Automatic radiotherapy treatment planning with deep functional reinforcement learning

Bin Liu[1], Yu Liu[1], Zhiqian Li[1], Jianghong Xiao [*2], and Huazhen Lin[1]

[1]Center of Statistical Research, School of Statistics, Southwestern University of Finance and Economics, Chengdu, China

[2]Radiotherapy Physics & Technology Center, Cancer Center, West China Hospital, Sichuan University, Chengdu, Sichuan, China

June 18, 2024

Abstract

Intensity-modulated radiation therapy (IMRT) is one of the most important modern radiotherapy techniques and is often modeled as an optimization problem. The objective function and constraints consist of multiple clinical requirements designed for different clinical settings. When a tightly constrained optimization problem has no solution,

[*]corresponding author:xiaojh@scu.edu.cn

the planner can empirically relax certain constraint parameters and re-solve the problem until a more satisfactory solution is obtained. This process is time-consuming and laborious. Several inverse planning studies have been devoted to automated radiotherapy planning schemes. Reinforcement learning has been used by many studies to model this process, but they suffer from two important drawbacks: 1) designing a sub-network for each organ, which makes it difficult to extend the model to other patients with a different number of organs. Clinically, it is common for different patients to have inconsistent numbers of organs considered for radiotherapy, even for the same type of cancer; 2) directly feeding low signal-to-noise DVH curves as states into the reinforcement learning network, which ignores its functional characteristics and leads to low training efficiency. In this study, within the framework of deep reinforcement learning, a DVH function-based embedding layer was designed to directly extract the effective information of DVH and allow different organs to share a strategic network. The test results on a dataset of 135 patients with cervical cancer find that our proposed model can be applied to radiotherapy planning in real-world scenarios.

1 Introduction

1.1 Treatment planning problem Statement

Cancer is a leading cause of death worldwide, accounting for approximately 10 million deaths in 2020 (nearly one in six deaths worldwide) [10]. Radiation therapy is one of the most essential treatments for cancer that uses beams of intense energy to control the planning target volume(s) (PTVs). The challenge lies in the presence of organs-at-risk (OARs). That is, the beams

not only kill cancer cells but also affect nearby healthy tissue (OARs). Inverse treatment planning, such as intensity-modulated radiation therapy (IMRT), is a critical way of modern radiation therapy.

1.2 Basic framework of IMRT

In practice, IMRT is often achieved via a challenging optimization process. The objective of the optimization typically contains several dose-level constrains to maximize the delivery of radiation of targets and minimize the hurts for OARs simultaneously. However, the formulation of the optimization problem of IMRT is typically based on a set of empirical parameters, such as relative importance weights and the level of the constraints for PTVs and OARs that aim to satisfy several clinical considerations. These empirical parameters are initialized by human planner but can be adjustable later. For a given parameter, we can solve the corresponding optimization via a modern treatment planning system (TPS). However, the current solution may contradict the clinical considerations. Then the planner has to adjust these parameters manually and let the TPS to re-optimize it. This trial-and-error process will repeat for many times until get a satisfied result. What's more, the quality of the resulting treatment plan depends on planners' experience and skill. Therefore, there is a strong desire to develop automatic methods for high-quality and efficient treatment planning.

1.3 Existing solutions and their shortages

Over the past years, plenty of efforts in automatic planning have been endeavored. A typical solution is to add an outer-loop optimization for parameter searching on top of the inner optimization. Some typical examples of the double-optimization are [12, 4, 11]. Recently, people are beginning to focus

on the prospect of modeling automatic treatment planning with deep learning [1]. The deep neural networks make the parameter searching process more flexible than the traditional methods. More specially, the trial-and-error process of human planner can be well imitated by the technique of deep reinforcement learning. In fact, studies have shown that deep reinforcement learning can be used to adjust the parameters of treatment plans[7, 3, 2]. For example, Shen et al.[8] propose a virtual therapy planner network (VTPN) based on deep reinforcement learning (DRL) to model the behavior of human planners in treatment planning parameters (TPPs) adjustment for prostate cancer. The limitations of the existing methods lie in the following two perspectives, 1) Existing methods design a separate sub-network for each organ [5, 3], which makes it difficult to apply their models to other cases because the number of organs involved is not necessarily the same for each patient, even those suffering from the same disease. 2) these methods ignore the functional characteristics of DVH that accept the original DVH image or DVH curve coordinates as input for the deep models, however, the signal-to-noise ratio of these inputs is very low, which increases the learning difficulty and thus limiting the number of organs they can plan. For example, as claimed in their discussion [7, 5], they only consider the bladder and rectum as OARs, yet clinical settings often have more than a dozen organs that need to be adjusted. Therefore, the generalization of multi-parameter TPPs adjustment on multiple organs has great development prospects.

1.4 Our solution

We present a deep reinforcement learning framework by considering the functional characteristics of DVH, named functional Automatic Treatment (or Automatic Treatment Planner) Parameters Adjustment Network (fATPAN),

trained to manipulate the Treatment Planning System (TPS) and adjust treatment planning parameters to generate high-quality plans. fATPAN can automatically generate adjusted prescription doses based on patient data, replacing human planners to obtain treatment plans that satisfy clinical requirements.

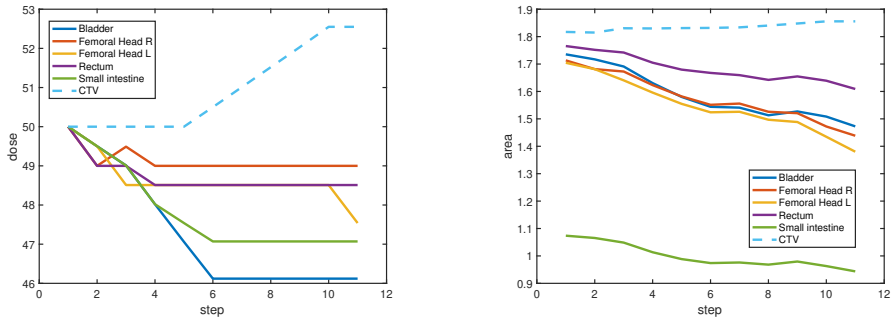
Usually, a patient's situation is highly variable. Even for different patients diagnosed with the same type of cancer, the area to be considered by the doctor for radiation therapy varies, because different patients have individual variables, such as the stage of tumor development and the size of their own body organs. The personality requirements make it difficult to apply a model based on a single organ corresponding to a sub-network to the clinic. To solve the above problem, fATPAN designed an organ-sharing tuned neural network, i.e., the DVH of all organs are input to the same network for training. When testing, the DVH of different organs of the patient can be directly input into the network to get the adjusted action prediction results. That is, we can no longer be used to limit the number of organs in the treatment planning. To further improve learning efficiency, fATPAN uses a functional decomposition layer to learn the feature of the DVH curves first. The signal-to-noise ratio of the input data is greatly improved by the functional mapping layer, which allows for a corresponding increase in model training efficiency and thus can be applied to real scenarios.

2 Results

2.1 Parameters adjusting process for IMRT

Figure 1(a) and 1(b) demonstrate the tuning process, with the clinical target volume (CTV) represented by a dashed line and the five organs-at-risk

(OARs) (Bladder, Femoral Head R/L, Rectum, Small Intestine) represented by solid lines. Initially, the proposed model aims to reduce the dose levels of the five OARs while maintaining the CTV dose at 50 Gy for the first five steps as shown in Figure 1(a). Starting from step 5, fATPAN begins to incrementally increase the dose level of the CTV until convergence at step 10. This alternating adjustment of the dose levels for the CTV and OARs is observed in the tuning process. Figure 1(b) showcases the changes in the area under the dose-volume histogram (DVH) curves (rewards), referred to as DVH AUC, for both the target and the five OARs. The five decreasing solid lines represent the changes in the DVH AUC for the five OARs. It can be observed that the Small Intestine receives the lowest dose level, while the Rectum initially has a DVH AUC of around 1.76, which gradually drops to 1.6 by the end of the iteration, satisfying the prescribed dose for the OARs. The DVH AUC for the target continuously increases, while the DVH AUC for the OARs decreases correspondingly, demonstrating the effectiveness and efficiency of the proposed model.



(a) dose adjustment (b) area of target and OARs

Figure 1: The Process of Treatment Planning.

Figure 2 provides a visual representation of the treatment planning process for a representative patient case using our model. The figure includes

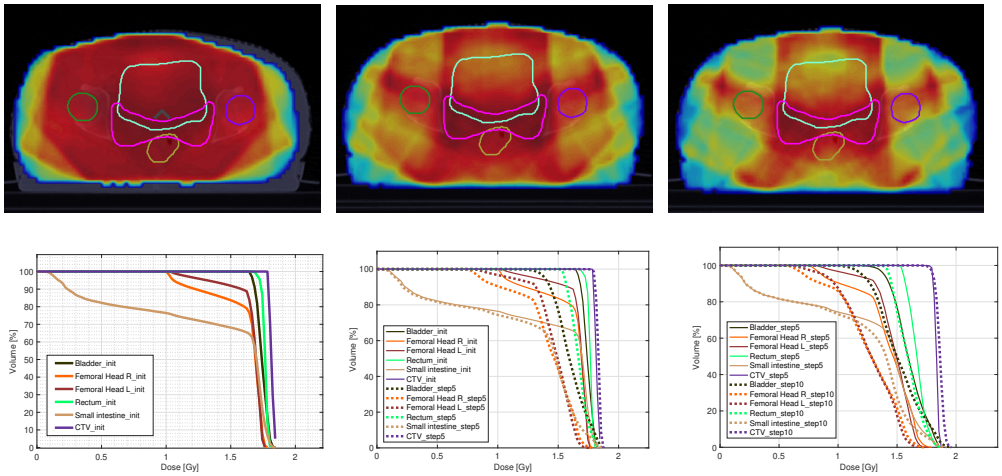


Figure 2: Variations of DVHs and dose distributions

three stages: the initial step, step 5, and step 10. The treatment plans are visualized through dose distribution heat maps and corresponding dose-volume histograms (DVHs). The heat maps display the dose distribution across the delineated areas, including the CTV and five OARs. The quality of the plans can be assessed by evaluating the dose parameter metrics for each structure in the dose distributions. The changes in dose distribution in these areas provide insights into the dynamic planning process.

The initial DVHs are shown in the bottom left panel of Figure 2, represented by solid lines. The corresponding initial dose distribution is displayed in the top left panel, with the segmentation curves delineated by clinicians. This visualization helps in understanding the initial state of the dose distribution. The bottom middle panel represents the results of step 5, and the bottom right panel represents the results of step 10. At step 5, the DVH changes are compared by plotting the initial step results (solid line) alongside the step 5 results (dashed line). It can be observed that the CTV curve was dragged to the right side, indicating an increase in dose level, while the OAR curves were dragged to the left, indicating a decrease in dose level. The

dose distribution map in the middle top panel reflects these changes, with a significant reduction in the dose level of the OAR areas and increased focus on the CTV. At step 10, the dose level gap between the CTV and OARs becomes more pronounced. The top right panel displays the overall dose distribution, which consistently satisfies the desired distribution determined by the clinician's delineation. By visually analyzing the changes in the dose distribution and comparing the DVHs, the effectiveness and progression of the treatment planning process can be evaluated.

Table 1 provides an overview of the key DVH parameters that clinicians are most interested in during the planning process. It includes both target structures, such as the planning target volume (PTV), CTV, and PTV rings, as well as OARs. For the PTV and CTV, the following parameters are evaluated:

- D95: The minimum absorbed dose that covers 95% of the volume of the region of interest.
- Dmean: The mean absorbed dose within the region of interest.
- CI (Conformity Index): A measure of how well the prescribed dose conforms to the target volume, taking into account the dose distribution outside the target.
- HI (Homogeneity Index): A measure of dose homogeneity within the target volume, indicating how evenly the dose is distributed within the target.

The PTV rings, which act as buffer tissues between the target and OARs, are generally evaluated based on their average dose. In this specific case, there are five PTV rings, labeled as Ring1PTV to Ring5PTV in Table 1.

Table 1 summarizes the average DVH parameters for 12 organs, including both target structures and OARs. These parameters are essential for quantitative comparisons and assessing the quality of the treatment plans. For the target structures (PTV and CTV), the D95 (minimum absorbed dose covering 95% of the volume) and Dmean (mean absorbed dose) are evaluated. From the provided information, it is observed that the D95 and Dmean values for the CTV increase from the initial step to step 5 and step 10. For example, the D95 of the CTV increases from 1.79 to 1.81 and 1.83 at step 5 and step 10, respectively. Similarly, the Dmean shows a similar increasing trend. It is noted that the final D95 value of 1.83 already satisfies the clinical standard. The changes in HI for the target structures are reported to be small, while the CI shows a significant improvement from the initial step to step 5.

For the OARs, two different categories are mentioned: serial organs and parallel organs. For serial organs such as the bladder, rectum, and small intestine, the DVH parameters that need to be controlled simultaneously are V30, V40, V45, and Dmean. The other category, parallel organs, includes Femoral Head R and Femoral Head L, where only the Dmean parameter is evaluated. Comparing the results at step 5 and step 10, it is observed that most DVH parameters for the OARs, except for V50 at step 10, are lower than or equal to the values at the initial step and step 5. For example, compared to the results at step 5, the V30 and V40 values of the bladder decreased by 1% and 8%, respectively, while the V30 and V40 values of the small intestine decreased by 1% and 4%, respectively. The mean dose of all OARs at step 10 was reduced compared to the results at step 5. Specifically, compared to the initial mean dose, the mean dose of the bladder, rectum, small intestine, femoral head R/L at step 10 was reduced by 0.177 Gy, 0.101 Gy, 0.129 Gy,

0.182 Gy, and 0.25 Gy, respectively. These quantitative comparisons provide insights into the improvement in treatment plan quality achieved by the proposed model, as evidenced by the changes in DVH parameters for both target structures and OARs.

2.2 Impact of the reward function on the model

According to the definition of the second reward, the reward changes only when the distribution of the high dose in the OAR changes. Therefore, when the target dose density distribution changes, it has a greater impact on the reward. Under the second reward function, the model would pay more attention to the dose variation of the target. In other words, when learning parameter adjustment, the decision network will first consider the target dose to meet the clinical requirements. We show in Figure 3 the patient's probability density function changes in the dosage distribution of each structure and the optimal dose area during the adjustment process. During the adjustment process, the dose distribution of OARs moved to the low dose area and became more uniform, while the dose distribution of the target moved to the high dose area and the dose value was concentrated near the prescription dose.

3 Discussion

The proposed automatic treatment planning framework was developed to mimic the behavior of human planners. It addresses two key challenges in automatic treatment planning. The first one is we broke through the restriction that the number of organs to be planned must be the same for different patients. In radiotherapy planning, even for patients with the same type of

Table 1: Summary of Interested DVH Parameters Changes (with standard deviation) to the CTV, PTV, and OARs.

ROI	parameter	init	Step5	Step10
CTV	D95(Gy)	1.794±0.003	1.81±0.014	1.83±0.028
	HI	0.024±0.003	0.029±0.007	0.03±0.232
	CI	0.104±0.013	0.46±0.007	0.448±0.201
	Dmean(Gy)	1.81±0.006	1.84±0.011	1.85±0.02
PTV	D95(Gy)	1.783±0.001	1.757±0.028	1.752±0.036
	HI	0.172±0.004	0.059±0.019	0.067±0.015
	CI	0.029±0.017	0.503±0.102	0.6±0.209
	Dmean(Gy)	1.807±0.004	1.823±0.011	1.832±0.024
Bladder	V30(%)	1±0	0.991±0.054	0.976±0.065
	V40(%)	1±0	0.915±0.159	0.749±0.16
	V50(%)	0.102±0.025	0.162±0.001	0.175±0.017
	Dmean(Gy)	1.744±0.177	1.631±0.056	1.567±0.078
Rectum	V30(%)	0.995±0.201	0.995±0.208	0.994±0.203
	V40(%)	0.991±0.202	0.989±0.222	0.932±0.172
	V50(%)	0.075±0.01	0.155±0.023	0.163±0.028
	Dmean(Gy)	1.759±0.346	1.7±0.342	1.658±0.306
Small intestine	V30(%)	0.703±0.201	0.687±0.208	0.675±0.203
	V40(%)	0.63±0.202	0.59±0.222	0.451±0.172
	V50(%)	0.021±0.01	0.027±0.023	0.033±0.028
	Dmean(Gy)	1.255±0.346	1.181±0.342	1.126±0.306
Femoral Head R	Dmean(Gy)	1.645±0.039	1.512±0.056	1.463±0.078
Femoral Head L	Dmean(Gy)	1.666±0.027	1.498±0.07	1.416±0.089
Ring1PTV	Dmean(Gy)	1.786±0.003	1.763±0.024	1.752±0.033
Ring2PTV	Dmean(Gy)	1.773±0.005	1.727±0.031	1.703±0.039
Ring3PTV	Dmean(Gy)	1.754±0.007	1.676±0.033	1.632±0.037
Ring4PTV	Dmean(Gy)	1.736±0.01	1.634±0.025	1.572±0.022
Ring5PTV	Dmean(Gy)	1.724±0.012	1.614±0.021	1.554±0.027

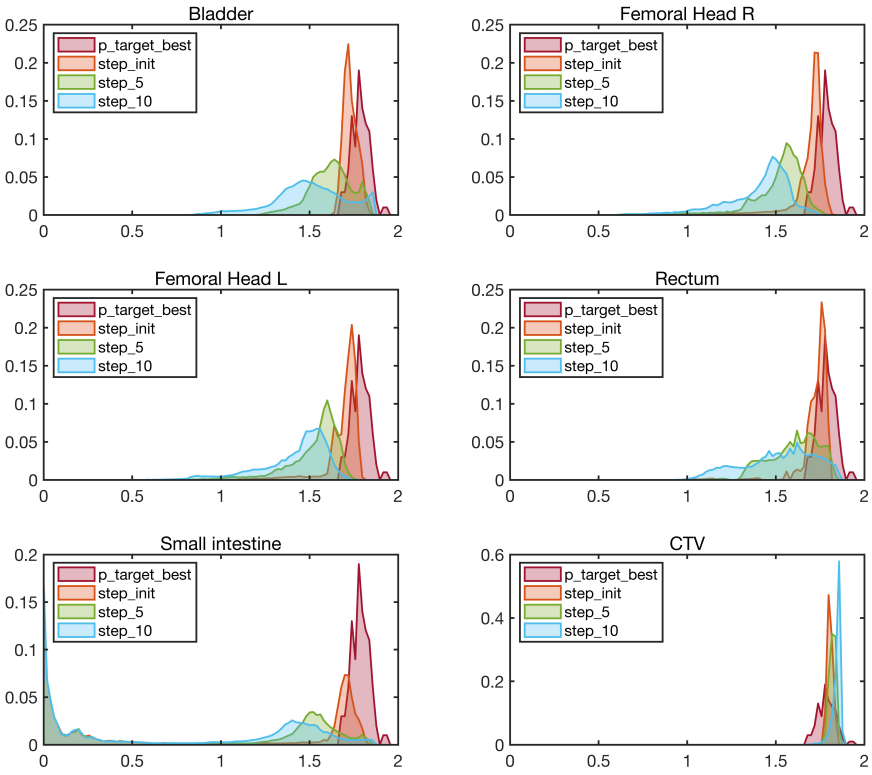


Figure 3: Variations dose distribution

cancer, the areas (organs) of radiotherapy to be considered will be different. However, traditional DRL-based models design an exclusive sub-network for each organ to be planned, which implies the model to be consistent with the training patient in the test patient, severely reducing the scalability of the model [9, 5, 3]. To improve the model scalability, we only use one network that is shared by all organs in fATPAN. During training, the functional features of DVH of different organs are fed into the network sequentially, and the rewards of each organ are accumulated in one round of iterations before the gradient is passed back. In this way, the proposed model can be applied to real clinical scenarios where the number of organs planned for radiotherapy need not be the same for different patients. That is, personal planning protocols can be implemented for different patients.

The proposed automatic treatment planning framework was developed to mimic the behavior of human planners. It addresses two key challenges in automatic treatment planning. First, the ability to handle varying numbers of organs in treatment planning is indeed an important challenge addressed by the proposed method. Traditional methods often assume a fixed number of organs and design separate sub-networks for each organ, making it difficult to generalize the model to patients with different organ configurations [9, 5, 3]. In contrast, the fATPAN framework overcomes this limitation by using an organ-sharing tuned neural network. It takes into account the DVH of different organs sequentially and accumulates the rewards of each organ within one round of iterations before the gradient is propagated back. This approach enables the model to be more adaptable and scalable, accommodating patients with varying organ structures and facilitating personalized treatment planning.

Second, we design a functional embedding layer to extract the input DVH

curve features, which is a valuable contribution to improving the efficiency and effectiveness of the training process. Traditional approaches often treat DVH curves as raw data and process them directly using convolutional operation [3, 5, 6]. While this approach may capture certain spatial patterns in the DVH curves, it overlooks the inherent functional characteristics of the DVH data. By incorporating a functional embedding layer, fATPAN takes advantage of the functional nature of DVH curves, enabling more effective feature extraction. This layer helps enhance the signal-to-noise ratio of the processed features, leading to improved learning efficiency and more accurate representations of the DVH data.

The functional data embedding layer in fATPAN facilitates efficient computation for complex tumor patients by capturing the functional aspects of the DVH curves, enabling the model to better understand the dose distribution patterns and make more informed decisions during the treatment planning process. This contributes to the overall effectiveness and efficiency of the automatic treatment planning framework.

Currently, automated radiotherapy planning algorithms proposed in the existing literature are far from clinical needs and are generally considered to be solved under a simplified problem only. For example, some of the existing methods require additional human intervention to narrow the state-action space [5]. And in [5], the authors only evaluate the case of one PTV and two OARs. This is because as the number of organs to be considered increases, the process of solving the optimization problem becomes more complex or even infeasible. In contrast, our algorithm makes the computational process efficient due to the two aforementioned strategies, thus simplifying the solution process. For example, the results demonstrate in Figure 1,2 and Table 1 are the treatment planning for cervical carcinoma patients which involve

12 areas. In the example of Figure 2, the results of 12 organs are actually processed, and only the results of the most critical 6 of them are kept in order to facilitate the presentation of the DVH curves. In Table 1, we give the results of a patient with 12 organs to be considered during treatment planning.

There are several limitations of our work. First, we know that radiotherapy planners are able to make planning across tumor categories for patients with different cancers. Although we have designed the organ-sharing TPPs network for patients from the same kind of cancer but with different numbers of organs, it is poorly trained in cases with different tumors. Second, the computational speed of our system is not fully satisfactory, taking 3 to 4 hours for a single test, which is close to the time of a human planner. Hence, it fails to reflect the advantages of the machine. The most time-consuming computational module is the interaction with the environment. During the interaction, a new optimization problem (from Pytorch) needs to be formed based on the action selection results predicted by DRL. Our model passes the constraint parameters of this new optimization problem to matRad (a MatLab-based TPS used in the environment) to perform the next step of the optimization solution.

4 Methods

4.1 Overall Framework

Figure 4 illustrates the overall framework. The proposed method contains three important modules, a functional embedding layer, a virtual treatment planner neural network (VTPNN), and an environment for DRL.

In clinical radiotherapy, radiotherapy planners often have to manually

adjust the dosage and weight to obtain a high-quality treatment plan. This trial-and-error process is time-consuming and labor-intensive. Therefore, automatic radiotherapy regimens with automatic dose and weight adjustments are needed by planners. We propose a deep reinforcement learning network, a functional Automatic Treatment Parameters Adjustment Network (fATPAN), trained to learn adjusting treatment parameters via an end-to-end algorithm called Actor to Critic(A2C) as shown in Figure 4. The trained network can automatically generate adjusted prescription doses based on patient data, replacing human planners to obtain high-quality treatment plans.

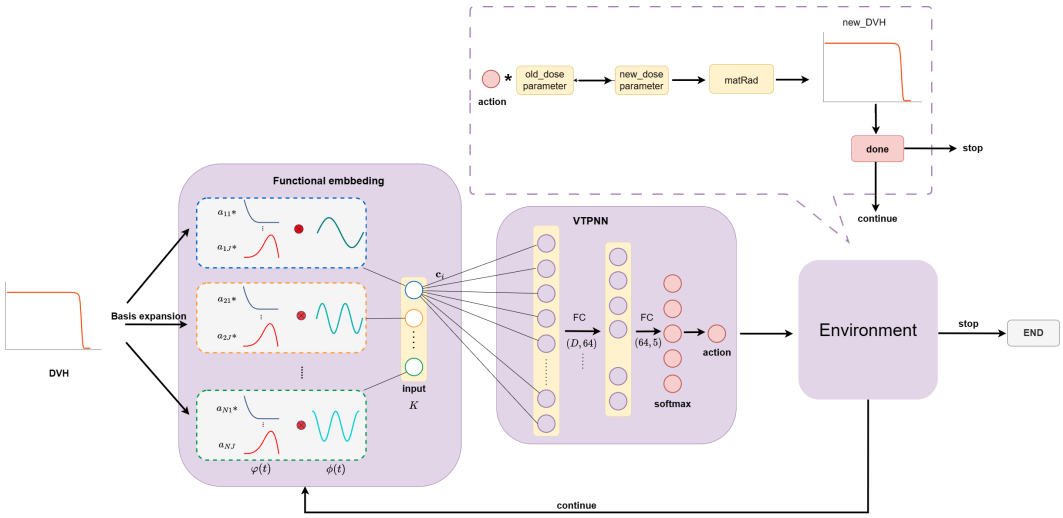


Figure 4: The overall framework of the proposed fATPAN.

4.2 Functional embedding layer

4.2.1 Dose-volume histogram

Dose-volume histograms (DVH) are a straightforward way to represent the dose distribution of each structure in the treatment area. Figure 5 shows a typical example of DVH for an OAR and a target in a treatment plan.

The horizontal axis of DVH represents the dose level, and the vertical axis represents a fraction of volume. For example, the point P of the target DVH in Figure 5 suggests that at least 60% volume of the target voxels receiving 1.8Gy or less dose level. Ideally, the DVH in the target area is vertical at the prescribed dose, indicating that all PTV voxels receive the prescribed dose, while the OAR is vertical at a relative dose of 0, indicating that the received dose is 0.

Oncologists are often willing to sacrifice some portion of an OAR close to the target area to achieve adequate tumor control probability. Therefore, OAR must be required to have at least a certain percentage of the dose below the specified level.

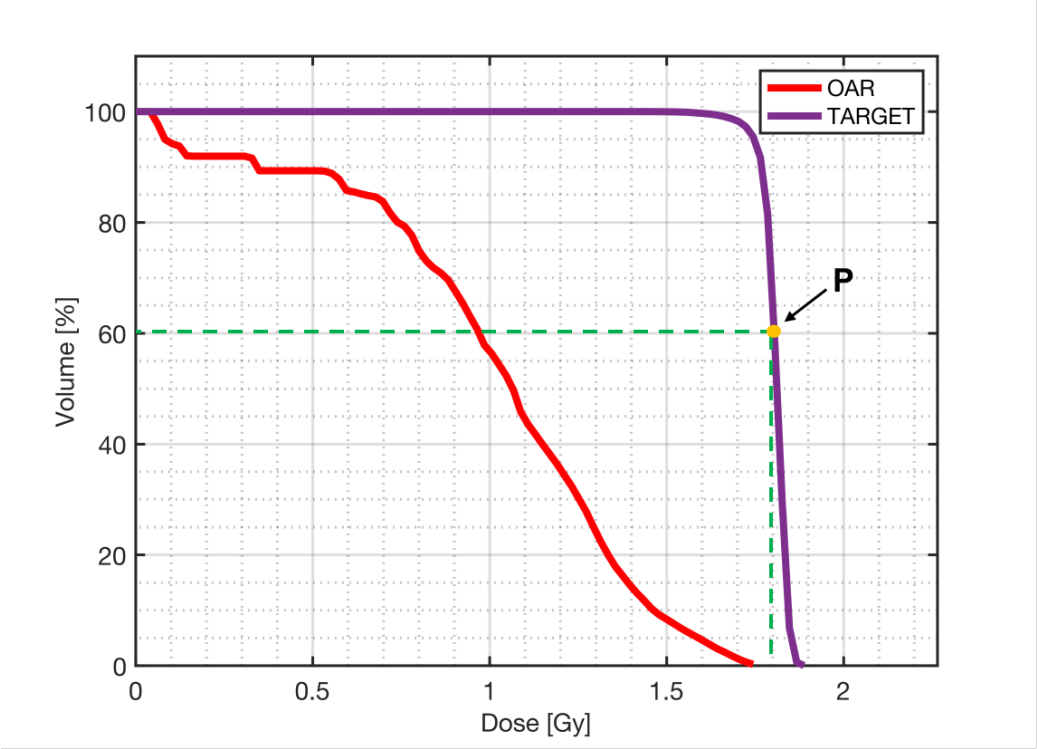


Figure 5: Solution quality is typically assessed by DVH

Figure 5 illustrates that the DVH observations $\mathbf{x}(t) \in [0, 1]$ are recorded

as a functional relationship between the fraction of volume $\mathbf{x}(t)$ and dose level t [Gy](or relative dose level).

4.2.2 Functional Embedding of DVH

Rather than feed the DRL model with DVH directly, we propose representing it into a finite-dimensional vector via a functional neural network first. We think the embedding vector is more appropriate to adapt to regular deep neural networks subsequently. In particular, for a DVH input $\mathbf{x}(t)$, its corresponding functional embedding $\mathbf{h} = (h_1, h_2, ..., h_D) \in \mathbb{R}^D$ can be calculated as follows,

$$h_i = \sigma\left(\int_{\mathcal{T}} \mathbf{w}_i(t)\mathbf{x}(t)dt + b_i\right) \tag{1}$$

where $i = 1, 2, ..., D$, D is the hidden dimension, and $\mathbf{w}_i(t)$ are functional weights. We further decompose $\mathbf{w}_i(t)$ with its basis representation, that is,

$$\mathbf{w}_i(t) = \sum_{k=1}^{K} c_{ik}\boldsymbol{\phi}_{ik}(t) \tag{2}$$

where $\boldsymbol{\phi}_{ik}(t)$ is the basis function of the functional weights $\mathbf{w}_i(t)$, which could be a Fourier basis or spline basis. K is the number of basis functions we choose. We select $\boldsymbol{\phi}_k(t)$ as Fourier basis,

$$\begin{aligned} \phi_0(t) &= \frac{\sqrt{2}}{\sqrt{2}} \\ \phi_{2k-1}(t) &= \frac{\sin(\frac{2\pi k}{T}t)}{\sqrt{\frac{T}{2}}} \\ \phi_{2k}(t) &= \frac{\cos(\frac{2\pi k}{T}t)}{\sqrt{\frac{T}{2}}} \end{aligned} \tag{3}$$

and set $K = 5$ as shown in Figure 6.

Therefore, we have

$$\begin{aligned} h_i &= \sigma\left(\int_{\mathcal{T}} \sum_{k=1}^{K} c_{ik}\boldsymbol{\phi}_{ik}(t)\mathbf{x}(t)dt + b_i\right) \\ &= \sigma\left(\sum_{k=1}^{K} c_{ik}\int_{\mathcal{T}} \boldsymbol{\phi}_{ik}(t)\mathbf{x}(t)dt + b_i\right) \\ &= \sigma\left(\mathbf{c}_i^{\mathrm{T}}\boldsymbol{\phi}_i + b_i\right) \end{aligned} \tag{4}$$

where

$$\boldsymbol{\phi}_i = \begin{bmatrix} \int_\tau \phi_{i1}(t)\mathbf{x}(t)dt \\ \int_\tau \phi_{i2}(t)\mathbf{x}(t)dt \\ \vdots \\ \int_\tau \phi_{iK}(t)\mathbf{x}(t)dt \end{bmatrix}$$

Our problem usually involves multiple organs, that is there are input DVHs. Suppose there are N organs required to adjust. For each DVH $\mathbf{x}_n(t), n = 1, 2..., N$, we perform a Fourier basis expansion, i.e., project $\mathbf{x}_n(t)$ onto several Fourier bases. That is, express each DVH as a linear combination of J Fourier bases as follows,

$$\begin{cases} \mathbf{x}_1(t) = \sum_{j=1}^{J} a_{1j}\varphi_j(t) \\ \mathbf{x}_2(t) = \sum_{j=1}^{J} a_{2j}\varphi_j(t) \\ \quad \vdots \\ \mathbf{x}_N(t) = \sum_{j=1}^{J} a_{Nj}\varphi_j(t) \end{cases}$$

where J is the number of the B-spline basis, we set $J = 35$ as shown in Figure 7, $a_{nj}, n = 1, 2, \cdots, N, j = 1, 2, ..., J$ is the corresponding coefficients of $\mathbf{x}_n$ with respect to the basis $\varphi_j(t)$. $\varphi_j(t)$ could be a Fourier basis or spline basis.

In practice, the integral $\int_\tau \phi_k(t)\mathbf{x}(t)dt, k = 1, 2, ..., K$ in Equation (4) can be approximated with numerical integration methods such as the composite

Simpson's rule.

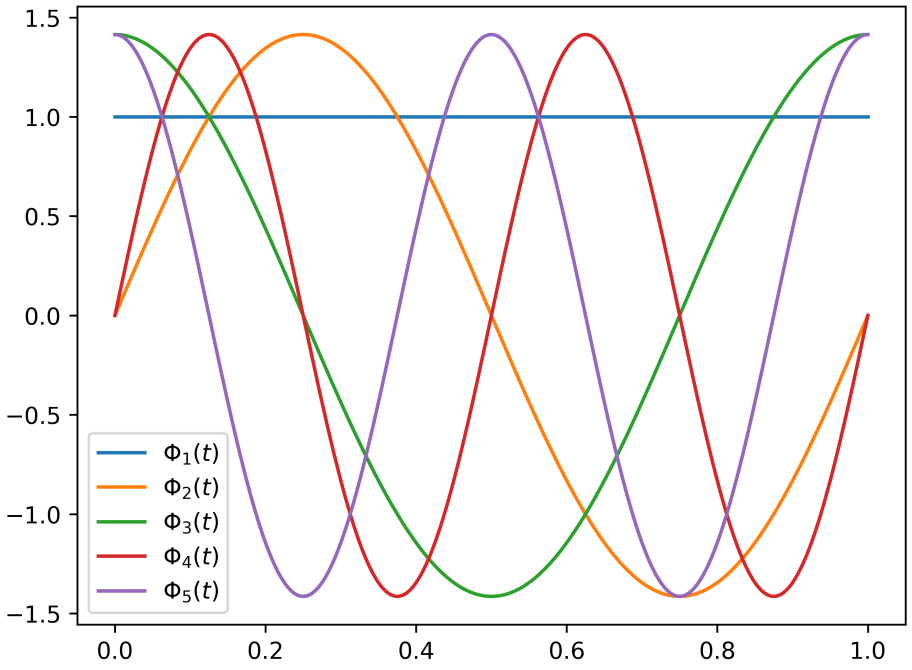


Figure 6: The 5 Fourier basis $\phi(t)$.

4.3 Treatment Planner Network

The virtual treatment planner network contains two parts, a three-layer Full-Connected layer and a normal random layer. The numbers of hidden neurons of the first part are as shown in Figure 4. The second part is Gaussian random layer. We assume that each TPP action follows a Gaussian distribution $a_i \sim \mathcal{N}(\mu, \sigma^2)$, and the parameters μ and σ^2 depend on the output feature $\mathbf{z} \in \mathbb{R}^d$ by the first part. One solution could be,

$$\mu(\mathbf{h}) = \mathbf{c}^\top \mathbf{h} + d \tag{5}$$

$$\log \sigma^2(\mathbf{h}) = \mathbf{e}^\top \mathbf{h} + f \tag{6}$$

where **h** is learned by two single-layer perception using feature **z**. After getting the distribution of actions, we can randomly sample a value from the distribution as the next adjustment action.The purpose of this is to make the output of the model more consistent with the actual process of dose adjustment by physicists by introducing continuous actions.

4.4 Environment

We built an interactive environment for reinforcement learning. The overall framework of the environment is as shown in the Top panel of Figure 4. The input of the environment module is the adjustment action of the constraint parameters that output from the previous virtual treatment planner network module. The resulting new constraint yields an updated optimization problem. Then the optimization engine computes a corresponding new DVH by solving this new optimization problem. We can compare the new DVH with the old DVH in the previous step to calculate the reward. The details of reward computation will be introduced in section 4.4.2.

4.4.1 Actor and critic (A2C) network

We implement the deep reinforcement learning in the framework of A2C network. The most important concept of A2C network is the state-value function, which is an expectation of the action-value $Q_\pi(s_t, a_t)$ with respect to action a_t,

$$V_\pi(s_t) = \mathbb{E}_A[Q_\pi(s_t, A)] = \sum_a \pi(a|s_t) Q_\pi(s_t, a), \tag{7}$$

where $A \sim \pi(\cdot|s_t)$, for a given policy function $\pi(\cdot|s_t)$, $V_\pi(s_t)$ evaluate the current state good or bad. In our case, the state s_t is the current DVH of each organ, and the action a_t is dose adjustment options, that is, **increase**

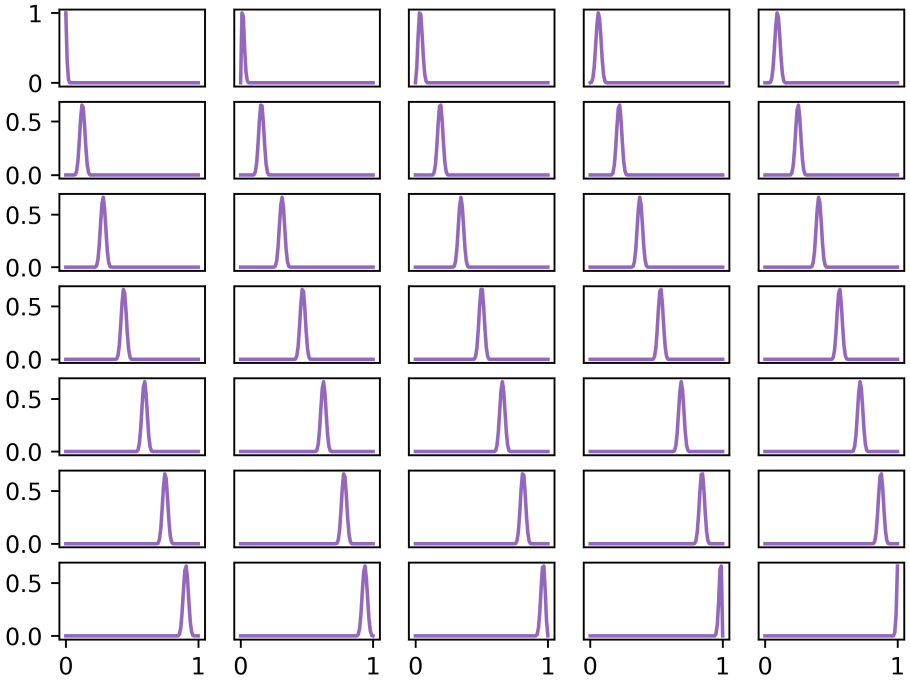


Figure 7: The 35 B-spline basis $\varphi(t)$.

or decrease its value by 2%, increase or decrease by 1%, and keep it unchanged. The expectation in Eq. (7) is in summation form rather than integral form because the action is discrete in our case.

The state space S (DVH) is huge, hence the Actor-Critic method tries to approximate the policy function $\pi(a|s)$ (actor) with a policy neural network $\pi(a|s_t;\Theta)$, where Θ is the parameter of the neural network. That is,

$$V_\pi(s_t;\Theta) = \sum_a \pi(a|s_t;\Theta) Q_\pi(s_t,a)$$

Therefore, we have a policy based training object to maximize

$$J(\Theta) = \mathbb{E}_S[V_\pi(S)] \tag{8}$$

Obviously, we can get the optimal via policy gradient ascent,

$$\begin{aligned}
\nabla_w J(\Theta) &= \frac{\partial J(\Theta)}{\partial \Theta} \\
&= \mathbb{E}_S[\sum_a \frac{\partial \pi(a|s;\Theta)}{\partial \Theta} Q_\pi(s,a)] \\
&= \mathbb{E}_S[\sum_a \pi(a|s;\Theta) \frac{\partial \log \pi(a|s;\Theta)}{\partial \Theta} Q_\pi(s,a)] \\
&= \mathbb{E}_S[\mathbb{E}_A[\frac{\partial \log \pi(a|s;\Theta)}{\partial \Theta} Q_\pi(s,a)]]
\end{aligned} \tag{9}$$

In practice, we use the following process to approximate the policy gradient. For an observed state s_t, we first randomly sample action a_t according to $\pi(\cdot|s_t;\Theta_t)$, then compute $Q_\pi(s_t,a_t)$, then the an approximate policy gradient is,

$$g(a_t,s_t;\Theta) \triangleq Q_\pi(s_t,a_t) \cdot \nabla_\Theta \log \pi(a_t|s_t;\Theta) \tag{10}$$

we can update the policy network with $g(a_t,s_t;\Theta)$ with gradient ascend.

As we stated before, we use a policy network $\pi(a|s_t;\Theta)$ to approximate the real policy function $\pi(a|s_t)$ (actor). Similarly, we approximate the value

function $Q_\pi(s,a)$ with a value network $q(s,a;\Phi)$ (critic), that is

$$V_\pi(s_t) = \sum_a \pi(a|s_t) Q_\pi(s_t,a) \approx \sum_a \pi(a|s,\Theta) q(s,a;\Phi)$$

where Φ is the parameter of the value network.

We update the actor-network $\pi(a|s,\Theta)$ using the policy gradient as introduced before and update the critic-network $q(s,a;\Phi)$ using the time difference (TD) algorithm as follow.

We first calculate $q(s_t,a_t;\Phi_t)$ and $q(s_{t+1},a_{t+1};\Phi_t)$ first, then calculate the TD target

$$y_t = r_t + \gamma \cdot q(s_{t+1},a_{t+1};\Phi_t) \tag{11}$$

where γ is a parameter. The corresponding loss function is the square error between $q(s_t,a_t;\Phi_t)$ and y_t,

$$L(w) \triangleq \frac{1}{2}(q(s_t,a_t;\Phi_t) - \hat{y}_t)^2 \tag{12}$$

That is, the critic-network $q(s,a;\Phi)$ can be updated with gradient descent algorithm.

We set training stop conditions based on clinical experience. For example, the algorithm stops updating when the D95 in the target area exceeds its prescribed dose and when the V50 of the OAR is higher than its prescribed dose.

4.4.2 Reward function

Based on the updated DVH, we can calculate the reward for the current iteration. The learning goal of the proposed model is to have the Target dose as high as possible and OAR dose as low as possible. From this perspective, we can consider the area change between the DVH curves of the targets and OARs as part of the reward. In addition, D95 on the DVH curve is an

important indicator on the clinician's attention, so we use the change in D95 as another source of reward. The final reward is,

$$Reward_{area} = \Delta D_{95} + \Delta(A_{Target} - A_{OARs}). \tag{13}$$

4.4.3 Optimization (matRad)

For IMRT, the planning process can be formulated as solving a set of optimization problems. This problem usually contains certain well-defined objective functions and constraints. In this work, we choose a overall optimization objective as a weighted summation of some individual components $f_n(w)$ as shown in Eq 14.

$$\begin{aligned} \min_{w \in \mathbb{R}^B} \quad & f(w) = \sum_n p_n f_n(w) \\ \text{subject to} \quad & c_l^k \leq c_k(w) \leq c, 0 > w \end{aligned} \tag{14}$$

the c_l^k and c_u^k in the constraints indicate lower and upper bounds on the k-th constraint function $c_k(w)$. The positivity constraint $0 \leq w$ ensures that only positive radiation fluences are considered.

In this work, the individual objective $f_n(w)$ and the corresponding constraints are chosen as follows,

$$f_{\min} DVH = \frac{1}{N_s} \sum\nolimits_{i \in s} \Theta(d_i - \hat{d}) \Theta(d_i - \hat{d}) \Theta(d_i - \hat{d})^2 \tag{15}$$

$$f_{max} DVH = \frac{1}{N_s} \sum\nolimits_{i \in s} \Theta(\hat{d} - d_i) \Theta(\hat{d} - d_i) \Theta(\hat{d} - d_i)^2 \tag{16}$$

$$c_{min} DVH = \frac{1}{N_s} \sum\nolimits_{i \in s} \Theta(d_i - \hat{d}) \tag{17}$$

$$c_{max}DVH = \frac{1}{N_s}\sum\nolimits_{i\in s}\Theta(\hat{d} - d_i) \tag{18}$$

Here,N_s is the number of voxels, d_i represents the dose in voxel i,and $\hat{d}$ is the prescribed dose that we need to fine-tune.$\Theta(x)$ is Heaviside function.
We used the Matrad toolkit to solve the above optimization problem.matRad is an open source, cross-platform radiotherapy planning toolkit written by MATLAB that enables automatic radiotherapy planning based on dose distribution. It adopts an algorithm based on flow mapping and uses interior point method to optimize (IPOPT).

matRad is an open-source software for radiation treatment planning of intensity-modulated photon, proton, and carbon ion therapy. It is entirely written in MATLAB. matRad is used to accurately simulate the impact of radiation on patient tissue, including the entire treatment planning workflow from setting treatment parameters and optimizing the plan to visualizing and evaluating the results, and finally generating an accurate radiotherapy plan. An independent modified treatment planning system (TPS) based on the radiation treatment planning toolkit matRad was used to evaluate our method. Patients' CT and structure files were imported to matRad. Then matRad generates the beam geometry based on the specified treatment plan parameters and computes the dose matrix for each radioactive source element to the target area and surrounding normal tissue. Finally, matRad optimizes the radiation dose according to the defined clinical objectives and constraints to find the optimal dose distribution.

4.5 Datasets

Five cervical cancer patients treated with IMRT in our institution were selected for this study. The clinical target volume(CTV) and OARs include

the bladder, rectum, bilateral femoral heads, and small intestine. The prescription was 50.40Gy(1.8Gy per fraction). All IMRT plans were optimized in MatRad.Equally spaced 5 coplanar photon beams were employed and optimized with IPOPT for all IMRT plans.

References

[1] J. Fan, J. Wang, Z. Chen, C. Hu, Z. Zhang, and W. Hu. Automatic treatment planning based on three-dimensional dose distribution predicted from deep learning technique. *Medical physics*, 46(1):370–381, 2019.

[2] W. T. Hrinivich and J. Lee. Artificial intelligence-based radiotherapy machine parameter optimization using reinforcement learning. *Medical physics*, 47(12):6140–6150, 2020.

[3] Y. Liu, C. Shen, T. Wang, J. Zhang, X. Yang, T. Liu, S. Kahn, H.-K. Shu, and Z. Tian. Automatic inverse treatment planning of gamma knife radiosurgery via deep reinforcement learning. *Medical Physics*, 49(5):2877–2889, 2022.

[4] R. Lu, R. J. Radke, L. Happersett, J. Yang, C.-S. Chui, E. Yorke, and A. Jackson. Reduced-order parameter optimization for simplifying prostate IMRT planning. *Physics in Medicine & Biology*, 52(3):849–870, jan 2007.

[5] C. Shen, L. Chen, Y. Gonzalez, and X. Jia. Improving efficiency of training a virtual treatment planner network via knowledge-guided deep reinforcement learning for intelligent automatic treatment planning of radiotherapy. *Medical Physics*, 48(4):1909–1920, 2021.

[6] C. Shen, L. Chen, and X. Jia. A hierarchical deep reinforcement learning framework for intelligent automatic treatment planning of prostate cancer intensity modulated radiation therapy. *Physics in Medicine & Biology*, 66(13):134002, jun 2021.

[7] C. Shen, Y. Gonzalez, P. Klages, N. Qin, H. Jung, L. Chen, D. Nguyen, S. B. Jiang, and X. Jia. Intelligent inverse treatment planning via deep reinforcement learning, a proof-of-principle study in high dose-rate brachytherapy for cervical cancer. *Physics in Medicine & Biology*, 64(11):115013, 2019.

[8] C. Shen, D. Nguyen, L. Chen, Y. Gonzalez, R. McBeth, N. Qin, S. B. Jiang, and X. Jia. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. *Medical Physics*, 47(6):2329–2336, 2020.

[9] C. Shen, D. Nguyen, L. Chen, Y. Gonzalez, R. McBeth, N. Qin, S. B. Jiang, and X. Jia. Operating a treatment planning system using a deep-reinforcement learning-based virtual treatment planner for prostate cancer intensity-modulated radiation therapy treatment planning. *Medical physics*, 47(6):2329–2336, 2020.

[10] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray. Global cancer statistics 2020: Globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: a cancer journal for clinicians*, 71(3):209–249, 2021.

[11] H. Wang, P. Dong, H. Liu, and L. Xing. Development of an autonomous treatment planning strategy for radiation therapy with effective use of population-based prior data. *Medical Physics*, 44(2):389–396, 2017.

[12] L. Xing, J. G. Li, S. Donaldson, Q. T. Le, and A. L. Boyer. Optimization of importance factors in inverse planning. *Physics in Medicine & Biology*, 44(10):2525–2536, sep 1999.