan in which no planning objective can be improved without negatively affecting another [19]. The advantage of a posteriori MCO is that the impact of altering competing planning objectives is directly observable. This allows the planner to compare the generated plans and select the best trade-offs to achieve a clinically optimal plan for the specific patient [18]. A posteriori MCO offers a balance between automated planning and manual decision-making, but selecting the best plan can still be complex and is dependent on the planner's experience. A posteriori MCO has been implemented in commercially available applications such as RayStation TPS [3].

Deep learning-based planning is another category of automated treatment planning. Deep learning models, such as U-Net, utilize deep neural network architectures to learn features from a training dataset [5]. These models automatically extract features from the input data, allowing more features to be considered in the prediction process [20]. Deep learning-based convolutional neural networks can be used to predict 3D dose distributions [21]. However, deep learning models require large datasets of high-quality treatment plans for effective training [5]. A common challenge in the training process is the limited availability and variability of data. To ensure model robustness, training datasets should incorporate data from multiple institutions [22].

Scripts can also be used to automate specific steps within the treatment planning process. Tasks such as setting optimization objectives, adjusting constraints, or modifying priorities can be automated by using scripts. The planning strategies and steps typically taken during manual planning can be simulated by using predefined rules and logic set by the planner [23]. While scripting is not a standalone auto-planning method, it can serve as a tool to enhance other planning approaches.

Each auto-planning method has its own strengths and limitations, but most still require manual adjustments to meet the prespecified constraints [5]. To leverage the advantages of various approaches, hybrid techniques have been proposed. These techniques combine elements from different automation methods to improve plan quality and efficiency. For example, the fine-tuning algorithm evaluated in this study builds upon an existing KBP model.

2.4 Preliminary Study

This study builds upon the work titled “Optimizing volumetric modulated arc therapy prostate planning using an automated Fine-Tuning process through dynamic adjustment of optimization parameters” [5]. The preliminary study, conducted at Jessa Hospital, aimed to introduce a novel approach to automate the fine-tuning of optimization parameters for VMAT prostate treatment. A fully automated hybrid technique was proposed through the development of a script that invokes the KBP model and applies the fine-tuning process without manual intervention.

The study retrospectively selected 25 prostate cancer patients treated between 2022 and 2023 from the clinical database. Each patient received a prescribed dose of 60 Gy to the prostate and 44 Gy to the seminal vesicles, delivered over 20 fractions. Anisotropic margins were applied from the CTV to the PTV, with a 6 mm margin laterally and an 8 mm margin in all other directions. Treatment was delivered using VMAT with two opposing full arcs. The beam energy was set to 6 MV, and collimator rotations were positioned at 30° and 330°, respectively. Dose calculations and optimizations were performed using the Acuros XB algorithm with a grid size of 0.25 cm.

A KBP model was developed using the RapidPlan^{TM} (RP) application within the Varian Eclipse TPS. The model was trained on a dataset consisting of 41 high-quality prostate VMAT plans from cases treated between April 2020 and July 2022. The training dataset included structures such as the PTV, bladder, rectum, bowel, and both femoral heads.

To fully automate the planning process, an in-house C# binary plug-in script was developed using the Eclipse Scripting API^{TM}. This script automates the different steps of the treatment planning process and consists of three main stages, as illustrated in Figure 1.

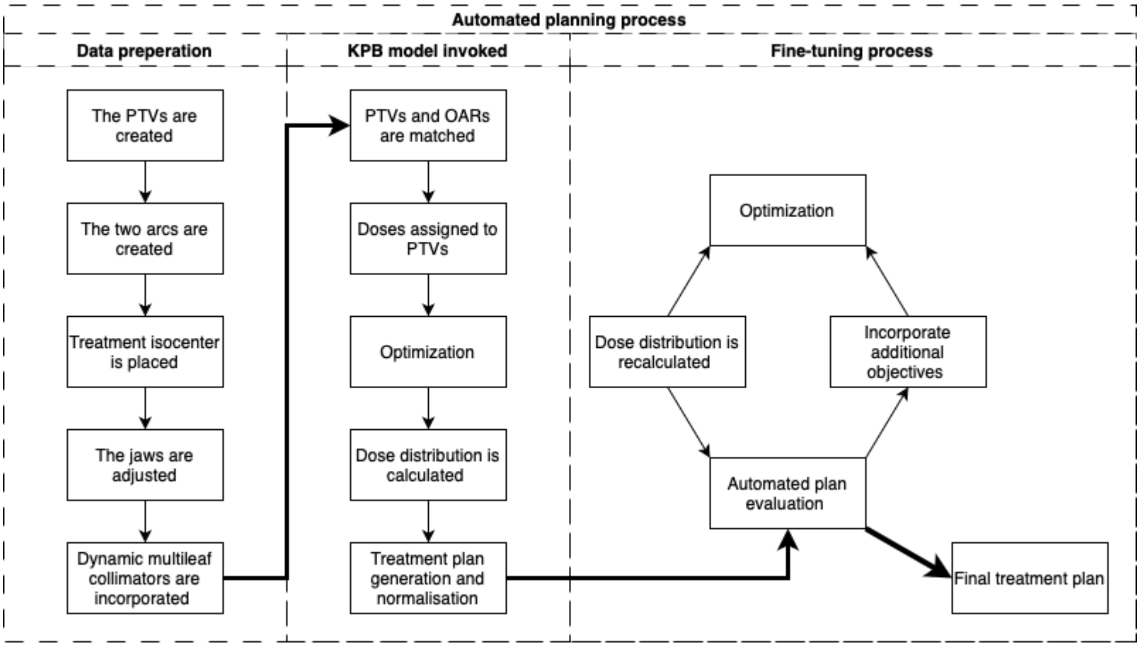
Automated planning process
Data preperation
KPB model invoked
Fine-tuning process
The PTVs are created
The two arcs are created
Treatment isocenter is placed
The jaws are adjusted
Dynamic multileaf collimators are incorporated
PTVs and OARs are matched
Doses assigned to PTVs
Optimization
Dose distribution is calculated
Treatment plan generation and normalisation
Optimization
Dose distribution is recalculated
Incorporate additional objectives
Automated plan evaluation
Final treatment plan

Figure 1: The fully automated planning process

In the first stage, the script generates all necessary optimization structures and treatment parameters required for planning. It creates the PTV-high and PTV-low structures and generates two arcs based on the previously described specifications. The treatment isocenter is placed at the center of the combined PTVs and the jaws are adjusted to be 5 mm from the total PTV. Dynamic multileaf collimators are incorporated in both arcs.

In the second stage, the script invokes the KBP model. The PTVs and OARs are matched with their corresponding structures, and the PTV-high and PTV-low are each assigned the prescribed dose in accordance with the treatment protocol. The KBP model estimates the dose-volume histogram (DVH) and the optimizer function is started. Once optimization is complete, the dose distribution is calculated, and a treatment plan is generated. The treatment plan is then normalized to ensure that the mean target volume matches the prescribed dose.

In the third stage, the plan is evaluated and, if necessary, the fine-tuning process is initiated. The generated plan is automatically evaluated using predefined constraints. If the initial plan meets all constraints, the script stops, and the plan is saved in the database. If the plan fails to meet all constraints, additional objectives are added to the optimizer. The optimization function is restarted, and the dose distribution is recalculated. The new treatment plan is re-evaluated and if the same constraint remains unmet, the priority value of the corresponding objective is increased by 10 units. If a new constraint is unmet, additional objectives are added to the optimizer. The fine-tuning process is repeated until the plan meets all constraints, or after completing 10 iterations. If, after 10 iterations, some constraints remain unmet, all generated plans are saved, allowing the planner to manually select the least unfavorable plan.

The script was tested using 25 prostate cases. Of these 25 cases, 10 treatment plans met all constraints using only the KBP model. The remaining 15 cases required the fine-tuning process, which addressed the unmet constraints in 12 of them. Specifically, the fine-tuning process addressed the unmet Dmax (body) constraint in 7 cases, the V60Gy (rectum) constraint in 7 cases, the V60Gy (bladder) constraint in 2 cases, and the D50% (PTV-high) constraint in 1 case. 3 cases failed to meet all constraints after 10 iterations.

This study demonstrated that the script effectively resolved the unmet constraints in a significant number of treatment plans. By enabling full automation of the planning process, the script minimized the need for manual intervention, reducing inter-operator variability. The script can run a list of multiple patients with only one mouse click. Moreover, the script is easy to integrate into the TPS and can run entirely in the background, making it an efficient and practical tool for clinical implementation.

This preliminary study was a feasibility study used to show the potential of the developed algorithm. The goal of this new study is to test the clinical validity of the algorithm by testing using a large number of previously treated patients and performing a statistical analysis.

3 Methods and Materials

3.1 Patient Selection

Prostate cancer patients treated between 2023 and 2025, were randomly selected from the clinical database. The prescribed dose was 60 Gy to the prostate, and 44 Gy to the seminal vesicles, delivered in 20 fractions. Anisotropic margins were applied from the CTV to the PTV, with a 6 mm expansion laterally and an 8 mm expansion in all other directions [5]. All patients were treated using VMAT with two opposing full arcs, a beam energy of 6 MV, and collimator rotations set to 30° and 330°, respectively.

3.2 Automated Planning

A new treatment plan was generated for each selected patient using the script developed in the preliminary study. The complete workflow used in this process is illustrated in Figure 2.

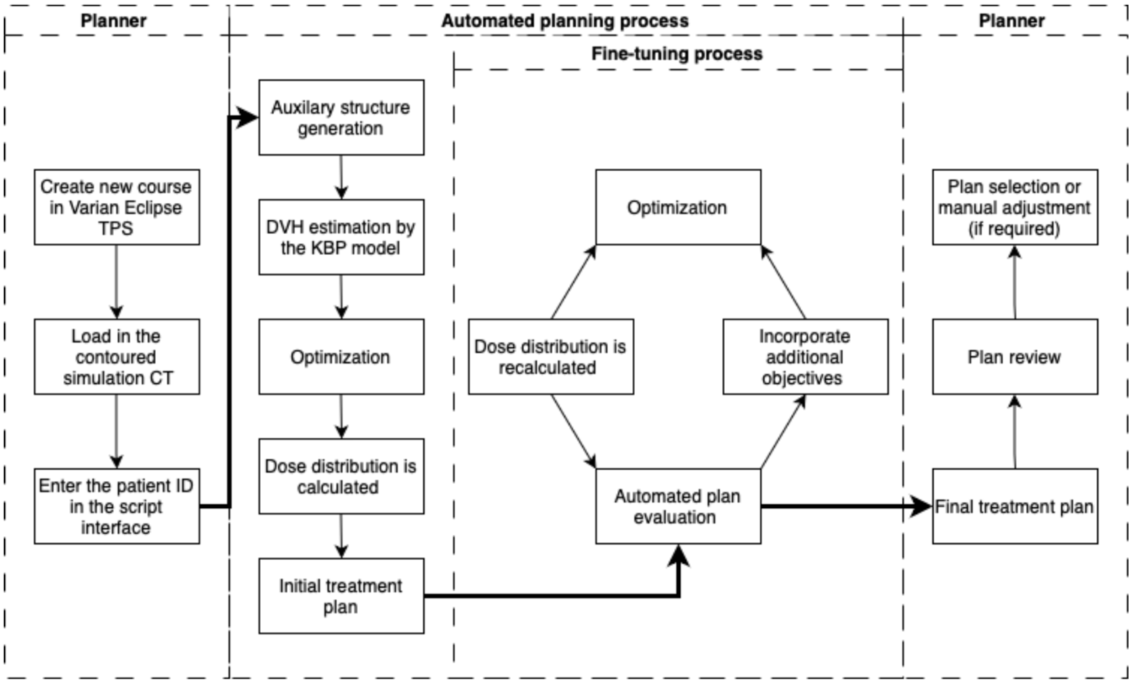
Planner
Automated planning process
Planner
Fine-tuning process
Create new course in Varian Eclipse TPS
Load in the contoured simulation CT
Enter the patient ID in the script interface
Auxilary structure generation
DVH estimation by the KBP model
Optimization
Dose distribution is calculated
Initial treatment plan
Optimization
Dose distribution is recalculated
Incorporate additional objectives
Automated plan evaluation
Plan selection or manual adjustment (if required)
Plan review
Final treatment plan

Figure 2: Workflow for generating new treatment plans

First, a new course was created for each patient, in which the contoured simulation CT was uploaded. The patient ID was then entered into the script interface and the script was started. The script first invoked the KBP model, which used patient-specific anatomy, the prescribed dose, and predefined dose constraints to generate the initial treatment plan. This treatment plan was then automatically evaluated with the preset dose constraints, shown in Table 1.

Table 1: Institute constraints used to evaluate the plans

Boost D99% (%)	> 90%
Boost D95% (%)	> 95%
Boost D50% (%)	> 100%
Boost D5% (%)	< 105%
Boost Dmax (%)	< 107%
Boost V107% (cc)	< 0.03 cc
Body Dmax (%)	< 107%
Body V107% (cc)	< 0.03 cc
Rectum V60Gy (cc)	< 1 cc
Rectum V50Gy (%)	< 22.2%
Rectum V40Gy (%)	< 37.7%
Rectum V30Gy (%)	< 56.7%
Rectum V26Gy (%)	< 68.2%
Rectum V20Gy (%)	< 85.2%
Rectum Dmean (Gy)	< 30 Gy
Left femoral V41Gy (%)	< 50%
Right femoral V41Gy (%)	< 50%
Bladder V63.6Gy (cc)	< 1 cc
Bladder V60Gy (%)	< 5%
Bladder V49Gy (%)	< 25%
Bladder V41Gy (%)	< 50%
Bladder V31Gy (%)	< 60%
Bowels V58.5Gy (cc)	< 1 cc
Bowels V41Gy (cc)	< 17 cc
Bowels V36Gy (cc)	< 195 cc

If the initial plan met all the constraints, the script was stopped, and the plan was saved. If the plan failed to meet all constraints, a fine-tuning process was started. The fine-tuning process incorporated additional objectives into the optimizer based on the unmet constraints. The script then iteratively adjusted the optimization objectives until all dose constraints were met or a maximum of 10 iterations was reached. The objectives and priorities added to the optimizer are listed in Table 2.

Table 2: Objective and priorities for each structure [5]

Type	ID	Objective type	Vol (%)	Dose (Gy)	Priority
Target	PTV-high (60 Gy)	Lower	100	58.8	120
		Upper	0	61.8	120
Target	PTV-low (44 Gy)	Lower	100	43.56	120
		Upper	0	61.8	120
Body	External	Upper	0	63.9	550
Organ	Bladder	Upper	0	63	150
		Upper	4.5	54	120
		Upper	22.5	44.1	100
		Upper	45	36.9	100
		Upper	54	27.9	100
Organ	Rectum	Upper	0	54	150
		Upper	20	45	100
		Upper	33.9	36	100
		Upper	51	27	100
		Upper	61.4	23.4	100
		Upper	76.7	18	100
		Mean	/	27	100
Organ	Bowel	Upper	V58.5Gy x 0.9	52.7	80
		Upper	V41Gy x 0.9	36.9	80
		Upper	V36Gy x 0.9	32.4	80
Organ	Femoral heads	Upper	45	36.9	50

All generated plans were saved, and two datasets were created for each parameter. The RP dataset contained the DVH parameters from the initial plan generated by the KBP model. The FT dataset contained the DVH parameters from the plan that met all the constraints. If a plan failed to meet all constraints after 10 loops, the DVH parameters from the last plan were included in the FT dataset.

3.3 Plan Evaluation

The fine-tuning algorithm was evaluated by comparing the treatment plans that were generated using the KBP model with the plans that were optimized using the fine-tuning script. Two datasets were analyzed during the evaluation. The first included all cases that required fine-tuning after the initial plan failed to meet all constraints. This dataset was used to assess the overall effect of the fine-tuning process. The second dataset consisted of only the data from plans that met all constraints with the fine-tuning process. The data of the plans that met all constraints with the KBP model and those that failed to meet all constraints after ten loops were excluded. This subset was used to evaluate the algorithm when it successfully addressed all constraints.

OAR sparing, target dose coverage, and complexity of the plans were used as evaluation metrics. Each parameter was analyzed using appropriate statistical tests, depending on the type of variable and the distribution. The statistical tests were performed in Excel using the Analysis ToolPak and the Real Statistics add-in package.

The first step of the data analysis was to apply descriptive statistics to summarize and visualize the collected data [24]. Both visual and numerical methods were used. The boxplot was used to visualize the data and to identify potential outliers and missing data. The mean, standard deviation, median, and interquartile range were calculated to summarize the central tendency and variability of the data. Descriptive statistics provided an initial idea of the distribution of each dataset and the difference between the datasets.

To determine if the difference between the FT plans and the RP plans was statistically significant, statistical tests were used. The statistical test used to compare the two datasets was chosen according to the characteristics of the collected data [25]. The statistical tests can be divided into parametric methods and nonparametric methods [26]. Parametric tests, such as the Student's t-test, are the most powerful statistical tests, but they are only reliable if the data meets the normality assumption [25]. When the data deviates from the normal distribution, parametric tests may lead to unreliable or misleading results. If the normality assumption was not met, a nonparametric equivalent was used. Nonparametric tests are more robust than parametric tests when the data does not follow the normal distribution because they do not make assumptions about the data distribution [25].

The two datasets analyzed in this study were the RP and the FT datasets. Each patient has a corresponding FT and RP plan, resulting in paired data. If the data met the normality assumption, a paired t-test was used to determine if the observed difference was statistically significant. If the t-test assumptions were violated, the Wilcoxon signed-rank test was used as a nonparametric equivalent. The workflow that was used for the statistical analysis of each parameter is shown in Figure 3.

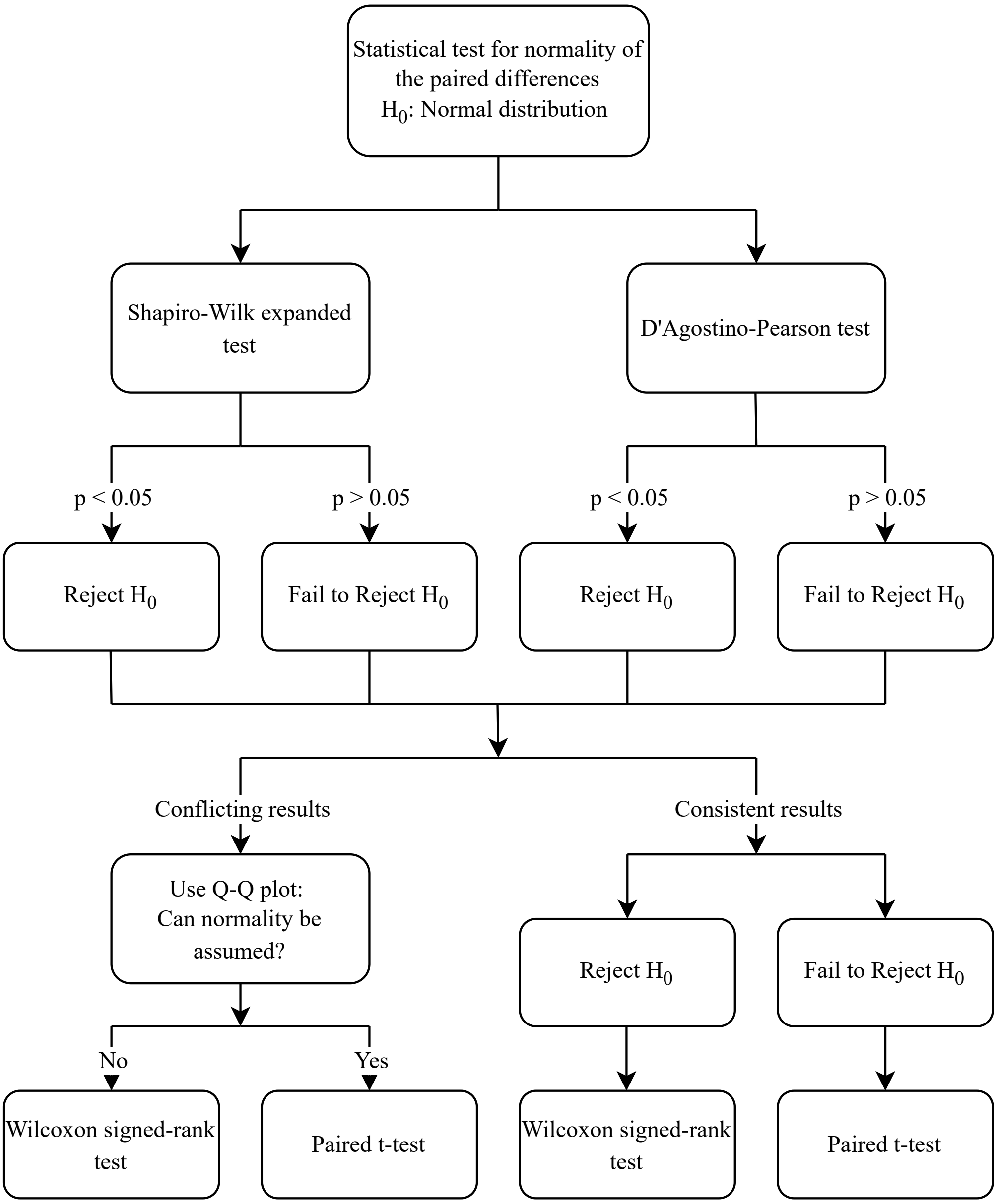

Figure 3: Workflow for selecting a statistical test

The first step in this workflow is to test for normality. Since the paired t-test uses the mean of the differences, the normality of the differences between FT and RP should be tested [27]. There are several methods for testing normality, which can be categorized into graphical methods and statistical tests [26]. In this study, methods from both categories were used. The Shapiro-Wilk expanded test and D'Agostino-Pearson test were used in combination with a normal quantile-quantile plot (Q-Q plot).

The Shapiro-Wilk expanded test compares the dataset against the normal distribution to test for normality [28]. The expanded version of the test is used because the original Shapiro-Wilk test is limited to datasets with a sample of size less than 50 [29]. The null hypothesis H_0 of the Shapiro-Wilk expanded test states that the data is normally distributed. The alternative hypothesis H_a states that the data is not normally distributed [30]. In this study, the test was performed with a significance level α of 0.05. The Shapiro-Wilk expanded test was performed using the Real Statistics add-in package. The null hypothesis was rejected if the p-value was smaller than 0.05. If the p-value was greater than 0.05, the null hypothesis was not rejected.

The second test used for testing normality is the D'Agostino-Pearson test. This test determines the skewness and kurtosis of the data and calculates how far the values differ from the value that is expected for the normal distribution [31]. The skewness of the data is a measure of symmetry. The kurtosis of the data is a measure of the sharpness or heaviness of the tails of a distribution. Both the skewness and the kurtosis have a value of 0 for the normal distribution [32]. As with the Shapiro-Wilk test, the null hypothesis H_0 assumes the data are normally distributed, and the alternative hypothesis H_a assumes they are not. The test was also conducted with a significance level α of 0.05 using the Real Statistics add-in. A p-value less than 0.05 resulted in the rejection of the null hypothesis, while a p-value greater than 0.05 indicated that the null hypothesis could not be rejected.

The normal Q-Q plot was used to visualize the distribution of the data and to test for normality. The observed quantiles and the expected quantiles were plotted against one another [32]. A distribution is considered to be approximately normal if all the points are on or near the reference line while deviations from the reference line indicate possible outliers [31].

To test the normality of the data, all three methods were used. If both the Shapiro-Wilk expanded test and the D'Agostino-Pearson test rejected the null hypothesis, a nonparametric statistical test was used to compare the FT and RP datasets. If both tests failed to reject the null hypothesis, a parametric test was used to compare the FT and RP datasets. If the Shapiro-Wilk expanded test and the D'Agostino-Pearson test had conflicting results, the Q-Q plot was used to visually assess if normality can be assumed.

The two-tailed paired t-test was used to compare the datasets when the paired differences between the FT and RP datasets follow an approximately normal distribution. The null hypothesis H_0 of this test states that there is no difference between the two paired datasets. This means that the population mean of the difference is $\mu_d = 0$. The alternative hypothesis H_a of this test states that μ_d has a value different from zero. Equation 1 presents the formula used to determine the test statistic t for the paired t-test.

$$t = \frac{\bar{x}_d - \mu_d}{s_d/\sqrt{n}} \tag{1}$$

In this formula, $\bar{x}_d$ is the mean of the differences between FT and RP, μ_d is the population mean of the difference, s_d is the standard deviation of the difference and n is the sample size [25]. The p-value is determined using the observed t-statistic, the degrees of freedom, and the significance level α. The null hypothesis was rejected if the p-value was smaller than 0.05. If the p-value was greater than 0.05, the null hypothesis could not be rejected. The two-tailed paired t-test was performed using the Real Statistics add-in package in Excel. The test was performed with a significance level α of 0.05.

The Wilcoxon signed-rank test was used to compare the datasets when the null hypothesis of the normality test was rejected. The null hypothesis H_0 of this test states that there is no significant difference between the two datasets. This means that the median of the differences is $M_d = 0$. The alternative hypothesis H_a states that there is a significant difference between the two datasets. The Wilcoxon signed-rank test uses the differences between the RP dataset and the FT dataset to calculate the test statistic. The absolute differences are ranked from smallest to largest and they are assigned a label according to the sign of the difference. Equations 2 and 3 represent the formulas used to calculate the rank sums for positive ranks and negative ranks, respectively.

$$W^+ = \sum(\text{ranks of positive differences}) \tag{2}$$

$$W^- = \sum(\text{ranks of negative differences}) \tag{3}$$

The test statistic W is the smallest of W^+ and W^-. The p-value is determined using the observed test statistic W, the sample size n, and the significance level α [33]. The null hypothesis was rejected if the p-value was smaller than 0.05. If the p-value was greater than 0.05, the null hypothesis could not be rejected. The Wilcoxon signed-rank test was performed using the Real Statistics add-in package in Excel with a significance level α of 0.05.

3.4 Statistical Power and Sample Size

When performing statistical tests, there are two types of errors that can lead to the wrong conclusion. A Type I error gives a false positive and leads to the rejection of the null hypothesis even though it is true [34]. The acceptable level of this error is denoted by the significance level α, resulting in a 5% chance of a Type I error in this study. A Type II error gives a false negative, which means the null hypothesis will not be rejected, even though it is false. The probability of a Type II error is given by β. The power of a statistical test is defined as $(1-\beta)$, representing the probability of rejecting the null hypothesis when it is false [35]. To ensure statistical validity of the results of this study, a high statistical power is required.

The power of the statistical tests was determined using an a priori power analysis. This method was used to calculate the minimum sample size that is required to achieve the desired power [36]. In this study, the minimum sample sizes were calculated for both the two-tailed paired t-test and the Wilcoxon signed-rank test using the G*Power software [35]. G*Power calculates the required sample size based on the selected test, desired power, significance level and the population effect size to be detected with a probability of $(1-\beta)$ [37]. The effect size is a standardized measure of the smallest difference to be detected by a statistical test [38].

The a priori power analysis tool in G*Power uses Cohen's conventions for interpreting effect sizes [37]. According to these guidelines, an effect size of 0.2 is considered small, 0.5 medium, and 0.8 large. In this study, an effect size of 0.5 was used to perform the a priori power analysis. This indicates that the difference to be detected with a probability of $(1-\beta)$ corresponds to half of the standard deviation [38].

For the two-tailed paired t-test, the required sample size was calculated with a desired power of 0.99, a significance level of 0.05, and an effect size of 0.5. For the Wilcoxon signed-rank test, G*Power uses an approximation by applying an additional parameter k. This parameter represents the asymptotic relative efficiency (ARE) compared to the paired t-test and is dependent on the selected parent distribution [37]. In this study, the "min ARE" option was used, representing the most conservative calculation. This choice gives the largest required sample size, ensuring sufficient statistical power.

3.5 Evaluation Parameters

To compare the quality and deliverability of the generated treatment plans, several evaluation parameters were used. First, the dose distribution to the target volumes and organs at risk were compared. For each relevant structure, the dose-volume parameters of the RP plan and the FT plan were compared. This allowed for a quantitative comparison of the clinical objectives.

To evaluate plan deliverability, monitor units (MU) and the complexity metric were used. The total number of monitor units was calculated for each plan and divided by the prescribed dose in Gray to obtain the monitor units per Gray. This parameter serves as a measure of delivery efficiency. An increase in the number of MU/Gy typically results in a lower gamma passing rate due to increased radiation leakage and internal scatter [39]. The gamma passing rate is a commonly used metric that quantifies the deviation between the calculated and the measured dose distribution [20]. A low gamma passing rate indicates a greater difference between the calculated and measured dose distribution.

The complexity metric, introduced by Younge et al. (2012) [40], was also used to evaluate plan deliverability. Equation 4 represents the formula used to calculate the complexity metric.

$$CM = \frac{1}{MU} \sum_{i=1}^{n} MU_i \text{ x } \frac{y_i}{A_i} \qquad (4)$$

Here, MU is the total number of monitor units in the arc, n is the number of control point apertures, MU_i is the number of monitor units delivered through aperture i, y_i is the aperture perimeter excluding the MLC leaf ends and A_i is the open area of aperture i. A higher CM value indicates a more complex plan [5].

4 Results

4.1 Minimum Sample Size

The minimum sample size required was calculated using G*Power software. Figures 4 and 5 illustrate the required sample size as a function of power for the two tailed paired t-test and the Wilcoxon signed-rank test, respectively.

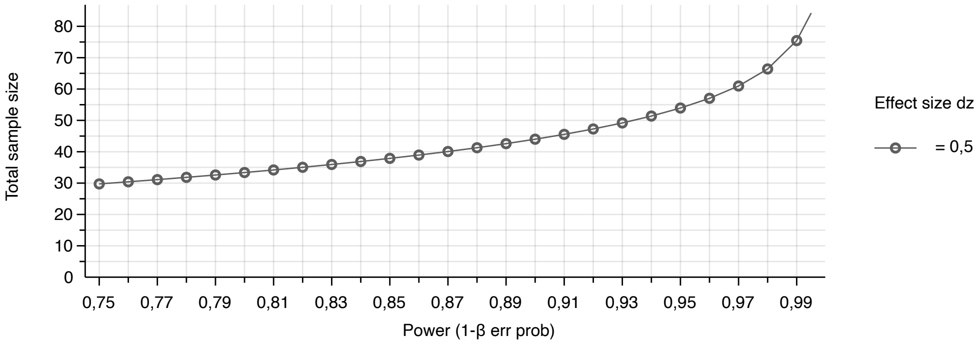

Figure 4: Sample size as a function of power for two-tailed paired t-test

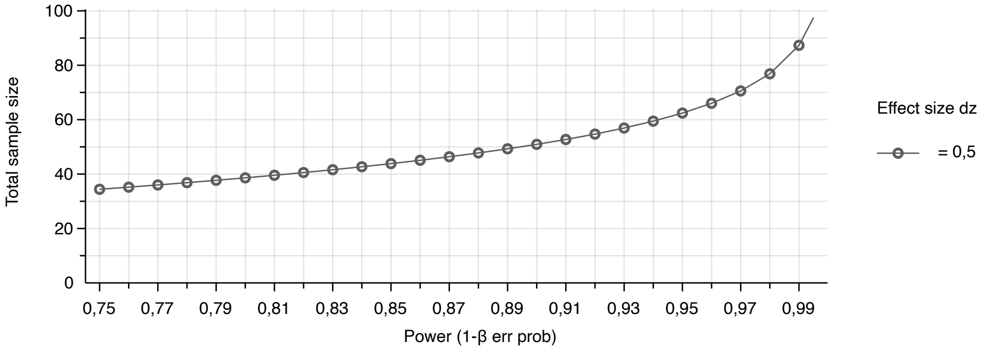

Figure 5: Sample size as a function of power for Wilcoxon signed-rank test

For the two-tailed paired t-test, the minimum required sample size to achieve a statistical power of 0.99 with a significance level of 0.05 and an effect size of 0.5 is 76. For the Wilcoxon signed-rank test, the minimum required sample size under the same conditions is 88.

4.2 Initial Results

Out of the 200 prostate cases that were used, 35 plans met all constraints using only the KBP model. The other 165 plans were optimized using the fine-tuning algorithm. The fine-tuning process addressed the unmet constraints in 99 cases. In 54 cases, the plan did not meet all constraints after ten loops and the process automatically stopped. In the 12 remaining cases, the fine-tuning process stopped before completing the ten loops, however, the plans did not meet all the constraints.

The fine-tuning process resolved the boost D99% constraint in 20 cases, the boost D95% constraint in 29 cases, the boost D50% constraint in 3 cases, the boost Dmax constraint in 40 cases, the body Dmax constraint in 63 cases, the rectum V60Gy constraint in 58 cases, the bladder V60Gy constraint in 15 cases and the bowels V58.8 constraint in 1 case.

The 12 cases that did not complete the fine-tuning process all failed to meet the boost D99% constraint. To address this problem, an additional constraint was incorporated into the script and the cases were reprocessed. In 5 of the cases, all constraints were met after fine-tuning. The other 7 failed to meet all constraints after completing 10 iterations. This resulted in a total of 61 cases out of the 165 that failed to meet all constraints and 104 cases that met all constraints after completing the fine-tuning process. The specific constraints that were resolved by the fine-tuning process are shown in Table 3.

Table 3: Constraints resolved by the fine-tuning process

Boost D99% (%)	25
Boost D95% (%)	33
Boost D50% (%)	3
Boost V107% (cc)	40
Body V107% (cc)	64
Rectum V60Gy (cc)	62
Bladder V60Gy (cc)	15
Bowels V58,5 (cc)	1

4.3 Full Dataset

To evaluate the overall impact of the fine-tuning process, all 165 cases that required fine-tuning were used. For cases that failed to meet all constraints after completing the fine-tuning process, the last plan was used in the FT dataset.

4.3.1. Normality Test

The first step of the statistical analysis was to test the normality of the data. Since the paired t-test uses the mean of the differences to determine if the observed difference is statistically significant, the normality of the differences in the data pairs is tested. A third dataset, containing the differences in the data pairs (RP – FT), was created to perform the normality tests. Table 4 provides an overview of the results of the normality tests for the DVH parameters of the full dataset.

Table 4: Normality test results for the DVH parameters of the full dataset

	Shapiro-Wilik		**D’Agostino-Pearson**	
	p-value	Reject H_0	p-value	Reject H_0
Boost D99% (%)	8.87E-06	yes	2.64E-04	yes
Boost D95% (%)	3.81E-08	yes	1.88E-09	yes
Boost D50% (%)	2.95E-05	yes	1.59E-05	yes
Boost D5% (%)	1.90E-10	yes	0.00E+00	yes
Boost Dmax (%)	0.00E+00	yes	0.00E+00	yes
Boost V107% (cc)	0.00E+00	yes	0.00E+00	yes
Body Dmax (%)	7.17E-03	yes	2.67E-02	yes
Body V107% (cc)	1.05E-12	yes	2.55E-15	yes
Rectum V60Gy (cc)	2.22E-16	yes	0.00E+00	yes
Rectum V50Gy (%)	1.11E-16	yes	0.00E+00	yes
Rectum V40Gy (%)	0.00E+00	yes	0.00E+00	yes
Rectum V30Gy (%)	0.00E+00	yes	0.00E+00	yes
Rectum V26Gy (%)	9.66E-15	yes	8.44E-15	yes
Rectum V20Gy (%)	0.00E+00	yes	0.00E+00	yes
Rectum Dmean (Gy)	0.00E+00	yes	0.00E+00	yes
Left femoral V41Gy (%)	0.00E+00	yes	0.00E+00	yes
Right femoral V41Gy (%)	8.57E-12	yes	8.19E-07	yes
Bladder V63,6Gy (cc)	1.26E-10	yes	3.59E-07	yes
Bladder V60Gy (%)	4.81E-13	yes	2.22E-16	yes
Bladder V49Gy (%)	4.44E-16	yes	0.00E+00	yes
Bladder V41Gy (%)	0.00E+00	yes	0.00E+00	yes
Bladder V31Gy (%)	0.00E+00	yes	2.21E-13	yes
Bowels V58,5Gy (cc)	0.00E+00	yes	0.00E+00	yes
Bowels V41Gy (cc)	8.87E-06	yes	2.64E-04	yes
Bowels V36Gy (cc)	3.81E-08	yes	1.88E-09	yes

The results of both the Wilcoxon signed-rank test and the D’Agostino-Pearson test indicate that the null hypothesis of normality should be rejected for all evaluated datasets. As a result, the Wilcoxon signed-rank test was used for further comparison of the DVH parameters between FT and RP in the full dataset.

For the complexity parameters, MU/Gy and average CM, the normality of the differences in the data pairs was tested using the same method. Table 5 provides an overview of the results of the normality tests for the complexity of the full dataset.

Table 5: Normality test results for the complexity parameters of the full dataset

	Shapiro-Wilk		**D’Agostino-Pearson**	
	p-value	Reject H_0	p-value	Reject H_0
MU/Gy	1.60E-05	yes	2.55E-01	no
Average CM	2.53E-04	yes	1.25E-06	yes

For the average CM, the results for both the Shapiro-Wilk test and the D'Agostino-Pearson test indicated that the null hypothesis of normality should be rejected. For the MU/Gy parameter, the results of the two tests were contradicting. The Q-Q plot, shown in Figure 6, was used to decide if the assumption of normality should be rejected.

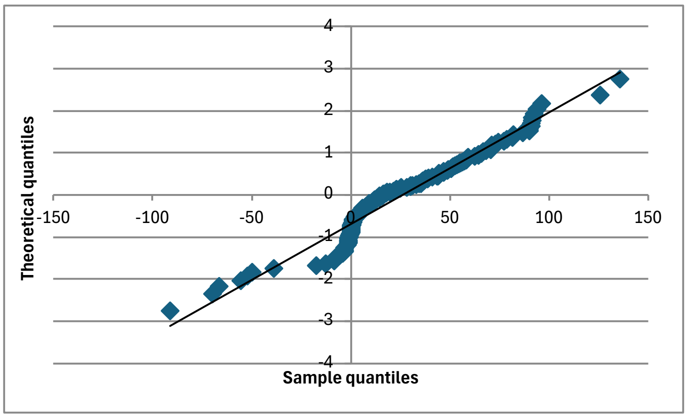

Figure 6: Q-Q plot of the differences in MU/Gy for the full dataset

The Q-Q plot for the MU/Gy differences shows that although most data points lie close to the reference line, there are notable deviations. These deviations, combined with the result of the Shapiro-Wilk test, support the rejection of the null hypothesis of normality. As a result, the Wilcoxon signed-rank test was used to compare MU/Gy between RP and FT plans, providing a more conservative approach.

4.3.2. PTV

The difference in dose coverage of the PTV between the initial KBP-generated plan (RP) and the optimized plan (FT) was compared using DVH parameters. The D99%, D95%, D50%, D5% and V107% were used in this comparison. Table 6 presents the results of the DVH analysis for the PTV, including mean values, standard deviations, relative difference and corresponding p-values from the statistical tests performed.

Table 6: PTV dose results for the full dataset

	FT	RP	p-value	Relative difference (%)
Boost D99% (%)	91.73±1.23	90.62±1.81	4.95E-16	1.23
Boost D95% (%)	95.70±0.58	95.14±0.10	1.99E-10	0.53
Boost D50% (%)	100.29±0.13	100.32±0.19	3.46E-01	-0.03
Boost D5% (%)	103.26±.25	103.57±0.40	2.18E-24	-0.30
Boost V107% (cc)	0.01±0.02	0.13±0.41	3.79E-31	-91.87

The results from Table 6 show that the Boost D99% (%) and Boost D95% (%) were significantly higher in the FT plans compared to the RP plans, indicating improved dose coverage. The Boost D5% (%) and Boost V107% (cc) were significantly lower in the FT plans, suggesting a reduction in hotspots and excessively high-dose volumes. There was no statistically significant difference between the FT and RP plans for the Boost D50% (%). Although the results were statistically significant (p-value < 0.05) for most metrics, the relative differences only indicated a meaningful improvement for the Boost V107% (cc) constraint, where FT plans show a 91.87% reduction.

4.3.3. OAR Sparing

The difference in OAR sparing between the initial plan, generated with the KBP model, and the optimized plan was compared using DVH parameters. Table 7 provides an overview of the results of the DVH analysis for the OARs.

Table 7: OAR dose results for the full dataset

	FT	RP	p-value	Relative difference (%)
Body V107% (cc)	0.02±0.03	0.19±0.47	1.34E-31	-87.12
Rectum V60Gy (cc)	1.03±0.60	1.26±0.63	3.07E-06	-18.12
Rectum V50Gy (%)	11.09±4.03	10.24±3.57	4.70E-23	8.29
Rectum V40Gy (%)	17.99±6.19	16.51±5.29	1.33E-38	8.93
Rectum V30Gy (%)	26.06±9.33	23.22±6.93	4.95E-39	12.23
Rectum V26Gy (%)	30.14±10.68	26.30±7.57	8.72E-45	14.59
Rectum V20Gy (%)	37.46±13.07	31.83±8.77	1.06E-39	17.71
Rectum Dmean (Gy)	20.67±4.62	18.89±3.58	8.52E-42	9.44
Left Femoral V41Gy (%)	0.00±0.05	0.03±.0.28	2.50E-01	-84.96
Right Femoral V41Gy (%)	0.02±0.05	0.03±0.20	8.30E-02	-77.46
Bladder V63,3 Gy (cc)	0.02±0.04	0.08±0.26	1.77E-11	-74.69
Bladder V60Gy (%)	3.18±1.20	3.96±2.12	8.33E-04	-19.69
Bladder V49Gy (%)	12.05±6.38	11.63±6.27	3.98E-13	3.65
Bladder V41Gy (%)	17.17±8.49	16.39±7.99	1.20E-19	4.76
Bladder V31Gy (%)	24.40±11.89	22.67±10.33	9.45E-35	7.60
Bowels V58,5Gy (cc)	0.05±0.23	0.07±0.38	1.00E+00	-35.21
Bowels V41Gy (cc)	1.64±3.79	1.63±3.84	8.47E-01	0.40
Bowels V36 Gy (cc)	2.51±5.43	2.56±5.50	4.86E-01	-2.26

The first constraint that was used is the Body V107% (cc). This constraint was significantly lower in the FT plans compared to the RP plans. This result indicates a significant reduction of approximately 87.1% in the volume of the body receiving more than 107% of the prescribed dose, suggesting that the fine-tuning process effectively limits excessive doses outside the PTV.

The Rectum V60Gy (cc) was significantly lower in the FT plans, with an average reduction of approximately 18.3% compared to the RP plans. This suggests improved high-dose sparing of the rectum. In contrast, all other rectum constraints, including V50Gy (%), V40Gy (%), V30Gy (%), V26Gy (%), V20Gy (%), and Dmean (Gy), were slightly but significantly higher in the FT plans. The relative increases ranged from approximately 5% to 18%. Overall, the fine-tuning process improved rectal sparing at high doses while maintaining acceptable increases at lower dose levels.

For both femoral heads, the V41Gy (%) values were lower in the FT plans compared to the RP plans, with relative reductions of approximately 85% for the left and 77% for the right femoral head. However, these differences were not statistically significant. Given the very low absolute values, these reductions are unlikely to be clinically meaningful.

For the bladder, the FT plans resulted in a significantly lower volume receiving high doses. Specifically, Bladder V63.3Gy (cc) was reduced by approximately 75%, and Bladder V60Gy (%) by about 20%, both showing statistically significant improvements. At lower dose levels, V49Gy (%), V41Gy (%), and V31Gy (%), the FT plans showed small but statistically significant increases of 3.65%, 4.76%, and 7.60%, respectively. These results indicate that the fine-tuning process effectively reduces high-dose exposure to the bladder while maintaining acceptable increases at lower dose levels.

For all bowel constraints, there were no statistically significant differences between FT and RP plans. The values for Bowels V58.5Gy (cc), V41Gy (cc), and V36Gy (cc) were very low in both groups, with p-values of 1.00, 0.847, and 0.486, respectively. Although minor relative differences were observed, the absolute differences are minimal.

4.3.4. Complexity

To evaluate the effect of the fine-tuning process on the plan complexity, the number of MU/Gy and the average CM are used. To visualize the difference between the RP plans and the FT plans, boxplots are used. The boxplot comparisons of the number of MU/Gy and the average CM are illustrated in Figure 7 and Figure 8, respectively.

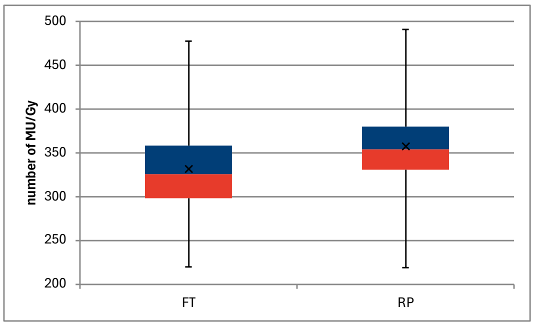

Figure 7: Boxplot comparison of the number of MU/Gy for the full dataset

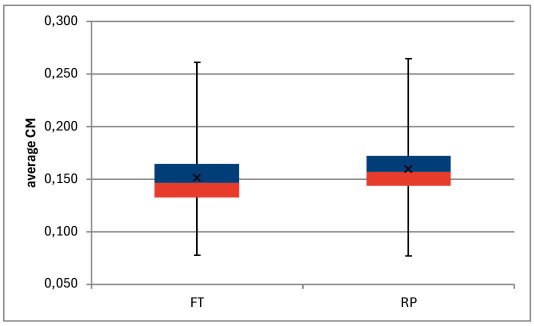

Figure 8: Boxplot comparison of the average CM for the full dataset

Figure 7 shows that the distribution of both groups was similar, but FT had a lower average and mean number of MU/Gy compared to RP. The boxplot comparison of the average CM in Figure 8 also indicated a similar distribution and a small reduction of average and mean in the FT plans. This reduction in both MU/Gy and average CM in the FT plans was confirmed by the results of the Wilcoxon signed-rank test, shown in Table 8.

Table 8: Complexity results for the full dataset

	FT	RP	p-value	Relative difference (%)
Average CM	0.151±0.03	0.160±0.03	8.42E-8	-5.4
MU/Gy	331.5±46.2	357.8±39.9	6.94E-17	-7.4

The Wilcoxon signed-rank test resulted in a p-value lower than 0.05 for both parameters, indicating a statistically significant difference. Specifically, the average CM was reduced by 5.4% and the MU/Gy by 7.4%. These reductions suggest that the fine-tuning process did not increase the plan complexity.

4.4 Successful Subset

To specifically assess the effect of the fine-tuning process on cases that successfully met all clinical constraints after fine-tuning, a subset of the data was analyzed. This subset included only the cases where the final fine-tuned plan met all predefined dose constraints.

4.4.1. Normality Test

The approach for testing the normality of the paired differences in the successful subset was the same as that used for the full dataset. The results of the normality tests for the DVH parameters are presented in Table 9.

Table 9: Normality test results for the DVH parameters of the successful subset

	Shapiro-Wilik		**D'Agostino-Pearson**	
	p-value	Reject H_0	p-value	Reject H_0
Boost D99% (%)	5.33E-06	yes	2.41E-05	yes
Boost D95% (%)	3.29E-07	yes	1.84E-06	yes
Boost D50% (%)	9.00E-05	yes	1.05E-03	yes
Boost D5% (%)	1.49E-03	yes	1.43E-03	yes
Boost Dmax (%)	0.00E+00	yes	0.00E+00	yes
Boost V107% (cc)	8.88E-16	yes	0.00E+00	yes
Body Dmax (%)	2.60E-05	yes	5.56E-05	yes
Body V107% (cc)	2.08E-11	yes	4.58E-13	yes
Rectum V60Gy (cc)	6.88E-15	yes	0.00E+00	yes
Rectum V50Gy (%)	4.44E-16	yes	0.00E+00	yes
Rectum V40Gy (%)	3.66E-15	yes	0.00E+00	yes
Rectum V30Gy (%)	1.44E-15	yes	0.00E+00	yes
Rectum V26Gy (%)	3.65E-14	yes	0.00E+00	yes
Rectum V20Gy (%)	0.00E+00	yes	0.00E+00	yes
Rectum Dmean (Gy)	0.00E+00	yes	0.00E+00	yes
Left femoral V41Gy (%)	9.50E-13	yes	1.11E-16	yes
Right femoral V41Gy (%)	2.66E-12	yes	5.14E-11	yes
Bladder V63,6Gy (cc)	2.12E-08	yes	1.52E-09	yes
Bladder V60Gy (%)	9.80E-11	yes	8.86E-14	yes
Bladder V49Gy (%)	4.04E-10	yes	2.21E-09	yes
Bladder V41Gy (%)	0.00E+00	yes	0.00E+00	yes
Bladder V31Gy (%)	2.00E-15	yes	1.48E-07	yes
Bowels V58,5Gy (cc)	2.22E-15	yes	8.76E-08	yes
Bowels V41Gy (cc)	5.33E-06	yes	2.41E-05	yes
Bowels V36Gy (cc)	3.29E-07	yes	1.84E-06	yes

For the successful subset, the results of both the Wilcoxon signed-rank test and the D'Agostino-Pearson test indicated that the null hypothesis of normality should be rejected for all evaluated datasets. Therefore, the Wilcoxon signed-rank test was used to compare the DVH parameters between FT and RP plans in the subset.

The same approach was used to test the normality of the differences in the data pairs for the complexity parameters, MU/Gy, and average CM. Table 10 provides an overview of the results of the normality tests for the complexity of the successful subset.

Table 10: Normality test results for the complexity parameters of the successful subset

	Shapiro-Wilk		**D'Agostino-Pearson**	
	p-value	Reject H_0	p-value	Reject H_0
MU/Gy	1.65E-06	yes	2.54E-02	yes
Average CM	4.25E-03	yes	2.12E-02	yes

For the average CM and the MU/Gy, the results for both the Shapiro-Wilk test and the D'Agostino-Pearson test indicated that the null hypothesis of normality should be rejected. As a result, the Wilcoxon singed-rank test was used to evaluate both complexity parameters.

4.4.2. PTV

The difference in dose coverage of the PTV between the initial KBP-generated plan (RP) and the optimized plan that met all constraints (FT) was compared using DVH parameters. The D99%, D95%, D50%, D5% and V107% were used in this comparison. Table 11 presents the results of the DVH analysis for the PTV, including mean values, standard deviations, relative differences, and corresponding p-values from the statistical tests performed.

Table 11: PTV dose results for the successful subset

	FT	**RP**	**p-value**	**Relative difference (%)**
Boost D99% (%)	91.80±0.83	91.14±1.52	1.15E-04	0.72
Boost D95% (%)	95.71±0.53	95.42±0.08	8.07E-03	0.30
Boost D50% (%)	100.29±0.13	100.30±0.19	5.40E-01	-0.01
Boost D5% (%)	103.25±0.21	103.46±0.28	2.23E-13	-0.20
Boost V107% (cc)	0.00±0.01	0.07±0.16	2.70E-18	-94.62

In the subset of successfully fine-tuned plans, the Boost D99% (%) and Boost D95% (%) were significantly higher in FT plans compared to RP plans, although the relative increases of 0.72% and 0.30% are minor. The Boost D5% (%) and Boost V107% (cc) values were significantly reduced, with relative decreases of 0.20% and 94.62%, respectively, indicating a reduction in hotspots and volumes receiving excessive doses. No statistically significant difference was found for the Boost D50% (%), which showed a negligible relative difference of -0.01%.

4.4.3. OAR Sparing

The difference in OAR sparing between the initial plans generated by the KBP model and the successfully fine-tuned plans that met all constraints was evaluated using DVH parameters. Table 12 summarizes the results of the DVH analysis for the OARs.

Table 12: OAR dose results for the successful subset

	FT	RP	p-value	Relative difference (%)
Body V107% (cc)	0.01±0.01	0.11±0.18	6.02E-20	-88.02
Rectum V60Gy (cc)	0.73±0.22	1.12±0.60	2.72E-13	-34.90
Rectum V50Gy (%)	9.80±3.68	9.39±3.42	1.64E-08	4.35
Rectum V40Gy (%)	16.19±5.26	15.53±5.24	1.52E-20	4.29
Rectum V30Gy (%)	23.17±7.92	22.05±6.90	1.47E-22	5.06
Rectum V26Gy (%)	26.70±8.55	25.06±7.54	1.01E-26	6.55
Rectum V20Gy (%)	33.04±10.28	30.65±8.63	1.09E-25	7.81
Rectum Dmean (Gy)	19.02±4.00	18.21±3.53	9.82E-24	4.44
Left Femoral V41Gy (%)	0.00±0.00	0.00±0.02	1.00E+00	-83.33
Right Femoral V41Gy (%)	0.01±0.05	0.02±0.15	6.88E-01	-67.80
Bladder V63,3 Gy (cc)	0.01±0.03	0.04±0.06	8.62E-06	-60.54
Bladder V60Gy (%)	2.81±1.07	3.11±1.58	3.15E-01	-9.59
Bladder V49Gy (%)	9.33±4.47	9.11±4.34	9.38E-06	2.41
Bladder V41Gy (%)	13.65±5.99	13.28±5.71	2.05E-09	2.75
Bladder V31Gy (%)	19.48±8.19	18.79±7.64	1.34E-17	3.66
Bowels V58,5Gy (cc)	0.01±0.09	0.01±0.11	1.00E+00	-5.30
Bowels V41Gy (cc)	1.06±2.34	1.08±2.38	3.15E-01	-1.36
Bowels V36 Gy (cc)	1.69±3.49	1.71±3.54	3.75E-01	-0.91

The Body V107% (cc) was significantly lower in the FT plans compared to the RP plans, corresponding to a relative reduction of 88.02%. This indicates a significant decrease in the volume of the body receiving more than 107% of the prescribed dose.

In the subset of successfully fine-tuned plans, all DVH parameters for the rectum showed statistically significant differences between the FT and the RP plans. Specifically, the V60Gy (cc) was reduced by 34.90% in FT plans compared to the RP plans, indicating improved sparing of the rectum at high dose levels. In contrast, lower dose parameters including V50Gy (%), V40Gy (%), V30Gy (%), V26Gy (%), V20 Gy (%), and the mean dose were higher in the FT plans. Although the differences were statistically significant, the relative increases for these parameters are small and acceptable.

No statistically significant differences were observed for either femoral head. The results showed relative differences of -83.33% for the left and -67.80% for the right femoral head, however, the absolute differences were negligible.

For the bladder, the V63.3 Gy (cc) was significantly reduced in the FT plans compared to the RP plans, with a relative decrease of 60.54%, indicating improved high-dose sparing. No statistically significant difference was observed for the V60Gy (%), despite a relative reduction of 9.59%. The lower dose parameters V49Gy (%), V41Gy (%), and V31Gy (%) were higher in the FT plans. Although these differences were statistically significant, the relative increases were small and acceptable.

For the bowels, no statistically significant differences were observed between the FT and RP plans for any of the evaluated dose parameters. The relative differences for V58.5Gy (cc), V41Gy (cc), and V36Gy (cc) were minimal, and the absolute values were low, indicating comparable dose exposure to the bowels in FT and RP plans.

4.4.4. Complexity

To evaluate deliverability of the plans that met all constraints after fine-tuning, the number of units per gray and a complexity metric are used. The boxplot comparisons of the number of MU/Gy and the average CM are illustrated in Figure 9 and Figure 10, respectively.

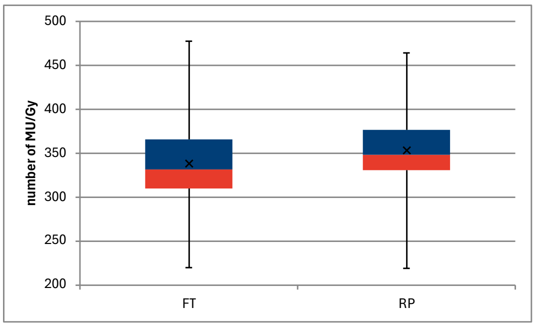

Figure 9: Boxplot comparison of the number of MU/Gy for the successful subset

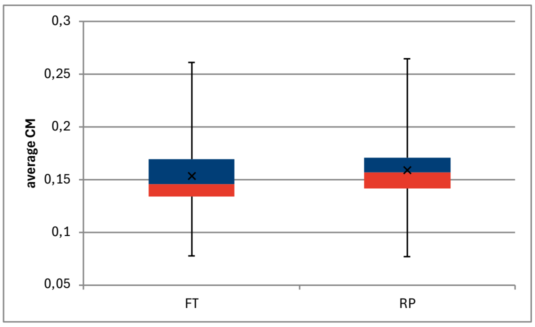

Figure 10: Boxplot comparison of the CM for the successful subset

Figure 9 shows that the median number of MU/Gy was lower in the FT plans compared to the RP plans. Both groups showed similar variability, but the FT plans had a slightly lower average. The boxplot comparison of the average CM, which is illustrated in Figure 10, showed a slightly lower mean and average in the FT plans compared to the RP plans. The reduction in both MU/Gy and average CM in the FT plans was confirmed by the results of the Wilcoxon signed-rank test, shown in Table 13.

Table 13: Complexity results for the successful subset

	FT	RP	p-value	Relative difference (%)
Average CM	0.154±0.03	0.159±0.03	6.52E-03	-3.3
MU/Gy	338.6±48.0	353.8±39.1	1.80E-09	-4.3

The p-value for both parameters was lower than 0.05, confirming a statistically significant difference between the FT and RP plans. The relative reductions for the average CM and the number of MU/Gy were 3.3% and 4.3%, respectively. These results indicate a small reduction in plan complexity after fine-tuning.

5 Discussion

The goal of this study was to evaluate a fully automated fine-tuning algorithm designed to improve prostate VMAT treatment plans initially generated using a KBP model. A retrospective analysis was performed using treatment data from 200 previously treated prostate cancer patients, randomly selected from the clinical database. For each patient, a new course was created in which the contoured simulation CT was uploaded. New treatment plans were generated by entering the patient ID into the script interface and running the script. The script first invoked the KBP model and then applied the automated fine-tuning process. Only those plans that failed to meet all constraints using only the KBP model, were included for evaluation.

For each DVH and complexity parameter, two datasets were generated. The RP dataset contained the values from the initial treatment plans produced by the knowledge-based planning (KBP) model. The FT dataset included the corresponding values from the final plans generated through the automated fine-tuning script. The effect of the fine-tuning process was assessed by comparing the RP and FT values for each parameter.

The fine-tuning algorithm was evaluated in two parts. First, the overall impact of the fine-tuning process was analyzed by including all treatment plans that required fine-tuning. For the plans that failed to meet all constraints, the FT values of the last generated plan were used for the evaluation. Secondly, a subset that only contained the cases where the fine-tuning process resolved all unmet constraints was evaluated. This successful subset was evaluated to specifically assess the effect of the fine-tuning process on successful cases.

During the initial evaluation of the generated plans, an unexpected issue with the script was discovered. The fully automated planning script was designed to result in one of three expected outcomes. A treatment plan either met all constraints using only the KBP model, the fine-tuning process resolved all unmet constraints, or the fine-tuning process failed to resolve all unmet constraints after completing 10 loops. The initial evaluation of the generated plans showed that in some cases, a fourth and unexpected outcome occurred. In these cases, the fine-tuning process was terminated prematurely, even though the treatment plan still failed to meet all constraints. Further investigation of these cases showed that all of them failed to meet the boost D99% constraint. This problem was revealed to be a result of an assumption that was made during scripting: that if a plan met the boost D95% constraint it would also meet the boost D99% constraint. To address this problem, an additional constraint for the boost D99% was implemented in the script, which successfully prevented this issue.

The discovery of this problem highlights the importance of performing a validation study on a large dataset of patients when developing automated planning techniques. Rare but clinically significant problems can go undetected when small datasets are used for testing. It also emphasizes the need for manual verification and evaluation by physicists, even in automated workflows.

Out of the 165 cases that required fine-tuning, the process resolved all unmet constraints in 104 of them. These cases were used to create the successful subset for the data analysis. The normality tests indicated that the null hypothesis of normality for the differences in the data pairs should be rejected for all DVH and complexity parameters. As a result, all statistical comparisons of the parameters in this subset were performed using the Wilcoxon signed-rank test.

In the successful subset, most of the DVH parameters that were analyzed showed a statistically significant difference between the RP and FT plans, although not all differences were meaningful. The most notable improvements were the reduction of the boost V107% (cc) and the body V107% (cc) constraints. These improvements indicate a reduction of hotspots in the PTV and a reduction in the volume of the body that receives more than 107% of the prescribed dose. For both the bladder and the rectum, the results showed improved sparing at high-dose levels and a slight increase at low-dose levels. Although these increases were statistically significant, the relative differences were small and acceptable and are likely a result of the trade-offs made during optimization to achieve an acceptable plan.

These results showed that the fine-tuning process was able to resolve all unmet constraints in most of the cases. Additionally, the fine-tuning also slightly decreased the complexity of the plans, as indicated by the reduced number of MU/Gy and average CM values. These results show that the fine-tuning process can improve plan quality without compromising dose accuracy for prostate cases.

The normality tests indicated that the null hypothesis of normality for the differences in the data pairs should also be rejected for all DVH and complexity parameters in the full dataset. As a result, all statistical comparisons of the parameters were performed using the Wilcoxon signed-rank test.

The analyses of the full dataset, consisting of all 165 cases that required fine-tuning, showed similar trends in trade-offs between the RP and FT plans. The most notable improvements were observed in the Boost V107% (cc) and the Body V107% (cc). The results of this dataset also showed improved sparing at high-dose levels and a slight increase at low-dose levels for the rectum and bladder. The comparison of the number of MU/Gy and the average CM showed a slight, but statistically significant reduction in plan complexity.

The results of the full dataset show that even if the fine-tuning process fails to solve all unmet constraints, it could still serve as an effective starting point for manual refinement. The 10 generated plans can be used by the physicists to get an idea of the possible trade-offs.

When interpreting the results, it is important to consider that unmet constraints in the delivered treatment plans were not used as exclusion criteria during patient selection. As a result, some of the selected cases included treatment plans that did not fully meet all predefined clinical constraints used to evaluate the fine-tuning process. It is possible that if only patients whose delivered plan met all constraints were used, the observed success rate of the fine-tuning process would have been higher.

A limitation of the fine-tuning process, and automated planning methods in general, is the inability to account for patient-specific clinical considerations that may influence planning decisions. The fine-tuning algorithm uses predefined constraints and objectives to optimize the treatment plan, but some patient-specific factors like age, limited lifetime expectancy, or secondary diseases are not taken into account. These case-by-case decisions are difficult to implement into planning tools and should be assessed by experienced clinicians.

The script is currently also limited to treatment planning for prostate cases following the planning protocol described in section 3.1. In future work, this approach of automated fine-tuning could be used for other planning protocols or other tumor sites. The fine-tuning algorithm could also be used to enhance other automated planning models like deep learning-based models.

Another promising area for further research is adaptive radiotherapy, where frequent anatomical changes due to organ motion or tumor regression are used to regularly update the treatment plan. The fine-tuning process could be applied in day-to-day planning to re-optimize a treatment plan based on updated images, addressing anatomical changes over the course of the treatment.

6 Conclusion

This study demonstrated the importance of performing a validation study and the continued need for manual verification and evaluation in automated workflows. The validation study showed that, after the implementation of the additional constraint, the fully automated fine-tuning algorithm can effectively improve prostate VMAT plans initially generated by a knowledge-based model. The algorithm successfully resolved unmet clinical constraints in most cases and led to small but consistent reductions in plan complexity and monitor units per Gray. These results highlight the potential of automated fine-tuning to enhance plan quality, increase efficiency, and reduce manual workload in clinical radiotherapy planning.

Bibliography

[1] IAEA, "Optimized Radiotherapy Approach Could Extend Treatment to 2.2 Million More Cancer Patients, IAEA Co-authored Report Finds," 30 September 2024. [Online]. Available: https://www.iaea.org/newscenter/pressreleases/optimized-radiotherapy-approach-could-extend-treatment-to-22-million-more-cancer-patients-iaea-co-authored-report-finds. [Accessed 22 May 2025].

[2] American Cancer Society, "Treatment types," [Online]. Available: https://www.cancer.org/cancer/managing-cancer/treatment-types.html. [Accessed 22 May 2025].

[3] L. Guangqi, W. Xin and M. Xuelei, "Artificial intelligence in radiotherapy," *Seminars in Cancer Biology,* vol. 86, pp. 160-171, 2022.

[4] 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.

[5] H. Cavus, T. Rondagh, A. Jankelevitch, K. Tournel, M. Orlandini, P. Bulens, L. Delombaerde, K. Geens, W. Crijns and B. Reniers, "Optimizing volumetric modulated arc therapy prostate planning using an automated Fine-Tuning process through dynamic adjustment of optimization parameters," *Physics and Imaging in Radiation Oncology,* vol. 31, no. 100619, 2024.

[6] M. Lederman, "The early history of radiotherapy: 1895-1939," *Int. J. Radiation Oncology Biol. Phys.,* vol. 7, no. 5, pp. 639-648, 1981.

[7] J. Thariat, J.-M. Hannoun-Levi, A. Sun Myint, T. Vuong and J.-P. Gérard, "Past, present and future of radiotherapy for the benefit of patients," *Nature Review Clinical Oncology,* vol. 10, pp. 52-60, 2013.

[8] C. A. Ravi, F. K. Keane, F. E. M. Voncken and C. R. Thomas Jr, "Contemporary radiotherapy: present and future," *The Lancet,* vol. 398, pp. 171-184, 2021.

[9] T. Landberg, J. Chavaudra, J. Dobbs, G. Hanks, K.-A. Johansson, T. Möller and J. Purdy, "Reports of the International Commission on Radiation Units and Measurements," *Journal of the ICRU,* vol. 26, no. 1, pp. 67-70, 1993.

[10] N. G. Burnet, S. J. Thomas, K. E. Burton and J. S. Jefferies, "Defining the tumour and target volumes for radiotherapy," *Cancer imaging,* vol. 4, pp. 153-161, 2004.

[11] J. Unkelbach, M. Alber, M. Bangert, R. Bokrantz, T. C. Y. Chan, J. O. Deasy, A. Fredriksson, B. L. Gorissen, M. van Herk, W. Liu, H. Mahmoudzadeh, O. Nohadani, J. V. Siebers, M. Witte and H. Xu, "Robust radiotherapy planning," *Physics in Medicine and Biology,* vol. 63, no. 22, 2018.

[12] S. Rana, "Intensity modulated radiotherapy versus volumetric intensity modulated arc therapy," *Journal of Medcal Radiation Sciences,* vol. 60, no. 3, pp. 81-83, 2013.

[13] C. Misher, "OncoLink," 14 March 2024. [Online]. Available: https://www.oncolink.org/cancer-treatment/radiation/introduction-to-radiation-therapy/radiation-therapy-treatment-process. [Accessed 23 April 2024].

[14] E. Weiss and C. F. Hess, "The Impact of Gross Tumor Volume (GTV) and Clinical Target Volume (CTV) Definition on the Total Accuracy in Radiotherapy," *Strahlentherapie und Onkologie ,* vol. 179, no. 1, pp. 21-30, 2003.

[15] L. Simon, C. Robert and P. Meyer, "Artificial intelligence for quality assurance in radiotherapy," *Cancer/Radiothérapie,* vol. 25, pp. 623-626, 2021.

[16] P. Mohandass, D. Khanna, B. Nishaanth, C. Saravanan, N. Bhalla, A. Puri and B. Mohandass, "Impact of three different matching methods on patient set-up error in X-ray volumetric imaging for head and neck cancer," *Reports of practical oncology and radiotherapy,* vol. 25, no. 6, pp. 906-912, 2020.

[17] S. Momin, Y. Fu, Y. Lei, J. Roper , J. D. Bradley, W. J. Curran, T. Liu and X. Yang, "Knowledge-based radiation treatment planning: A data driven method survey," *Journal of applied medical physics,* vol. 22, no. 8, pp. 16-44, 2021.

[18] K. L. Moore, "Automated Radiotherapy Treatment Planning," *Seminars in Radiation Oncology,* vol. 29, pp. 209-218, 2019.

[19] I. Foster, E. Spezi and P. Wheeler, "Evaluating the Use of Machine Learning to Predict Expert-Driven Pareto-Navigated Calibrations for Personalised Automated Radiotherapy Planning," *Applied Sciences,* vol. 13, no. 4548, 2023.

[20] A. F. Osman and N. M. Maalej, "Applications of machine and deep learning to patient-specific IMRT/VMAT quality assurance," *Journal of applied clinical medical physics,* vol. 22, no. 9, pp. 20-36, 2021.

[21] A. F. I. Osmans, N. M. Tamam and Y. A. M. Yousif, "A comparative study of deep learning-based knowledge-based planning methods for 3D dose distribution prediction of head and neck," *Journal of applied clinical medical physics,* vol. 24, no. 9, pp. 57-63, 2023.

[22] R. R. Savjani, M. Lauria, S. Bose, J. Deng, Y. Yuan and V. Andrearczyk, "Automated Tumor Segmentation in Radiotherapy," *Seminars in Radiation Oncology,* vol. 32, pp. 319-329, 2022.

[23] C. Ling, X. Han, Z. Peng, H. Xu, J. Chen, J. Wang and W. Hu, "A hybrid automated treatment planning solution for esophageal cancer," *Radiation Oncology,* vol. 14, no. 232, 2019.

[24] B. R. Overholser and K. M. Sowinski, "Biostatistics Primer: Part I," *Nutrition in Clinical Practice,* vol. 22, no. 6, pp. 629-635, 2007.

[25] B. R. Overholser and K. M. Sowinski, "Biostatistics Primer: Part 2," *Nutrition in Clinical Practice,* vol. 23, no. 1, pp. 76-84, 2008.

[26] S. W. Lee, "Methods for testing statistical differences between groups in medical research: statistical standard and guideline of Life Cycle Committee," *Life Cycle,* vol. 2, no. 1, 2022.

[27] C. Zaiontz, "Real Statistics Using Excel," [Online]. Available: https://real-statistics.com/students-t-distribution/problems-data-t-tests/. [Accessed 12 February 2025].

[28] M. Saculinggan and E. Amor Balase, "Empirical Power Comparison Of Goodness of Fit Tests for Normality In The Presence of Outliers," *Journal of Physics: Conference Series,* vol. 435, no. 012041, 2013.

[29] C. Zaiontz, "Shapiro-Wilk Original test," Real Statistics Using Excel, [Online]. Available: https://real-statistics.com/tests-normality-and-symmetry/statistical-tests-normality-symmetry/shapiro-wilk-test/. [Accessed 25 February 2025].

[30] S. G. Kwak and S.-H. Park, "Normality Test in Clinical Research," *Journal of Rheumatic Diseases,* vol. 26, no. 1, pp. 5-11, 2019.

[31] D. ÖZTUNA, A. H. ELHAN and E. TÜCCAR, "Investigation of Four Different Normality Tests in Terms of Type 1 Investigation of Four Different Normality Tests in Terms of Type 1 Error Rate and Power under Different Distributions Error Rate and Power under Different Distributions," *Turkish Journal of Medical Sciences,* vol. 36, no. 3, pp. 171-176, 2006.

[32] P. Msihra, 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, pp. 67-72, 2019.

[33] R. Shier, "The Wilcoxon Signed Rank Test," 2004. [Online]. Available: https://www.statstutor.ac.uk/resources/uploaded/wilcoxonsignedranktest.pdf. [Accessed 22 February 2025].

[34] C. Zaiontz, "Null and Alternative Hypothesis," [Online]. Available: https://real-statistics.com/hypothesis-testing/null-hypothesis/. [Accessed 22 February 2025].

[35] F. Faul, Erdefelder, Edgar, A.-G. Lang and A. Buchner, "G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior Research Methods,* vol. 39, no. 2, pp. 175-191, 2007.

[36] C. Zaiontz, "Statistical Power and Sample Size," [Online]. Available: https://real-statistics.com/hypothesis-testing/statistical-power/. [Accessed 22 february 2025].

[37] Heinrich Heine Universität Düsseldorf, "G*Power," 1 June 2023. [Online]. Available: https://www.psychologie.hhu.de/fileadmin/redaktion/Fakultaeten/Mathematisch-Naturwissenschaftliche_Fakultaet/Psychologie/AAP/gpower/GPowerManual.pdf. [Accessed 9 November 2024].

[38] C. Zaiontz, "Effect size," [Online]. Available: https://real-statistics.com/hypothesis-testing/effect-size/. [Accessed 22 february 2025].

[39] E. Timakova and S. F. Zavgorodni, "Effect of modulation factor and low dose threshold level on gamma pass rates of single isocenter multi-target SRT treatment plans," *Journal of Applied Clinical Medical Physics,* vol. 25, no. 9, 2024.

[40] K. C. Younge, M. M. Matuszak, J. M. Moran, D. L. McShan, B. A. Fraass and D. A. Roberts, "Penalization of aperture complexity in inversely planned volumetric modulated arc therapy," *Medical Physics,* vol. 39, no. 11, pp. 7160-7170, 2012.

[41] E. Marschall and T. Waqanika, "Checking normality in Excel," 2017. [Online]. Available: https://www.sheffield.ac.uk/media/30651/download?attachment. [Accessed January 07 2025].