ADVANCING THE RADIATION ONCOLOGY CLINIC WITH MOTION MANAGEMENT AND AUTOMATIC TREATMENT PLANNING

By

DAMON ANTON SPROUTS

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

August 2022

Supervising Committee:

Yujie Chi, Supervising Professor

Georgios Alexandrakis

Mingwu Jin

Amir Shahmoradi

Dedicated to:

My Grandma Chantal Gavaldon

ACKNOWLEDGEMENTS

It has been a long and fun ride over the past five years. I came to the University of Texas at Arlington with the dream of being a fully American Board of Radiology (ABR) board-certified Medical Physicist. I can truly say that everyone I have met along my journey has helped not just in my research purse but my ultimate career goals.

The first person I would like to thank is my supervising professor Dr. Yujie Chi, who gave a chance to a Bioengineering student that she didn’t know much about to do research in her lab -- Physics Research & Techniques in Radiation Medicine (PRTRM) lab. She guided me to work in a brand-new research direction -- deep learning application in radiation medicine. Her consist patience, understanding and guidance during my PhD training are very valuable. I would like to thank my graduate advisor Dr. Georgios Alexandrakis for helping me navigate working between two departments and helping me figure out who did Medical Physics research at UTA that would help me realize my career goals. I would like to thank the rest of my dissertation committee Dr. Mingwu Jin and Dr. Amir Shahmoradi for taking time out of their busy schedules to guide and help me during this time. I would also like to thank Dr. Chenyang Shen who helped me by showing me the rope of deep learning and was there if I had any questions for better understanding.

I would like to thank all my colleagues at UTA not just in Physics but also in Bioengineering that were there along the way helping me keep in focusing on what I wanted to do and supporting me when I felt I was going nowhere in my research or writing, and I particularly would like to thank Youfang Lai, Harsh Arya, Ananta Chalise, Shiwei Zhou and Marcos Guillen.

ABSTRACT

Advancing the Radiation Oncology Clinic with Motion Management and Automatic Treatment Planning

Damon Anton Sprouts

The University of Texas at Arlington, 2022

Supervising Professor(s): Yujie Chi

The leading cause of premature death (death under the age of 70) is cancer. The top five cancers for both male and female are: lung, colorectum, pancreas, breast cancer, and prostate. In 2020 there was an estimated 19.3 million new cases with an estimated 9.9 million deaths. The cancer burden is expected to grow to 28.4 million by the year 2040. Surgery, chemotherapy, and radiotherapy are the three pillars in the modern clinic for cancer treatment. In radiotherapy, ionizing radiation particles can travel through the patient body, deposit energy along the way and damage the DNA Structure. There needs to be a balance between killing tumor cells and sparing healthy tissue.

Intensity Modulated Radiation Therapy (IMRT) made it possible to better focus ionizing radiation deposition to tumors by using multi-leaf collimators (MLC). Stereotactic body radiotherapy (SBRT) further differentiated the radiation response between tumors and normal tissues with delivering a much higher dose per fraction with fewer fractions than conventional radiotherapy for tumors that are sensitive to fractionation. Yet, due to the complex procedure in

radiation clinic, including imaging, planning, treatment simulation, and patient setup, the effectiveness of IMRT and SBRT could be hindered. Some hindering factors include organ motion that introduce large uncertainties between dose delivered and dose planned, treatment planning hinder by the quality of each treatment plan being heavily depending on the time and skill of the human planner.

In our research, we retrospectively investigated the inter-fractional and intra-fractional motion for patient data collected from a clinic trial in high-risk prostate cancer SBRT. Our investigation revealed that the relative inter-fractional pelvic to prostate motion has a small impact on the pelvic target dose coverage when the patient was set up with prostate site aligned. This was mainly due to the restrict bladder filling protocol before treatment. As for the intra-fractional prostate motion, on average the dose to the prostate dropped by 6.5% because of the motion, which indicated the importance of effective motion intervention during treatment.

We also investigated the possibility of automating the treatment planning procedure with reinforcement learning technique. With the use of MLC, treatment planning for radiotherapy is a challenging task as it requires to solve an inverse optimization problem which contains millions of possible solutions. Consequently, the current treatment planning process heavily relies on human labor to tune the planning parameters, which can be tedious, not easily reproducible and time consuming. With the rapid development of Artificial Intelligence (AI), there have been increasing efforts to automate the treatment planning process. One attracting AI technique, named reinforcement learning (RL), enabled the possibility to build a virtual treatment planner (VTP) that can mimic the human-like decision making to tune the treatment planning parameters. In this dissertation, I investigated two types of RL technique, named Q Learning and Actor &

Critic techniques for automatic treatment planning. Q Learning is a value-based learning and I applied it to construct a VTP that can operate an in-house dose-volume constrained treatment planning system (TPS) for prostate cancer IMRT treatment planning. The VTP was successfully trained with 10 prostate cancer patient cases and tested with additional 50 cases. One problem for the Q learning is that it is hard to be trained when facing complex treatment planning task as it does not specify an exploration mechanism. To solve the problem, I implemented the Actor & Critic algorithm, which specifies the exploration by the action probabilities of the actor. The preliminary results indicated that the Actor/Critic network is more powerful in generating high-quality treatment plans. Furthermore, to enable the selection of action from a continuous space (to continuously tune a treatment planning parameter), I am in the process of implemented the Proximal Policy Optimization 2 (PPO2) for automatic treatment planning, in the hope that it can provide more powerful tuning of the treatment planning parameters.

Overall, I expect my work in motion management and automatic treatment planning would lead to a technique advancement in radiation clinic for cancer treatment.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

In the 2021 estimates on global mortality data, more than three quarters of all premature deaths were caused by noncommunicable diseases (NCDs) (Bray et al., 2021). According to Pan American Health Organization (PAHO), NCD refers to diseases that are caused by an acute infection that results in long-term health care. Some diseases in this group are cancers, cardiovascular disease, diabetes, and chronic lung illnesses. According to GLOBOCAN estimates 2020, cancer, with 19.3 million new cases worldwide in 2020, is either 1st or 2nd leading cause of death before the age of 70 in 112 of 183 countries, while another 23 countries have cancer ranked 3rd or 4th of the leading cause of death (Figure 1.1) (Sung et al., 2021).

Consequently, to improve the quality of life of the human society, there is a high need to improve the cancer patient survival rate and/or improve the patient quality of life with effective treatment methods.

In modern clinic of cancer treatment, conventional treatment modalities include surgery, chemotherapy, radiotherapy, etc. (Prevention, 2021). Among them, radiotherapy that uses ionizing radiation to damage cancer cells, is a critical pillar. It is reported that more than 1/2 of cancer patients have experienced radiotherapy in their process of illness (Baskar et al., 2012). In our research, we are interested in improving cancer treatment through technique advancement in radiotherapy. I will discuss the principle and the status of radiotherapy, and our work in the field of radiotherapy with prostate cancer as the testbed in the rest of this dissertation content.

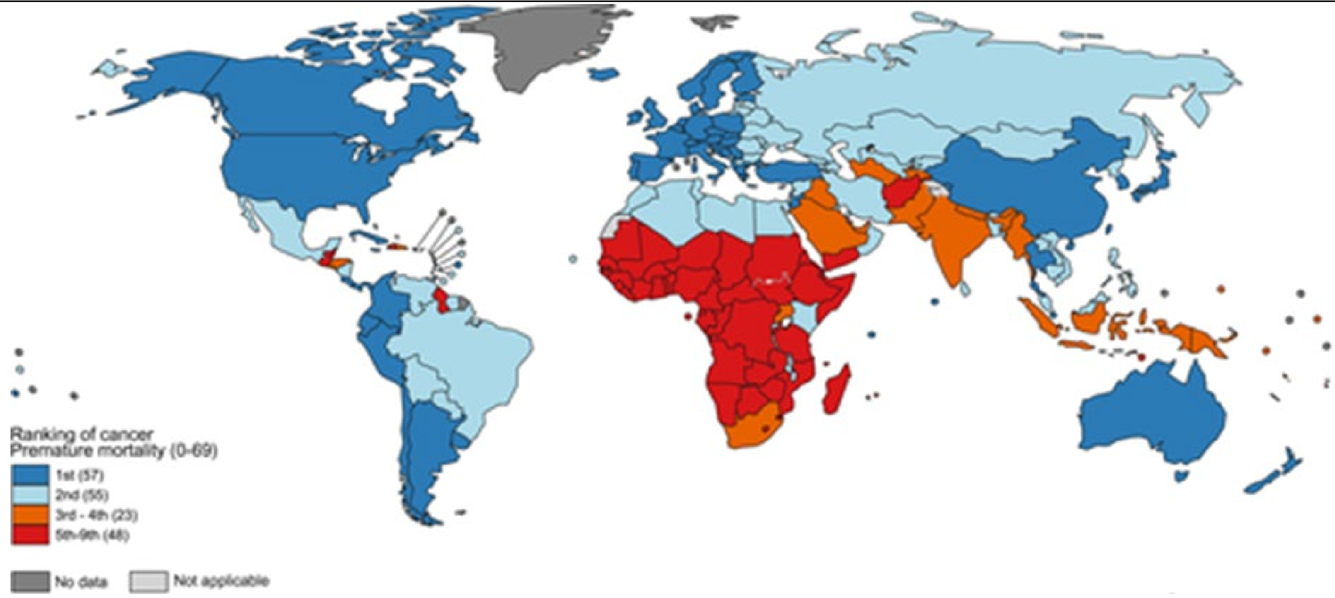

Figure 1.1. National Ranking of Cancer as a Cause of Death at Ages <70 Years in 2019. The numbers of countries represented in each ranking group are included in the legend. Source: World Health Organization. (Sung et al., 2021).

1.1 Physical and biological background of radiotherapy

Radiotherapy is using high energy photons or charged particles to kill cancers (Prevention, 2021). More than 90% of radiotherapy in the US is by x-ray beams, which are delivered by a linear accelerator (LINAC). X-rays were discovered in November 8,1895 by Wilhelm Conrad Roentgen, and they were used in breast cancer treatment in the following January (Society, 2001).

X-rays can interact with human body through electromagnetic interactions (Attix, 2004). There are five diverse types of interactions that can happen depending on the energy of the incident photon, which are: Rayleigh scattering, Photoelectric effect, Compton effect, Pair production and Photodisintegration(Attix, 2004).

The most common of the five is Compton effect which describes the collision between an incident photon and a free electron. A free electron is the one that is not tightly bounded to the atom nucleus. During this interaction both the photon and the electron are scattering. The scatter photon then moves on to interact with more electrons, but with less energy since with each interaction the photon loses energy that is imparted onto the electron. If the energy imparted to the electron is greater than the binding energy of the atom then the electron will be known as an ionizing electron. The energy range for the incident photon that Compton effect dominates is 25 keV - 30 MeV (Jerrold T. Bushberg, 2012).

In radiotherapy the therapeutic photon energy range is 6 - 20 MeV. In this range, except for the dominant Compton scatter effect, there are another two interactions that can happen: photoelectric effect and pair production. Photoelectric interaction happens when the incoming photon interacts with tightly bound electron and the photon energy was fully transferred to the electron. This process dominates photon-tissue interactions when the photon energy is in the range of 10 - 25 keV (Jerrold T. Bushberg, 2012). As for pair production, it describes the creation of a position and an electron through the energy loss of the photon. The threshold energy for pair production is 1.022 MeV. It dominates the interaction process when the photon energy is greater than 30 MeV (Jerrold T. Bushberg, 2012).

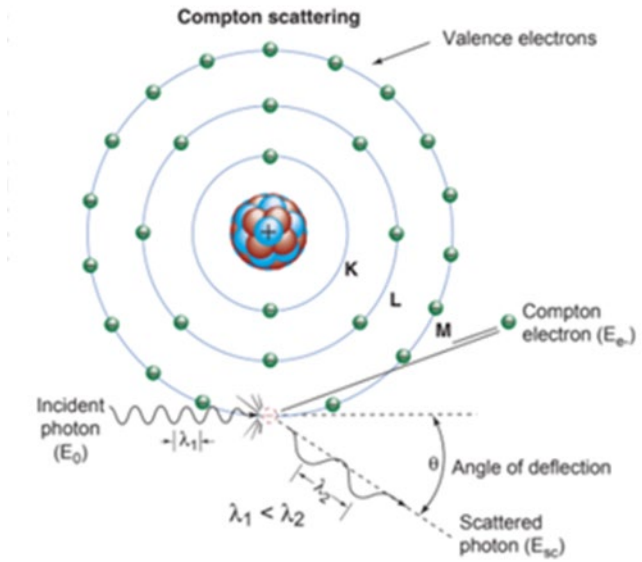

Figure 1.2: Compton Scattering Interaction (The Essential Physics of Medical Imaging, Bushberg et al, 2012)

After the energy is deposited in the body, it can also induce chemical species. When the interaction process happens inside a cellular nucleus, it can damage the cellular DNA (deoxyribonucleic acid) by direct and indirect action that leads to cell death. Direct action refers to the energy deposition hits the DNA molecule directly and disrupts the molecular structure. This leads to cell damage where mutation can be generated or even cell death (Desouky et al., 2015). Indirect action represents that the radiation hits water molecules and generates free radicals that interact with the DNA structure instead of the particles themselves (Eric Hall, 2012).

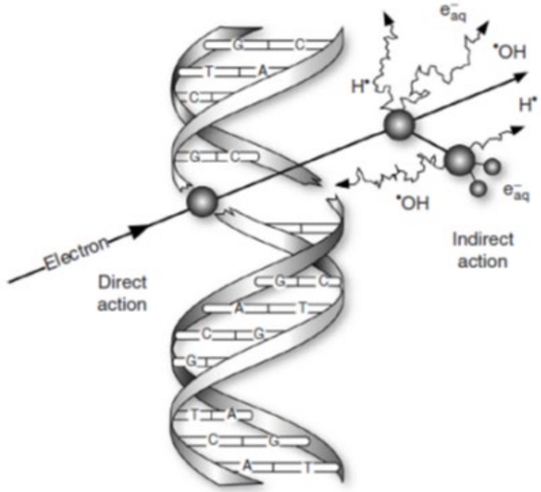

Figure 1.3: Direct and Indirect Action(The Essential Physics of Medical Imaging, Bushberg et al, 2012)

The guiding principle in radiotherapy is to differentiate cell survival rate between healthy tissue and tumor tissue. As x-ray distributes energy deposition events along its entire path, the focus of radiotherapy can't just be on inducing sufficient cell death in the tumor. It also needs to consider protect the nearby organs at risk (OARs). There are 4 R's principles of radiobiology to guide the tradeoff of the two: Repopulation, Redistribution, Repair and Reoxygenation (Eric Hall, 2012). Repopulation allows for the normal tissue surrounding the tumor cells to progressive through the cell cycle and generate new cells that replace those that die during the radiation treatment fraction. Redistribution is when the tumor cells progress through their cell cycle and have more cells in the radio sensitive phase known as mitosis (Eric Hall, 2012). This is part of the cell phase when the cell is actively dividing and has less protection from radiation. Repair is like repopulation that it is there to help the normal tissue to survive the radiation treatment. This time it allows damaged cells to try to repair any damage that happens during the treatment. Finally, the last R is reoxygenation. Tumors are hypoxic by nature. There needs to be a proficient level of oxygen in the cell to make radiation the most effect. The previous fraction dose to the tumor cells generates an increase in the level of oxygen in the surviving tumor cells. That makes the tumor itself more radiosensitive.

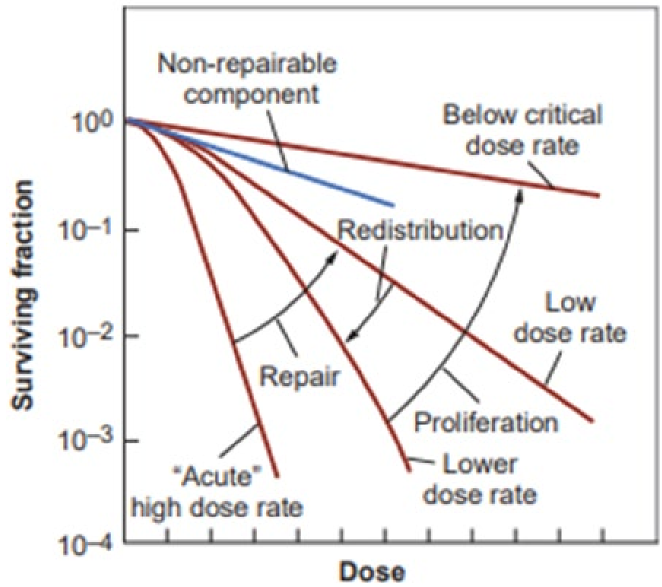

Figure 1.4: Dose rate effect due to the 4 R's (Radiobiology for the Radiologist, Hall et al, 2012)

Upon the four R's, principle, in radiation clinic, the concept of fractionation has been applied to realize the goal of differentiating the cellular response between tumors and normal tissues. Along the practice, there have been three different types of fractionation developed: conventional, hyper, and hypo. Conventional fractionation is the standard one, which delivers 1.8 Gy to 2 Gy for 5 sessions per week (Eric Hall, 2012). Hyper fractionation increases the treatment from once per day to twice per day with less dose in the range of 0.6- 1.2 Gy per fraction for early responding tissue and 0.1 Gy to 0.5 Gy per fraction for late responding tissues (Eric Hall, 2012). Hypo fractionation allows the dose to increase to anything greater than 2.2 Gy per fraction and then entire treatment fractions are much fewer than conventional treatment (Eric Hall, 2012).

The model that represents this principle is the α/β model where: α represents cells killed by a single incident particle while the β represents cell kills through multiple hits, i.e. two particles on different tracks cause damage to the DNA. Cell survival fraction (SF) can mathematically be represented by the following equations:

$$SF = \exp(-\alpha D - \beta D^2)$$

$$SF^n = \exp\big(-n(\alpha D - \beta D^2)\big)$$

$$E = -\ln(SF^n) = n(\alpha D + \beta D^2)$$

$$BED = \frac{E}{\alpha} = nD\left(1 + \frac{D}{\alpha/\beta}\right) \tag{1}$$

The first equation determines what the SF is for one single dose, the following determines what the SF is for the whole treatment by incorporating the number of fractions. The third equation is the Equivalent dose of the fractional scheme. BED stands for Biological Equivalent Dose by which different fractionation schemes can be compared. The nD is the total dose for the whole treatment. The second component $(1 + \frac{D}{\alpha/\beta})$ is the relative effectiveness of the chosen fractionation scheme (Eric Hall, 2012).

The ratio of α/β has impact on what type of fractional scheme that is best for the targeted tumor cells. If it is a high ratio, then it is more sensitive to the radiation and less sensitive to fractionation. Ideal for this would be a hypo-fraction scheme. These types of tissue are known as early responding tissue (ERT) (Desouky et al., 2015; Eric Hall, 2012). If the ratio is low, then the tumor is less sensitive to radiation and would be more sensible to increasing the number of fractions. This would be classified as a late responding tissue (LRT). Prostate cancer is an ERT that why SBRT which is a form of hypo-fractionation is a great treatment technique.

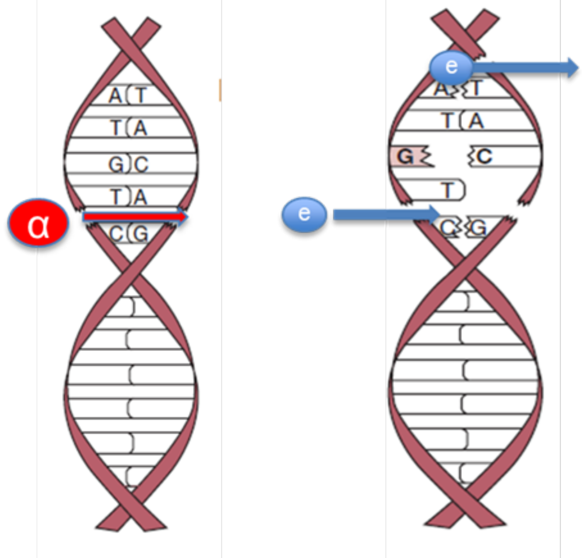

Figure 1.5: Representation of α and β damage on the DNA structure

1.2 Radiotherapy Treatment

Radiotherapy is the use of high-energy radiation to kill cancer cells and shrink tumors (Institute, 2019). Radiation can be one of these three categories: external beam, internal, and systemic radiation (Institute, 2019). The first step of radiotherapy treatment is the diagnosing and staging phase whereas the name suggests that it is to test and determine how far the cancer has progressed. For prostate cancer, one testing can be prostate-specific antigen (PSA). PSA is secreted by normal and cancerous cells. Some of it is also located in the blood. The threshold that determines if a male has a chance of having cancer is 4 ng/ml of blood (Society, 2022). After a male has been diagnosed with prostate cancer then a computerized tomography (CT) will be done to grade the tumor. Figure 1.6 shows what a CT simulator looks like in the radiation clinic. A physician will also grade the prostate cancer using the Gleason score, which is a grading system on how abnormal the prostate cancer is (Society, 2022).

Figure 1.6. Computed Tomography (Treatment Planning in Radiation Oncology, Kahn et al, 2012)

During the CT scanning that information is also sent to the treatment planning system (TPS). Where a dosimetrist can contour all the OARs and volume of interest and set the beams as determined by the chosen radiation therapy techniques. CT is the most used imaging modality for treatment planning. CT provides both the anatomy of the patient and the electron density of different tissues in the body(Damon Sprouts, 2017). The electron density of the surrounding tissue is important to the dose calculation that is done in the TPS. Each position of the couch placement represents a 2D slice of the body and the images obtained during the couch motion produce the 3D image.

1.2.1 Radiotherapy Techniques and Treatment Planning

Take prostate cancer as an example. There are many ways that a Radiation Oncologist

(RadOnc) could decide to plan on how to treat prostate cancer, which is divided into three distinct categories: Brachytherapy, systemic radiation therapy and external beam radiotherapy. The first one brachytherapy is when a radioactive source is implanted directly into the prostate (Society, 2022). The radioactive sources used are Iodine-125, palladium-103 and iridium-192 (Burger, 2003). In systemic radiation therapy, a radioactive substance is connected to a monoclonal antibody that allows for the source to pass through the blood stream and diffuse into the tissue. The last one is external beam therapy, where the radiation is delivered from outside of the body. External beam radiotherapy is most commonly delivered using a machine known as a linear accelerator (Linac). Linac generates electrons for thermionic emission by heating a filament. Once the electrons are generated they are accelerated by the accelerating waveguide. The electrons are bend by 270°, because the patient is located below the electron gun. After the bending the electrons

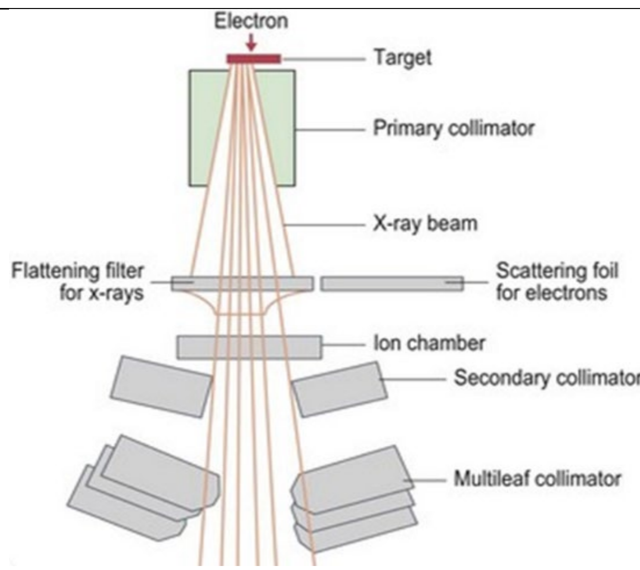

Figure 1.7: Treatment Head schematic

enter the treatment head. Figure 1.7 shows what the head is with the seven components that help shape the beam to better cover the target volumes and spare the surrounding OARs.

Techniques for beam delivery in external beam radiotherapy include three-dimensional conformal radiation therapy (3D-CRT), Intensity-Modulated Radiation Therapy (IMRT) and Image guided radiation therapy (IGRT). 3D-CRT is a procedure that uses a computer to generate a 3D model of the tumor volume (Faiz M. Kahn, 2012). That allows for more accurate delivery of high dose to the tumor volume. IGRT allows you to use CT to image the tumor location before or during treatment to verify the location of the tumor. The below figure 1.8: is a representation of an on-rail CT system that allows for imaging of the tumor before and after treatment.

Figure 1.8: Kahn's example of on CT on rails

IMRT is a sophisticated form of external beam technique that allows for more precise, intense, and effective doses of radiation by knowing what dose you want in the primary target volume (PTV) and working its way back to adjust the beam intensity to match the desire dose (Faiz M. Kahn, 2012) this is known as inverse treatment planning. A major component of the linac' s head that allows for this is known as multi-leaf collimator (MLC). MLC became available in the 1990's. The below figure shows how the MLC can contour better around the tumor than the secondary collimator. This is achieved by pixelating the MLC position. MLC will shape the beam around the tumor and these pixelized beams are called beamlets. Beamlet weights can be changed to change the dose in the PTV and the OARs (Faiz M. Kahn, 2012).

Figure 1.9: Image of MLC leaves shape around tumor

Lastly there is Stereotactic Body Radiation Therapy (SBRT) that is like IMRT. Both can use many angles of the gantry to conform the dose distribution into the tumor site. The biggest difference is that SBRT uses a hypo-fractional scheme while IMRT using a conventional dose approach.

Once a treatment technique is chosen, the next step is to perform treatment planning which involves sending the pre-treatment CT to the TPS and setting the beam configuration to the desire position and contour the structure of interest: for instance OARs (bladder, rectum, penile bulb, left and right femurs, etc.…), gross tumor volume (GTV), clinical tumor volume (CTV), and the PTV. GTV is the palpable disease what can be felt or see in an image. CTV is the margin extended from the GTV that accounts for the clinical uncertainties like: beam alignment, patient positioning, organ motion, and organ deformation (Faiz M. Kahn, 2012).

Treatment planning is the most crucial part of radiotherapy once the pre-treatment imaging is uploaded to the TPS. It can be divided into 7 key steps for proper and safe planning: Positioning, Immobilization, Localization, Field Selection, Dose Distribution, Calculation and Verification. For standard positioning, the patient can be prone (on the front side) or supine (on back). Key point for good patient positioning is that it is reproducible and minimize movement of the patient. Immobilizations are devices that can help with positioning, for instance in prostate cancer a commonly used device is called a Pelvic Form Modified. Localization is where the CT images come into play. It allows for the delineation of the target volumes and OARs in respect to external surface. Field selection greatly depends on the type of cancer being treated. It can range for one field to multiple fields. Beam modification is shaping or blocking the beam intensity. For instance, using MLC to shape the beam around the tumor is one way of beam modification. Dose distribution is the dose desired to distribute inside the tumor target and avoided to the organs at risk, following

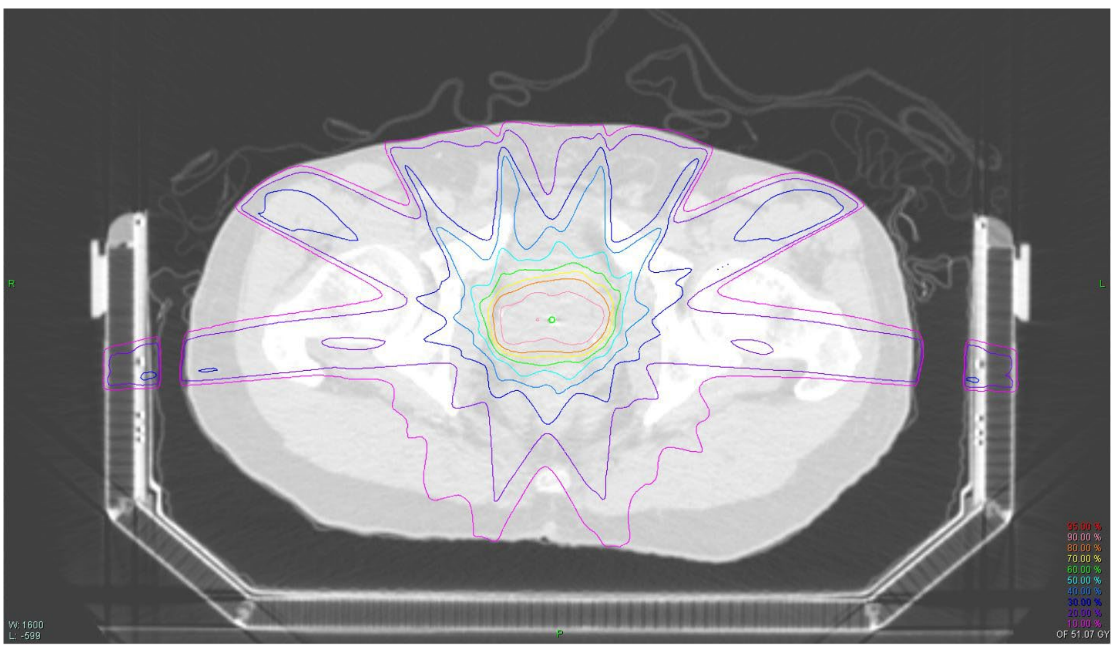

Figure 1.10: Example of Isodose Line in the body for a SBRT prostate patient

the prescription by the physician and the guideline from International Commission on Radiological

Protection (ICRP). Dose distribution can be visualized through Isodose distribution lines. With a desired dose distribution, treatment planning system (TPS) then started the reverse treatment planning based on the treatment planning parameters given by a human planner to obtain the beam intensity and the MLC motion sequence. After that, it would compute the dose inside the patient body with the obtained MLC positions and beam intensity. This calculation would be repeated until a clinic acceptable plan was generated. The final step in the treatment planning is known as verification. This is done before the patient treatment to verify the treatment is deliverable and the planned dose is correct. It typically can be realized through iso-dose verification, film dose verification or software-based dose verification.

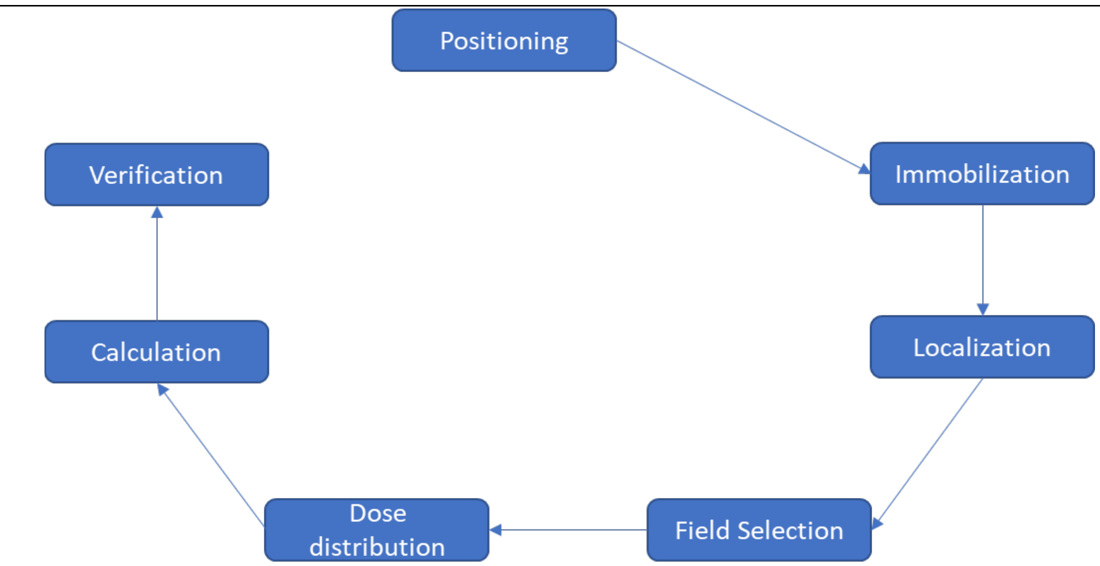

Figure 1.11: Flow Chart of what makes up Treatment Planning

1.3 Target Motion in Radiation Therapy

One of the current challenges in the radiation clinic is motion management due to the uncertainties in treatment planning, patient setup and organ motion. There are 5 main different target motions that are of concern: breathing, cardiac cycle, and different filling levels at separate treatment fractions, soft tissue irregular motion and patient setup (Faiz M. Kahn, 2012). The two that generated the biggest motion is the breathing cycle which affects the tumors in the thoracic cavity but can be less extent to tumors in the abdominal cavity. Clinic will have different gating polices depending on how much the tumor moves during the breathing cycle. The breathing cycle is represented by a sinusoidal like graph.

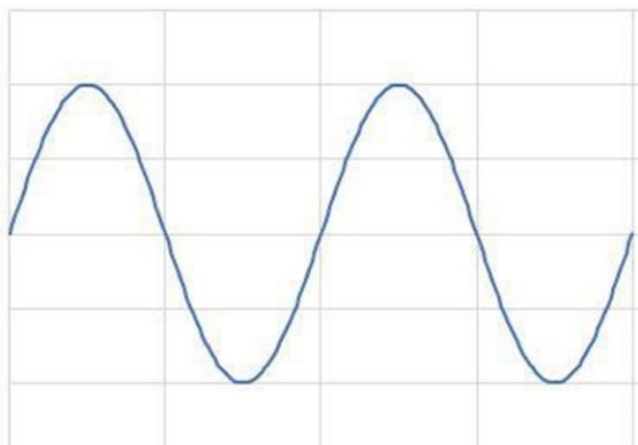

Figure 1.12: Example of a Breathing cycle

Patient setup can also play a huge role in the motion of the tumor. There are two different types of motion: inter-fractional which is the patient setup in between fraction deliver, and intra-fractional which is the motion during the time the fraction is delivered to the patient.

Motion management is a critical aspect in radiotherapy for all techniques, but it is even more critical in the SBRT. As mentioned earlier, SBRT is a hypo-fractional treatment technique that is it deliver higher dose then conventional treatment technique with less fraction. Even through the dose being deliver is higher in the tumor, this change doesn't affect the radiosensitive of the surrounding healthy tissue. If there is any shrinkage or motion of the tumor there will be a higher dose in tissues that can't handle it, so this would increase toxicities.

Within those two types of motion there are two motion that contribute to the total motion: deformation and rigid motion. Deformation motion can be defined as changes in the organ or tumor as reference to the center of gravity this means the changes of shape itself during treatment which can mean shrinking or enlarging during the delivery of treatment. Deformation needs to be considered when there is a large shape change during the motion. For example during the breathing cycle, the lung volume at the different phases can vary a lot, which could impact the accuracy of dose computation and delivery. In the case of prostate cancer, deformation doesn't play a big role since its shape doesn't change as greatly as lung tumors. Instead, we typically consider its rigid motion. Rigid motion of the tumor is the rotational and transitional movement of the tumor during treatment.

In the case of prostate cancer, the rigid motion of the tumor is visible by implanting three fiducials by a needle using ultrasound. This is an internal surrogate for motion of the tumor. The reason being is that cone beam CT is hard to distinguish the difference between tissue since what governs the difference between the tissue is the electron density while all tissues have similar densities.

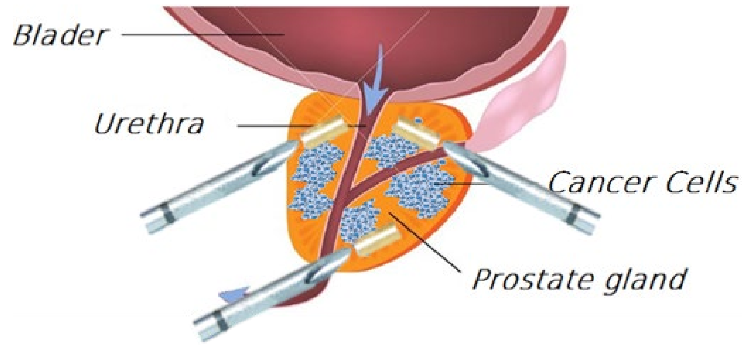

Figure 1.13: Example how Fiducals are implanted

1.4 Clinic automation

Clinic automation has become a key focus of research in medical physics to more accurately and efficiently accomplish tasks in the clinic with the rapid development of Artificial Intelligence.

AI describes when a machine mimics cognitive functions that humans associate with other human minds, such as learning and problem solving. Two examples of AI technique are machine learning (ML) and deep learning (DL). ML is as a series of algorithms that analyze data, learn from it and make informed decisions based on those learned insights. As shown in Figure 1.14. ML typically extracted features from an input, applied complex algorithms to analyze the features and made prediction. The learning process is to minimize the error between the prediction and the ground truth. Different from ML, DL doesn't need to extract features, but uses a multi-layered structure of algorithms called the neural network to directly learn and solve problems from input data.

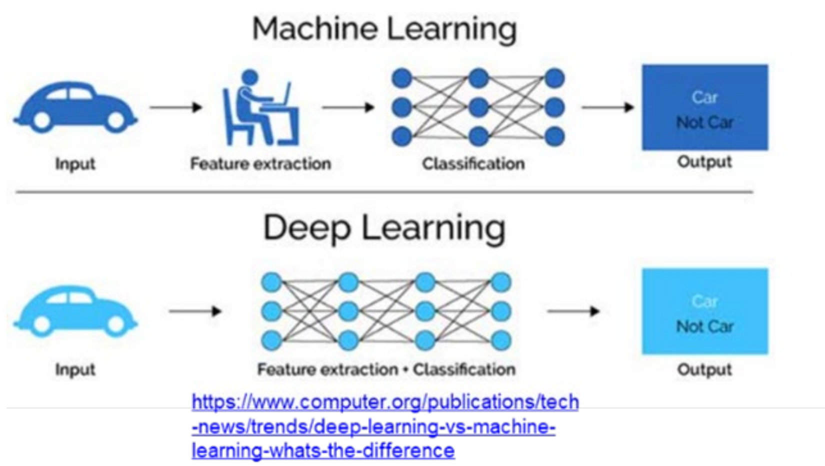

Figure 1.14: Representation of both ML and DL

There are currently both ML and DL in the clinic already that ranges for treatment planning, classification of the disease area, contouring of lesions. AI is not just in therapeutic it is also in diagnostic: it is used to get rid of artifacts and streaks, generate a CT from an MRI images, extracting breathing motion to apply breathing compensation.

In my research, I am interested in automating the treatment planning process in radiation clinic. As described above, treatment planning in modern radiation clinic is to solve an inverse optimization problem, which can be tedious and time consuming. There are multiple efforts to automate this process with AI technique. These include the knowledge-based planning (KBP) method (Chang et al., 2016; Chanyavanich et al., 2011; Fogliata et al., 2014; Hussein et al., 2016; Kubo et al., 2017; Wang et al., 2017), the multicriteria optimization (MCO) method (Chen et al., 2012; Craft et al., 2012; Thieke et al., 2007), the protocol-based automatic iterative optimization (PB-AIO) approach (Wang et al., 2012; Xhaferllari et al., 2013; Yan et al., 2003; Zhang et al., 2011), etc. Key to the KBP method is to utilize historically achieved, high-quality treatment plans to predict an achievable dose in a new patient of a similar population, or to generate a better starting point for a human planner to start with (Chang et al., 2016; Chanyavanich et al., 2011; Fogliata et al., 2014; Hussein et al., 2016; Kubo et al., 2017; Wang et al., 2017). In this method, the quality of the newly generated plan can highly depend on the historical plans or the anatomy similarity between the two sets of patients. Moreover, the predicted dose is not guaranteed to be achievable and further adjustments of the plan configurations from a human planner might be required (Hussein et al., 2018). Central to the MCO method is the concept of the 'pareto optimal solution', which denotes a plan that cannot be further improved for a given objective without degrading one or multiple other objectives (Chen et al., 2012; Craft et al., 2012; Thieke et al., 2007). Yet, a 'pareto optimal solution' may not be clinically desired. It may need to generate many plans before a clinic acceptable plan can be selected or interpolated, which can be computational resource or manual interaction demanding. As for the PB-AIO approach, script-based or fuzzy-logic-based automatic adjustments of the optimization objectives and constraints are established to gradually improve the plan quality to clinic acceptable level (Wang et al., 2012; Xhaferllari et al., 2013; Yan et al., 2003; Zhang et al., 2011). A concern of this approach is that it is not easy to optimize the parameter

Figure 1.15: Projection image matching example: (a) query image, (b) best match (MI =2.8), (c) poor match (MI=1.6), (d) worst match (MI=0.8). (Chanyavanich et al., 2011)

adjustment process, such that the planning efficiency may not be assured (Hussein et al., 2018).

In my study, I have been focusing on a different AI technique, reinforcement learning, for automatic treatment planning. In the following, I will briefly introduce the reinforcement learning and the specific algorithms that were used in my research study.

1.4.1 Reinforcement Learning

There are many different types of learning that pertain to AI. They can be categorized as learning problem, hybrid learning, statistical inference, and learning technique based types. The group we focused on is the learning problem based, which contains supervised, unsupervised, and reinforcement learning (Li, 2017). Supervised learning is trained with labeled data. The models need to determine the mapping function to map the input variable to the output variable. Useful in classification and regression type problems. Unsupervised learning infers patterns from unlabeled input data. The main goal of this type of learning is to find the structure and patterns form the input data. Its ability to find patterns in the data itself. This is useful in clustering and association problems. The last one of the three is reinforcement learning that trains in a trial-and-error method and has proven to have the ability to think like a human. It first became popular in AlphaGo algorithm back in 2017 that was proven to beat world champion level Go players. Go is an ancient Chinese strategy that one needs not just to know their move, but what would be their opponent next move will be. Being so proficient in Go is why we decided to apply this algorithm in the treatment planning.

There are five core structures in reinforcement learning (RL): agent, environment, state, action, and reward(Li, 2017). The agent is the entity that performs actions in the environment and gets rewarded based on the action it takes. Environment is the scenario that an agent must face.

State refers to the current situation by the environment that the agent can act against. Action is the change to the current state by the agent. A reward is the immediate return given to an agent when it performs an action that is negative or positive by the results of the action.

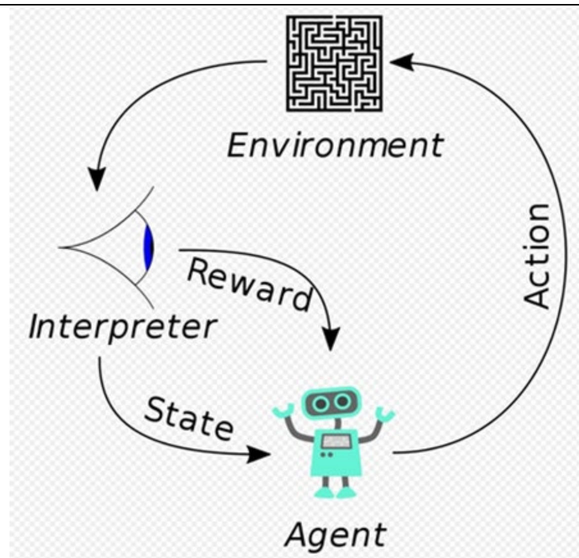

Figure 1.16: Layout for Reinforcement Learning

There are three separate ways that RL can learn: Model, Value, and Policy. Model based create a virtual model of each environment and the agent learns to perform in that specific environment. The model is updated often and is the one method of learning that has the best sample efficient but takes the longest to run. Value-based is known as off-policy where it maximizes a value function. The action that is chosen is the one that will result in the best change in the environment. Policy-based learn the policy function that maps state to action. The subgroup in RL that this paper will focus on are Q learning and Actor/Critic techniques.

1.4.2 Q Learning

Q learning is a model-free RL algorithm which doesn't need to use transition probability distribution to learn. Instead, it is a value-based algorithm. It learns by the way of the state-action value or Q value. To make a decision, the agent needs to follow a policy which is mapping a state to the probabilities of selecting each possible action given that selected state. Q Learning uses off-policy. These types of algorithms don't consider past state-action decisions, but instead the agent constantly performs the action that it believes will yield the highest reward. This process is called exploitation of the environment, but the agent still needs to explore the environment to make more informed decisions. This trade-off process is governed by epsilon-greedy policy. Epsilon is a parameter in the range of 0 and 1. In the training process, if epsilon is higher than a randomly sampled number, a random action is chosen. Otherwise, a greedy action or the highest reward action will be selected.

Markov Decision Process (MDP) is the mathematical framework for modeling the decision made by the agent. It allows for the assumption that the probability for an event to occur only depends on the current state. Yet the transition probability is hard to be known in advance. Hence, to estimate the likely transition from one state to another, it is reasonable to observe multiple transition episodes and take an average. To let the agent effectively learn which state or which transition is good, it is important to give the agent proper feedback after it made a decision, which is known as the reward r.

Bellman equation is one of the central elements of many RL algorithms that decomposes the value function into two parts, the immediate reward plus the discounted future values for finding the optimal Q value. Value function estimates how good it is to perform a given action in a given space by the agent. The below equation illustrations the concept mathematically The LHS Q(s,a) is the new Q value for the current state and action. The RHS Q(s,a) is the current Q value. α is the learning rate which determines how often the Q values are updated, and the range is [0-1]. 0 means it won't update the Qvalue which in turn no learning is happening, while 1 means learning will happen quickly. γ is the discount factor it has the same range as α, but it models the fact that future rewards are worth less than immediate rewards. max_{α} is the maximum reward that can be attainable in next state. This is the reward for the optimal action that led to the next state.

$$Q(s,a) = Q(s,a) + \alpha[r + \gamma max_{\alpha} Q(s', a') - Q(s,a)] \quad (3)$$

1.4.4 Deep Q Learning (DQL)

Neural Network (NN) is a computing system that is model off human neural network hence the name. Policy or value in RL sometimes is complex and nonlinear. Applying a NN to QL allows for powerful representation of value function for QL. Another benefit of having a NN is that the RL algorithm would only need a state to generate as many Q value as

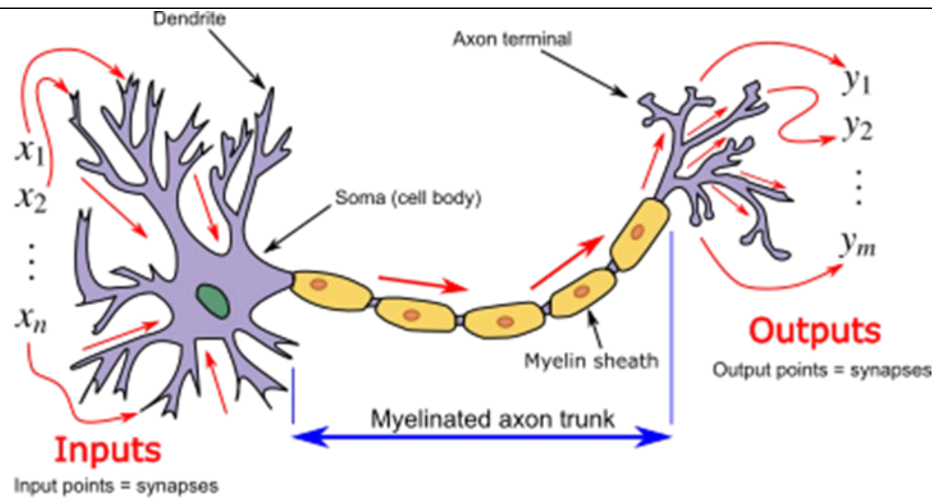

Figure 1.17: Human Neural compared to NN

needed. As you can see by the above figure the generalization of the NN takes the input in the front as synapses and structure of the inside mimic the Myelinated axon trunk and the outpoint or action that the network takes is the Output synapses.

1.4.4 Actor-Critic Network

The second type of algorithm that was implemented in this paper is Actor and Critic network (ACN). The difference between Q learning and ACN is that in Q learning, there is one network to determine both the action and how well that action was (Silver et al., 2016). In the case of can, there are two networks. One is policy-based (Actor), determining what action to pick in a giving situation. The other is Critic, which is a value based and tells how well the chosen action was (Timothy P.Lillicrap, 2016). Then the value function will send the temporal difference (TD) error into the policy function (Silver et al., 2016), so that the actor knows how well it did. TD is also known as the advantage which is the difference between the actual reward and the expected reward.

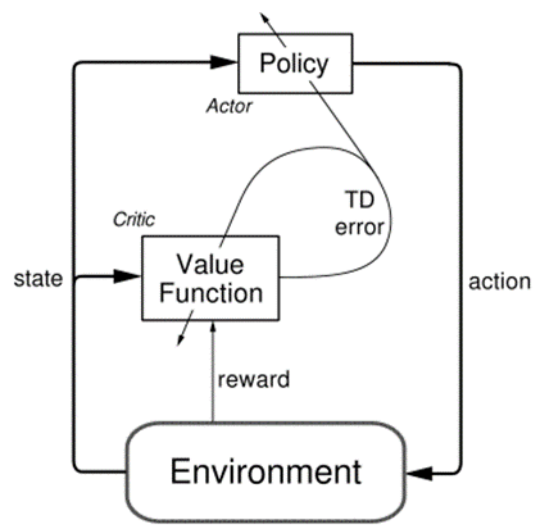

Figure 1.18: Actor and Critic flowchart

1.5 Dissertations compositions

The main goal of this Dissertations was to advance the radiation oncology in two ways: Motion Management in prostate cancer treatment and applying reinforcement learning to automatic treatment planning. Chapter describes the work on motion management, Chapter 3 is about the work using Deep Q Learning for automatic treatment planning. Chapter 4 gives the further development with Actor/Critic network for treatment planning, while Chapter 5 gives the conclusion and future work.

Chapter two

Dosimetric impact of inter-fractional and intra-fractional target motion in high-risk prostate cancer stereotactic body radiation therapy

Yujie Chi[1, **], Damon Sprouts[1], Li Wang[2], Nima Hassan Rezaeian[4], Ming Yang[4], Raquibul Hannan[4], Xun Jia[3, 4]

[1]Department of Physics, University of Texas at Arlington, Arlington, TX 76019

[2]Department of Mathematics, University of Texas at Arlington, Arlington, TX 76019

[3]innovative Technology Of Radiotherapy Computation and Hardware (iTORCH) Laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390

[4]Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75390

2.1 Introduction

Stereotactic body radiation therapy (SBRT) (Lotan et al., 2006; Timmerman et al., 2007) has demonstrated its efficacy of tumor eradication in the treatment of low- to intermediate-risk prostate cancer (PCa) (Azzam et al., 2015; Chen et al., 2013; Cihan, 2018; Park et al., 2018; Ray, 2011). Yet, its safety and feasibility for high-risk PCa group is still at its early-stage investigation (Gonzalez-Motta & Roach III, 2018; Hannan et al., 2021; Kishan & King, 2019; Mesci et al., 2021). Different from the low- and intermediate-risk PCa groups, the high-risk PCa patients are associated with a destined local or systematic recurrence after local therapy (Chang et al., 2014). Consequently, it can be challenging to optimize SBRT fractionation, targeting, doses, etc., for effective tumor control. There have been multiple SBRT strategies under exploration, ranging from SBRT boost (Anwar et al., 2016; H. J. Kim et al., 2017; Lin et al., 2014; Mercado et al., 2016; Miralbell et al., 2010) to SBRT monotherapy (Alayed et al., 2018; Bauman et al., 2015; Bolzicco et al., 2013; Davis et al., 2015; Hannan et al., 2021; Janowski et al., 2014; Kang et al., 2011; Kotecha et al., 2016; Lee et al., 2014; Murthy et al., 2018; Musunuru et al., 2018; Pinitpatcharalert et al., 2019; Ricco et al., 2016), and from prostate (and seminal vesicle) local therapy (Alayed et al., 2018; Bolzicco et al., 2013; Davis et al., 2015; Janowski et al., 2014; Kang et al., 2011; Kotecha et al., 2016; Lee et al., 2014; Mercado et al., 2016; Ricco et al., 2016) to pelvic lymph nodal (PLN) involvement (Alayed et al., 2018; Anwar et al., 2016; Bauman et al., 2015; H. J. Kim et al., 2017; Lin et al., 2014; Miralbell et al., 2010; Murthy et al., 2018; Musunuru et al., 2018; Pinitpatcharalert et al., 2019). Among them, SBRT monotherapy with PLN irradiation is of especial interest, considering its overall short treatment duration and distant target coverage (Alayed et al., 2018; Bauman et al., 2015; Murthy et al., 2018; Musunuru et al., 2018;

Pinitpatcharalert et al., 2019).

In SBRT for high-risk PCa treatment with PLN involvement, the target includes both the prostate and PLN. Proper image guidance technique is central to ensure dosimetric coverage to these targets. A commonly used practice is to set up the patient prior to treatment delivery under cone beam CT (CBCT) to align the prostate target with the planned geometry. However, due to independent motion of the pelvic region relative to the prostate, this patient positioning strategy may lead to large inter-fractional geometry uncertainty and hence degradation of dosimetric coverage of the PLN target (Baker & Behrens, 2016; Huang et al., 2015; Kershaw et al., 2018; Kishan et al., 2015; Tyagi et al., 2019). As for the prostate target, it is known to be subject to intra-fractional motion, which may also affect the dosimetric coverage, even under a precise setup of the prostate target prior to treatment delivery (Dang et al., 2018; Kang et al., 2011; J. H. Kim et al., 2017; Wu et al., 2013; Zhu et al., 2009).

It is hence of critical importance to quantify and characterize the prostate and PLN motion and its impact on the target dosimetric coverage in SBRT monotherapy for high-risk PCa treatment. This could help the development of effective motion management strategies (Franz et al., 2014; Keall et al., 2015; Patrick Kupelian et al., 2007; Liu et al., 2010; Poulsen et al., 2010; Su et al., 2011; Zhu et al., 2009). The consequent target motion mitigation will positively contribute to the SBRT dose escalation studies, which will be correlated to an improved local control (Greco et al., 2020; Line Krhili et al., 2019). It can also help reduce organ toxicities, and hence enhance the patient's health-related quality of life (Line Krhili et al., 2019).

Over the years, extensive studies have been devoted to quantifying motions of different types in PCa SBRT. Yet, there have been limited reports regarding the inter- and intra- fractional pelvic-prostate relative motion in the context of SBRT for high-risk PCa treatment (Kishan et al., 2015; Tyagi et al., 2019). At our institution, we have an ongoing phase I clinical trial (ClinicalTrials.gov: NCT02353819) to study the safety of dose escalation in SBRT for high-risk PCa treatment. The purpose of this study is to analyze the organ motion data collected in this trial to quantify and characterize the inter-fractional pelvic-to-prostate relative motion and intra-fractional prostate motion, as well as the dosimetric impacts.

2.2 Methods and Materials

2.2.1 Patients and treatment planning

At our institution, we have an ongoing Phase I clinical trial on the use of SBRT for high-risk PCa treatment. The high-risk PCa in this trial was identified as prostate specific antigen (PSA) ≥ 20 ng/mL, or grade group ≥ 4, or American Joint Committee on Cancer clinical/radiographic (AJCC) stage ≥ T3. The goal of this trial was to determine the maximum tolerated dose for the prostate and pelvic regions. It is composed of three-level dose escalations. Ten high-risk PCa patients involved in the lowest dose level were included in the current study. Institutional board review approval was in place for all image and dose analyses described in this paper.

For all patients, three fiducial markers were implanted into the prostate site, and a hydrogel spacer (SpaceOAR; Boston Scientific, Marlborough, MA) was injected peri-rectum. A planning CT simulation scan was performed on a Philips CT Big Bore scanner (Philips Medical Systems, Boston, MA) with a 2-mm-thick slice. A diagnostic multi-parametric magnetic resonance image (MRI) and a planning MRI on the prostate site were obtained on a Philips MRI scanner (Philips Medical Systems, Boston, MA). Combining all three sets of images, a radiologist with 15 yearsexperience in prostate imaging annotated the intra-prostatic lesion(s) (Hannan et al., 2021).

The planning target volume (PTV) for the prostatic lesion was then formed by a direct 0~3 mm expansion. After that, the prostate and proximal 1.0 cm of seminal vesicles were delineated. An expansion of 3 mm based on it gave the prostate PTV. As for PLN, its clinical target volume $(CTV)_N$ was obtained by applying the Radiation Therapy Oncology Group contouring atlas. Based on it, the PLN PTV was generated with a 5 mm margin. Organs at risk (OARs), including the femur head, bladder, rectum, sigmoid, bowel bag, urethra, etc., were all contoured. The prescription dose was given in five fractions, with 10 Gy/fraction to the prostatic lesion, 9.5Gy/fraction to the prostate and 4.5 Gy/fraction to the PLN, respectively. For each target, treatment plan achieved minimally 95% of the PTVs covered by the prescription dose.

2.2.2 Image acquisition, patient setup and radiation delivery

Patients were immobilized with the stereotactic body frame. Before the CT scan and each treatment, a strict bladder filling protocol was followed (Hannan et al., 2021). At treatment positioning stage, one set of cone beam CT (kV-CBCT) was acquired for the prostate site using the kilovoltage imaging system equipped on a Varian Truebeam linear accelerator (LINAC). A second set of CBCT was acquired for the pelvic region by shifting the couch inferiorly by 10 cm. The field of review (FOV) on patient lateral and longitudinal directions (for each CBCT) were 45 cm and 16 cm, respectively. The reconstructed CBCT images had 1.17 mm pixel spacing and 2 mm slice thickness. The patient was set up by matching the positions of fiducial markers in the prostate on the CBCT of the prostate site to those on the planning CT.

The treatment plan was delivered in four full arcs with volumetric modulating radiotherapy (VMAT), with each arc lasting about one minute. During the treatment delivery, kV triggered images were acquired every three seconds. After treatment, the planning CT, contours, treatment plan, kV-CBCT images and kV triggered projection images were collected for further analysis.

2.2.3 Inter-fractional motion analysis

For each patient, after obtaining the two sets of pre-treatment CBCT images, we merged them into an extended-view CBCT based on the known 10 cm couch shift. We then calculated the inter-fractional pelvic-to-prostate relative motion in a two-stage registration process. First, we registered the prostate in the extended-view CBCT to that in the planning CT by matching the fiducial markers. Only translational shifts were allowed, to follow the clinical practice. Second, due to limited visibility of the PLN target on the CBCT images, we used the pelvic bones (Netter & Colacino, 1989) as a surrogate to approximate the relative pelvic-to-prostate motion. Specifically, on the CBCT images that were already registered with planning CT for the prostate target, we selected those slices containing the pelvic bones, and then aligned the bones to those on the planning CT. Again, only translational shifts were considered in this bone-alignment process.

With the relative motion between the two targets calculated, we correspondingly shifted the contours of the CTV and PTV of the pelvic target (CTV_N and PTV_N) on the planning CT to obtain those on the CBCT as $CTV_{N, CBCT}$ and $PTV_{N, CBCT}$. We computed the overlap between the $CTV_{N, CBCT}$ volume and the planning PTV_N volume as V_{over}= ($CTV_{N, CBCT}$ inside PTV_N)/ $CTV_{N, CBCT}$ to check the tolerance of the PTV margin under the inter-fractional motion. We further estimated the impact on the dose coverage of the PLN target. As such, we assumed the dose distribution in the 3D space was unchanged from the treatment plan and obtained the dose distribution to the CTV_N at its new position $CTV_{N, CBCT}$. This was repeated for each fraction. We calculated the delivered dose to the CTV_N by accumulating doses over all fractions. The delivered dose was compared to the treatment planning dose using metrics of $D_{95\%}$ (the minimum dose delivered to 95% of the PTV volume) and $V_{100\%}$ (the volume receiving the prescription dose).

2.2.4 Intra-fractional motion analysis

To estimate the intra-fractional prostate motion, we segmented the fiducial markers using an auto-segmentation algorithm developed by Mao *et al.*(Mao et al., 2008) from the in-treatment kV triggered projections. With the 2D marker positions obtained, we retrospectively reconstructed the 3D intra-fractional prostate motion via the Projection Marker Matching Method (PM^3) (Chi et al., 2017). The 3D motion reconstruction error was also estimated while the detail of the estimation method and estimation results were given in Appendix A.

After obtaining the intra-fractional prostate motion, we estimated its impact on the dose coverage of the prostate target (CTV of prostate target, i.e. CTV_P) and intra-prostatic lesion target (PTV of lesion target, i.e. PTV_{pl}). Again, we assumed that the prostate motion does not affect 3D dose distribution in the space. At every moment of a triggered image, we obtained the dose to the targets by fetching the dose in the original plan at the new CTV position, form the dose distribution per projection. Summing over all the moments of triggered images for one fraction yielded the delivered dose distribution per fraction. We also summed over the dose for one patient yielding the delivered dose distribution per patient. For each dose distribution (dose-per-projection, dose-per-fraction and dose-per-patient), we compared the dose distributions with treatment planning dose using metrics of $D_{95\%}$ and $V_{100\%}$.

2.2.5 Statistical analysis

Spearman's rank correlation was used to test the correlation of the inter-fractional motion amplitude with the PLN target dose coverage reduction, and that between the intra-fractional motions and the change of prostate target dose coverage.

2.3 Results and Discussion

2.3.1 Inter-fractional pelvic-prostate motion

Figure 2.1 depicts the geometric differences between pelvic bones in the CBCT (magenta) and those in the planning CT (green) for a representative case. A mismatch of bones was clearly observed in Figure 2.1(a) for point cloud visualization and Figure 2.1(b) for slice-by-slice comparison. Here, the CBCT image were already registered with CT at the prostate target, so the difference presents the relative motion between the pelvic region and the prostate. Applying a translational shift, the bone structures were aligned well with each other as shown in Figures 2.1(c)-(d).

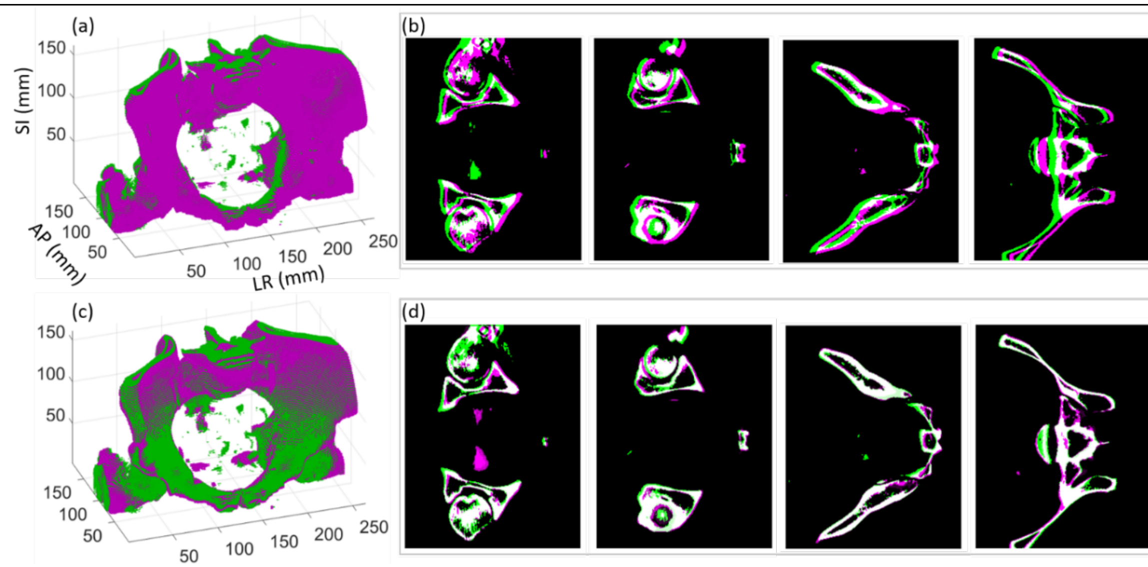

Figure 2.1: The visualization depicting the differences for pelvic bones between prostate-registered CBCT and planning CT with a blending of magenta for CBCT and green for CT. (a) Point clouds before registration, (b) the corresponding slice-by-slice comparison along SI direction, (c) point clouds after registration using translational transformation, and (d) the

corresponding slice-by-slice comparison along SI direction.

We obtained the relative pelvic-prostate motion for 45 treatment fractions that had the CBCT images available. Among them, 44 fractions had satisfying alignment results between pelvic bones in CBCT images and those in CT images. Yet, there was one fraction associated with obvious mismatch after alignment. We excluded it from the subsequent statistical analysis, yet we discussed this special fraction in the discussion section. We analyzed the motion magnitudes along LR, AP and SI directions and the 3D motion magnitude, and summarized the mean, standard errors of the mean (SEMs) and ranges of the motion magnitudes in Table 2.1. The average motion magnitudes over fractions along LR, AP and SI directions were 1.8 mm, 3.3 mm and 0.8 mm, respectively, while the maximal motion magnitudes along the three directions were 4.2 mm, 8.1 mm and 2.2 mm. The average and maximal 3D motion magnitudes were 4.1 mm and 9.2 mm.

Based on the motion matrix, we obtained $CTV_{N,\,CBCT}$ and computed the volume overlap V_{over} between $CTV_{N,\,CBCT}$ and planning PTV_N. Averaging over fractions, 98.6% of $CTV_{N,\,CBCT}$ volume was within the volume of the planning PTV_N. The corresponding SEM was 0.3% and the range was 90.0% - 100.0%. Four out of 44 fractions had$V_{\text{over}} < 95\%$, which were 94.7%, 93.6%, 93% and 90%, respectively.

The inter-fractional motion was found to have a relatively small impact on the CTV_N dose coverage. Averaging over fractions, $V_{100\%}$ and $D_{95\%}$ dropped by 0.9% and 0.3% respectively. Among the 44 treatment fractions, there was only one fraction with $V_{100\%}$ dropping more than 5% (which was 9.6%). Spearman correlation test showed that the drop-off of $V_{100\%}$ significantly negatively correlated to the motion magnitudes along LR and AP directions ($r = -0.56$ and $-$

Table 2.1. Statistical summary of inter-fractional pelvic-prostate motion and the corresponding impacts on dose distributions of pelvic target in the form of $D_{95\%}$ and $V_{100\%}$.

	Motion (mm)				Dose metrics (%) of PLN target	
	LR	AP	SI	3D	$D_{95\%}$(motion)/$D_{95\%}$(plan)	$V_{100\%}$(motion)/$V_{100\%}$(plan)
Mean	1.8	3.3	0.8	4.1	99.7	99.1
SEM	0.2	0.3	0.1	0.3	0.1	0.2
range	0.0-4.2	0.2-8.1	0.0-2.2	1.3-9.2	95.7-100.5	90.4-100.2

Abbreviations: LR = left-right, AP = anterior-posterior, SI = superior-inferior, SEM = standard error of the mean, PLN = pelvic lymph node.

0.55 respectively and $p < 0.001$), which did not reach statistical significance in the SI direction ($r = -0.19$ and $p = 0.01$).

2.3.2 Intra-fractional prostate motion

Three-thousand nine-hundred eighty-one kV triggered projections were available for these 10 patients over 44 fractions. In Figure 2.2, we showed the intra-fractional prostate motion for a patient treatment fraction, which lasted for about 4 minutes. The motion was small for most of the treatment moments, yet there was persisting large motion (≥ 3 mm) along AP and SI directions for the last ~15 projection moments (~45 seconds).

The absolute intra-fractional prostate motion for all patients was summarized in Table 2.2. The average magnitude of this motion was 2.5 mm. The average magnitudes over projection moments in the LR, AP and SI directions were 1.0 mm, 1.4 mm and 1.4 mm, respectively. Nine hundred forty-two projections (23.7%) from 35 fractions had a relatively large motion range (3D motion magnitude > 3 mm). When averaging the motion over fractions, the average 3D motion magnitude and average motion magnitudes along LR, AP and SI directions were 2.6 mm, 1.0 mm, 1.4 mm and 1.4 mm respectively. Eight out of 44 fractions from 6 patients had an average 3D motion magnitude larger than 3 mm. When averaging over patients, 2 out of 10 patients had an average 3D motion magnitude larger than 3 mm.

Table 2.2. Statistical summary of intra-fractional prostate motion magnitudes and the corresponding impacts on dose distributions of prostate target in the form of $D_{95\%}$ and $V_{100\%}$.

		Motion (mm)				Dose metrics (%) of prostate (lesion) target	
		LR	AP	SI	3D	$D_{95\%}$(motion)/$D_{95\%}$(plan)	$V_{100\%}$(motion)/$V_{100\%}$(plan)
Overall (3981†)	Mean	1.0	1.4	1.4	2.5	98.3 (99.7)	96.8 (96.9)
	STD	1.4	1.1	1.2	1.7	5.6 (0.9)	6.9 (6.1)
	Range	0.0-15.5	0.0-7.0	0.0-7.5	0.1-15.5	52.5-100.1(89.7-101.5)	52.9-100.5(65.2-104.7)

Abbreviations: LR = left-right, AP = anterior-posterior, SI = superior-inferior, STD = standard deviation.

†: the number of trigger projection images.

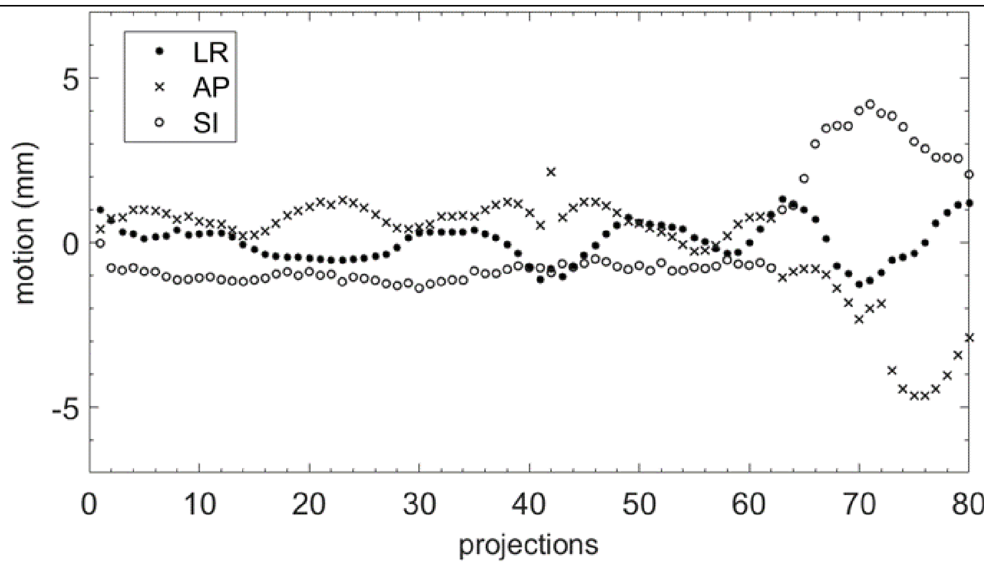

Figure 2.2: The 3D intra-fractional prostate motion along LR, AP and SI directions obtained via the PM^3 method for a patient treatment fraction (4 arcs with 20 projections per arc).

For the entire 3981 projection moments from the 10 patients, the intra-fractional motion introduced a drop-off of 1.7% and 3.2% for $D_{95\%}$ and $V_{100\%}$ of the prostate target on average. It introduced a drop-off of 0.3% and 3.1% for $D_{95\%}$ and $V_{100\%}$ of the intra-prostatic lesion target on average. The largest drop-off of $D_{95\%}$ and $V_{100\%}$ for the prostate target was 47.5% and 47.1%, which were 10.3% and 34.8% for the lesion target. The corresponding transient 3D motion magnitude was 15.1 mm. As for the accumulated dose per fraction, 5 and 7 out of 44 fractions were correlated with a drop-off of $D_{95\%}$ and $V_{100\%}$ larger than 5% for the prostate target, which were 0 and 10 fractions for the intra-prostatic lesion target. The largest drop-off for the two metrics were 29.3% and 28.8% for the prostate target and 3.0% and 27.3% for the lesion target. The average 3D motion magnitude for the corresponding fraction was 10.1 mm. As for the accumulated dose per patient, 2 out of 10 patients had that dose drop-off larger than 5% for the prostate target, which were 11.1% for $D_{95\%}$ and 15.9% for $V_{100\%}$ for one patient, and 8.5% for $D_{95\%}$ and 10.2% for $V_{100\%}$ for the other patient. The corresponding average 3D motion for the two patients were 5.2 mm and 4.2 mm respectively. Only the former patient also had a drop-off of $V_{100\%}$ for the lesion target larger than 5% (8.7% specifically). The less sensitivity of $D_{95\%}$ to the motion magnitude for the lesion target could be partially explained by the relatively small volume of the intra-prostatic lesion target compared to the prostate target and the small dose difference between the two targets (50 Gy/fraction and 47.5 Gy/fraction, respectively.

Spearman correlation test showed that the drop-off of $D_{95\%}$ of the prostate target moderately negatively correlated to the overall 3D motion magnitude and motion magnitude along AP direction ($r = -0.39 \text{ and} - 0.44$ respectively, $p < 0.001$), which did not reach statistical significance in the LR and SI directions ($r = -0.12 \text{ and} - 0.04$, $p \leq 0.01$). The drop-off of

$V_{100\%}$ showed a similar correlation to the intra-fractional motion magnitude.

2.4 Discussion

In one of the patients at one treatment fraction, it was not possible to align the pelvic bones on the CBCT image to that on the planning CT only using the translational rigid registration with adequate accuracy (mismatch > 3 mm for most bone structures). A significant tilt was observed between the two bone structures. We then applied the rotation-allowed point-cloud registration of iterative closest point (ICP) algorithm (Besl & McKay, 1992; Chen & Medioni, 1992), which was able to align the bone structures well. After the re-registration, we found that fraction was associated with $V_{100\%}$ dropping by 2.8% and $D_{95\%}$ dropping by 1.3%, which was dropping by 11.1% and 5.3% under the defect translational rigid registration. We concluded that for most inter-fractional PLN-prostate motion, a translational rigid registration could describe the motion well under the surrogate of the pelvic bone structure. Yet, under some circumstances, for example, inappropriate setup where the patient has non-negligible bone tilt, rotational motion should be considered.

It is worth mentioning that uncertainties may exist in the estimation of the impact of inter-fractional and intra-fractional motions on target dose coverage. In the inter-fractional PLN dose estimation, we directly mapped the planning dose on CT to CBCT based on the motion matrix. A more accurate way should consider the relative geometry change between the patient and the treatment beam. However, we argue that our estimation should still be effective to reflect the motion impact. For small motion range, the impact of the relative patient-beam geometry change should be very small. As for large motion, no matter considering the relative geometry change or not, the dose coverage for the PLN target should have a significant change. In the intra-fractional prostate dose estimation, we used the planning dose for a whole fraction to estimate the dose delivered at each projection moment after considering motion. It ignored the influence of interactions between the multi leaf collimator (MLC) and the prostate motion at that moment, which may affect the dose estimation accuracy. Yet, we expect statistically our dose estimation is still reasonable and we will investigate the impact of the interaction of MLC and prostate motion in our future work.

Besides our study, the inter-fractional PLN motion was also studied by Kishan et. al. (Kishan et al., 2015) and Tyagi et. al. (Tyagi et al., 2019) for high-risk PCa with IMRT/SBRT treatment There are two significant differences between our findings and that in the two studies. First, the average translational motion magnitude along SI direction in our study is much smaller than those from the other two studies, which is $\leq$ 1 mm from our study comparing to 3.4 mm in (Kishan et al., 2015) and -2.8 mm in (Tyagi et al., 2019). Second, the inter-fractional motion induced dose coverage drop-off expressed in $V_{100\%}$ or $D_{95\%}$ from our study is also much smaller. On average, $V_{100\%}$ dropped by ~1% in our study, which is 7.4% in Kishan et. al.'s study (Kishan et al., 2015). $D_{95\%}$ dropped by <0.5% in our study, while it is 4.4% in Tyagi et. al.'s study (Tyagi et al., 2019). The difference could be partially explained by the strict bladder filling protocol and the relatively large PTV margin (5-7 mm compared to 4-5 mm) applied in our study.

There were multiple studies for the intra-fractional prostate motion tracking for PCa IMRT treatment (Keall et al., 2016; P. Kupelian et al., 2007; Su et al., 2011), while our study is the first report regarding the motion in PCa SBRT treatment. There are two consistent findings for our study with previous reports. In our study, the percentage of time for prostate motion $\geq$ 3 and 5 mm was 24% and 8%, which is similar to that reported in (Su et al., 2011) (20% and 6%, respectively) with Calypso tracking system and that in (Keall et al., 2016) (18% for motion $\geq$ 3 mm) with KIM tracking method. We observed similar motion patterns as that reported in study (P. Kupelian et al., 2007) for the motion tracked with Calypso system and patients from five medical centers that is the motion was unpredictable and varied from persistent drift to transient rapid motion. In our study, motion $\geq$ 3 and $\geq$ 5 mm for cumulative durations of at least 30 s were during 41% and 16% of all treatment fractions, quite comparable to that observed in study of (P. Kupelian et al., 2007) (41% and 15% respectively). One difference between our study and published literatures was that a significant positive correlation was found for prostate motion along AP and SI directions in (Su et al., 2011), whereas the correlation was moderate in our study (spearman □ =0.22).

We also studied the motion impact on the dose coverage of prostate target. For transient motion magnitude as large as 15 mm, $D_{95\%}$ and $V_{100\%}$ could drop significantly by 47.5% and 47.1% respectively. As for the motion impact per fraction, a relatively large 3D motion of 10.1 mm was observed for one fraction, which resulted the dose drop-off by 29.3% and 28.8% for the prostate target. This finding was consistent with that reported in Azcona *et. al.*'s study (Azcona et al., 2014), where $V_{100\%}$ of the prostate target dropping below 60% for one trajectory was observed (dropped by more than 40%). In addition, for the first time, we reported the intra-fraction prostate motion impact on the dose coverage of the intra-prostatic lesion target. The drop-off of $V_{100\%}$ for the lesion target was similar to that for the prostate target. The drop-off of $D_{95\%}$ for the lesion target was not that significant as in this specific study, the dose difference between the lesion and prostate targets was within 5%. It can be expected that the larger the dose difference between the two targets, the more sensitive of $D_{95\%}$ reduction was to the prostate motion magnitude. Overall, all studies consistently indicated the importance of treatment intervention (e.g. monitoring, gating, adaptive re-planning, etc.) to reduce large intra-fractional prostate motion.

2.5 Conclusion

We have demonstrated that with applying a relatively large pelvic node PTV margin and following a strict bladder filling protocol, the inter-fractional relative pelvic-prostate motion had a limited impact on the dose coverage of pelvic nodes, even under SBRT treatment for high-risk PCa. We have also showed that the intra-fractional prostate motion was small in most of the treatment duration period. However, both persistent large drift and transient large prostate movement (3D motion magnitude $\geq$ 3 mm) was observed in a portion of patient treatment fractions. The large motion magnitude was significantly correlated to the drop-off of dose coverage on the prostate target, indicating the importance of treatment intervention on intra-fractional prostate motion.

Chapter 3

The Development of a Deep Reinforcement Learning Network for Dose-Volume-Constrained Treatment Planning in Prostate Cancer Intensity Modulated Radiotherapy

Damon Sprouts[1], Yin Gao[2], Chao Wang[2], Xun Jia[2], Chenyang Shen[2,a], Yujie Chi[1,a]

[1] Department of Physics, the University of Texas at Arlington, Arlington, TX 76019,USA

[2] innovative Technology Of Radiotherapy Computation and Hardware (iTORCH) laboratory, Department of Radiation Oncology, University of Texas Southwestern Medical Center, Dallas, TX 75287, USA

ABSTRACT

Although commercial treatment planning systems (TPSs) can automatically solve the optimization problem for treatment planning, human planners need to define and adjust the planning objectives/constraints to obtain clinically acceptable plans. Such a process is labor-intensive and time-consuming. In this work, we show an end-to-end study to train a deep reinforcement learning (DRL) based virtual treatment planner (VTP) that can behave like a human to operate a dose-volume constrained treatment plan optimization engine following the parameters used in Eclipse TPS for high-quality treatment planning. We considered the prostate cancer IMRT treatment plan as the testbed. The VTP took the dose-volume histogram (DVH) of a plan as input and predicted the optimal strategy for constraint adjustment to improve the plan quality. The training of VTP followed the state-of-the-art Q-learning framework. Experience replay was implemented with epsilon-greedy search to explore the impacts of taking different actions on a large number of automatically generated plans, from which an optimal policy can be learned. Since a major computational cost in training was to solve the plan optimization problem repeatedly, we implemented a graphical processing unit (GPU)-based technique to improve the efficiency by 2-fold. Upon the completion of training, the established VTP was deployed to plan for an independent set of 50 testing patient cases. Connecting the established VTP with the Eclipse workstation via the application programming interface, we tested the performance the VTP in operating Eclipse TPS for automatic treatment planning with another two independent patient cases. Like a human planner, VTP kept adjusting the planning objectives/constraints to improve plan quality until the plan was acceptable or the maximum number of adjustment steps was reached under both scenarios. The generated plans were evaluated using the ProKnow scoring system. The

mean plan score (± standard deviation) of the 50 testing cases were improved from 6.18 ± 1.75 to 8.14 ± 1.27 by the VTP, with 9 being the maximal score. As for the two cases under Eclipse dose optimization, the plan scores were improved from 8 to 8.4 and 8.7 respectively by the VTP. These results indicated that the proposed DRL-based VTP was able to operate the in-house dose-volume constrained TPS and Eclipse TPS to automatically generate high-quality treatment plans for prostate cancer IMRT.

3.1 INTRODUCTION

Intensity modulated radiation therapy (IMRT) has been widely used in modern clinic for cancer treatment (Bortfeld, 2006). IMRT holds the potency to deliver a high therapeutic dose to the tumor volume while sparing the nearby organs at risk (OARs). Consequently, it provides the possibility to improve the local tumor control as well as to retain the patients' quality of life (Cho, 2018).

One critical component affecting the effectiveness of IMRT is the quality of IMRT treatment plan, which defines the beam characteristics needed to achieve the desired radiation dose distribution. This is often formulated as an optimization problem in a multi-objective function and automatically solved by the modern treatment planning systems (TPSs) (Intensity Modulated Radiation Therapy Collaborative Working Group, 2001). However, depending on the specific objectives and constraints applied to define the plan optimization problem, the generated plan may not be satisfying. To obtain a clinically acceptable plan, it typically requires a human planner to repetitively observe intermediate optimization results and adjust the objectives/constraints to improve the plan quality. Such a planning process can be labor intensive and time consuming.

Hence, the final plan quality could highly depend on the planning experience and the planning time attained by the planner (Atun et al., 2015). A fully-automatic treatment planning system that can automatically adjust the plan objectives/constraints for high-quality IMRT treatment planning is then critical to advance the radiation clinic using IMRT for cancer treatment.

To date, there are multiple techniques developed to automate the treatment planning process (Hussein et al., 2018). These include the knowledge-based planning (KBP) method (Chang et al., 2016; Chanyavanich et al., 2011; Fogliata et al., 2014; Hussein et al., 2016; Kubo et al., 2017; Wang et al., 2017), the multicriteria optimization (MCO) method (Chen et al., 2012; Craft et al., 2012; Thieke et al., 2007), the protocol-based automatic iterative optimization (PB-AIO) approach (Wang et al., 2012; Xhaferllari et al., 2013; Yan et al., 2003; Zhang et al., 2011), etc. Key to the KBP method is to utilize historically achieved, high-quality treatment plans to predict an achievable dose in a new patient of a similar population, or to generate a better starting point for a human planner to start with (Chang et al., 2016; Chanyavanich et al., 2011; Fogliata et al., 2014; Hussein et al., 2016; Kubo et al., 2017; Wang et al., 2017). In this method, the quality of the newly generated plan can highly depend on the historical plans or the anatomy similarity between the two sets of patients. Moreover, the predicted dose is not guaranteed to be achievable and further adjustments of the plan configurations from a human planner might be required (Hussein et al., 2018). Central to the MCO method is the concept of the 'pareto optimal solution', which denotes a plan that cannot be further improved for a given objective without degrading one or multiple other objectives (Chen et al., 2012; Craft et al., 2012; Thieke et al., 2007). Yet, a 'pareto optimal solution' may not be clinically desired. It may need to generate many plans before a clinic acceptable plan can be selected or interpolated, which can be computational resource or manual interaction demanding. As for the PB-AIO approach, script-based or fuzzy-logic-based automatic adjustments of the optimization objectives and constraints are established to gradually improve the plan quality to clinic acceptable level (Wang et al., 2012; Xhaferllari et al., 2013; Yan et al., 2003; Zhang et al., 2011). A concern of this approach is that it is not easy to optimize the parameter adjustment process, such that the planning efficiency may not be assured (Hussein et al., 2018).

Most recently, along with the rapid development of deep learning (Krizhevsky et al., 2012) and reinforcement learning (Sutton & Barto, 2018), a new architecture named "intelligent automatic treatment planning (IATP) framework" has been put forward (Shen, Chen, Gonzalez, et al., 2021; Shen, Chen, & Jia, 2021; Shen et al., 2019; Shen et al., 2020). In IATP framework, an intelligent virtual treatment planner (VTP) is constructed to operate the in-house TPS like a human planner to generate high-quality treatment plans. Specifically, Shen *et. al.* introduced the deep neural network-based reinforcement learning (Mnih et al., 2015) to automate the weighting parameter tuning in inverse treatment planning with a proof-of-principle study in high dose-rate brachytherapy for cervical cancer (Shen et al., 2019), and then extended the principle to external radiotherapy with developing a virtual treatment planner (VTP) for prostate cancer IMRT planning (Shen et al., 2020). The VTP-based treatment planning was proved to be able to generate high quality treatment plans with a relatively high efficiency.

Yet, in these studies, the in-house developed TPS was relatively simple in the aspects of objective functions and adjusted parameters, compared to that employed in the commercial TPS. It brought concerns that the concept of VTP-based treatment planning may not work for complex TPS, for example, the commercial TPS, where dose-volume constraints were typically applied. Recently, a reinforcement-learning based Eclipse treatment planning was tested to be effective to generate treatment plans for pancreas stereotactic body radiation therapy, yet much more efforts are still needed to investigate the effectiveness of IATP-based automatic treatment planning for broad clinical applications. Hence, it is desired to implement an in-house developed complex TPS, such as a dose-volume constrained TPS, to comprehensively investigate the effectiveness of VTP-based treatment planning with a goal that once the VTP architecture was tested to be effective, it could be easily adapted to operate a commercial TPS.

In this work, we implemented a dose-volume constrained TPS following the parameters used in Eclipse TPS for prostate cancer IMRT. We especially designed an end-to-end VTP neural network to operate the developed TPS. We trained and tested the VTP on two different sets of patient cases. We then connected the established VTP with the Eclipse workstation via the application programming interface (API) and tested the performance of the VTP-based Eclipse automatic treatment planning with another two independent patient cases. We found that the established IATP framework could operate the in-house dose-volume constrained TPS and Eclipse TPS for successful treatment planning in prostate cancer IMRT. We reported the method, results and discussions in the following sections.

3.2 METHODS AND MATERIALS

3.2.1 The overall architecture of the IATP framework

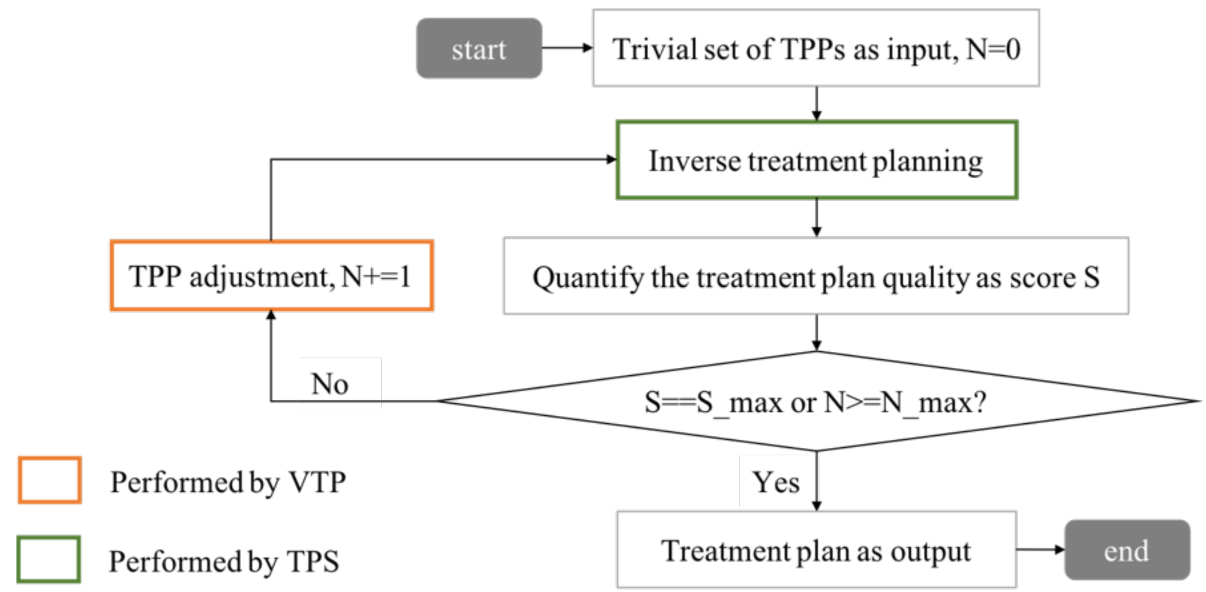

Figure 3.1: Flowchart of the intelligent automatic treatment planning (IATP) framework. VTP: virtual treatment planner. TPS: treatment planning system.

We illustrate the overall architecture of the proposed IATP framework in Figure 3.1. As is shown, the TPS started the inverse treatment planning with a trivial set of treatment planning parameters (TPPs). The quantification system then quantified the quality of the produced treatment plan as a numerical score S. If S was lower than the predefined maximum plan score, the VTP would observe the DVH of the current treatment plan and decided how to adjust the TPPs. After that, the VTP would perform the inverse treatment planning again under the updated TPPs. The process was repeated until a satisfying treatment plan was obtained or the VTP reached its maximum iteration for the TPP tuning. Compared to the conventional human-planner-based treatment planning, the IATP framework features the automatic decision-making process for TPP adjustment with the VTP system.

To establish an IATP framework suitable to operate a commercial TPS, we need an especial design of the in-house TPS system and the VTP network. The details of the IATP framework were discussed in the following subsections 2.2-2.6.

3.2.2 The inverse treatment planning optimization algorithm

We developed an in-house dose-volume constrained TPS following the detailed documentation of the plan optimization method for Eclipse TPS (Varian, 2014). We especially considered the following features for IMRT treatment planning in Eclipse TPS: 1) upper and lower constraints (each constraint contains volume, thresholding dose and priority) to optimize the dose distribution inside the planning target volume (PTV), 2) upper constraints for the OARs, 3) dose-volume-histogram (DVH)-based optimization, and 4) the dose deposition coefficient matrix. With considering points 1)-3), we formed the objective function as follows:

$$\min\left[\frac{1}{2}\left\|Mx-d_p\right\|_-^2+\frac{\lambda}{2}\left\|\left(Mx-td_p\right)_{V_{PTV}}\right\|_+^2+\sum_i\frac{\lambda_i}{2}\left\|(M_ix-t_id_p)_{V_i}\right\|_+^2\right], \qquad (2)$$

$$\text{s.t.}\, x\geq 0,\ D_{95\%}(Mx)=d_p.$$

Eq. (1) contains three terms: the first term $\|\cdot\|_-$ is the standard l_2 norm that computes only the negative elements. It requires the dose deposited to the PTV (i.e. Mx) no lower than the prescription dose d_p. Meanwhile, we have $D_{95\%}(Mx)=d_p$ as the hard lower-constraint that 95% of the PTV volume receives a dose no lower than the prescription dose. $\|\cdot\|_+$ in the second and third terms is the standard l_2 norm that computes only the positive elements. V_{PTV}, td_p and λ in the second term are the percent volume of PTV, the upper thresholding dose and the priority factor, which together serve as the upper constraint for the PTV. Similarly, V_i, $t_i d_p$ and λ_i in the third term form the upper constraint for the i^{th} OAR. In addition, M and M_i are the dose deposition coefficient matrices for the PTV and i^{th} OAR, respectively, which specified the dose delivered to each voxel inside the patient body from each beamlet under a unit output. In this work, they were extracted from the Eclipse TPS and stored in sparse matrix format. $x \geq 0$ is the beam fluence map to optimize.

It is worth mentioning that in the iterative optimization process to solve Eq. (1), V_{PTV} and V_i always refer to those voxels having higher dose than those non-selected voxels, following the idea of DVH-based treatment optimization. In all, we have the lower constraint for PTV as a hard constraint in our objective function and those upper constraints for PTV and OARs (λ, λ_i, t, t_i, V_{ptv} and V_i) as free treatment planning parameters (TPPs) that will be tuned by the VTP.

We took the prostate cancer IMRT as the testbed and considered cases with one target (prostate) and two critical OARs (the bladder and the rectum) in this work. We then had nine TPPs to tune in the treatment planning process: λ, λ_{bladder}, λ_{rectum}, t, t_{bladder}, t_{rectum}, V_{ptv}, V_{bladder} and V_{rectum}. With a given set of TPPs, the optimization problem of the prostate IMRT treatment planning was solved using the alternating direction method of multipliers (ADMM) (Boyd, 2010), which alternatively updated the fluence map by enforcing the gradient of the objective function close to zero in each step.

3.2.3 The virtual treatment planner network

After building the dose-volume constrained TPS, we employed the deep reinforcement learning (DRL) (Mnih et al., 2015) for the VTP development, which is a combination of the deep neural network (Krizhevsky et al., 2012) and the reinforcement learning (Sutton & Barto, 2018).

Specifically, under the framework of reinforcement learning, we considered the entire treatment planning process as tasks that the agent (the VTP) interacted with the environment (the TPS) in a sequence of observations (intermediate treatment plans), actions (TPP adjustments) and rewards (changes in the planning score). Noting that the TPP adjustment in one step could impact the decision making in future steps, we made the standard assumption that the total reward at time t as $R(t) = r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \ldots + \gamma^{T-t} r_T = \sum_{t'=t}^{T} \gamma^{t'-t} r_{t'}$. Here, r_i ($i = t, t+1, \ldots, T$) was the reward at step i, $\gamma \in [0, 1]$ was the discount factor for future rewards and T was the terminate step for the planning process. We then had the goal for the VTP that it could select those actions to maximize the future rewards. We applied the optimal action-value function (Q value function) from the Q-learning algorithm (CJ Watkins, 1992) to represent the maximum expected return at step t as

$$Q^*(s, a) = \max_{\pi} \mathbb{E}(R(t) | s_t = s, a_t = a, \pi). \qquad (2)$$

Here, π represents a policy mapping state s to action a.

Since we do not know the exact form of the Q value function, we applied the deep neural network (DNN) architecture to parametrize it, considering the flexibility of hyper-dimensional representation of DNN (Mnih et al., 2015). The specific DNN architecture used in this work was illustrated in Figure 3.2. Specifically, we had one DNN network to be responsible for one TPP tuning. With 9 TPPs, we created a VTP network composed of 9 subnetworks (Figure 3.2(a)). All subnetworks shared the same architecture, which was illustrated in Figure 3.2(b). As is shown, it contained four batch normalization layers, seven Leaky Rectified Linear Unit (LeakyReLU) layers, four 1D convolutional layers, four 1D max-pooling layers and one flatten layer, which is different from that used in previous work (Shen et al., 2020).

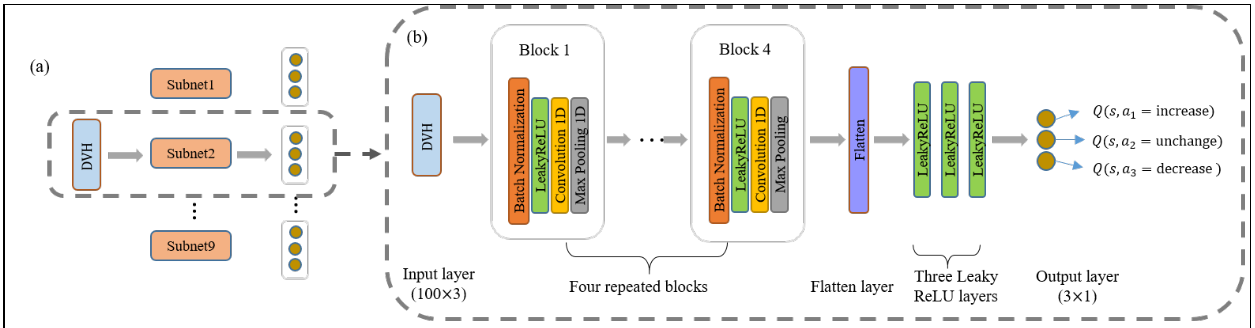

Figure 3.2: (a) The architecture of the deep neural network for the virtual treatment planner, which was composed of nine subnetworks. (b) The structure of a representative subnetwork, which contains 20 hidden layers.

Noting that the in-house TPS was DVH based, we sampled the DVH curves for the PTV and OARs to generate the input state for the VTP. We had three Q-values as outputs for each subnetwork, which corresponded to action options of increasing, decreasing or retaining the TPP magnitude, respectively. We empirically selected the numerical values for the TPP adjustments (Table 3.1). We expected that the specific value selections would not affect the VTP performance, but only the convergence speed. After obtaining the Q values from all nine subnetworks, the action resulting the highest Q value would be selected and fed into the TPS for treatment plan optimization.

Table 3.1: The empirical magnitude changes for different TPPs in step j based on their values in step $j-1$ for different action types.

	$\lambda^{j}_{\text{PTV,OAR}}$	t^{j}_{PTV}	t^{j}_{OAR}	V^{j}_{PTV}	V^{j}_{OAR}
action 1	$1.65 * \lambda^{j-1}_{\text{PTV,OAR}}$	$min(1.01 * t^{j-1}_{\text{PTV}}, 1.2)$	$min(1.25 * t^{j-1}_{\text{OAR}}, 1)$	$min(1.4 * V^{j-1}_{\text{PTV}}, 0.3)$	$min(1.25 * V^{j-1}_{\text{OAR}}, 1)$
action 2	$\lambda^{j-1}_{\text{PTV,OAR}}$	t^{j-1}_{PTV}	t^{j-1}_{OAR}	V^{j-1}_{PTV}	V^{j-1}_{OAR}
action 3	$0.6 * \lambda^{j-1}_{\text{PTV,OAR}}$	$max\,(0.91 * t^{j-1}_{\text{PTV}}, 1)$	$0.6 * t^{j-1}_{\text{OAR}}$	$0.6 * V^{j-1}_{\text{PTV}}$	$0.8 * V^{j-1}_{\text{OAR}}$

To reflect the effect of VTP operations on plan quality improvement, it was reasonable to compute the reward r as the difference of the plan qualities after and before the TPP adjustment by the VTP. That is, $r = \varphi(s') - \varphi(s)$. Here, we used ProKnow (ProKnow Systems, Sanford FL, USA) for prostate cancer IMRT plan to obtain $\varphi(s)$. Relevant to the treatment planning optimization algorithm as stated in section 2.1, nine clinical criteria in the ProKnow scoring system was used in this study: $D_{\text{PTV}}(0.03\text{ cc})$, V_{bladder} (80 Gy), V_{bladder} (75 Gy), V_{bladder} (70 Gy), V_{bladder} (65 Gy), V_{rectum} (75 Gy), V_{rectum} (70 Gy), V_{rectum} (65 Gy), and V_{rectum} (60 Gy) with 79.5 Gy the prescription dose to 95% volume of the PTV. For each treatment plan to be evaluated, it could receive a score $c_i \in [0, 1]$ for criterion i, following the same rule as defined in Table 3.1 of reference (Shen, Chen, Gonzalez, et al., 2021). Hence, the total score that the plan could receive was $\varphi(s) = \sum_{i=1}^{9} c_i$, the maximum of which was 9 and the minimum was 0. It is worthy to mention that we didn't employ the ProKnow score for $D_{95\%}$ due to that we set $D_{95\%} = 79.5$ Gy as a hard constraint for the PTV optimization in our optimization engine.

3.2.4 Training of the VTP network

The goal of training the established VTP network was to determine the parameters (weights) θ such that $Q(s,a;\theta) \approx Q^*(s,a)$. Following the idea of Bellman equation, the optimal value function $Q^*(s,a)$ could be rewritten as

$$Q^*(s,a) = \mathbb{E}_{s'}(r + \gamma \max_{a'} Q^*(s',a')), \quad (3)$$

where the next state s' was formed by taking an action a for current state s, while the corresponding reward was r. We then obtained the optimal value for the (s,a) pair via taking the action a' that maximized Q^* for s'. Consequently, we could train the Q-network by adjusting θ_i at iteration i to reduce the mean square error in the Bellman equation, forming the loss function $L_i(\theta_i)$ at iteration i as

$$L_i(\theta_i) = \mathbb{E}_{s,a,r,s'}\left(r + \gamma \max_{a'} Q(s',a';\theta_i^+) - Q(s,a;\theta_i)\right)^2$$
$$= \mathbb{E}_{s,a,r,s'}(y - Q(s,a;\theta_i))^2. \quad (4)$$

Here, $y = r + \gamma \max_{a'} Q(s',a';\theta_i^+)$ was the approximate target value for $Q^*(s,a)$ at iteration i. θ_i^+ were the network parameters for the target Q-network. Once θ_i^+ was fixed, the loss function $L_i(\theta_i)$ was well defined and could be solved via the stochastic gradient decent method (LeCun et al., 1998). After that, θ_i^+ could be updated based on θ_j $(j \leq i)$ such that we could alternatively optimize the Q-network and target Q-network. To reduce the potential divergence or oscillation for the update of the target Q-network ($Q(\theta^+)$), we only updated θ^+ every N steps with each update a clone of θ from previous Q-network. To make the VTP training efficient, we employed the experience replay method (Lin, 1992) for the updates of θ_i at each iteration i, as it was known to break potential correlations among observation sequences. Specific to our problem, we had $e_t =$

(s_t, a_t, s_{t+1}, r_t) representing the experience acquired at step t with observing an initial treatment plan s_t, applying TPP adjustment a_t to the TPS system, generating a new treatment plan s_{t+1} and obtaining a reward r_t. As e_t was continuously generated during the training process, we could create a replay memory to place them as $D = \{e_1, e_2, \ldots, e_t, \ldots\}$. Each time to update θ_i, we randomly sampled a minibatch of experiences with size L_m from D and applied it to solve Equation (4). Here, the size of D was fixed as L_D. When D was full, we would continue to pop-in those newly generated $e_t's$ and pop-out those oldest elements. The minibatch size was set to be L_M, with $L_M < L_D$. The specific values of L_D and L_M were manually tuned via observing the network training performance. To balance the exploration and exploitation process for effective Q-learning, we employed the ε greedy policy in the VTP network training. Specifically, at the initial training stage, the agent didn't have much experience to learn from and hence we set a relatively large ε ($\varepsilon = 0.999$) to allow it actively explore the state-action space with randomly choosing a TPP adjustment option for the next-step treatment planning. Along with the training time elapsed, the agent accumulated more and more experience from which it could exploit optimal strategies. We then gradually reduced ε with setting its value at the Nth episode as $\varepsilon_N = 0.999/(0.01 * N_{\text{episode}} + 1)$.

3.2.5 Improving training efficiency with Graphical Processing Unit parallel computing

It took time to train the established VTP considering that it contained nine subnetworks. To improve the training efficiency, except for employing the replay memory strategy, we also applied the cProfile technique (Python build-in module) to analyze the run time for each individual step. We found that the most time-consuming portion was relevant to the operation of compressed sparse matrices, including the multiplication of a compressed sparse column (CSC) or row (CSR) with a vector, the column indexing of a CSR, etc. In our algorithm, main sparse matrices were the dose deposition coefficient matrices M and M_i as denoted in Eq. (1), which were frequently operated during the treatment plan optimization process. Hence, to further improve the network training efficiency, we boosted the TPS system via employing the Graphical Unit Processing (GPU) parallel computing technique upon the Nvidia CUDA platform. To support the Pythonic access to the Nvidia's CUDA parallel computation, we utilized the PYCUDA API (application programming interface).

3.2.6 Case studies and evaluations under the in-house TPS and Eclipse TPS

We collected 64 patient cases with prostate cancer IMRT. They were divided into three groups: 10 for training, 2 for verification and the rest 52 for testing. The Q-network was built upon the TensorFlow platform in Python language. The in-house developed TPS was constructed on top of the CUDA platform with PYCUDA technique. The entire algorithm was executed on a GPU server with 8 Intel Xenon 2.30 GHz CPU processors, 32GB memory, and 8 Tesla V100-SXM2 GPU Cards.

The Q-network was trained for 200 episodes with each episode containing a maximum of 30 steps. At the beginning of each episode, we initialized the treatment plan for each training patient case with 7 beam angles and a uniform fluence map for each beam angle. It was then fed into the in-house developed TPS system with a trivial TPP setting (all TPPs $= 1$ except for $V_{\text{ptv}} = 0.1$) to generate an initial treatment plan. We then sampled a random number $\zeta \in [0, 1]$. When $\zeta > \varepsilon$, the DVH of the initial plan would be fed into the Q-network. The Q-network made a TPP adjustment decision and received a corresponding reward. Otherwise, a TPP adjustment option would be randomly picked up from the available TPP adjustment pool. After that, the new TPP was fed into the TPS system for the next-round of treatment plan optimization. This process was repeated until reaching a maximal time step of 30 or a maximal planning score of 9.

Meanwhile, the obtained TPP adjustment experience was placed into the replay memory for the update of the Q-network and target Q-network following the method discussed in section 2.3.

After training each episode with ten patient cases, we verified the obtained network with the two verification cases. With obtaining promising results from both training and verification process, we comprehensively tested the network with fifty testing cases. Lastly, to test whether the VTP developed and trained based on the in-house TPS was effective to operate a commercial TPS, we connected the trained VTP with Eclipse research workstation via API and enabled the VTP-guided Eclipse treatment planning for the rest 2 testing cases. We quantified the plan quality for all patient cases with the ProKnow score system.

3.3 RESULTS

3.3.1 Results for the training and verification cases

The optimal performance of the developed VTP was found at episode 190. The corresponding hyperparameter settings were listed in Table 3.2.

Table 3.2: The hyperparameters and their values used to train the VTP.

Hyperparameter	**Val**	**Description**
learning rate	$1\text{x}\ 10^{-5}$	The learning rate used by the VTP
minibatch size	16	The number of training samples that are used to update θ_i in Equation (4)
target update frequency	500	The frequency with which the target parameters θ^{+} are updated
discount factor	0.7	Discount factor γ used by the Q learning
initial exploration	0.999	Initial value of ε from ε-greedy exploration
final exploration	0.333	Final value of ε from ε-greedy exploration
replay memory	125000	The number of state action pairs that are stored
number of episodes	200	Total number of training episodes
number of steps	30	Maximum number of time steps in each episode

In Figure 3.3, we showed a representative case to illustrate how the VTP iteratively observed a treatment plan DVH generated by the in-house TPS and made a TPP adjustment decision in the network training process. As is shown, the plan DVH at the beginning step failed to satisfy six out of eight criteria for OARs, resulting in a low initial plan score of 3. The VTP observed the plan and decided to lower the threshold dose value for the bladder (t_{BLA} in Figure 3.3(e)), which produced a plan with a better bladder sparing in step 1 (Figure 3.3(b)). It then lowered the threshold dose value and volume for rectum in the subsequent few steps, resulting in a treatment plan with a good sparing for both bladder and rectum but an overdose in PTV in step 8 (Figure 3.3(c)). The VTP then continuously boosted the weight for PTV and finally generated a plan with a full plan score of 9 in step 14 (Figure 3.3(d)).

Statistically, the average and standard deviation of the initial plan scores over the 10 training patient cases was 5.51 ± 2.16. After the VTP guided treatment planning with the in-house TPS, the average and standard deviation of the final plan scores was 8.35± 2.59. Six of them reached the maximum score of 9. In addition, the two verification cases had an initial score of 4.5 ± 1.50 and ended up with a final score of 8.69 ± 0.27. These results indicated that the VTP agent was trained as expected.

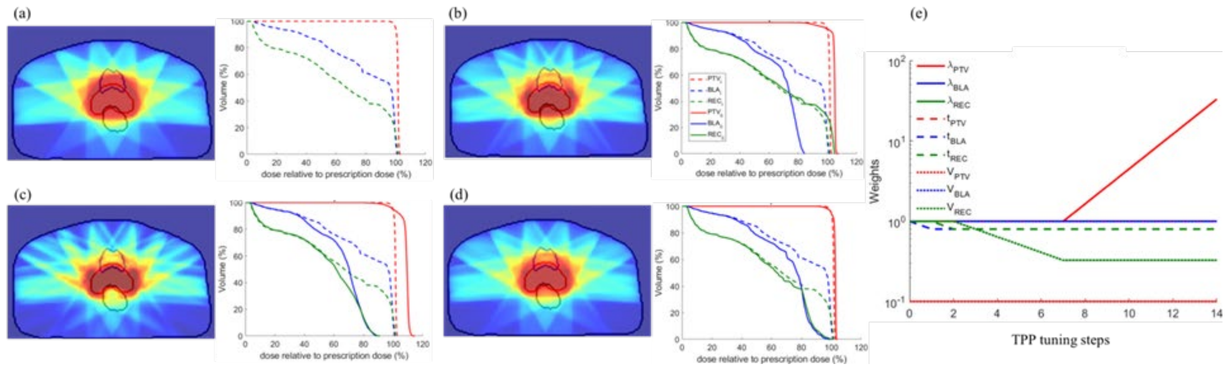

Figure 3.3: The illustration of the VTP-based treatment planning process for a representative training patient case. (a)-(d): the dose fluence maps and DVHs for the treatment plan before TPP adjustment, after one, eight and fourteen steps of TPP adjustments by the VTP, respectively. (e) The specific TPP adjustment made by the VTP. Here, 'BLA' means bladder and 'REC' represents for rectum.

3.3.2 Results for testing cases under the in-house TPS and Eclipse TPS

In Figure 3.4, we illustrated the VTP based treatment planning for a representative testing patient case. From Figure 3.4(a), before the VTP based treatment planning, a portion of bladder and rectum volumes were exposed to the high prescription dose, resulting in a low initial plan score of 4.71. The VTP then decided to decrease the PTV volume that received a dose larger than the prescription dose (V_{ptv} in Eq. (1)) and decreased the threshold dose value for the rectum (t_{rectum} in Eq. (1)) in steps 1-2 (Figure 3.4(e)). It resulted in an effective dose sparing in the rectum while the bladder dose was still high (Figure 3.4(b)). The VTP then decided to decrease the threshold dose for the bladder (t_{bladder} in Eq. (1)) in step 3, which effectively reduced the dose exposure to rectum and bladder but at the expense of overdosing to PTV (Figure 3.4(c)). The VTP then gradually increased the weighting factor of PTV overdose term (λ in Eq. (1)) in the steps 4-9, ending up with a nearly optimal treatment plan (plan scores 8.95 out of 9)

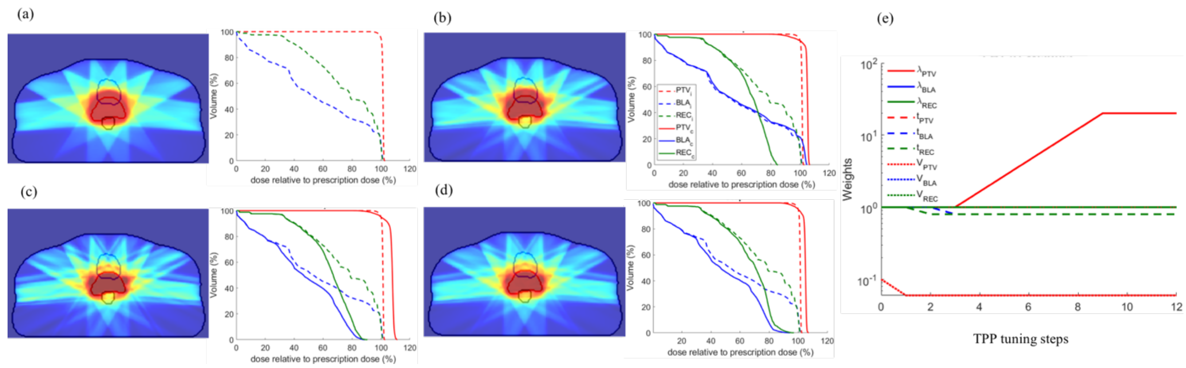

Figure 3.4: The illustration of the VTP-based treatment planning process for a representative testing patient case. (a)-(d): the dose fluence maps and DVHs for the treatment plan before TPP adjustment, after two, three and nine time-steps of TPP adjustments by the VTP, respectively. (e) The specific TPP adjustment made by the VTP. Here, 'BLA' means bladder and 'REC' represents for rectum.

We then analyzed the statistical distributions of the plan scores before and after the VTP based treatment planning for all 50 testing cases. Specifically, we divided the initial treatment plans into 8 categories. For the first 7 categories, the treatment plans satisfied $a \leq$ plan score$<b$, with $a = 2, 3, \ldots, 8$ and $b = a + 1$. As for the last category, the treatment plans were with a plan score equaling 9. We then performed the analysis in two ways. In the first analysis, we tracked the score changes after the treatment planning for all 8 categories, computed the average and standard deviation for each category before and after the treatment planning and showed the result in Figure 3.5(a). As is shown, after the VTP guided treatment planning, the average plan scores were significantly improved for the first seven categories, which remained the same for the last category (maximal plan score). This behavior indicated that the trained VTP was effective in operating the dose optimization engine in generating high-quality treatment plans even for those cases with

relatively low initial plan scores. In the second analysis, we divided the final treatment plans into another 8 categories based on their own plan scores. We counted the total case numbers belonging to each category for both initial and final treatment plans and plotted them side by side in Figure 3.5(b). As is shown, before the VTP based treatment planning, most patient cases had a plan score between 5 and 6. After the plan optimization, the majority ended up with a score 8 and above. Both distributions showed the capability of our trained VTP in performing high-quality treatment planning for prostate cancer IMRT. Overall, the average and standard deviations of the 50 cases were 6.18 ± 1.75 and 8.14 ± 1.27 before and after the VTP based treatment plan optimization.

We also analyzed the dose volume distributions of the 50 testing patient cases following the ProKnow score system. The results were listed in Table 3.3. As is shown, compared to the initial treatment plans, the average percent volumes exposing to doses ≥75, 70 and 65 Gy for bladder and that exposing to dose ≥75, 70, 65 and 60 Gy for rectum have all been significantly reduced after the VTP based treatment planning. On the other hand, the average percent volume of bladder exposing to doses ≥80 Gy and the average minimum dose that 0.03 cm^3 of PTV was exposed were slightly increased, but still well below the criterion values. This dose volume distribution of the testing patient cases indicated that the trained VTP was able to make effective TPP adjustment decisions that could maximize its reward (planning score).

In addition, for all 50 testing patient cases, it took the trained VTP engine less than 1 minute to generate the finally optimized treatment plan per patient case with the in-house TPS system. In comparison, it took an experienced human planner around 3 minutes to complete the same planning process with an average score ~8.5 (Shen, Chen, & Jia, 2021), which indicated the high efficiency of the VTP-guided treatment planning.

As for the VTP-guided Eclipse treatment planning, the DVHs of the initial, intermediate and final treatment plans and the corresponding TPP adjustment process for one patient case were illustrated in Figure 3.6. As is shown, Eclipse generated the initial treatment plan under trivial TPP settings (Figure 3.6(b) step 0). The plan suffered from hot PTV coverage and scored at 8 under the ProKnow score system. In the subsequent VTP-guided Eclipse treatment planning, the VTP observed the intermediate treatment plans through the established API and decided to reduce the priority of rectum dosing in steps 1-5. The Eclipse inverse treatment planning under the updated TPPs helped reduce the dose to OARs yet it did not help improve the plan score significantly. At step 6, the VTP decided to reduce the upper dose limit of PTV, which helped improve the plan score to 8.7 out of a full score of 9. As for the other patient case, it also started at a plan score of 8 and was improved to 8.4 after VTP-guided treatment planning. Observing the entire treatment planning process for both patient cases, the VTP-based automatic TPP adjustments were quite reasonable and helped improve the plan qualities. It indicated that the VTP established upon the in-house dose-volume constrained TPS was also effective in operating Eclipse TPS for high quality treatment planning for prostate cancer IMRT.

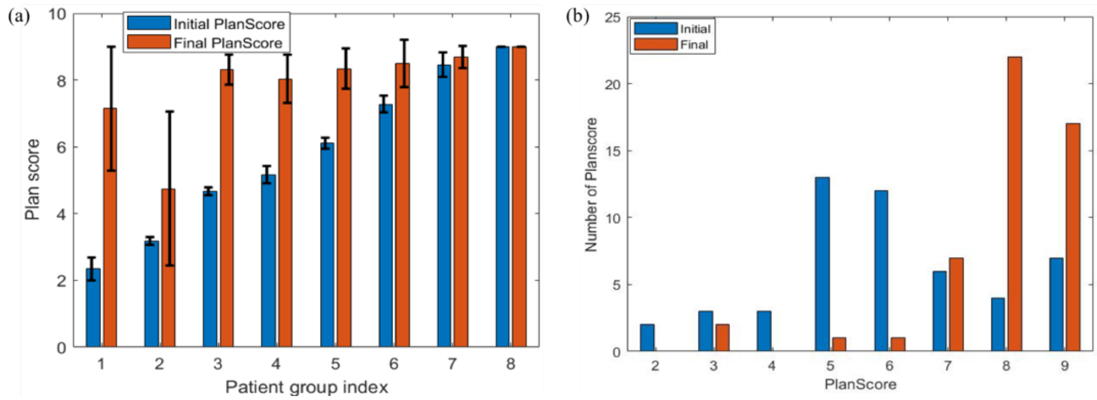

Figure 3.5: The plan score distributions for the 50 testing patient cases before and after the VTP guided treatment planning. (a) The 50 cases were clustered into 8 groups based on their initial plan scores. For each group, the mean plan scores before and after the VTP based planning were represented by the heights of the blue and red bars. The standard deviations were plotted in black. (b) The number of patient cases within different plan score ranges (e.g., '3' means a plan score larger or equal than 3 and smaller than 4) before and after VTP based treatment planning (in blue and red color, respectively).

Table 3.3: The mean and standard deviation (std.) of the dose-volume values for the 50 testing cases. “Criterion” means the requirement from the ProKnow score system. “Before” and “After” represent treatment plans obtained before and after the VTP guided treatment planning.

		Bladder				Rectum				PTV
		V(80 Gy)	V(75 Gy)	V(70 Gy)	V(65 Gy)	V(75 Gy)	V(70 Gy)	V(65 Gy)	V(60 Gy)	D(0.03 cc)
Criterion		<20 %	<30 %	<40 %	<55 %	<20 %	<30 %	<40 %	<55 %	<87.12 Gy
Before	Mean	2.4 %	19.8 %	24.0 %	26.1 %	26.6 %	34.5 %	39.1 %	42.9 %	80.8 Gy
	Std.	2.4 %	8.6%	10.3 %	10.5 %	13.5 %	14.5 %	14.8 %	15.5 %	0.24 Gy
After	Mean	5.9 %	**12.5 %**	**15.5 %**	**18.5 %**	**5.2 %**	**7.5 %**	**12.7 %**	**29.8 %**	85.0 Gy
	Std.	5.6 %	9.1%	10.3 %	**9.5 %**	**10.0 %**	**12.4 %**	**12.4 %**	**13.1 %**	3.4 Gy

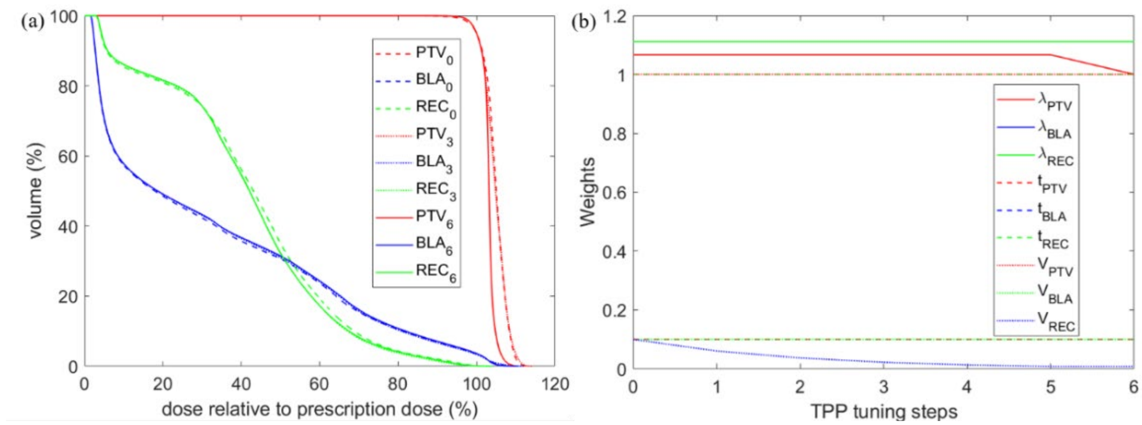

Figure 3.6: The illustration of the VTP-guided Eclipse treatment planning process for a representative testing patient case. (a): DVHs for the treatment plans obtained before, after three-steps and after six-steps of TPP adjustments. (b) The specific TPP adjustment made by the VTP.

3.4 Time performance

As mentioned in the method section, when we implemented the in-house TPS on the CPU platform, sparse matrix operations were extremely time-consuming in the entire VTP guided treatment planning. We then reimplemented the TPS system on the CUDA platform via the PYCUDA technique. We compared the time performance of the VTP training before and after the PYCUDA acceleration and showed the results for the top 5 most time-consuming steps in Figure 3.7. As is shown, time cost for all sparse-matrices-correlated operations ('CSC matvec', 'CSR matvec', 'CSR column index2', 'CSR column index1') were significantly reduced with the PYCUDA acceleration. As for the item relevant to Tensorflow operation ('TF SessionRunCallable'), its execution time was not affected as expected. Overall, the PYCUDA technique improved the running efficiency of the in-house TPS by around 7.1-fold. It reduced the VTP training time from ~80 hours to ~40 hours, increasing the efficiency by about 2-fold. These results indicated that the PYCUDA technique effectively improved the VTP training efficiency.

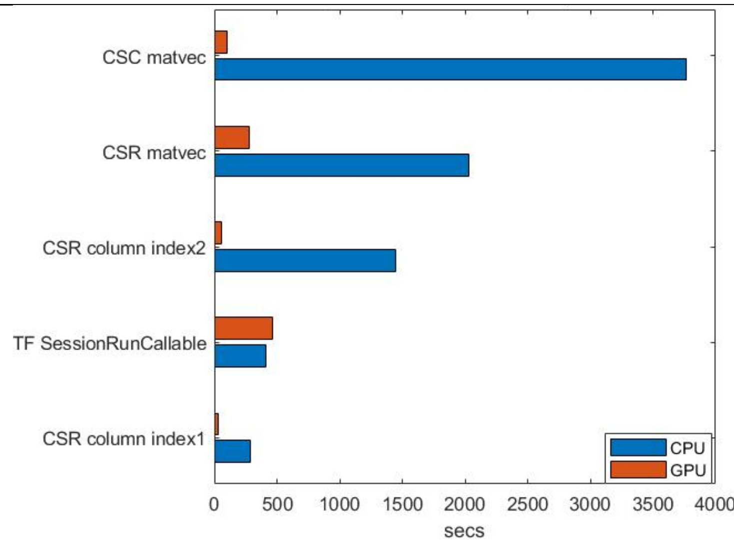

Figure 3.7. The time performance for the top 5 most time-consuming functions in the Q-network training process with incorporating the CPU-based (blue) and PYCUDA-accelerated (red) treatment planning optimization engine, respectively. Here, 'CSC (CSR) matvec' represents the multiplication between a CSC (CSR) matrix with a vector, 'CSR column index' stands for the column indexing of the CSR matrix, and 'TF SessionRunCallable' is a Tensorflow operation.

3.5 DISCUSSION

We successfully trained a deep reinforcement learning based VTP that could operate both in-house dose-volume constrained TPS and Eclipse TPS for automatic treatment planning in prostate cancer IMRT. We used the ProKnow scoring system to quantify the treatment plan quality and to generate the reward for the VTP-based TPP adjustment. We applied the replay memory, the ε-greedy policy and the PYCUDA technique for effective and efficient VTP training. Among them, the PYCUDA technique successfully reduced the VTP training time from ~80 hours to ~40 hours. After the VTP was trained for 200 episodes with 10 patient cases, we tested it with another 50 patient cases. On average, it took the trained VTP less than 1 minute to operate the in-house TPS to generate a final treatment plan for each case, while it took an experienced human planner around 3 minutes to complete the same planning process (Shen, Chen, & Jia, 2021). The average plan score was improved from 6.18 to 8.14 (full score of 9). The effectiveness of the trained VTP in operating Eclipse TPS for automatic treatment planning was also tested with another two independent cases through API connection. The corresponding plan scores were successfully improved from 8 to 8.4 and 8.7 respectively.

It is worth mentioning that in the dose-volume constrained TPS, we had three adjustable constraints for each organ. With two OARs and one PTV considered in this work, nine adjustable constraints were available while the adjustment decision for each constraint was made by an independent set of deep-neural network. Compared to our previous work that was built upon dose constrained TPS (Shen et al., 2020), the networks employed in this work were almost doubled. Although it is more challenging to train a bigger network, with more TPPs to choose from, the newly-established VTP could make a better TPP adjustment decision in each step and hence be more efficient in obtaining a high-quality treatment plan once well-trained. More importantly, we found the VTP established upon the dose-volume constrained TPS was also effective in operating Eclipse TPS for treatment planning without any further tuning of the network. This inspired us to consider the dose-volume constrained TPS a good approximator of the commercial TPS and develop new intelligent networks upon the dose-volume constraint TPS before adapting to commercial TPS, as it is much more convenient to access an in-house TPS than a commercial TPS.

Except for the above success, we also noticed several limitations in our current work. One problem was that a small portion of patient cases with a low starting plan score (2-4) were not effectively improved after the VTP based treatment plan optimization (Figures 3.4(a) and 3.4(b)). One possible reason was that in our current employment of the replay memory technique, the memory buffer was always updated with most recent experiences without differentiating their levels of importance. This could make the agent insufficient in learning those rarely appearing but important TPP adjustment experiences. A potential solution was to employ a more complex case-sampling strategy from the replay memory. In this way, the agent could more frequently 'saw' those rarely-appeared but importance experiences and rapidly learnt to make optimal TPP adjustment decisions when facing challenging cases. It is our next step work to explore this technique to improve the VTP performance.

In addition, as discussed in our previous publication (Shen, Chen, & Jia, 2021), under the current IATP framework, the network parameters could increase quickly along with the increase of the number of TPPs. When the set of TPPs was large enough, the training of the network could be extremely challenging and time consuming. To solve this problem, one way was to reduce the network parameters via employing a hierarchy DRL that decomposed the TPP decision process into three subnetworks, which has been realized in (Shen, Chen, & Jia, 2021). Another possible way was to split the treatment planning goal into a sequence of less-challenging sub-goals. We then could organize a multi-level network with each subnetwork only targeting on the corresponding sub-goal. In this way, each subnetwork was expected to be less complex and easy to reach convergence. Specifically, we could employ the hierarchy actor-critic network, inspired by the work of (Levy, Konidaris, et al., 2017; Levy, Platt, et al., 2017). We will explore this possibility in our future work.

3.6 CONCLUSION

We successfully implemented DRL intelligence with Q-learning technique to operate an in-house dose-volume constrained TPS for high-quality treatment planning in prostate cancer IMRT. The established DRL network was also found to be considerably effective in operating commercial Eclipse TPS for high-quality treatment planning. In both situations, the DRL network was able to make reasonable parameter-adjustment decisions when facing given intermediate treatment plans. We consider the in-house dose-volume constrained TPS a good approximator for commercial TPS, which provides a convenient environment to test newly-developed intelligent treatment planning architectures before adapting to commercial TPS.

CHAPTER 4

Applying a Policy and Value Based Network for a More Efficient Treatment Planning

4.1 INTRODUCTION

In our previous work, we established a deep reinforcement learning (DRL) network that could operate a dose-volume-constrained treatment planning system for IMRT(Damon Sprouts, 2022). The DRL network was Q learning that is value-based reinforcement learning. This work has shown great promise in generating useable treatment plan for the clinic. Not only in plans that started off with poor OARs sparing but it also shows improvement in actual treatment plans in the clinic (Damon Sprouts, 2022). Even with this performance it still couldn't investigated all the possible treatment planning parameters (TPPs) at once. The network itself could only discretely tune and needed 9 subnetworks one for each of the possible TPPs. There were 3 volumes that were used Planning Target Volume (PTV), Bladder, and Rectum. These 3 volumes each had 3 parameters with 3 different type actions. It took about 3 days to reach an episode that had acceptable weights and bias for the network. That could allow for the virtual treatment planner (VTP) to generate acceptable plans. Given that the previous paper was only using a simple prostate testbed and didn't consider all the actual organs at risk (OARs) and target volumes that are needed for a real prostate cancer treatment. (Damon Sprouts, 2022; Shen, Chen, Gonzalez, et al., 2021; Shen, Chen, & Jia, 2021; Shen et al., 2020). The current network to run all the needed contours would add a lot more computation time for the network to achieve an acceptable weight and would decrease the overall efficiency of the network. In this paper we propose using another reinforcement learning algorithm that has shown great promise in not just training in continuous space, but also being able to train with looking at all possibilities. That algorithm is Actor & Critic network (ACN) has two main components to it: Actor which is responsible for choosing the action and Critic which tells how well that state is mapped to the selected action. It gives instance feedback to the agent when making decision, which helps make the training more efficient and quicker as compared to the previous effort (Damon Sprouts, 2022). Applying an ACN into our intelligent automatic treatment planning (IATP) framework would allow for an even more human-like tuning process. That give the agent a chance to look at all TPPs before deciding on how to tune the parameters.

In this study we follow the same principle of end-to-end VTP neural network that can operate a dose-volume constrained TPS.

4.2 METHODS AND MATERIALS

4.2.1 Optimization engine and Treatment Planning Parameters

Similar to our previous paper, we train the VTP on an in-house developed dose-volume constrained TPS that follow the detailed documentation of the plan optimization method for Eclipse TPS. The inverse plan optimization engine TPS by adjusting TPS by solving the below fluence map optimization.

$$\min\left[\frac{1}{2}\left\|Mx-d_p\right\|_{-}^{2}+\frac{\lambda}{2}\left\|\left(Mx-td_p\right)_{V_{PTV}}\right\|_{+}^{2}+\sum_{i}\frac{\lambda_i}{2}\left\|(M_ix-t_id_p)_{V_i}\right\|_{+}^{2}\right], \qquad (1)$$

$$\text{s.t.}\, x\geq 0,\; D_{95\%}(Mx)=d_p.$$

$\|\cdot\|_{-}$ and $\|\cdot\|_{+}$ are l_2 norms that only computed for the negative and positive elements, respectively $x \geq 0$ gives the beam fluence map to be determined, while Mx, M_ix indicates the dose deposition matrices for PTV and i^{th} OARs. λ and λ_i are the weighting factors that can penalize the overdose region for the PTV and i^{th} OARs and lastly t and t_i are the dose threshold for PTV and

OAR. Alternating direction method of multipliers (ADMM) was how the above optimization problem was solved.

4.2.2 Actor and Critic VTP network

The new network is a four-layer dense connected with a rectified linear unit(ReLU) as the activation function. This allows for only positive value to pass from layer to layer. This design is for both the actor and critic network beside the final output layer. The number of output nodes and what type of activation function layer that is used before the output layer is the different between the two networks. The actor has 27 possible actions as its output that corresponds to the nine TPPs and the three actions that can happen. The activation function for the output layer is a softmax that gives the probability for that action to be chosen. The critic only as one output that is the Q value for the quality of that action.

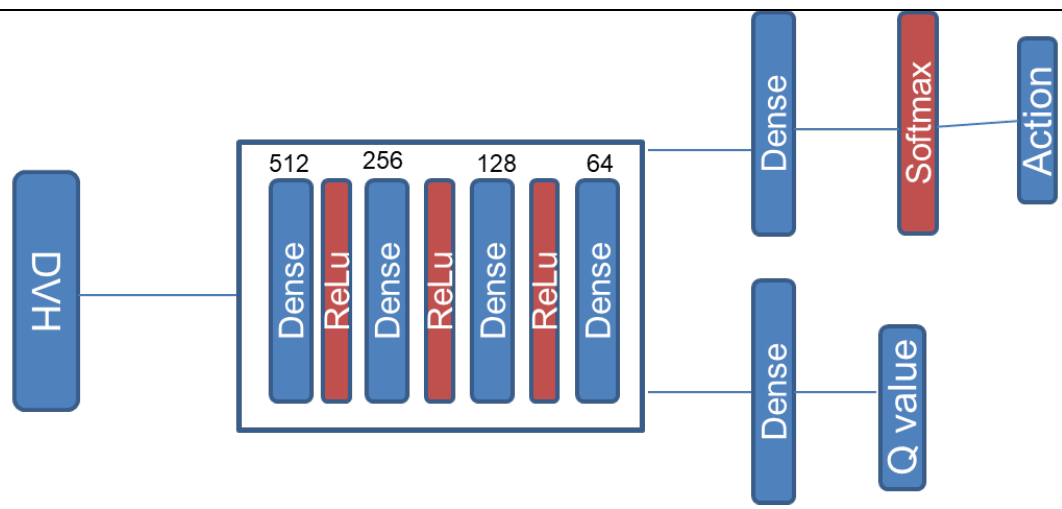

Figure 4.1: Layout of the Actor Critic Network

As in the above figure shows the actor will decide how and what TPPs to tune and then the critic will determine how well that action was by taking in the current state and reward generating a Q value. The TD error will be feed to the actor to know how well the action. TD is known as temporal difference that compares the actual reward to what the estimated reward is. It keeps the policy function separate from the value function(Shablabh Bhatnagar, 2009). Also known as advantage the loss function wants the advantage to be zero. This means that the actor is picking the best action to generate the best state. There are two common ways that ACN can be layout for our project we used the one above another option would have been keeping input and decoder sperate. A key difference between previous papers that use DQN is that ACN doesn't need replay buffer to store the state, action, reward, and next state pair, but does its learning from the previous pair at every next step. This means that the model is actively learning from all steps and not randomly selected pair from the replay memory buffer.

Another adding feature to the IATP was the additional of using gymAI to the framework which is usually used for testing reinforcement algorithms. We used it in this project to better organize the running of the ACN environment where we combine both the TPS with the reward function. The VTP can operate the TPS much in the same way like in the previous papers. Where the whole process of treatment planning is a task that the agent completes by interacting with the TPS and looking at intermediate DVHs and if it meets the stopping criteria of a perfect planScore then it will stop.

4.2.3 Training VTP Actor Critic

The training of the network still follows a ten-patient case, after running for 5 episodes then it would be validated with two patients. If good scoring validation cases were founded, then it would be tested with 50 never before seen cases. The loss function that governs the learning of Actor and Critic was a combined of Actor and Critic Loss function.

$$L_{actor} = -\sum_{t=1}^{T} log\pi_{\theta}(\, a_t|s_t)[G(s_t, a_t) - \mathrm{V}_{\theta}(s_t)]$$

The above equation is the loss function for the actor. The first component is that of the policy is the probability of a_t generate s_t. Where t is the current timestep during the running of the network. $G(s_t, a_t)$-V(s_t) is the pervious mention advantage The G component is what generate from the actor network it is the expected return from the network and V(s_t) is the results from the critic component after it has been parameterized by Ө. Just like the DQN it following bellman equation. Finally, the negative term is there to make sure that it maximizes the probabilities of the actions yielding higher rewards by minimizing the total loss of the network.

$$L_{critic} = {td_{target}}^2 - {Q_{value}}^2$$

For the critic loss function it is the squared difference of the actual Q value that the critic network generated minus what is expected that is represented by the td_{target} . TD is the reward from the previous DVHs minus the current DVHs' planscore plus the discount factor of how to weight future value times the Qvalue predicted for the next state.

$$L_{total} = L_{actor} + L_{critic}$$

The network will combine the two losses and take the total loss is what updates the network. The entire algorithm was run on a GPU server with 8 Intel Xenon 2.30 GHz CPU processors, 32 GB memory, and 8 Tesla V100-SXM2 GPU Cards. The TensorFlow 2.62 was used to take advantage of gradient tape, so that sess run didn't need to be run every time.

4.3 RESULTS

4.3.1 Results from VTP Actor Critic Network

Below are the results for the two validation cases that run every 5 epochs to see how the training, Figure 4.2 shows the scoring from the initial to the last step of the tunning. Planscore is the normal scoring from ProKnow IQ while PlanScore_fine is a special weighted to favor the PTV.

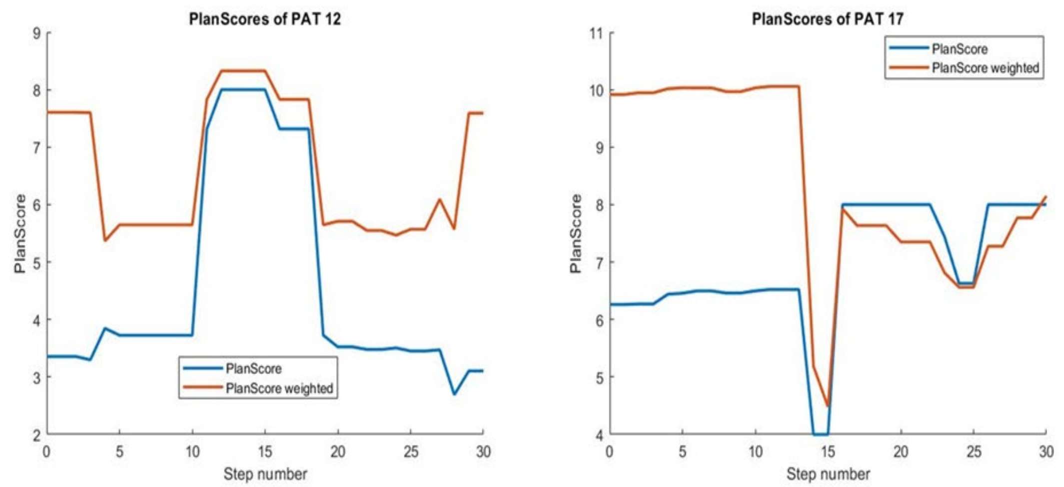

Figure 4.2: Showing the PlanScore and PlanScore_fine of the vaildation cases

PAT 12 is one of the more difficult cases for the VTP to tune do to the fact that it has high overlapping region between the OARs and the PTV. For ACN it takes less epochs to generate a PAT 12 planScore of eight in the previous paper it took longer for the network to genereate a plan at this level.While PAT 17 is more of an average reprentation of the patient datasets. Another thing different with these results of planScore for this low of an epoch there is a lot more difference between the steps.

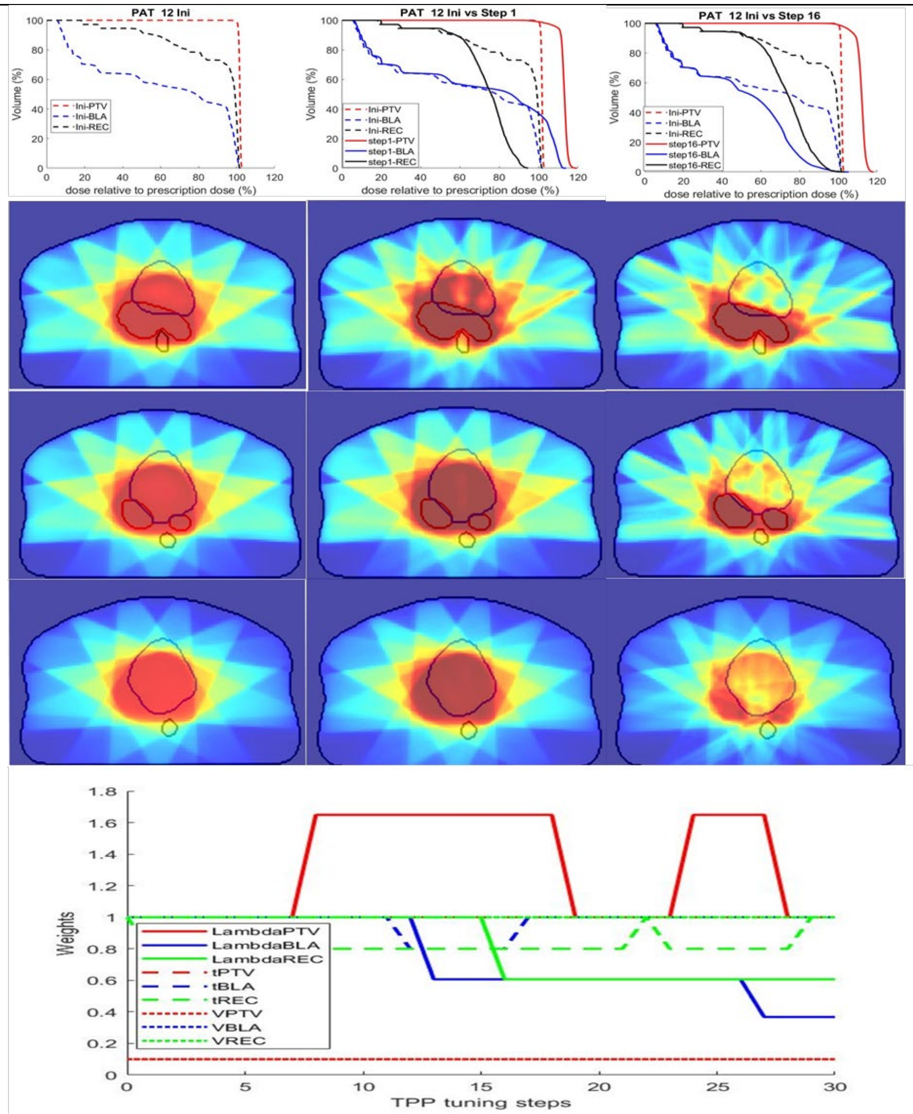

Figure 4.3: The resulting intermediate step DVHs and the weight tunning for Patient 12

In the first couple of steps VTP determines that the main concern for PAT 12 is dealing with the rectum which out of the two OARs is the most radiosenitive and has the strongest dose constriants against it. Once the VTP puts the extra dose in the bladder it becomes consider of the bladder which is reasonable with a human planner. That OARs should be the tune first and avoid as much as possible. In the corresponding figure it can be see that the high dose get concentrated into the rectum from the first column to the second column.

As for PAT 17 the VTP tune lambda PTV first which help with getting the PTV from over dosing, and it has show in previous iterations of the IATP to be an very important weight to be tune, but it didn't take many steps before the agent become more concern for the OARs. It is still balancing act between the two OARs which one the agent wants to tune. The resulting dosemap show how the VTP wants to save the bladder but end up putting too much dose in the rectum, so to overcome this the next steps were to push more dose back into the bladder to resolve the high dose.

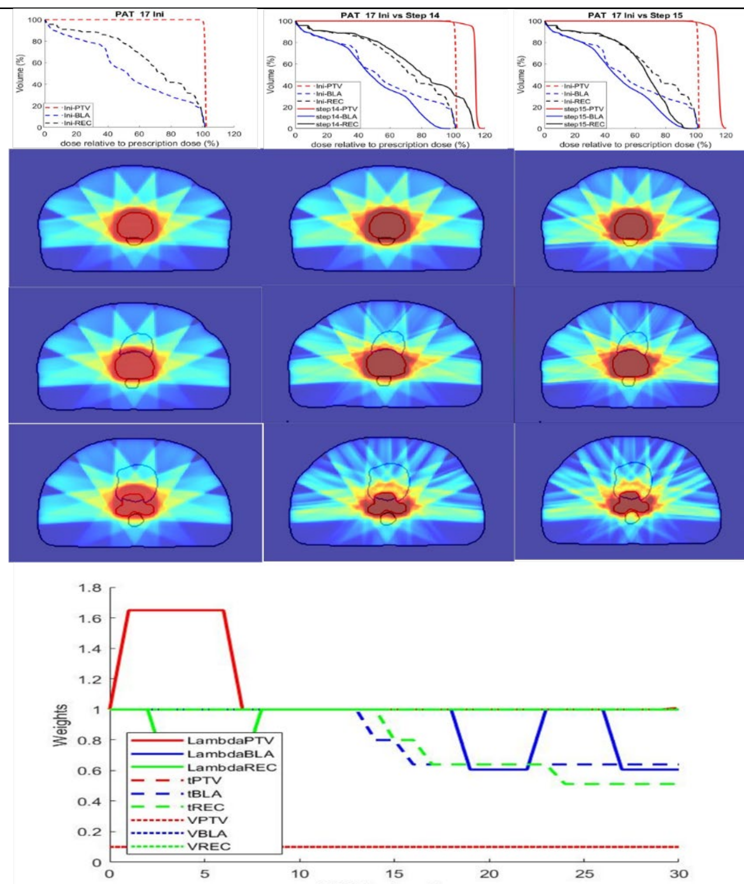

Figure 4.4: The resulting intermediate step DVHs and the weight tunning for Patient 17

Chapter 5

SUMMARY AND FUTURE WORK

Cancer being the leading cause of early death that being cause of death before the age of 70.. There needs to be efficient motion management and reproducible treatment planning for the Radiation Oncology clinic. In order to advance the clinic we investigated motion management of prostate cancer in late stage cancer when the diseases has spread to the lymph nodes. To make the treatment planning more efficient and reproducible we investigate in three different types of deep reinforcement learning: deep Q learning and actor critic.

The first direction of efficient motion management was looking into SBRT monotherapy PLN. SBRT has been used for years in the early-stage prostate where the disease is all located in the prostate, and it was easier to send higher dose without worry about motion due to filling policy of the instructions and using immobilization device. Once the monitoring of motion was moved to the PLN region there needing to be a way to measure the motion of intra-fractional and inter-fractional. Inter-fractional motion has to do with the motion difference between the fraction given i.e., day to day motion. This was the motion of the hip bones since they were used for surrogate of the PLN. Intra-fractional motion has to do with the motion in treatment was given. This motion is the motion of the prostate due to the filling of the bladder or rectum. The motion was track by applying a rotation and translational matrix due to the prostate and hip bones respectfully and recorded what the dose coverage was to the PLN.

We prove that applying a large pelvic node PTV margin and following what the bladder filing protocol that your institution has in place. The inter-fractional pelvic-prostate motion had limited impact on the dose coverage of the pelvic nodes even when using SBRT treatment of the late-stage prostate cancer. For the intra-fractional prostate motion in most case there was a small drift, but in those cases that showed a motion management greater than 3mm there was a significant drop off in dose.

The second direction of this dissertation was automating the treatment planning process for IMRT prostate cancer. Being a simple testcase that both currently A.I. software and dosimetrist could treatment plan easy making it a good test bed. In this direction we also used two different algorithms: Deep Q Learning which is value-based learning that takes maximum output value and does the corresponding action. The second algorithm we use was the Actor Critic method that has both policy and value-based learning. This type of learning learns the state-action value or Q value. It acts by choosing the best action in the state or highest Q value. The actor part of the network uses policy-based which learn the stochastic policy function that maps the state to the chosen action. The critic part of the network uses value-based like deep Q learning. Even though we investigated two different deep reinforcement algorithms the framework which is IATP was the same. That means the TPSs, and reward functions were the same.

We implement a DRL intelligence with Q learning technique to operate an in-house developed dose-volume constrained TPS for useable treatment plans in the prostate cancer IMRT. The trained VTP shown that it was effective in operating commercial Eclipse TPSs. The train VTP was able to make reasonable adjustments in both the in-house and commercial TPSs. In this direction we also moved the in-house developed dose-volume constrained from the CPU to the GPU by using PYCUDA. This allows for an eight-fold increases in the running of the TPS and a two-fold increases in the whole training process.

In the case of applying the actor critic network we were able to decrease the total number of subnetworks from 9 to only 1 network. This change allows for the agent be able to tune and look at all the TPPs at once. This has show promise on allow the agent to get feedback quicker and find the optimal weight for the VTP. The ACN takes the discrete tuning of the network and makes the action into continuous environment.

Now IATP is operating under ACN instead of DQN. This is allowing for the implementation of hierarchal ACN (HACN). HACN would be a simple three-layer hierarchy that takes the initial DVHs and try to achieve s subgoal at each layer to more efficiently handle the increase number of contours for IMRT prostate cancer.

Alayed, Y., Cheung, P., Vesprini, D., Liu, S., Chu, W., Chung, H., Musunuru, H. B., Davidson, M., Ravi, A., & Ho, L. (2018). SABR in High-Risk Prostate Cancer: Outcomes From 2 Prospective Clinical Trials With and Without Elective Nodal Irradiation. *International Journal of Radiation Oncology* Biology* Physics*.

Anwar, M., Weinberg, V., Seymour, Z., Hsu, I. J., Roach, M., 3rd, & Gottschalk, A. R. (2016). Outcomes of hypofractionated stereotactic body radiotherapy boost for intermediate and high-risk prostate cancer. *Radiat Oncol*, *11*, 8. https://doi.org/10.1186/s13014-016-0585-y

Attix, F. H. (2004). *Introduction to Radiological Physics and Radiation Dosimetry*. WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim.

Atun, R., Jaffray, D. A., Barton, M. B., Bray, F., Baumann, M., Vikram, B., Hanna, T. P., Knaul, F. M., Lievens, Y., Lui, T. Y., Milosevic, M., O'Sullivan, B., Rodin, D. L., Rosenblatt, E., Van Dyk, J., Yap, M. L., Zubizarreta, E., & Gospodarowicz, M. (2015). Expanding global access to radiotherapy. *Lancet Oncol*, *16*(10), 1153-1186. https://doi.org/10.1016/S1470-2045(15)00222-3

Azcona, J. D., Xing, L., Chen, X., Bush, K., & Li, R. (2014). Assessing the dosimetric impact of real-time prostate motion during volumetric modulated arc therapy. *Int J Radiat Oncol Biol Phys*, *88*(5), 1167-1174. https://doi.org/10.1016/j.ijrobp.2013.12.015

Azzam, G., Lanciano, R., Arrigo, S., Lamond, J., Ding, W., Yang, J., Hanlon, A., Good, M., & Brady, L. (2015). SBRT: An Opportunity to Improve Quality of Life for Oligometastatic Prostate Cancer. *Front Oncol*, *5*, 101. https://doi.org/10.3389/fonc.2015.00101

Baker, M., & Behrens, C. F. (2016). Determining intrafractional prostate motion using four dimensional ultrasound system. *BMC Cancer*, *16*, 484. https://doi.org/10.1186/s12885-016-2533-5

Baskar, R., Lee, K. A., Yeo, R., & Yeoh, K. W. (2012). Cancer and radiation therapy: current advances and future directions. *Int J Med Sci*, *9*(3), 193-199. https://doi.org/10.7150/ijms.3635

Bauman, G., Ferguson, M., Lock, M., Chen, J., Ahmad, B., Venkatesan, V. M., Sexton, T., D'Souza, D., Loblaw, A., Warner, A., & Rodrigues, G. (2015). A Phase 1/2 Trial of Brief Androgen Suppression and Stereotactic Radiation Therapy (FASTR) for High-Risk Prostate Cancer. *Int J Radiat Oncol Biol Phys*, *92*(4), 856-862. https://doi.org/10.1016/j.ijrobp.2015.02.046

Besl, P. J., & McKay, N. D. (1992). Method for registration of 3-D shapes. Sensor fusion IV: control paradigms and data structures,

Bolzicco, G., Favretto, M. S., Satariano, N., Scremin, E., Tambone, C., & Tasca, A. (2013). A single-center study of 100 consecutive patients with localized prostate cancer treated with stereotactic body radiotherapy. *BMC Urol*, *13*, 49. https://doi.org/10.1186/1471-2490-13-49

Bortfeld, T. (2006). IMRT: a review and preview. *Phys Med Biol*, *51*(13), R363-379. https://doi.org/10.1088/0031-9155/51/13/R21

Boyd, S. (2010). Distributed Optimization and Statistical Learning via the Alternating Direction Method of Multipliers. *Foundations and Trends® in Machine Learning*, *3*(1), 1-122. https://doi.org/10.1561/2200000016

Bray, F., Laversanne, M., Weiderpass, E., & Soerjomataram, I. (2021). The ever-increasing importance of cancer as a leading cause of premature death worldwide. *Cancer*, *127*(16), 3029-3030. https://doi.org/10.1002/cncr.33587

Burger, J. (2003). Radioactive sources in brachytherapy. *Radiat Oncol*, *37*(2), 127-131.

Chang, A. J., Autio, K. A., Roach, M., 3rd, & Scher, H. I. (2014). High-risk prostate cancer-classification and therapy. *Nat Rev Clin Oncol*, *11*(6), 308-323. https://doi.org/10.1038/nrclinonc.2014.68

Chang, A. T. Y., Hung, A. W. M., Cheung, F. W. K., Lee, M. C. H., Chan, O. S. H., Philips, H., Cheng, Y. T., & Ng, W. T. (2016). Comparison of Planning Quality and Efficiency Between Conventional and Knowledge-based Algorithms in Nasopharyngeal Cancer Patients Using Intensity Modulated Radiation Therapy. *Int J Radiat Oncol Biol Phys*, *95*(3), 981-990. https://doi.org/10.1016/j.ijrobp.2016.02.017

Chanyavanich, V., Das, S. K., Lee, W. R., & Lo, J. Y. (2011). Knowledge-based IMRT treatment planning for prostate cancer. *Med Phys*, *38*(5), 2515-2522. https://doi.org/10.1118/1.3574874

Chen, L. N., Suy, S., Uhm, S., Oermann, E. K., Ju, A. W., Chen, V., Hanscom, H. N., Laing, S., Kim, J. S., Lei, S., Batipps, G. P., Kowalczyk, K., Bandi, G., Pahira, J., McGeagh, K. G., Collins, B. T., Krishnan, P., Dawson, N. A., Taylor, K. L., . . . Collins, S. P. (2013). Stereotactic body radiation therapy (SBRT) for clinically localized prostate cancer: the Georgetown University experience. *Radiation Oncology*, *8*, 58. https://doi.org/10.1186/1748-717X-8-58

Chen, W., Unkelbach, J., Trofimov, A., Madden, T., Kooy, H., Bortfeld, T., & Craft, D. (2012). Including robustness in multi-criteria optimization for intensity-modulated proton therapy. *Physics in Medicine Biology*, *57*(3), 591.

Chen, Y., & Medioni, G. (1992). Object modelling by registration of multiple range images. *Image vision computing*, *10*(3), 145-155.

Chi, Y., Rezaeian, N. H., Shen, C., Zhou, Y., Lu, W., Yang, M., Hannan, R., & Jia, X. (2017). A new method to reconstruct intra-fractional prostate motion in volumetric modulated arc therapy. *Phys Med Biol*, *62*(13), 5509-5530. https://doi.org/10.1088/1361-6560/aa6e37

Cho, B. (2018). Intensity-modulated radiation therapy: a review with a physics perspective. *Radiat Oncol*

J, *36*(1), 1-10. https://doi.org/10.3857/roj.2018.00122

Cihan, Y. (2018). The role and importance of SBRT in prostate cancer. *Int Braz J Urol*, *44*(6), 1272-1274. https://doi.org/10.1590/S1677-5538.IBJU.2018.0484

CJ Watkins, P. D. (1992). *Q-learning Machine learning*

Craft, D. L., Hong, T. S., Shih, H. A., & Bortfeld, T. R. (2012). Improved planning time and plan quality through multicriteria optimization for intensity-modulated radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, *82*(1), e83-e90.

Damon Sprouts, U. S., Mauro Tambasco,Peter Blomgren, Laura Cervino. (2017). Comparison of Device-Based and Deviceless 4DCT Reconstruction

Damon Sprouts, Y. G., Chao Wang, Xun Jia, Chenyang Shen, Yujie Chi. (2022). The Development of a Deep Reinforcement Learning Network for Dose-Volume-Constrained Treatment Planning in Prostate Cancer Intensity Modulated Radiotherapy.

Dang, A., Kupelian, P. A., Cao, M., Agazaryan, N., & Kishan, A. U. (2018). Image-guided radiotherapy for prostate cancer. *Transl Androl Urol*, *7*(3), 308-320. https://doi.org/10.21037/tau.2017.12.37

Davis, J., Sharma, S., Shumway, R., Perry, D., Bydder, S., Simpson, C. K., & D'Ambrosio, D. (2015). Stereotactic Body Radiotherapy for Clinically Localized Prostate Cancer: Toxicity and Biochemical Disease-Free Outcomes from a Multi-Institutional Patient Registry. *Cureus*, *7*(12), e395. https://doi.org/10.7759/cureus.395

Desouky, O., Ding, N., & Zhou, G. (2015). Targeted and non-targeted effects of ionizing radiation. *Journal of Radiation Research and Applied Sciences*, *8*(2), 247-254. https://doi.org/10.1016/j.jrras.2015.03.003

Edward Melian, G. S. M., Zvi Fuks, Steven A. Leibel, Anita Niehaus, Helen Lorant, Micheal Zelefsky, Bernard Balwin, Gerald Kutcher (1997). Variation in Prostate Position Quantitation and implications for three-dimensional Conformal Treatment Planning. *Internal Journal Radiation Oncology Biology 38*(1), 77-81.

Eric Hall, A. J. G. (2012). *Radiobiology for the Radiologist*. LIPPINCOTT WILLIAMS&WILKINS, a WOLTERS KLUWER.

Faiz M. Kahn, B. J. G. (2012). *Treatment Planning in Radiation Oncology* LIPPINCOTT WILLIAMS & WILKINS, a WOLTERS KLUWER.

Fogliata, A., Belosi, F., Clivio, A., Navarria, P., Nicolini, G., Scorsetti, M., Vanetti, E., & Cozzi, L. (2014). On the pre-clinical validation of a commercial model-based optimisation engine: application to volumetric modulated arc therapy for patients with lung or prostate cancer. *Radiother Oncol*, *113*(3), 385-391. https://doi.org/10.1016/j.radonc.2014.11.009

Franz, A. M., Haidegger, T., Birkfellner, W., Cleary, K., Peters, T. M., & Maier-Hein, L. (2014). Electromagnetic Tracking in Medicine-A Review of Technology, Validation, and Applications. *Ieee Transactions on Medical Imaging*, *33*(8), 1702-1725. <Go to ISI>://WOS:000340237800012

Gonzalez-Motta, A., & Roach III, M. (2018). Stereotactic body radiation therapy (SBRT) for high-risk prostate cancer: Where are we now? *J Practical radiation oncology*, *8*(3), 185-202.

Greco, C., Pares, O., Pimentel, N., Louro, V., Morales, J., Nunes, B., Vasconcelos, A. L., Antunes, I., Kociolek, J., Stroom, J., Viera, S., Mateus, D., Cardoso, M. J., Soares, A., Marques, J., Freitas, E., Coelho, G., & Fuks, Z. (2020). Target motion mitigation promotes high-precision treatment planning and delivery of extreme hypofractionated prostate cancer radiotherapy: Results from a phase II study. *Radiother Oncol*, *146*, 21-28. https://doi.org/10.1016/j.radonc.2020.01.029

Hannan, R., Salamekh, S., Desai, N. B., Garant, A., Folkert, M. R., Costa, D. N., Mannala, S., Ahn, C., Mohamad, O., & Laine, A. (2021). SAbR for High-Risk Prostate Cancer—A Prospective Multilevel MRI-Based Dose Escalation Trial. *International Journal of Radiation Oncology* Biology* Physics*.

Huang, C. Y., Tehrani, J. N., Ng, J. A., Booth, J., & Keall, P. (2015). Six degrees-of-freedom prostate and lung tumor motion measurements using kilovoltage intrafraction monitoring. *Int J Radiat Oncol Biol Phys*, *91*(2), 368-375. https://doi.org/10.1016/j.ijrobp.2014.09.040

Hussein, M., Heijmen, B. J. M., Verellen, D., & Nisbet, A. (2018). Automation in intensity modulated radiotherapy treatment planning-a review of recent innovations. *Br J Radiol*, *91*(1092), 20180270. https://doi.org/10.1259/bjr.20180270

Hussein, M., South, C. P., Barry, M. A., Adams, E. J., Jordan, T. J., Stewart, A. J., & Nisbet, A. (2016). Clinical validation and benchmarking of knowledge-based IMRT and VMAT treatment planning in pelvic anatomy. *Radiother Oncol*, *120*(3), 473-479. https://doi.org/10.1016/j.radonc.2016.06.022

Institute, N. C. (2019). *Radiation Therapy to Treat Cancer*. Retrieved April from https://www.cancer.gov/about-cancer/treatment/types/radiation-therapy

Intensity Modulated Radiation Therapy Collaborative Working Group. (2001). Intensity-modulated Radiotherapy: Current Status and issues of Interest. *International Journal of Radiation Oncology* Biology* Physics*, *51*(4), 880-914.

Janowski, E., Chen, L. N., Kim, J. S., Lei, S., Suy, S., Collins, B., Lynch, J., Dritschilo, A., & Collins, S. (2014). Stereotactic body radiation therapy (SBRT) for prostate cancer in men with large prostates

(>/=50 cm(3)). *Radiation Oncology*, *9*, 241. https://doi.org/10.1186/s13014-014-0241-3

Jerrold T. Bushberg, J. A. S., Edwin M. Leidholdt, John M. Boone (2012). *The Essential Physics of Medical Imaging* LIPPINCOTT WILLIAMS & WILKINS, a WOLTERS KLUWER

Kang, J. K., Cho, C. K., Choi, C. W., Yoo, S., Kim, M. S., Yang, K., Yoo, H., Kim, J. H., Seo, Y. S., Lee, D. H., & Jo, M. (2011). Image-guided stereotactic body radiation therapy for localized prostate cancer. *Tumori*, *97*(1), 43-48. https://www.ncbi.nlm.nih.gov/pubmed/21528663

Keall, P. J., Aun Ng, J., O'Brien, R., Colvill, E., Huang, C.-Y., Rugaard Poulsen, P., Fledelius, W., Juneja, P., Simpson, E., Bell, L., Alfieri, F., Eade, T., Kneebone, A., & Booth, J. T. (2015). The first clinical treatment with kilovoltage intrafraction monitoring (KIM): A real-time image guidance method. *Medical physics*, *42*(1), 354-358. https://doi.org/doi:http://dx.doi.org/10.1118/1.4904023

Keall, P. J., Ng, J. A., Juneja, P., O'Brien, R. T., Huang, C. Y., Colvill, E., Caillet, V., Simpson, E., Poulsen, P. R., Kneebone, A., Eade, T., & Booth, J. T. (2016). Real-Time 3D Image Guidance Using a Standard LINAC: Measured Motion, Accuracy, and Precision of the First Prospective Clinical Trial of Kilovoltage Intrafraction Monitoring-Guided Gating for Prostate Cancer Radiation Therapy. *International Journal of Radiation Oncology Biology Physics*, *94*(5), 1015-1021. https://doi.org/10.1016/j.ijrobp.2015.10.009

Kershaw, L., van Zadelhoff, L., Heemsbergen, W., Pos, F., & van Herk, M. (2018). Image Guided Radiation Therapy Strategies for Pelvic Lymph Node Irradiation in High-Risk Prostate Cancer: Motion and Margins. *Int J Radiat Oncol Biol Phys*, *100*(1), 68-77. https://doi.org/10.1016/j.ijrobp.2017.08.044

Kim, H. J., Phak, J. H., & Kim, W. C. (2017). Prostate-specific antigen kinetics following hypofractionated stereotactic body radiotherapy boost and whole pelvic radiotherapy for intermediate- and high-risk prostate cancer. *Asia Pac J Clin Oncol*, *13*(1), 21-27. https://doi.org/10.1111/ajco.12472

Kim, J. H., Nguyen, D. T., Huang, C. Y., Fuangrod, T., Caillet, V., O'Brien, R., Poulsen, P., Booth, J., & Keall, P. (2017). Quantifying the accuracy and precision of a novel real-time 6 degree-of-freedom kilovoltage intrafraction monitoring (KIM) target tracking system. *Phys Med Biol*, *62*(14), 5744-5759. https://doi.org/10.1088/1361-6560/aa6ed7

Kishan, A. U., & King, C. R. (2019). Optimizing Selection of Patients for Prostate SBRT: Overview of Toxicity and Efficacy in Low, Intermediate, and High-Risk Prostate Cancer. In *Stereotactic Radiosurgery for Prostate Cancer* (pp. 1-16). Springer.

Kishan, A. U., Lamb, J. M., Jani, S. S., Kang, J. J., Steinberg, M. L., & King, C. R. (2015). Pelvic nodal dosing with registration to the prostate: implications for high-risk prostate cancer patients receiving stereotactic body radiation therapy. *Int J Radiat Oncol Biol Phys*, *91*(4), 832-839.

https://doi.org/10.1016/j.ijrobp.2014.11.035

Kotecha, R., Djemil, T., Tendulkar, R. D., Reddy, C. A., Thousand, R. A., Vassil, A., Stovsky, M., Berglund, R. K., Klein, E. A., & Stephans, K. L. (2016). Dose-Escalated Stereotactic Body Radiation Therapy for Patients With Intermediate- and High-Risk Prostate Cancer: Initial Dosimetry Analysis and Patient Outcomes. *Int J Radiat Oncol Biol Phys*, *95*(3), 960-964. https://doi.org/10.1016/j.ijrobp.2016.02.009

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*, 1097-1105.

Kubo, K., Monzen, H., Ishii, K., Tamura, M., Kawamorita, R., Sumida, I., Mizuno, H., & Nishimura, Y. (2017). Dosimetric comparison of RapidPlan and manually optimized plans in volumetric modulated arc therapy for prostate cancer. *Phys Med*, *44*, 199-204. https://doi.org/10.1016/j.ejmp.2017.06.026

Kupelian, P., Willoughby, T., Mahadevan, A., Djemil, T., Weinstein, G., Jani, S., Enke, C., Solberg, T., Flores, N., & Liu, D. (2007). Multi-institutional clinical experience with the Calypso System in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy. *International Journal of Radiation Oncology* Biology* Physics*, *67*(4), 1088-1098.

Kupelian, P., Willoughby, T., Mahadevan, A., Djemil, T., Weinstein, G., Jani, S., Enke, C., Solberg, T., Flores, N., Liu, D., Beyer, D., & Levine, L. (2007). Multi-institutional clinical experience with the Calypso System in localization and continuous, real-time monitoring of the prostate gland during external radiotherapy. *International Journal of Radiation Oncology Biology Physics*, *67*(4), 1088-1098. https://doi.org/10.1016/j.ijrobp.2006.10.026

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Lee, S. W., Jang, H. S., Lee, J. H., Kim, S. H., & Yoon, S. C. (2014). Stereotactic body radiation therapy for prostate cancer patients with old age or medical comorbidity: a 5-year follow-up of an investigational study. *Medicine (Baltimore)*, *93*(28), e290. https://doi.org/10.1097/MD.0000000000000290

Levy, A., Konidaris, G., Platt, R., & Saenko, K. (2017). Learning multi-level hierarchies with hindsight. *arXiv preprint arXiv:.00948*.

Levy, A., Platt, R., & Saenko, K. (2017). Hierarchical actor-critic. *arXiv preprint arXiv:.00948*, *12*.

Li, Y. (2017). Deep Reinforcement learning: An overview. *arXiv preprint*.

Lin, L.-J. (1992). *Reinforcement learning for robots using neural networks*. Carnegie Mellon University.

Lin, Y. W., Lin, L. C., & Lin, K. L. (2014). The early result of whole pelvic radiotherapy and stereotactic body radiotherapy boost for high-risk localized prostate cancer. *Front Oncol*, *4*, 278. https://doi.org/10.3389/fonc.2014.00278

Line Krhili, S., Crehange, G., Albert-Dufrois, H., Guimas, V., Minsat, M., & Supiot, S. (2019). [Moderate or extreme hypofractionation and localized prostate cancer: The times are changing]. *Cancer Radiotherapie*, *23*(6-7), 503-509. https://doi.org/10.1016/j.canrad.2019.07.139 (Hypofractionnement modere ou extreme et cancers prostatiques localises : les temps sont en train de changer.)

Liu, W., Qian, J. G., Hancock, S. L., Xing, L., & Luxton, G. (2010). Clinical development of a failure detection-based online repositioning strategy for prostate IMRT-Experiments, simulation, and dosimetry study. *Medical Physics*, *37*(10), 5287-5297. <Go to ISI>://WOS:000283483700016

Lotan, Y., Stanfield, J., Cho, L. C., Sherwood, J. B., Abdel-Aziz, K. F., Chang, C. H., Forster, K., Kabbani, W., Hsieh, J. T., Choy, H., & Timmerman, R. (2006). Efficacy of high dose per fraction radiation for implanted human prostate cancer in a nude mouse model. *J Urol*, *175*(5), 1932-1936. https://doi.org/10.1016/S0022-5347(05)00893-1

Lu, X. Q., Shanmugham, L. N., Mahadevan, A., Nedea, E., Stevenson, M. A., Kaplan, I., Wong, E. T., La Rosa, S., Wang, F., & Berman, S. M. (2008). Organ deformation and dose coverage in robotic respiratory-tracking radiotherapy. *Int J Radiat Oncol Biol Phys*, *71*(1), 281-289. https://doi.org/10.1016/j.ijrobp.2007.12.042

Mao, W., Wiersma, R. D., & Xing, L. (2008). Fast internal marker tracking algorithm for onboard MV and kV imaging systems. *Med Phys*, *35*(5), 1942-1949. https://doi.org/10.1118/1.2905225

Mercado, C., Kress, M. A., Cyr, R. A., Chen, L. N., Yung, T. M., Bullock, E. G., Lei, S., Collins, B. T., Satinsky, A. N., Harter, K. W., Suy, S., Dritschilo, A., Lynch, J. H., & Collins, S. P. (2016). Intensity-Modulated Radiation Therapy with Stereotactic Body Radiation Therapy Boost for Unfavorable Prostate Cancer: The Georgetown University Experience. *Front Oncol*, *6*, 114. https://doi.org/10.3389/fonc.2016.00114

Mesci, A., Isfahanian, N., Dayes, I., Lukka, H., & Tsakiridis, T. (2021). The Journey of Radiotherapy Dose Escalation in High Risk Prostate Cancer; Conventional Dose Escalation to Stereotactic Body Radiotherapy (SBRT) Boost Treatments. *Clinical Genitourinary Cancer*.

Miralbell, R., Molla, M., Rouzaud, M., Hidalgo, A., Toscas, J. I., Lozano, J., Sanz, S., Ares, C., Jorcano, S., Linero, D., & Escude, L. (2010). Hypofractionated boost to the dominant tumor region with intensity modulated stereotactic radiotherapy for prostate cancer: a sequential dose escalation pilot study. *Int J Radiat Oncol Biol Phys*, *78*(1), 50-57. https://doi.org/10.1016/j.ijrobp.2009.07.1689

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., & Ostrovski, G. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*(7540), 529-533.

Murthy, V., Gupta, M., Mulye, G., Maulik, S., Munshi, M., Krishnatry, R., Phurailatpam, R., Mhatre, R., Prakash, G., & Bakshi, G. (2018). Early Results of Extreme Hypofractionation Using Stereotactic Body Radiation Therapy for High-risk, Very High-risk and Node-positive Prostate Cancer. *Clin Oncol (R Coll Radiol)*, *30*(7), 442-447. https://doi.org/10.1016/j.clon.2018.03.004

Musunuru, H. B., D'Alimonte, L., Davidson, M., Ho, L., Cheung, P., Vesprini, D., Liu, S., Chu, W., Chung, H., Ravi, A., Deabreu, A., Zhang, L., Commisso, K., & Loblaw, A. (2018). Phase 1-2 Study of Stereotactic Ablative Radiotherapy Including Regional Lymph Node Irradiation in Patients With High-Risk Prostate Cancer (SATURN): Early Toxicity and Quality of Life. *Int J Radiat Oncol Biol Phys*, *102*(5), 1438-1447. https://doi.org/10.1016/j.ijrobp.2018.07.2005

Netter, F. H., & Colacino, S. (1989). *Atlas of human anatomy*. Ciba-Geigy Corporation.

Nwankwo, O., Mekdash, H., Sihono, D. S., Wenz, F., & Glatting, G. (2015). Knowledge-based radiation therapy (KBRT) treatment planning versus planning by experts: validation of a KBRT algorithm for prostate cancer treatment planning. *Radiat Oncol*, *10*, 111. https://doi.org/10.1186/s13014-015-0416-6

Park, Y., Park, H. J., Jang, W. I., Jeong, B. K., Kim, H. J., & Chang, A. R. (2018). Long-term results and PSA kinetics after robotic SBRT for prostate cancer: multicenter retrospective study in Korea (Korean radiation oncology group study 15-01). *Radiat Oncol*, *13*(1), 230. https://doi.org/10.1186/s13014-018-1182-z

Pinitpatcharalert, A., Happersett, L., Kollmeier, M., McBride, S., Gorovets, D., Tyagi, N., Varghese, M., & Zelefsky, M. J. (2019). Early Tolerance Outcomes of Stereotactic Hypofractionated Accelerated Radiation Therapy Concomitant with Pelvic Node Irradiation in High-risk Prostate Cancer. *Advances in Radiation Oncology*.

Poulsen, P. R., Cho, B., Sawant, A., & Keall, P. J. (2010). Implementation of a new method for dynamic multileaf collimator tracking of prostate motion in arc radiotherapy using a single kV imager. *International Journal of Radiation Oncology* Biology* Physics*, *76*(3), 914-923.

Prevention, C. f. D. C. a. (2021). *How is Prostate Cancer treated?* Retrieved April 10 from https://www.cdc.gov/cancer/prostate/basic_info/treatment.htm

Ray, C. (2011). Long-term outcomes of SBRT in low-risk prostate cancer. *Nat Rev Urol*, *8*(4), 174.

https://doi.org/10.1038/nrurol.2011.32

Ricco, A., Manahan, G., Lanciano, R., Hanlon, A., Yang, J., Arrigo, S., Lamond, J., Feng, J., Mooreville, M., Garber, B., & Brady, L. (2016). The Comparison of Stereotactic Body Radiation Therapy and Intensity-Modulated Radiation Therapy for Prostate Cancer by NCCN Risk Groups. *Front Oncol*, *6*, 184. https://doi.org/10.3389/fonc.2016.00184

Shablabh Bhatnagar, R. S., Mohammad Ghavamzadeh, Mark Lee. (2009). Natural Actor-Critic Algorithms. *Automatica*, *45*(11).

Shen, C., Chen, L., Gonzalez, Y., & Jia, X. (2021). Improving Efficiency of Training a Virtual Treatment Planner Network via Knowledge-guided Deep Reinforcement Learning for Intelligent Automatic Treatment Planning of Radiotherapy. *Med Phys*. https://doi.org/10.1002/mp.14712

Shen, C., Chen, L., & Jia, X. (2021). A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. *Phys Med Biol*, *66*(13). https://doi.org/10.1088/1361-6560/ac09a2

Shen, C., Gonzalez, Y., Klages, P., Qin, N., Jung, H., Chen, L., Nguyen, D., Jiang, S. B., & Jia, X. (2019). Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Phys Med Biol*, *64*(11), 115013. https://doi.org/10.1088/1361-6560/ab18bf

Shen, C., Nguyen, D., Chen, L., Gonzalez, Y., McBeth, R., Qin, N., Jiang, S. B., & Jia, X. (2020). Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. *Med Phys*, *47*(6), 2329-2336. https://doi.org/10.1002/mp.14114

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*(7587), 484-489. https://doi.org/10.1038/nature16961

Society, A. C. (2022). *Test to Diagnose and Stage Prostate Cancer*. Retrieved April 2 from https:\\www.cancer.org\cancer\prostate-cancer\detection-diagnosis\how-diagnosed.html

Society, A. P. (2001). *This Month in Physics History (Roentgen's Discovery of X-rays)*. Retrieved April 10 from

Su, Z., Zhang, L. S., Murphy, M., & Williamson, J. (2011). Analysis of Prostate Patient Setup and Tracking Data: Potential Intervention Strategies. *International Journal of Radiation Oncology Biology*

Physics, *81*(3), 880-887. https://doi.org/10.1016/j.ijrobp.2010.07.1978

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., & Bray, F. (2021). Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries. *CA Cancer J Clin*, *71*(3), 209-249. https://doi.org/10.3322/caac.21660

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Thieke, C., Küfer, K.-H., Monz, M., Scherrer, A., Alonso, F., Oelfke, U., Huber, P. E., Debus, J., & Bortfeld, T. (2007). A new concept for interactive radiotherapy planning with multicriteria optimization: first clinical evaluation. *Radiotherapy Oncology (Williston Park)*, *85*(2), 292-298.

Timmerman, R. D., Kavanagh, B. D., Cho, L. C., Papiez, L., & Xing, L. (2007). Stereotactic body radiation therapy in multiple organ sites. *J Clin Oncol*, *25*(8), 947-952. https://doi.org/10.1200/JCO.2006.09.7469

TImothy P.Lillicrap, J. J. H., Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Sliver & Daan Wierstra. (2016). Continuous Control with Deep Reinforcement Learning. *arXiv preprint arXiv:1509.02971*.

Tyagi, N., Hipp, E., Cloutier, M., Charas, T., Fontenla, S., Mechalakos, J., Hunt, M., & Zelefsky, M. (2019). Impact of daily soft-tissue image guidance to prostate on pelvic lymph node (PLN) irradiation for prostate patients receiving SBRT. *J Appl Clin Med Phys*, *20*(7), 121-127. https://doi.org/10.1002/acm2.12665

Varian. (2014). Eclipse Photon and Electron Algorithms Reference Guide.

Vitali Moiseenko, M. L., Sarah Kristensen, Gerald Gelowitz, Eric Berthelet. (2007). Effect of bladder filling on doses to prostate and organs at risk: a treatment planning. *Journal of Applied Clinical Medical Physics* *8*(1), 55-68.

Wang, J., Hu, W., Yang, Z., Chen, X., Wu, Z., Yu, X., Guo, X., Lu, S., Li, K., & Yu, G. (2017). Is it possible for knowledge-based planning to improve intensity modulated radiation therapy plan quality for planners with different planning experiences in left-sided breast cancer patients? *Radiat Oncol*, *12*(1), 85. https://doi.org/10.1186/s13014-017-0822-z

Wang, W., Purdie, T. G., Rahman, M., Marshall, A., Liu, F.-F., & Fyles, A. (2012). Rapid automated treatment planning process to select breast cancer patients for active breathing control to achieve cardiac dose reduction. *International Journal of Radiation Oncology* Biology*Physics*, *82*(1), 386-393.

Wu, Q. J., Li, T., Yuan, L., Yin, F. F., & Lee, W. R. (2013). Single institution's dosimetry and IGRT analysis of prostate SBRT. *Radiation Oncology*, *8*, 215. https://doi.org/10.1186/1748-717X-8-215

Xhaferllari, I., Wong, E., Bzdusek, K., Lock, M., & Chen, J. Z. (2013). Automated IMRT planning with regional optimization using planning scripts. *Journal of applied clinical medical physics*, *14*(1), 176-191.

Yan, H., Yin, F.-F., Guan, H.-q., & Kim, J. H. (2003). AI-guided parameter optimization in inverse treatment planning. *Physics in Medicine Biology*, *48*(21), 3565.

Zhang, X., Li, X., Quan, E. M., Pan, X., & Li, Y. (2011). A methodology for automatic intensity-modulated radiation treatment planning for lung cancer. *Physics in Medicine Biology*, *56*(13), 3873.

Zhu, X., Bourland, J. D., Yuan, Y., Zhuang, T., O'Daniel, J., Thongphiew, D., Wu, Q. J., Das, S. K., Yoo, S., & Yin, F. F. (2009). Tradeoffs of integrating real-time tracking into IGRT for prostate cancer treatment. *Physics in Medicine and Biology*, *54*(17), N393-N401. <Go to ISI>://WOS:000269074500021

PEER-REVIEWED JOURNAL PUBLICATIONS

- Chi, Y., Sprouts, D., Wang, L., Rezaeian, N., Yang, M., Hannan, R., Jia, X., Dosimetric impact of inter-fractional and intra-fractional target motion in high-risk prostate cancer stereotactic body radiation. Under review by Co-Author
- Sprouts, D., Gao, Y., Wang, C., Jia, X., Shen, C., Chi, Y., The Development of a Deep Reinforcement Learning Network for Dose-Volume-Constrained Treatment Planning in Prostate Cancer Intensity Modulated Radiotherapy. *Biomedical Physics & Engineering Express,* Accepted
- Sprouts, D., Chi, Y., The Development of an Actor-Critic Network based Virtual Treatment Planner for Automatic Treatment Planning in Prostate Cancer Intensity Modulated Radiotherapy. Under preparation

PRESENTATIONS

- Oral presentation in 2021 Virtual AAPM annual meeting
- Oral presentation in 2020 Virtual AAPM annual meeting
- Poster presentation in 2019 SWAAPM annual meeting at Grapevine