A Reinforcement Learning Approach for Finding Optimal Policy of Adaptive Radiation Therapy Considering Uncertain Tumor Biological Response

Saba Ebrahimi, Gino J. Lim*

Department of Industrial Engineering, University of Houston, 4800 Calhoun Road, Houston, TX 77204,

Abstract

Recent studies have shown that a tumor's biological response to radiation varies over time and has a dynamic nature. Dynamic biological features of tumor cells underscore the importance of using fractionation and adapting the treatment plan to tumor volume changes in radiation therapy treatment. Adaptive radiation therapy (ART) is an iterative process to adjust the dose of radiation in response to potential changes during the treatment. One of the key challenges in ART is how to determine the optimal timing of adaptations corresponding to tumor response to radiation. This paper aims to develop an automated treatment planning framework incorporating the biological uncertainties to find the optimal adaptation points to achieve a more effective treatment plan. First, a dynamic tumor-response model is proposed to predict weekly tumor volume regression during the period of radiation therapy treatment based on biological factors. Second, a Reinforcement Learning (RL) framework is developed to find the optimal adaptation points for ART considering the uncertainty in biological factors with the goal of achieving maximum final tumor control while minimizing or maintaining the toxicity level of the organs at risk (OARs) per the decision-maker's preference. Third, a beamlet intensity optimization model is solved using the predicted tumor volume at each adaptation point. The performance of the proposed RT treatment planning framework is tested using a clinical non-small cell lung cancer (NSCLC) case. The results are compared with the conventional fractionation schedule (i.e, equal dose fractionation) as a reference plan. The results show that the proposed approach performed well in achieving a robust optimal ART treatment plan under high uncertainty in the biological parameters. The ART plan outperformed the reference plan by increasing the mean biological effective dose (BED) value of the tumor by 2.01%, while maintaining the OAR BED within +0.5% and reducing the variability, in terms of the interquartile range (IQR) of tumor BED, by 25%.

Keywords: Reinforcement Learning, Radiotherapy, Biological Tumor Response, Adaptive radiation therapy.

1. Introduction

Cancer is one of the primary health problems in the world, and it is the second leading cause of death in the United States (Siegel et al., 2019). Radiation therapy (RT) is a common treatment modality for cancer

*Corresponding author: Gino J. Lim

Email addresses: sebrahimi@uh.edu (Saba Ebrahimi), ginolim@uh.edu (Gino J. Lim)

patients. External beam radiotherapy is one of the most commonly used types of RT, in which ionizing radiation (e.g., photon, proton, etc.) goes through a particular part of a patient's body to eradicate tumor cells (Lim et al., 2020). Unfortunately, it also damages healthy organs around the tumor called organs at risk (OARs). A high dose of radiation is required to control tumor cell growth while minimizing the radiation exposure to the OARs. The prescribed RT dose is usually delivered in multiple fractions to achieve tumor control while enabling OARs cells repair. In conventional RT, the patient receives an equal amount of radiation dose in each fraction according to the treatment plan which was developed based on the computed tomography (CT) images (Zaghian et al., 2017).

Recent studies showed that both tumor and OARs cells change dynamically during treatment, and their biological responses to radiation also change over time. Yet, the conventional plans mostly do not fully consider the dynamic nature of biological processes (Bai et al., 2020). Healthy cell repair, reoxygenation and repopulation of tumor cells, and radio-sensitivity are important biological factors in controlling the efficiency of fractionated RT (Hall and Giaccia, 2006). These factors play a significant role in the tumor and/or OAR response to RT treatment (Withers, 1975; Steel et al., 1989). Several studies have shown the possibility of achieving better treatment outcomes by modifying the amount of dose per fraction in fractionated RT based on the tumor's biological response. Information from biological images have been used in several studies to find a biologically conformal nonuniform dose (Lawrence et al., 2008; South et al., 2008; Kim et al., 2012).

Most researchers have considered biological factors in modeling tumor response to radiation during the treatment. The linear-quadratic (LQ) model (Douglas and Fowler, 2012; Fowler, 1989) is one of the common radiation response models in fractionated RT. More comprehensive tumor response models have also been developed considering different biological factors, such as tumor repopulation (Fowler, 2001; Saberian et al., 2016; Bortfeld et al., 2015). Other factors in RT planning include redistribution, repair of sublethal damage, and reoxygenation (Brenner et al., 1995; Yang and Xing, 2005). Furthermore, Jeong et al. (2013) a tumor response model to assimilate hypoxia and proliferation interplay. Their model considers three cell compartments (i.e., proliferating, intermediate, and hypoxic) according to different levels of available oxygen and glucose. Similarly, OAR repair and radio-sensitivity should also be considered to measure the OAR's biological response to radiation in the RT treatment planning (Khaled and Held, 2012).

The biological response to radiation varies from one patient to another (Bibault et al., 2013). Also, radio-sensitivity and tumor proliferation are associated with the tumor cell cycle phase and gene-level activities (Scott et al., 2017). Therefore, personalized RT treatment planning has attracted researcher's attention in biological RT planning. Various models have been proposed to incorporate the biological response in RT treatment planning. Dynamic programming has been used in several studies to account for dynamic tumor response during RT and finding tumor fractionation dose (Wein et al., 2000), OAR repair (Kim et al., 2009), tumor repopulation (Bortfeld et al., 2015; Ghate, 2011) and tumor shrinkage (Unkelbach et al., 2014). However, a comprehensive model incorporating important biological factors in a dynamic treatment framework has not been well studied.

The literature shows that the adaptive radiation therapy (ART) method improves treatment quality in terms of normal-tissue sparing and tumor cell reduction (van de Schoot et al., 2017; Ramella et al., 2017;

Belfatto et al., 2016; Zhu et al., 2011) as well as treatment cost and time (Dial et al., 2016; Veresezan et al., 2017). An ideal approach to consider dynamic tumor changes will be to take images of the patient at every visit, update tumor contours, and revise the treatment plan if a significant change was observed in the tumor geometry. However, daily imaging may not be useful in practice because any changes over the span of a day may not be significant enough to modify the existing treatment plan. More importantly, imaging the patient at every visit during the treatment period can be costly, time-consuming, and prone to human errors. Therefore, a trade-off must be made considering tumor geometry change, costs, timing, and the recommended number of adaptive plans. This is the primary motivation for finding the optimal timing for adaptation to improve the clinical feasibility of ART.

Several approaches have been proposed to optimally determine the frequency of adaptation during the treatment based on the latest tumor geometry information, focusing on target-volume reduction (Saka et al., 2011; Guckenberger et al., 2011; Belfatto et al., 2016) and the amount of dose per volume received in the tumor (Zheng et al., 2015; Lee et al., 2014; Berkovic et al., 2015; Zarepisheh et al., 2014). Most studies suggest that the optimal time for adaptation is when an adequate target volume reduction is observed and maintained (Saka et al., 2011; Guckenberger et al., 2011; Veresezan et al., 2017). However, there are some conflicting reports regarding the occurrence time of the largest tumor volume reduction and the best time to adapt the plan to the tumor volume changes for different patients with different biological response characteristics. Therefore, treatment should be customized for each patient.

Recent studies have proposed machine learning (ML) techniques to predict radiation therapy outcomes (Kawata et al., 2017; Zhang et al., 2015), identify patients who would benefit from ART (Surucu et al., 2016), and determine the ideal time for adaptation in ART (Berkovic et al., 2015; Guidi et al., 2016). ML techniques can help identify patients who will have high tumor volume reduction during RT and select them for ART by predicting the tumor regression during the course of treatment. Reinforcement learning (RL) is a machine learning algorithm that features modeling of sequential data based on the interactions between an agent and an environment. RL can be a good alternative to dynamic programming when there is a high level of uncertainty in sequential decision-making problems in different fields such as finance (Ashrafi and Thiele, 2021), healthcare (Ling et al., 2017), and robotics (Kober et al., 2013) because it can generate a robust and risk-averse solution by incorporating the uncertainty in its environment. Deep reinforcement learning (DRL) algorithms have been applied to find the best policy (a sequence of decisions) in many diverse fields such as robotics (Kober et al., 2013), computer vision (Mnih et al., 2013), energy (Glavic et al., 2017; Wen et al., 2015), and healthcare (Ling et al., 2017; Tseng et al., 2017).

DRL approaches have been successfully used in many applications in the healthcare domain such as treatment regime development (Fox and Wiens, 2019; Tejedor et al., 2020; Yu et al., 2019a), automated medical diagnosis (Liu et al., 2019; Stember and Shalu, 2020), resource scheduling and healthcare management systems (Yu et al., 2019b; Coronato et al., 2020). Several studies have developed DRL models to select the best treatment policy for some critical diseases such as cancer (Tseng et al., 2017), sepsis (Futoma et al., 2018; Petersen et al., 2018), diabetes (Fox and Wiens, 2019; Tejedor et al., 2020), and human immunodeficiency virus (HIV) (Yu et al., 2019a) with the goal of improving the long-term treatment outcome for the patients. These studies show the need for developing dynamic treatment regimes and sequential clinical

decision-making approaches (Naeem et al., 2020).

El Naqa et al. (2016) investigated the feasibility of RL for two-stage adaptive radiation therapy using a simplified Q-learning algorithm with linear regression considering clinical covariant history as states and tumor control probability as the reward function. Their results demonstrated the promising feasibility of RL models in adaptive radiation therapy. However, more advanced nonlinear models are needed to be able to address biological aspects of multi-stage ART planning. Later, Jalalimanesh et. al. (2017) developed an agent-based model to simulate the tumor growth during radiation therapy and used a tabular Q-learning algorithm to find the optimal RT plan. Their results suggested that the agent-based approach combined with RL is useful for simulating and optimizing Rt plans. However, they did not consider the uncertain biological response of the tumor and OAR cells in their model. Moreover, the tabular Q-learning cannot map high-dimensional state space due to the complexity. Also, finding the optimal action based on Bellman's equation is hard when we have stochastic and non-linear dynamics in the decision-making environment (Li et al., 2020). Alternatively, using a neural network to map input states to (action, Q-value) pairs can help to handle high-dimensional state space, uncertainty in tumor response dynamics, and nonlinear rewards (Mnih et al., 2013). Tseng et al. (2017) explored the feasibility of using deep reinforcement learning (DRL) based on historical treatment plans for automated knowledge-based ART for NSCLC patients. They proposed a three-component neural networks framework consisting of a generative adversarial network (GAN) to learn patients' characteristics, a deep neural network (DNN) to estimate transition probabilities, and a deep Q-network (DQN) to find the optimal action. The results of their study show that DRL can be used to achieve clinically acceptable results for knowledge-based ART while maximizing tumor local control. Their proposed approach can be useful if one has access to large-scale historical patients' data and the certain value of tumor and OAR dosimetric and biological parameters are known. However, collecting such large-scale data for each cancer site needs significant time and effort, and it is not accessible for everyone and all cancer cases. Moreover, the accuracy of the data set may not be guaranteed and prone to human errors. On the other hand, extracting the patients' characteristics from a large-scale data set is a time-consuming process and prone to overfitting (Yousefi et al., 2017).

Therefore, this paper introduces a novel biological response model that incorporates tumor cell death, repopulation, reoxygenation, radio-sensitivity for tumor cells, and healthy tissues cell repair. The proposed model is used to predict the radiation response of the tumor and OARs during the course of treatment. Using the tumor response model, an automated optimization framework is proposed by combining Reinforcement Learning (RL) and optimization method to find the optimal adaptation points for ART and dynamically adapt the plan to the tumor's uncertain biological response over time.

The contributions of this paper are as follows: (1) A biological-based treatment planning framework is proposed such that it not only controls the biological aspect of the treatment and incorporates the tumor biological response uncertainty, but also ensures the dose-volume requirements and clinical limits of the treatment without the need of dealing with complex optimization models; (2) The proposed reinforcement learning framework for ART planning can help the decision-maker to achieve a robust solution under high levels of uncertainty in the biological parameters while reducing the variability in the solution and improving the control on the worst-cases which minimizes the undesirable effects of worst-cases on the treatment

outcome; (3) Using the proposed comprehensive biological response model, the tumor volume regressions can be estimated without taking significant time and effort to collect large-scale datasets and avoid the need for expensive CT images. Also, an ART treatment plan can be determined in a shorter time compared to employing imaging information for the clinical implementation of ART considering the patient wait time and data collection time; (4) This approach enables the physicians to find an appropriate personalized ART policy in terms of fraction dose and timing of the adaptations using the volumetric and biological information to adapt the plan to the updated patient anatomy. It also can be used to identify patients who would benefit from ART as an alternative to the conventional equal-dose plan;, and (5) The proposed approach is flexible enough to support a wide range of treatment objectives and preferences based on different decision-makers for various cancer types.

The rest of this paper is organized as follows. Section 2 explains how the temporal evolution of the tumor due to the radiation response is modeled. We then develop an RL framework for ART policy decision-making and discuss the associated mathematical formulations. Section 3 provides the sensitivity analysis of the model and the results from our experimental study using clinical lung cancer patient data. We conclude the paper in Section 4.

2. Methodology

2.1. Problem description

The goal of this paper is to find the optimal policy for ART (i.e., the optimal timing of adaptation and the associated dose at each adaptation point) considering biological uncertainties to improve the quality of treatment in terms of tumor control and OAR sparing. First, a novel biological response model is introduced to estimate the tumor volume regressions with zero or minimal imaging during the treatment. Second, we propose an automated framework that combines RL and optimization methods. In this framework, the RL uses the biological response model to estimate the tumor volume regression during radiation therapy considering uncertainty on the biological parameters. Then, adaptation points and their associated radiation doses are determined by finding the actions corresponding to the maximum RL reward function value. Third, beamlet intensities are optimized to satisfy dose-volume requirements according to the updated tumor volume prediction at each adaptation point. Therefore, our approach will find a robust optimal ART treatment plan that is biologically and clinically acceptable. The list of sets, parameters, and variables used in the proposed tumor response model and RL algorithm are summarized in Table 1 and Table 2, respectively.

2.1.1. ART environment, states, actions, and reward for the RL framework

2.2. Dynamic tumor response model

This section introduces a model that incorporates the temporal evolution of the tumor responding to radiation. Two cell compartments are considered in this study, namely, proliferating and hypoxic. This classification is consistent with the study by Jeong et al. (2013) with one exception that we merged intermediate and hypoxic compartments into one called hypoxic to be accounted for reoxygenation and hypoxia. The proliferating compartment contains cells that have adequate principle nutrients (i.e., glucose and oxygen) and are in the proliferating phase. The hypoxic compartment comprises cells without enough nutrients. Most of

Table 1: Notations used in the dynamic biological response model and biological metrics

Notation	Description
Sets	
I	Set of treatment sessions (decision epochs)
T	Tumor structure
ϕ	OAR structure
Dynamic variables	
v_i	Tumor volume after delivering fraction $i \in I$
u_i	Number of viable tumor cells after delivering fraction i
w_i	Number of dead tumor cells after delivering fraction i
m_i	Number of doomed tumor cells after delivering fraction i
d_i	Amount of dose in fraction i
Parameters	
N	Total number of treatment sessions
t_i	Time gap between fraction i and $i-1$
τ_g	The repopulation parameter
τ_d	Tumor decay parameter
τ_r^ϕ	OAR repair parameter
τ_g^ϕ	OAR repopulation parameter
OER	Reoxygenation parameter (Oxygen Enhancement Ratio)
α_p^T	Linear tumor radio-sensitivity parameter of LQ model in proliferating phase
α_h^T	Linear tumor radio-sensitivity parameter of LQ model in hypoxic phase
β_p^T	Quadratic tumor radio-sensitivity parameter of LQ model in proliferating phase
β_h^T	Quadratic tumor radio-sensitivity parameter of LQ model in hypoxic phase
α^ϕ	Linear radio-sensitivity parameter of OAR
ρ	Ratio of dead cells at each stage
Biological metrics	
BED_i^T	Cumulative biological effective dose of the tumor after delivering fraction i
BED_i^ϕ	Cumulative biological effective dose of the OAR after delivering fraction i
SF_i	Total surviving fraction of the tumor after delivering fraction i

the cells in the hypoxic component are starving and extremely hypoxic. Presumably, cells in the hypoxic compartment cannot proliferate and the starving cells can die without being further exposed to radiation (necrotic cell death due to starving (Wouters, 2009)). Only a fraction of cells in the proliferating compartment are in the cell cycle and can proliferate. Therefore, we also incorporate three sub-compartments into each compartment to track different cell conditions (i.e., reoxygenation, cell-kill, cell decay caused by starving, cell cycle effect) during the course of treatment: (1) viable cells in the cell cycle, (2) doomed cells that are not in the cell cycle, and (3) dead cells that are hypoxic and are in the decay process caused by cellular necrosis, apoptosis, metastasis, and cell migration. Figure 1a shows the visualization of assumed tumor cell compartments and sub-compartments.

Doomed cells are the middle sub-compartment between metabolically active cells (intermediate) and hypoxic cells. Because of the increased mitotic cell death, a proliferation of doomed cells does not have much impact on the number of cells in the hypoxic compartment (Jeong et al., 2013). Hence, we assumed that doomed cells do not proliferate. However, they could move move to the viable sub-compartment and

Table 2: RL algorithm notations

Notation	Description
Sets	
S	Set of possible states
A	Set of possible actions
RL components	
R	Immediate reward function
P	Transition probability
Q	Q-value function
Q^*	Optimal Q-value function
s_i	State of the system at step i
a_i	Decision set at step i
r_i	Immediate reward at step i
γ	discount factor for long-term reward
DDQN components	
w	DQN weights
w^-	Target network weights
L	Loss function
$N_{\mathcal{D}}$	Replay memory capacity
N_e	The total number of episodes
N_e	The total number of steps (treatment sessions)
ξ	Exploration decay rate

proliferate if they receive enough oxygen. In contrast, doomed cells can move to the dead sub-compartment if they receive enough radiation. At each stage of treatment, a proportion of viable cells can be moved into the doomed and/or dead sub-compartment as a result of the radiation exposure. Figure 1b shows the redistribution of tumor cells sub-compartments after receiving the radiation dose.

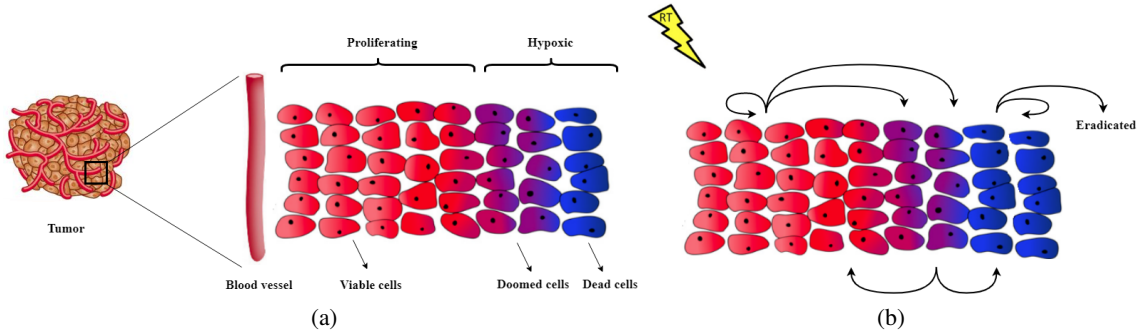


Figure 1: (a) Visualization of two tumor cells compartments and three sub-compartments (i.e., viable, doomed, and dead) (b) Redistribution of tumor cells sub-compartments after receiving radiation dose

Our proposed model considers proliferation and hypoxia as two important factors in tumor radiation response and repopulation. The traditional L-Q model does not account for the necrotic cell death and exponential decay of the dead cells. Based on our model, only a fraction of lethally damaged cells are truly dead. The remaining cells are doomed, which means they are metabolically active without having enough oxygen

to proliferate and are radio-resistant (Jeong et al., 2013). As the doomed cells are hypoxic, higher doses of radiation are needed to kill these cells. This characteristic of hypoxic cells can be incorporated into the response model by adding the Oxygen Enhancement Ratio (OER). The OER indicates the required extra dose to achieve the same level of cell-kill/cell-survival of the hypoxic cell compared to the non-hypoxic (normoxic) cells (Carlson et al., 2006). The oxygen level can vary based on the distance from the tumor to blood vessels and blood vessel damage during radiation therapy (Paul-Gilloteaux et al., 2017). We assume that surviving doomed cells can receive oxygen and return back to the proliferating phase as alive cells (re-oxygenation of hypoxic cells). Necrotic cell loss is assumed to follow an exponential decay with parameter τ_d. Also, we consider an exponential tumor growth with parameter τ_g. Based on these assumptions, the number of cells in each sub-compartment at each time epoch can be calculated for viable (u_i), dead (w_i), and doomed (m_i) cells:

$$u_{i+1} = u_i \cdot exp(-\alpha_p^T d_i - \beta_p^T d_i^2) \cdot exp(\frac{t_i}{\tau_g}) + m_i \cdot exp(-\alpha_p^T \frac{d_i}{OER} - \beta_p^T \frac{d_i^2}{OER^2}), \tag{1}$$

$$\begin{aligned} w_{i+1} = \rho\, u_i \cdot (1 - exp(-\alpha_p^T d_i - \beta_p^T d_i^2)) \cdot exp(-\frac{t_i}{\tau_d}) \\ + m_i \cdot (1 - exp(-\alpha_p^T \frac{d_i}{OER} - \beta_p^T \frac{d_i^2}{OER^2})) \cdot exp(-\frac{t_i}{\tau_d}) + w_i \cdot exp(-\frac{t_i}{\tau_d}), \end{aligned} \tag{2}$$

$$m_{i+1} = (1 - \rho)\, u_i \cdot (1 - exp(-\alpha_p^T d_i - \beta_p^T d_i^2)). \tag{3}$$

Where d_i is the dose for fraction i and t_i is the time gap between fraction $i-1$ and i. Tumor volume changes during RT can be estimated using cell compartment distribution at each time epoch and the initial total number of tumor cells. In addition to the viable cells, doomed and dead cells (hypoxic cells) were also included in the total tumor volume calculation (Jeong et al., 2017; Unkelbach et al., 2014). Therefore, the tumor volume at each time epoch can be estimated as

$$v_{i+1} = u_{i+1} + w_{i+1} + m_{i+1}. \tag{4}$$

The increased radioresistance of hypoxic cells compared to proliferating cells can be quantified as $\alpha_h^T = \alpha_p^T/OER$ and $\beta_h^T = \beta_p^T/OER^2$ (Carlson et al., 2006). Hence, we propose the response models for the three cells types as

$$u_{i+1} = u_i \cdot exp(-\alpha_p^T d_i - \beta_p^T d_i^2) \cdot exp(\frac{t_i}{\tau_g}) + m_i \cdot exp(-\alpha_h^T d_i - \beta_h^T d_i^2), \tag{5}$$

$$\begin{aligned} w_{i+1} = \rho\, u_i \cdot (1 - exp(-\alpha_p^T d_i - \beta_p^T d_i^2)) \cdot exp(-\frac{t_i}{\tau_d}) + m_i \cdot (1 - exp(-\alpha_h^T d_i - \beta_h^T d_i^2)) \cdot exp(-\frac{t_i}{\tau_d}) \\ + w_i \cdot exp(-\frac{t_i}{\tau_d}), \end{aligned} \tag{6}$$

$$m_{i+1} = (1 - \rho)\, u_i \cdot (1 - exp(-\alpha_p^T d_i - \beta_p^T d_i^2)). \tag{7}$$

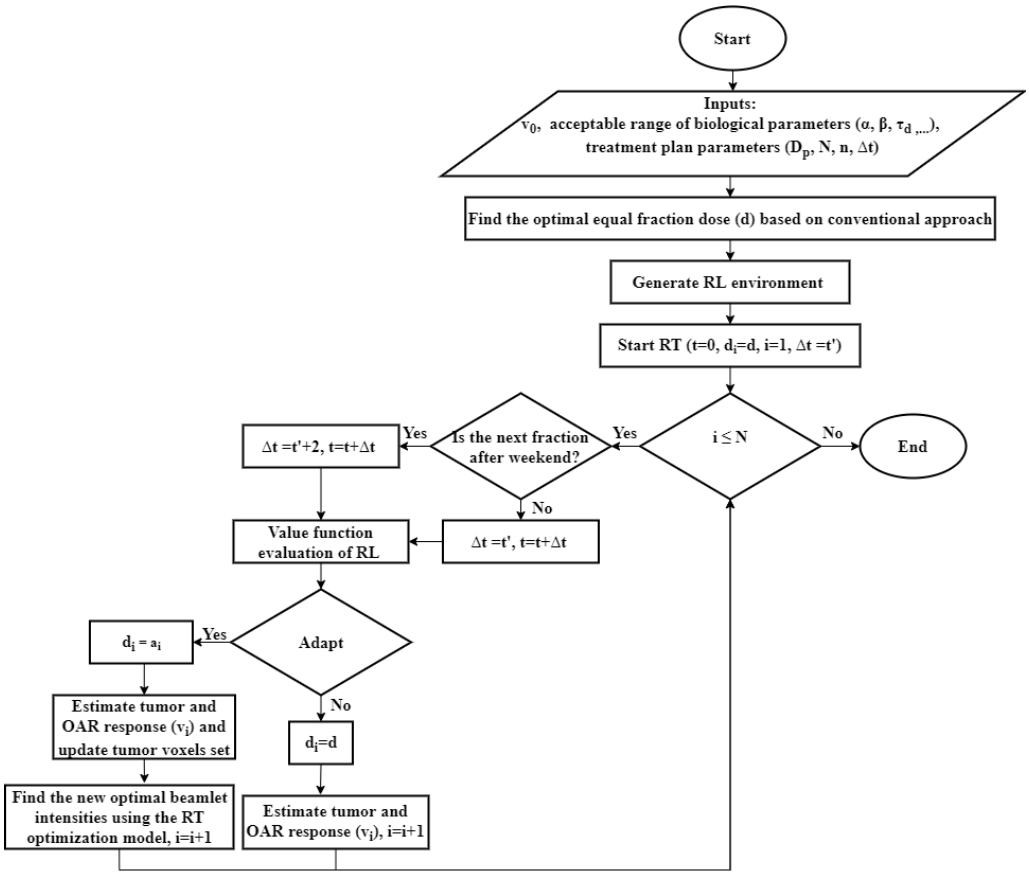


Figure 2: Flowchart of the proposed algorithm

2.3. Reinforcement learning framework for the ART problem

Our aim is to develop an automated framework that combines RL and optimization methods, in which the adaptation points are found using the RL based on the biological effects and responses of the tumor and OAR to radiation using the proposed response model. Figure 2 shows the entire process of the proposed algorithm. The goal of the agent in RL is to take actions that maximize the expected value of a predefined reward function. The RL environment can be described by the various states. The agent receives a reward (r_t) according to the selected decision being made under a specific state (s_t), which leads to the next state (s_{t+1}). Using this feedback mechanism between the state and its corresponding reward, the agent can optimize its subsequent strategy for future actions. Figure 3 shows the RL procedure and its components.

We propose a general RL framework based on a Markovian environment generated by the dynamic tumor response model to find the optimal policy of ART in which the environment can be customized based on the tumor biological parameters and patient information. The RL agent learns by training the RL model for each patient with an specific cancer type. As it gets trained over time for many patients, its ability to find optimal actions gets improved accordingly for future patients with the same cancer type. Instead of taking images frequently to detect tumor volume changes during the treatment, our approach enables us to estimate the tumor volume regressions based on the proposed tumor response model used in the RL environment quantification. These estimates can be validated and/or corrected using a limited amount of imaging data.

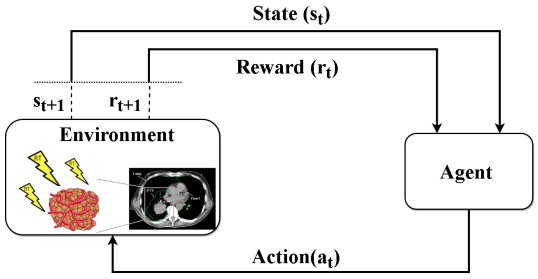


Figure 3: Reinforcement Learning Procedure

Our approach is a continuous adaptation protocol on a weekly/daily basis, and we aim to find the optimal number of adaptations, the corresponding times, and the radiation dose to be used until the next adaptation point.

The environment in the proposed RL algorithm is a virtual environment to simulate ART treatment planning considering tumor volume changes using the proposed dynamic tumor response model. After the execution of the selected action, the agent obtains information on the next state and its corresponding reward value. At each time stage, a number of scenarios for tumor biological factors are generated. For each scenario, the tumor response to radiation and its immediate reward is calculated based on the current state and the action taken to determine the best action and next state. The set of actions includes possible decisions such as dose *increase* ($+\Delta$)*, maintain, decrease* ($-\Delta$) as a result of the plan adaptation, where Δ is an amount of dose deviation from the conventional prescription dose d. For a given cancer type and its clinical protocols for treatment, some key input parameter values for reinforcement learning method can be given to the planner such as daily fractional dose d, Δ, dose lower-bound ($\underline{d}$), and dose upper-bound ($\bar{d}$). In some cases, the action sets may include more than the three actions, which could be multiples of Δ. Therefore, Algorithm 2 (see Appendix A) is used to construct an appropriate action set.

Our aim here is to find the optimal action at each time epoch (e.g., beginning of each week) to determine optimal adaptation points. We begin by showing in Theorem 1 the existence of an optimal dose d_i^* of each fraction for a concave reward function on the closed interval $[\underline{d}, \bar{d}]$.

Theorem 1. *For a concave reward function at each fraction,* $R(d_i)$*, there exists an optimal fractional dose* $d_i^* = \operatorname{argmax} R(d_i)$*, where* $d_i \in [\underline{d},\ \bar{d}]$.

$$\begin{cases} d_i^* = \underline{d}, & \text{if } \frac{\partial R(d_i^*)}{\partial d_i} < 0, \\ d_i^* \in [\underline{d},\ \bar{d}], & \text{if } \frac{\partial R(d_i^*)}{\partial d_i} = 0, \\ d_i^* = \bar{d}, & \text{if } \frac{\partial R(d_i^*)}{\partial d_i} > 0. \end{cases}$$

Proof. See Bolzano-Weierstrass Theorem (Bazaraa et al., 2013). □

Using Theorem 1, one can find the relation between the optimal fractional dose d_i^* and the equal fraction dose $d \in [\underline{d},\ \bar{d}]$ of the conventional reference plan. Hence, the corresponding action can be *increase, maintain, or decrease*, and the range of d_i^* can be determined by Corollary 1.1.

Corollary 1.1. *The relation between the optimal fractional dose, d_i^*, to the conventional reference dose, $d \in [\underline{d},\ \bar{d}]$, can be determined by the gradient, $\frac{\partial R(d_i)}{\partial d_i}$, of the reward function at $d_i = d$ as follows*

$$\begin{cases} \textit{if } \frac{\partial R(d_i)}{\partial d_i}|_{d_i=d} < 0, & \underline{d} \leq d_i^* < d, \\ \textit{if } \frac{\partial R(d_i)}{\partial d_i}|_{d_i=d} = 0, & d_i^* = d, \\ \textit{if } \frac{\partial R(d_i)}{\partial d_i}|_{d_i=d} > 0, & d < d_i^* \leq \bar{d}. \end{cases}$$

Proof. The proof is trivial following Theorem 1. □

Few performance measures are commonly used to evaluate an RT treatment plan including BED, tumor control probability (TCP) and normal tissue control probability ($NTCP$). BED is a measure to estimate the amount of radiation damage received in any structure. A higher tumor BED is known to give better tumor control. In contrast, a lower OAR BED is desirable to have lower OAR toxicity. Therefore, we developed a multi-stage optimization model, in which the biological response of the treatment is defined based on the BED of the tumor and OAR. At each adaptation point determined by the RL (i.e., at stage k), the following optimization problem is used to find the optimal fraction dose of the stage (d_k).

$$\max \quad \sum_{i=1}^{k} BED_i^T(d_i) \tag{8}$$

$s.t.$

$$\sum_{i=1}^{k} BED_i^{\phi}(\gamma\ d_i) \leq BED_k^{\phi_{Ref}}(\frac{k}{N}\ \gamma\ D_{pres}) \tag{9}$$

$$\sum_{i=1}^{k} BED_i^{T}(d_i) \geq BED_k^{T_{Ref}}((1-\frac{k}{N})\ D_{pres}^L) \tag{10}$$

$$SF_k\ SF_{N-k}^{Ref} \leq \varepsilon \tag{11}$$

$$SF_{N-k}^{Ref} = SF((1-\frac{k}{N})\ D_{pres}) \tag{12}$$

$$SF_k = SF(d_{k-1}) \cdot SF(d_k) \tag{13}$$

$$d_l \leq d_k \leq d_u. \tag{14}$$

The objective function (8) maximizes the total BED on the tumor (T) by delivering $d_1, d_2, ...,$ and d_k at stages $1, 2, ..., k$. Constraint (9) controls the BED deviations from the conventional reference plan (i.e., an equal fraction dose $d_i = d, \forall i \in \{1, 2, ..., k\}$) for the OAR biological tolerance to achieve the same or better OAR (ϕ) toxicity. Constraint (10) sets a lower-bound for tumor BED based on a set of biological parameters and delivered dose using the lower bound of the prescription dose to be accounted for the required clinical tumor BED. Constraint (11) is to ensure that tumor cells will be completely eradicated at the end of treatment even if we continue the rest of the treatment with the conventional plan (i.e., $d_i = d, \forall i \in \{N-K, ..., N\}$). Where, SF_k is the total surviving fraction at the end of stage k by delivering $d_1, d_2, ...,$ and d_k at stages $1, 2, ..., k$, and SF_{N-k}^{Ref} is the the total surviving fraction after $N-k$

fractions based on the reference plan (labeled as Ref). Constraints (12) and (13) represent the calculation of SF_{N-k}^{Ref} and SF_k, respectively. The total survival fraction at stage k is calculated based on the total dose delivered to the patient by the end of stage k (i.e., $\sum_{i=1}^{k} d_i$) which is equivalent to multiplications of surviving fractions due to each fraction dose $d_i \forall i \in \{1, 2, \ldots, k\}$ (McMahon, 2018). Finally, constraint (14) ensures that the amount of the fraction dose is within its lower and upper bounds.

We can formulate a reward function based on the proposed biological optimization model by relaxing constraints and penalizing weighted constraint violations. Our proposed reward function of RL is defined as follows:

$$\begin{aligned} R(s_i, a_i) = \lambda_1 \, BED_i^T(s_i, a_i) - \lambda_2 \, \left(BED_i^{T_{Ref}}(s_i) - BED_i^T(s_i, a_i)\right)^+ \\ - \lambda_3 \, \left(SF_i(s_i, a_i) \, SF_{N-i}^{Ref}(s_i, a_i) - \varepsilon\right)^+ - \lambda_4 \, \left(BED_i^{\phi}(s_i, a_i) - BED_i^{\phi_{Ref}}(s_i)\right)^+, \end{aligned} \tag{15}$$

where $BED_i^T(s_i, a_i)$ and $BED_i^{\phi}(s_i, a_i)$ are the cumulative BED after taking action $a_i \in \{A\}$ (delivering d_k dose) in state $s_i \in \{S\}$ at each time stage i and $(.)^+$ is a sign function defined as

$$(x)^+ = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}$$

The BED estimate of the tumor is calculated considering the surviving fraction of viable and doomed cells, tumor repopulation of viable cells, and decay of dead cells. Hence, the proposed formulation for the BED of the tumor at each time stage is

$$\begin{aligned} BED_i^T(d_i) = -\left(\frac{u_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha \, \tau_g} + \left(\frac{w_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha \, \tau_d} + \left(\frac{u_{i-1}}{v_{i-1}}\right)\left(d_i + \frac{d_i^2}{\alpha/\beta}\right) \\ + \left(\frac{m_{i-1}}{v_{i-1}}\right)\left(\frac{d_i}{OER} + \frac{d_i^2}{OER^2 \, \alpha/\beta}\right). \end{aligned} \tag{16}$$

Theorem 2. *There exists a lower bound (d_l) on d_i such that $BED_i^T(d_i)$ is a non-negative and monotonically increasing function for all $d_i \geq d_l, \forall i = 1, \cdots, N$.*

- *If* $\frac{u_{i-1}}{w_{i-1}} \leq \frac{\tau_g}{\tau_d}$, *then* $d_l = 0$ *and* $BED_i^T(d_i) \geq 0$,
- *If* $\frac{u_{i-1}}{w_{i-1}} > \frac{\tau_g}{\tau_d}$, *then* $d_l = \frac{-b+\sqrt{\delta}}{2a} > 0$, *and* $BED_i^T(d_i) > 0$
  where,
  $a = \frac{OER^2 \, u_{i-1}+m_{i-1}}{OER^2 \, \alpha/\beta \, v_{i-1}}$,
  $b = \frac{OER \, u_{i-1}+m_{i-1}}{OER \, v_{i-1}}$,
  $c = -\left(\frac{u_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha \, \tau_g} + \left(\frac{w_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha \, \tau_d}$,
  $\delta = b^2 - 4ac$.

Proof. See Appendix B. □

We consider OAR repair and repopulation, which are major biological factors affecting the response to radiation on healthy tissues. The following equation is used to capture the OAR's biological response to

radiation during the RT treatment:

$$v_{i+1}^{\phi} = v_i^{\phi}\ exp(-\alpha^{\phi} d_i^{\phi} - \beta^{\phi} {d_i^{\phi}}^2)\ exp(\frac{t_i}{\tau_g^{\phi}})\ exp(\frac{t_i}{\tau_r^{\phi}}), \tag{17}$$

using the general BED formulation, the BED of an OAR can be calculated as

$$BED_i^{\phi} = \left(d_i^{\phi} + \frac{{d_i^{\phi}}^2}{\alpha^{\phi}/\beta^{\phi}} \right) - \frac{\Delta t_i}{\tau_g^{\phi}\alpha^{\phi}} - \frac{\Delta t_i}{\tau_r^{\phi}\alpha^{\phi}}. \tag{18}$$

We assume that the OAR receives a heterogeneous dose with a sparing factor of θ, which indicates the ratio of the average dose received by the OAR to the average dose received by the tumor ($d_k^{\phi} = \theta\ d_k$).

2.3.1. Deep double Q-learning network for RL training

The training data is a tuple of $\{S, A, R\}$ in a finite horizon (treatment duration) and the goal is to develop an optimal policy (sequence of decision rules) for ART to maximize the long-term reward which is defined based on RT performance metrics (e.g., BED, SF) for the treatment outcome. Therefore, the effect of the actions is evaluated not only based on the immediate reward but also the long-term or subsequent rewards. The value function $V(s)$ presents the value of a state which is defined as the total expected reward starting from the state expressed as *Q-value function* $Q(s, a)$ of

$$Q(s,a) = E\left[R_t + \gamma R_{t+1} + \gamma^2 R_{t+2} + ...|s,a\right] = E_{s'}\left[R_t + \gamma Q(s',a')|s,a\right], \tag{19}$$

the Q-value function can be used to find the optimal value function $Q^*(s, a)$ as

$$Q^*(s,a) = E_{s'}\left[R_t + \gamma\ max\ Q^*(s',a')|s,a\right]. \tag{20}$$

Q-learning is a common approach to find the optimal Q-values in RL. Recent studies by Google DeepMind have shown that the Q-function can be evaluated efficiently using a deep Q-network (DQN) that provides a stable solution to deep value-based RL (Mnih et al., 2013, 2015). Several studies showed that the DQN algorithm achieved better computational performance than the Q-learning algorithm in complex RL environments (Li et al., 2020; Liu et al., 2021). The time complexity is sublinear in the length of the state period (i.e., the number of steps per episode multiplied by the total number of episodes), and the space complexity is sublinear in the number of state space, action space, and steps per episode (Leem et al., 2020, Liu et al., 2021). Since the same weights are used to estimate both target and Q-values in the DQN algorithm, both Q-values and target values are shifting and there is a big correlation between the target network and the output weights that are changing for training. So, we consider the idea of fixed Q-targets introduced by DeepMind and employ a separate network with a fixed parameter (w^-) for estimating the target values, and update the target network at every τ steps based on the current DQN parameters. Therefore, we will have a more stable learning process. An over-estimation of Q-values at the early stages of the training can be an issue in Q-learning. Hence, we use the Double DQN structure (Van Hasselt et al., 2016) to handle the problem. We use two networks to separate the selection of an action from the target Q-value generation during the

learning process to reduce a false positive error that resulted from a noisy Q-value. Therefore, the DQN network selects the best action first to take for the next state (i.e, the action with the highest Q-value). Then, the target network calculates the target Q-value according to the action taken at the next state. This results in a faster training and more stable learning process without increasing the computational complexity. The loss function can be calculated as

$$L(w) = E\left[\left(R + \gamma\, max\, Q(s', a', w^{-}) - Q(s, a, w)\right)^2\right], \tag{21}$$

where the Q-value function is evaluated for each action using a DQN with weights w as $Q(s, a, w)$ and the maximum possible Q-value for the next state is calculated based on the immediate reward R and the discounted maximum Q-value among all possible actions from the next state obtained from target network weights w^{-} as $Q(s', a', w^{-})$. The entire learning process is summarized in Algorithm 1.

Algorithm 1 Reinforcement Learning Training Process

Initialize: replay memory buffer $\mathcal{D}$ to capacity $N_{\mathcal{D}}$, network weights (w_1), target network ($w_1^{-} \leftarrow w_1$), the environment, exploration decay rate ξ, and episode counter $e = 0$;
for $e \leq N_e$ **do**
  Reset the environment; $t = 0$; observe the first state
  for step $t \leq N$ **do**
    Increase the exploration decay rate ($\xi \leftarrow \xi + \Delta_{\xi}$)
    Use Epsilon Greedy Strategy with probability ϵ to select a random action a_t
    Otherwise, select action $a_t = \mathrm{argmax}_a\, Q(s_t, a; w)$
    Execute action a_t; Calculate reward r_t; and observe next state s_{t+1}
    Store the transition (s_t, a_t, r_t, s_{t+1}) in the replay memory $\mathcal{D}$
    Sample a random minibatch of the transitions (sj, a_j, r_j, s_{j+1}) from $\mathcal{D}$
    if the episode ends at next state $(j + 1)$ **then**
      Set target $\hat{Q} = r_j$
    else
      Set $\hat{Q} = r_j + \gamma\, Q(s_{j+1}, \mathrm{argmax}_{a'}\, Q(s_{j+1}, a'; w), w^{-})$
    Perform a gradient descent step with loss $\left(\hat{Q} - Q(s_j, a_j; w)\right)^2$
    Every τ steps $(e.N) + t > \tau$ reset the target network weights ($w_t^{-} \leftarrow w_t$)
end

2.4. RT optimization model

Using the proposed RL approach, we can find the optimal adaptation points to improve the biological response of the tumor and OARs by maximizing the reward function. The agent's action is based on the biological-based reward function, but the treatment plan also needs to meet physical dose requirements for the clinical purpose. Since it is far too complicated to consider all aspects of an RT treatment (i.e., biological and physical) in one reward function, an optimization model is proposed to primarily control the dose-volume clinical requirements. Once the adaptation points are determined using the RL approach, a beamlet optimization model is solved to satisfy the dose-volume constraints.

This optimization model will be adapted at each adaptation point based on the corresponding predicted tumor volume. For this purpose, the tumor response model can be used to estimate tumor volume changes. The tumor volume change ratio at each stage $r_k^T(d_k)$ is a function of a delivered dose at the stage (d_k) and

can be calculated as $r_k^T(d_k) = \frac{v_k(d_k)}{v_{k-1}(d_{k-1})}$. Therefore, at each adaptation point determined by RL (i.e., at stage $i = k$), the following optimization model is solved to find the optimal beamlet intensities.

$$\min \quad \sum_{s\in\{T\cup S\}} \frac{C_s}{|V_s^k|} \sum_{v\in V_s^k} D_v^k \tag{22}$$

$s.t.$

$$D_v^k \le U_v^k \qquad \forall\, v \in V_s,\ s \in \{T \cup S\}\,, \tag{23}$$

$$D_v^k \ge L_v^k \qquad \forall\, v \in V_s,\ s \in \{T\}\,, \tag{24}$$

$$D_v^k = \sum_{b\in B} \Delta_{v,k,b}\, w_b, \qquad \forall\, v \in V_s,\ s \in \{T \cup S\}\,, \tag{25}$$

$$w_b \ge 0, \qquad \forall\, b \in B. \tag{26}$$

3. Numerical experiments and results

3.1. Experiment setup

We evaluate the proposed tumor response model using simulation and compare the result with the conventional LQ response model. Then, a sensitivity analysis is performed to explore the effect of variability in the corresponding parameters and evaluate the observations based on clinical practices. Since tumor growth and radio-sensitivity are two main biological factors in determining tumor radiation response, we consider four types of tumors based on the range of radio-sensitivity parameter (α) and tumor growth factor (τ_g). Figure 4 shows tumor volume after delivering 2 Gy radiation dose to a tumor with the initial volume of 10,000 voxels considering $\alpha \in [0.03, 0.0365]$ Gy^{-1} and $\tau_g \in [0.5, 15]$ days. We categorize the tumor response types into four groups and their corresponding ranges of model parameters as summarized in Table 3 (Yang and Xing, 2005; Saberian et al., 2016; Uzan and Nahum, 2012; Van Leeuwen et al., 2018). Case I and Case II refer to the fast-growing tumors with low levels of radio-sensitivity, and high radio-sensitivity (i.e., Early responding tumors), respectively. Case III and Case IV are for slow-growing tumors with low radio-sensitivity (i.e., Late responding tumors) and high radio-sensitivity (i.e., Intermediate to late responding tumors), respectively.

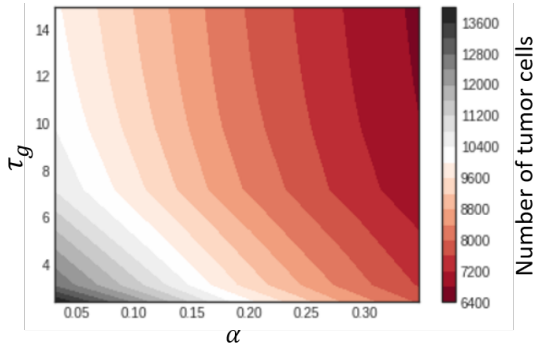


Figure 4: Number of tumor cells after delivering 2 Gy radiation to a tumor with the initial volume of 10,000 voxels based on the range of $\alpha \in [0.03, 0.0365]$ Gy^{-1} and $\tau_g \in [0.5, 15]$ days

Table 3: Tumor volume cases used in the sensitivity analysis of tumor response model

	Low radio-sensitivity ($\alpha \in [0.03, 0.25]$ Gy^{-1})	High radio-sensitivity ($\alpha \in [0.25, 0.365]$ Gy^{-1})
Fast growing tumor ($\tau_g \in [0.5, 7]$ days)	Case I	Case II
Slow growing tumor ($\tau_g \in [7, 60]$ days)	Case III	Case IV

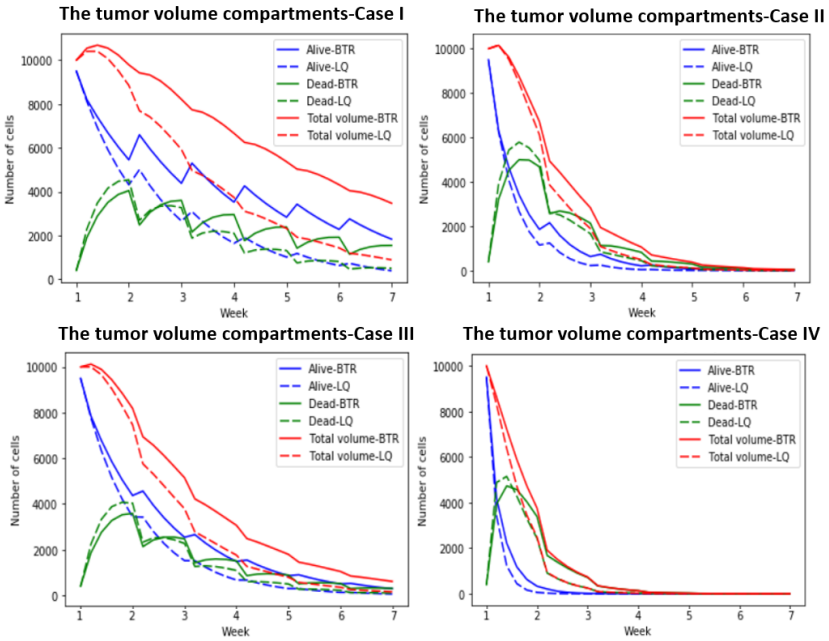


Figure 5: Radiosensitivity effect on the tumor response curves based on the biological tumor response (BTR) model and LQ model

3.2. Sensitivity analysis of the tumor response model

3.2.1. The effect of biological parameters

We study the effect of radio-sensitivity (α) and the tumor growth factor (τ_g) on tumor volume response considering four tumor cases (see Table 3). The following parameters were selected based on the ranges in Table 3: $\tau_d = 4$ days, $\rho = 0.8$, $\alpha/\beta = 10$ Gy, $OER = 1.8$ for all cases, $\alpha = 0.135$ Gy^{-1} and $\tau_g = 5$ days for Case I, $\alpha = 0.282$ Gy^{-1} and $\tau_g = 5$ days for Case II, $\alpha = 0.135$ Gy^{-1} and $\tau_g = 15$ days for Case III, $\alpha = 0.282$ Gy^{-1} and $\tau_g = 15$ days for Case IV. Experimental results of the proposed tumor response model will be compared with those of the conventional LQ response model. As it is common in practice, we assumed that the treatment plan is to deliver five equal fractional doses of 2 Gy in 6 weeks and no treatment will be given on the weekends. For each case, we simulated the weekly tumor response based on the LQ model and the proposed response model. Figure 5 shows the tumor response curves and cumulative tumor cell-kill rate compared with the conventional LQ response model for each tumor case. As shown in Figure 5, both models behaved similarly for Case IV which is assumed to have the highest α and τ_g values. The difference between the two models is more noticeable when the tumor is less radio-sensitive (e.g., Case I and Case III), and the impact of tumor growth and reoxygenation (which is not considered in the LQ model) became more apparent. The results indicate that our tumor response model is more sensitive to tumor reoccurrence risk and shows more realistic results than the LQ model for less radio-sensitive tumors

(e.g., Case I and Case III).

The best treatment outcome (i.e., no more remaining tumor cells after the fourth week of treatment) was observed in Case IV, which is the most radio-sensitive case with slow proliferation. Increasing the proliferation rate (Case II) resulted in a slightly worse treatment outcome by taking six weeks to eradicate the whole tumor. However, we observed that a lower tumor radio-sensitivity corresponded to a less desirable treatment outcome for both slow proliferating (Case III) and fast proliferating (Case I) tumors. As expected, the worst treatment outcome was observed in Case I, which has the lowest radio-sensitivity with fast tumor cell proliferation. We consider an exponential tumor growth with a constant rate of $1/\tau_g$, which can be

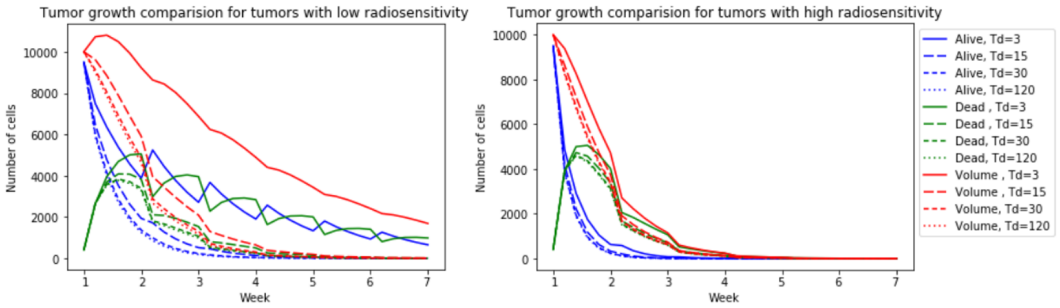


Figure 6: Tumor growth parameter effect on the tumor response curves based on the proposed tumor response model

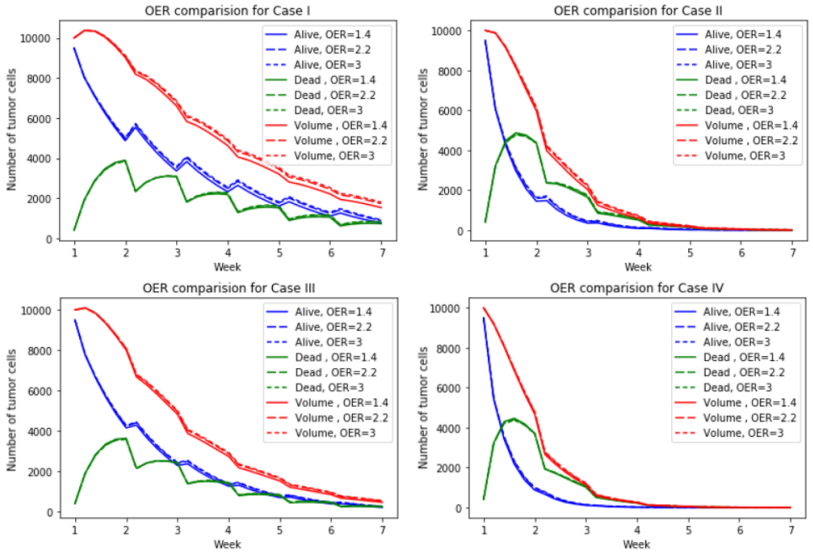


Figure 7: OER parameter effect on the tumor response curves based on the proposed tumor response model

determined based on the tumor doubling time ($\tau_g = T_d/Ln(2)$). A higher tumor growth rate (τ_g) would lead to a higher tumor doubling time and slower tumor growth. One test case has a high radio-sensitivity ($\alpha = 0.282$), while another one with low radio-sensitivity ($\alpha = 0.135$). The tumor response was simulated using four different settings of T_d at 3, 15, 30, and 120 days with the same treatment plan (five equal fractional doses of 2 Gy during a period of six weeks). Figure 6 plots the tumor response curves for four assumed tumor cases with different values of T_d. The tumor response curves of both cases show that the

tumor volume and the number of viable tumor cells are higher for lower T_d values (i.e., fast-growing tumors). The effect of changing T_d is more noticeable for less radio-sensitive cases as expected based on previous observations. The radio-resistant case with the lowest doubling time or fastest proliferation ($T_d = 3$ days) resulted in the highest final number of remaining tumor cells, which means that the number of repopulated tumor cells was higher than the number of cells killed by RT.

Increasing the value of OER means the hypoxia effect becomes more severe for the tumor; hence, the tumor would be less radio-sensitive. This effect of changing the OER parameter on the tumor volume regressions is explored based on the proposed tumor response model (see Figure 7). As seen in the figure, changing the OER value did not affect the tumor response curves for radio-sensitive tumors (Case II and Case IV). This is because the value of the radio-sensitivity parameter (α) is still high even in the hypoxic phase. However, a higher value of OER can lead to a larger volume of the tumor at the end of treatment.

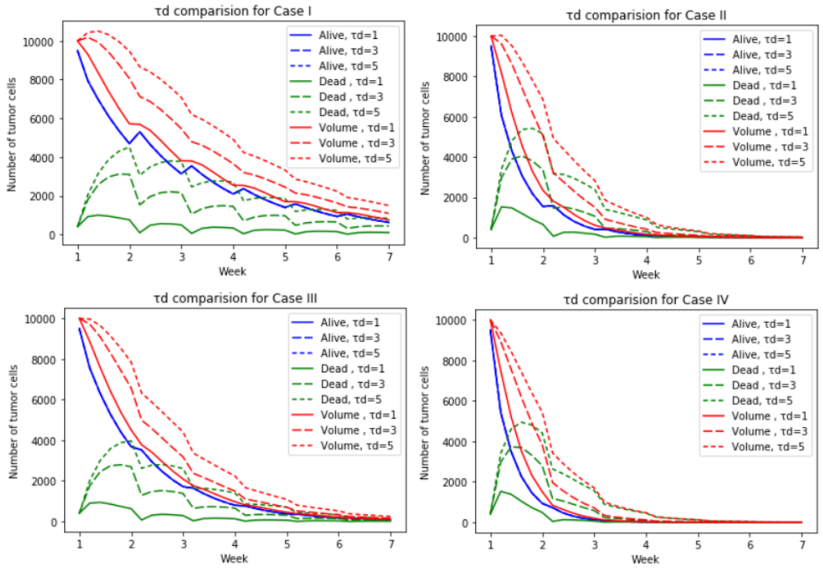


Figure 8: Decay parameter effect on the tumor response curves from the proposed tumor response model.

We further investigate the impact of changing the tumor decay parameter (τ_d) on the radiation response of the tumor. Three common τ_d values were used in this experiment, and Figure 8 shows the results. The figure shows that changing the value of τ_d only affected the number of dead cells which are in the process of decay. Having a longer tumor decay time corresponded to less decay of dead cells. As a result, the total number of tumor cells will increase. We can also see that changing τ_d affected the tumor response for all cases in the same way.

We assumed the same α/β ratio in all previous analyses and we investigated the sensitivity of the tumor response model to α/β values for tumors with high and low radio-sensitivity. Figure 9 shows the simulated tumor radiation response at four different α/β values of 2, 4, 8, and 10 Gy for high and low radio-sensitive tumors. The results showed that a higher α/β was associated with the reduction in the tumor volume regression rate during the course of treatment. This declining pattern was more noticeable for low radio-sensitive tumors. However, it showed a small effect on the tumors with higher radio-sensitivity.

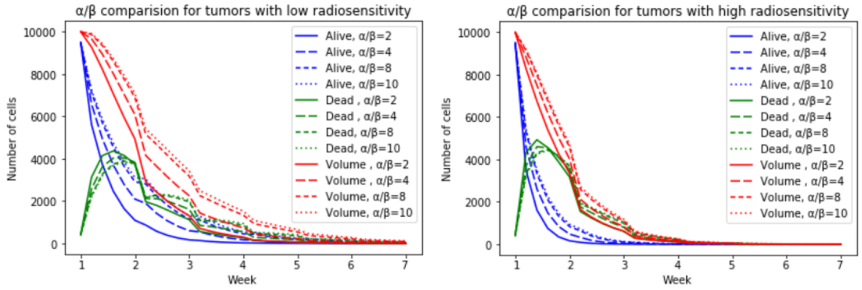


Figure 9: α/β parameter effect on the tumor response curves based on the proposed tumor response model

3.2.2. The effect of variable fractionation scheme

The variable dose fractionation is another important factor affecting the tumor biological response during radiation therapy treatment. We investigated the impact of having different fractionation plans on the tumor BED, OAR BED, and tumor cell killing rate for the four cancer cases. We assumed that the treatment protocol is to deliver a 60 Gy prescription dose that is delivered five fractions per week for 6 weeks. We considered one plan with an equal fraction dose of $d = 2$, and two other plans with the variable fraction dose $d \in \{1.8, 2, 2.2\}$ with approximately the same OAR toxicity ($BED^{\phi} \pm 1\%$) as an equal dose plan. Table 4 summarizes the weekly dose per fraction and the values of BED^{T}, BED^{Φ}, and tumor cell killing rate ($1 - SF$) for the four tumor cases. As shown in the table, the BED^{Φ} for both variable plans were the same, and it was increased by 0.3% compared to the equal dose plan. The value of BED^{T} was also increased for Plan (a) by 0.50% and decreased by 0.44% for Plan (b), on average. Furthermore, the tumor cell killing rate (1-Surviving fraction) was the highest for Plan (a) and the lowest for Plan (b) for all cases. This difference was more noticeable for cases with a worse treatment outcome (e.g., Case I). The results suggest that the variable fractionation can change the treatment outcome in terms of the tumor BED. The treatment outcome can be improved by changing the fractionation scheme specifically for the cases with a low BED.

3.3. RL environment generation and variability analysis

The environment of the RL algorithm is an ART treatment planning environment that includes all possible ART based on all possible tumor volume changes scenarios. At each time stage, a number of tumor volume cases and their associated BEDs and surviving fractions based on the state and action in the previous time stage are calculated and they are used to determine the state and immediate reward. To incorporate the variability inherent in biological parameters within the RL environment, we choose a set of values for each parameter based on its possible range at each episode for a specific cancer site. As a result, the RL agent can see all possible values of parameters, and a robust action can be taken accordingly.

To better understand the existing biological uncertainty in tumor radiation response, the RL environment was generated for the four general tumor cases. A set of values for each parameter was chosen randomly using a uniform distribution. Since the tumor response model is sensitive to α and τ_g values, five random values representing variations of each parameter were generated. Two values for T_d are considered because the variability within the possible range of T_d was lower than α and τ_g. Since the observed effect of OER

Table 4: Weekly dose per fraction, tumor BED, OAR BED and tumor cell killing rate ($1 - SF$) based on each plan for the four tumor cases

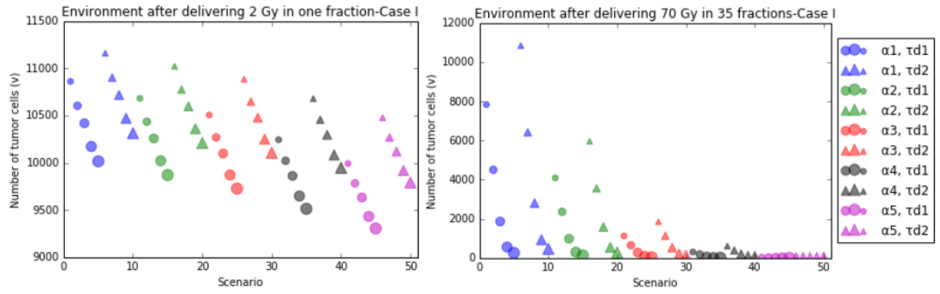
	Weekly dose per fraction (Gy)								
	W1	W2	W3	W4	W5	W6	BED^T	BED^Φ	$1 - SF(\%)$
Case I									
Equal dose plan	2	2	2	2	2	2	55.61	23.02	97.70
Variable dose plan (a)	2	2.2	2.2	1.8	1.8	2	55.94	23.20	97.92
Variable dose plan (b)	2	2	1.8	1.8	2.2	2.2	55.25	23.20	97.28
Case II									
Equal dose plan	2	2	2	2	2	2	57.15	23.02	99.42
Variable dose plan (a)	2	2.2	2.2	1.8	1.8	2	57.40	23.20	99.50
Variable dose plan (b)	2	2	1.8	1.8	2.2	2.2	56.93	23.20	99.28
Case III									
Equal dose plan	2	2	2	2	2	2	72.77	23.02	99.42
Variable dose plan (a)	2	2.2	2.2	1.8	1.8	2	73.17	23.20	99.48
Variable dose plan (b)	2	2	1.8	1.8	2.2	2.2	72.32	23.20	99.33
Case IV									
Equal dose plan	2	2	2	2	2	2	78.44	23.02	99.90
Variable dose plan (a)	2	2.2	2.2	1.8	1.8	2	78.72	23.20	99.91
Variable dose plan (b)	2	2	1.8	1.8	2.2	2.2	78.31	23.20	99.90

on tumor radiation response was negligible, we assumed $\alpha/\beta = 10$ and $OER = 1.8$ which are the most commonly reported values.

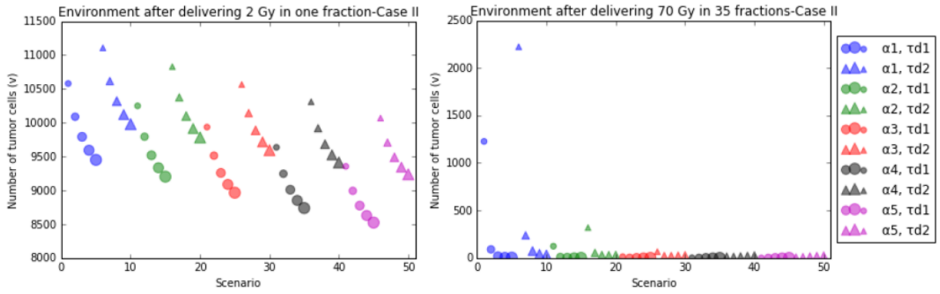
Figure 10 shows various results on the number of remaining tumor cells (tumor volume) in the RL environment. For each tumor case, two sets of plots are made, in which the left one shows the results after delivering the first fraction of 2 Gy, and the subfigure on the right shows the remaining tumor cells after 35 fractions are fully delivered. The number of remaining tumor cells for each value of α is shown by different colors, where the blue color and purple color represent the lowest and the highest values of α, respectively. The circles are markers of fast tumor decay ($\tau_d = 2$ days) and triangles are for slow tumor decay ($\tau_d = 6$ days). A different marker size corresponds to different values of τ_g such that larger values of τ_g are shown with larger marker size.

As shown in the figure, the variability in residual tumor volume is higher for less radio-sensitive cases (Case I and Case III) compared to more radio-sensitive cases (Case II and Case IV). This confirms the previous observation that the tumor response model is highly sensitive to the radio-sensitivity parameter (α). Therefore, for each tumor case, there might be a threshold for α such that any α value larger than the threshold will result in the total removal of tumor cells. Furthermore, a lower value of τ_g leads to a higher number of remaining tumor cells and higher variability among the scenarios.

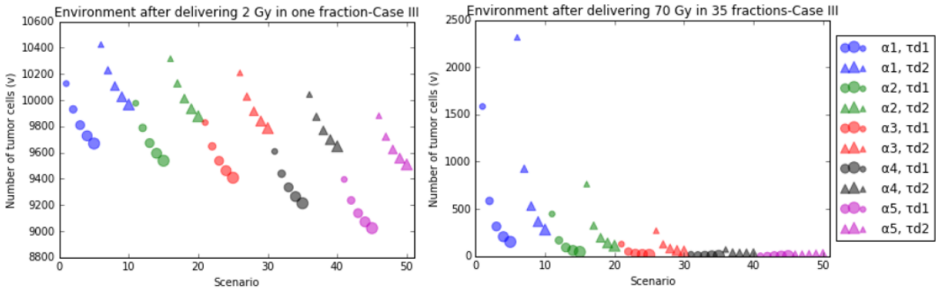
Moreover, having a smaller value of tumor decay (τ_d) resulted in a faster tumor volume reduction at each treatment stage. Overall, we observed that the variability among scenarios was similar with a different range of tumor volumes after delivering the first fraction. In assessing the final residual tumor volume, Case 1 resulted in the worst treatment outcome as it is a case of low radio-sensitivity and a high rate of proliferation. As expected, Case IV resulted in the best outcome and it is associated with high radio-sensitivity and slow proliferation. Furthermore, Case III resulted in a slightly worse treatment outcome than Case II, which



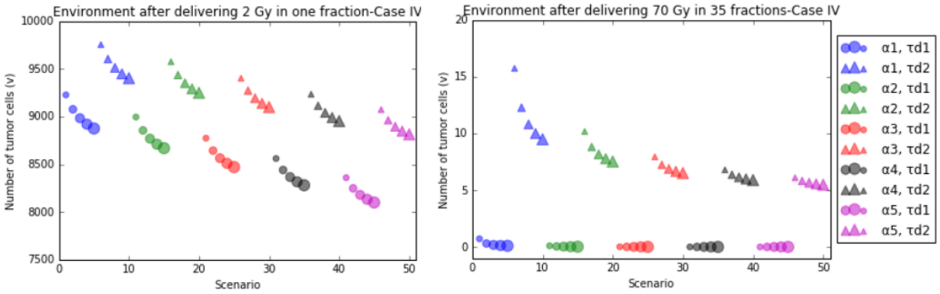
(a) Case I



(b) Case II



(c) Case III



(d) Case IV

Figure 10: Tumor volume variability in the RL environment for four tumor cases

indicates that the response model is more sensitive to radio-sensitivity than the tumor growth.

3.4. A case study on a clinical lung cancer cancer case

We evaluated the performance of our proposed RT treatment planning framework on an actual clinical non-small cell lung cancer (NSCLC) case obtained from the MD Anderson Cancer Center (MDACC),

Houston, TX. The patient went through a four-dimensional CT imaging as a part of a routine treatment simulation before starting the radiation therapy. A physician manually contoured the target volume and healthy structures on axial slices of the planning CT images. The anatomy was discretized into voxels of 2.5 mm (L) $\times$ 2.5 mm (W) $\times$ 2.5 mm (H). Table 5 lists the organs of interest, voxel counts of each organ, and the prescribed treatment protocol and requirements.

Table 5: Organs of interest, voxel counts of each organ, and Dose-volume requirements for the volumes of interest

Structure	Structure Type	Number of Voxels	Dose Requirements
Planning target volume (PTV)	Target	59,030	Volume receiving at least the prescription dose: $\geq 95\%$ Prescription: 70 Gy in 35 fractions
Heart	OAR	43,180	Volume receiving doses higher than 45 Gy: $\leq$65%
Total lung	OAR	287,616	Volume receiving doses higher than 20 Gy: $\leq$45%

We made the following assumptions to construct the RL environment. Following the clinical protocol, five fractionated radiation doses will be delivered to the target each week, skipping the treatment during the weekends to allow healthy tissues to recover. A total of 35 fractions, $N = 35$, will be delivered to complete the treatment. The optimal plan allows the total combined dose to be in the range between $d_l = 68$ Gy and $d_u = 72$ Gy (Roach et al., 2018) based on the tumor biological characteristics. We assumed the set of possible actions (i.e., fraction dose) of $A = \{1.8, 2.0, 2.2\}$ to include a 0.2 Gy deviation from the conventional fraction dose, which is reasonable in the ART fractionation scheme (Sonke et al., 2019; Roach et al., 2018).

Increasing the tumor's BED while keeping the OAR BED at a safe level can help improve tumor control without elevating the OAR toxicity. Therefore, our goal is to increase the BED of the tumor compared to the one from the reference plan, while keeping the OAR BED at most $+5\%$ from the reference plan. Also, the surviving fraction was capped at 0.01% to ensure the elimination of all tumor cells. The corresponding penalty coefficients in the reward function were determined by manual adjustments to achieve the desired goal of the treatment plan according to the treatment planner's preference.

Ranges of biological parameters for lung cancer were chosen based on the literature to set up the RL environment for tumor response during the treatment period. We assumed an exponential tumor growth rate of $\tau_g \in [10, 60]$ days (El Sharouni et al., 2003; Uzan and Nahum, 2012; Nahum and Uzan, 2012) and a tumor decay factor of $\tau_d \in [2, 6]$ days (Watanabe et al., 2016; Jeong et al., 2017), the uncertain ranges of $\alpha_p \in [0.20, 0.365]$ Gy^{-1} (Jeong et al., 2017; Uzan and Nahum, 2012) and $\alpha/\beta \in [4, 10]$ Gy (Santiago et al., 2016; Stuschke and Pöttgen, 2010; Jeong et al., 2017; Van Leeuwen et al., 2018). The standard value of oxygen enhancement ratio for lung cancer was considered as a constant value of $OER = 1.7$.

The total lung radiation toxicity is one of the most important metrics in determining the quality of the RT plan for lung cancer. In this study, we assumed the total lung as an OAR with $\alpha = 0.3$ Gy^{-1} and $\alpha^\phi/\beta^\phi = 3$ Gy (Seppenwoolde et al., 2003; Nahum and Kutcher, 2007; Bortfeld et al., 2015), OAR repopulation rate $\tau_g^\phi = 15$ days, and repair rates $\tau_r^\phi = 3.5$ days (Yang and Xing, 2005; Saberian et al., 2016) in the RL environment. We also considered a constant OAR dose sparing factor of $\theta = 0.7$ (Bortfeld et al., 2015).

At each episode of the training process, three values for α and τ_g, and two values for τ_d are randomly

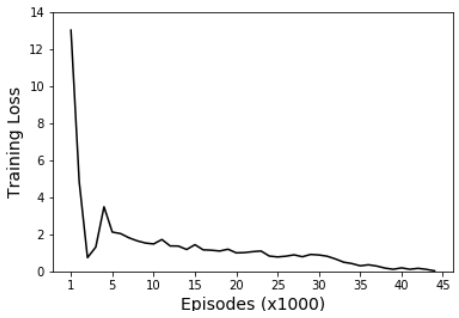


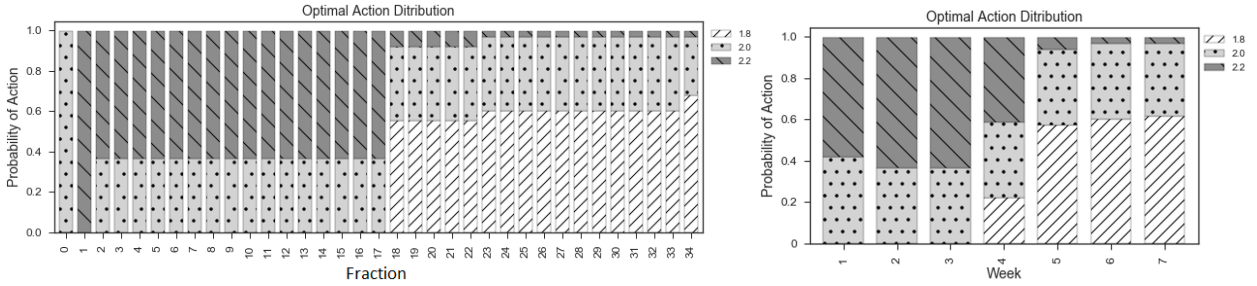
Figure 11: Training loss value in every 1000 episodes

selected from the defined ranges, which resulted in generating 50 tumor volume cases. At each time stage i within an episode (i.e., each fraction), 50 tumor volume cases and their associated $BED(s_i, a_i)$ and surviving fraction based on the current state s_i and selected action a_i were calculated and used in determining the state and immediate reward.

To find the optimal adaptation points, we need to determine the optimal action at each time epoch (i.e., the beginning of each week). Hence, we trained the RL model using Algorithm 1 to find the optimal Q-value function $Q^*(s, a)$, as a prediction of the reward function, with a minimum loss function value $L(w)$ in Equation (21). The model with the minimum loss value was achieved at 44000^{th} episode of training, which is the last episode before over-fitting happens. The network loss within an episode in every 1000 iterations is depicted in Figure 11. As shown in the figure, the network training loss has fluctuations at the beginning of training and it decreased as the episode increased. This behavior is reasonable as the RL model tries to learn over time to predict the Q-value function with the highest precision.

To find the optimal fractionation schedule considering the uncertainty in biological parameters, five hundred scenarios are generated based on different sets of selected biological parameters (i.e.,α, τ_g, τ_d) of a lung cancer patient. The best trained RL model is used to find the optimal fractionated plans among those scenarios. An optimal action (i.e., dose) was taken to obtain the optimal plan per fraction for each scenario. Figure 12a shows the distribution of optimal actions at each fraction. Note that a weekly fractionation is typically used for ART in clinical practice. Hence, we determined the distribution of optimal actions for each week by concatenating distributions of five fractions within a week (see Figure 12b). As shown in Figure 12, the most probable policy is to increase the fraction dose at the beginning of the treatment then decrease the dose at some point.

Using the weekly distribution of optimal actions, the expected dose amount over 500 scenarios for each week, the sum of the probability of each dose in action set multiplied by the dose, was calculated to determine the optimal fractionation scheme (i.e., $d_i = P(d_i = 2.2) \times 2.2 + P(d_i = 2) \times 2 + P(d_i = 1.8) \times 1.8$). Since changing the weekly fractionated dose by a small amount is not feasible in clinical practice, we considered week i as an adaptation point if the difference between the current dose regime and the projected dose for week i is greater than a threshold (i.e., $|d_i - d_{i-1}| \geq 0.03$ Gy). Once the adaptation points are identified, the ART weekly dose fractionation scheme is determined in such a way that the new weekly dose amount will be the average dose over all previous weekly doses since the last adaptation point. For example, the first identified adaptation point occurs at week 4 and the optimal ART



(a) Distribution of optimal actions for 35 fractions

(b) Distribution of optimal actions for 7 weeks

Figure 12: Distribution of optimal actions. (a) Distribution of optimal actions for 35 fractions. (b) Distribution of optimal actions for 7 weeks of treatment.

plan dose for the first three weeks is 2.13 which is calculated by averaging the optimal dose in week 1-3 (i.e., $(2.12+2.13+2.13)/3$). We used the conventional fractionation schedule with an equal weekly fractionated dose as a reference plan to compare with our proposed plan. Table 6 presents the weekly fractionation for the lung cancer patient based on the expected value of actions over the week, corrected values for ART, and the reference plan. The optimal ART schedule (*ART(1)*) is to have an initial radiation dose of 2.13 Gy for the first three weeks, and decrease it to 2.04 Gy at week four, and finally drop the dose to 1.86 Gy for the rest of the treatment. This implies that delivering a higher radiation dose at the beginning of the treatment will cause more damage to the tumor cells and this will change the dose requirement for the rest of the treatment to be lower.

Table 6: Weekly dose per fraction and total dose for the generated ART plan and the reference plan

	Weekly dose per fraction (Gy)							
	W1	W2	W3	W4	W5	W6	W7	Total Dose
Optimal fractionated plan	2.12	2.13	2.13	2.04	1.89	1.85	1.85	70.05
Optimal ART plan	2.13	2.13	2.13	2.04	1.86	1.86	1.86	70.05
Reference plan	2.00	2.00	2.00	2.00	2.00	2.00	2.00	70.00

The total dose of the proposed plan is slightly higher than the reference plan. This is because delivering a higher amount of radiation dose in a treatment will likely increase the total biological effective dose of the tumor and OARs cells. However, we can only increase the total dose by a certain amount (i.e., at most $+5\%$ of the reference plan) to maintain the desired range of OAR BED based on the assumed preferences. Figure 13 shows the box plot of the weekly tumor volume as a percentage of initial tumor volume among the assumed scenarios which has a decreasing trend during the course of treatment.

3.4.1. Plan evaluation: biological comparisons

To quantify the extent of potential biological benefits of the proposed approach, the final surviving fraction (SF) of the tumor, biological effective dose (BED) of the tumor and the OAR for reference plan, and the optimal ART plan were compared. Table 7 summarizes biological metrics resulted from the experiments on various scenarios for the reference plan and the proposed ART plan. The ART plan outperformed the reference plan for the tumor in terms of the BED by increasing the mean value by 2.01%, while the BED

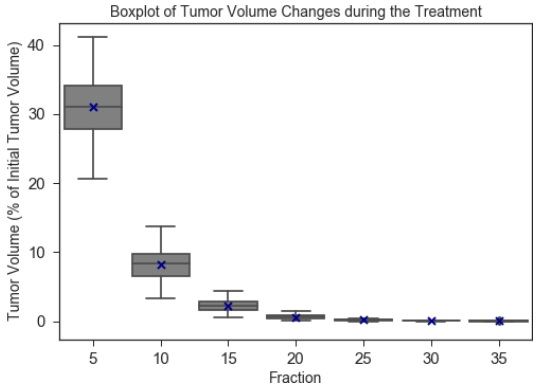


Figure 13: Box plot of the weekly tumor volume as a percentage of initial tumor volume among the assumed scenarios

of the OAR was slightly increased by 0.49%. Similarly, the ART plan significantly outperformed the reference plan on the probability of SF ($P(SF < 0.01\%)$), a 42.31% improvement. Both plans worked well in controlling the surviving fraction of the tumor.

Table 7: Biological measures for the optimal ART plan and the reference plan

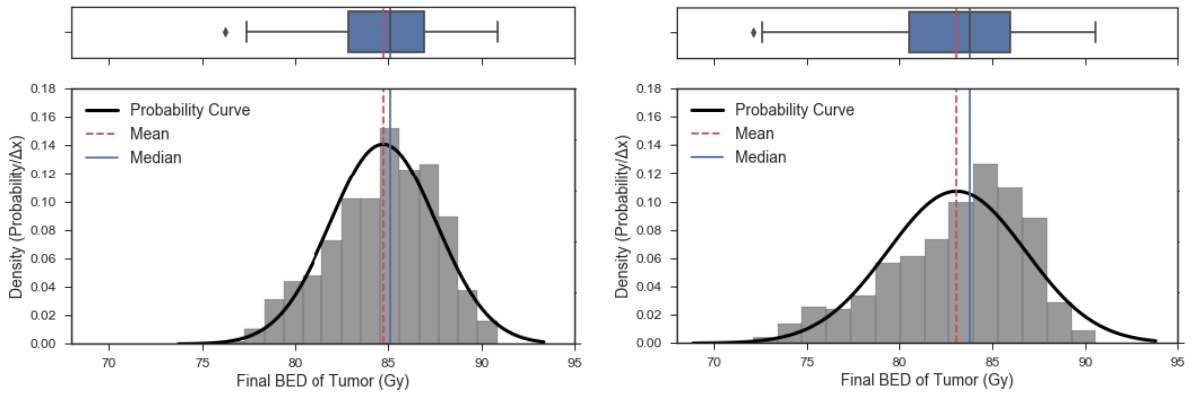
	BED^T	BED^Φ	$1 - SF(\%)$	$P(SF < 0.01\%)$
Optimal ART plan	84.74	26.63	99.98	0.37
	(95% CI [84.48,85.01])			
Reference plan	83.07	26.50	99.97	0.26
	(95% CI [82.74, 83.41])			

Measuring the outcome solely based on the mean BED value may hide some individual worst-case scenarios whose values are much lower than the desired value for the tumor, which can lead to undesirable effects on the treatment outcome. Therefore, it is important to reduce the variability of tumor BED values under different scenarios, where the RL framework can help address the issue.

Figure 14 shows the histogram, the estimated probability density function, and the box plot of the final BED of the tumor for each of the plans corresponding to 500 scenarios. The sum of the probability densities is equal to 1. The optimal ART plan resulted in the final BED^T distribution, which appears to follow a normal distribution with small variance and higher values around the average BED. As a comparison, the reference plan's distribution has a wider variance and is skewed left. Also, the ART plan has a shorter left tail distribution compared to the reference plan. This means that the ART plan will result in a smaller number of undesirable worst cases for BED^T. It indicates that the ART plan is more likely to produce a treatment plan with a greater final tumor BED than the reference plan.

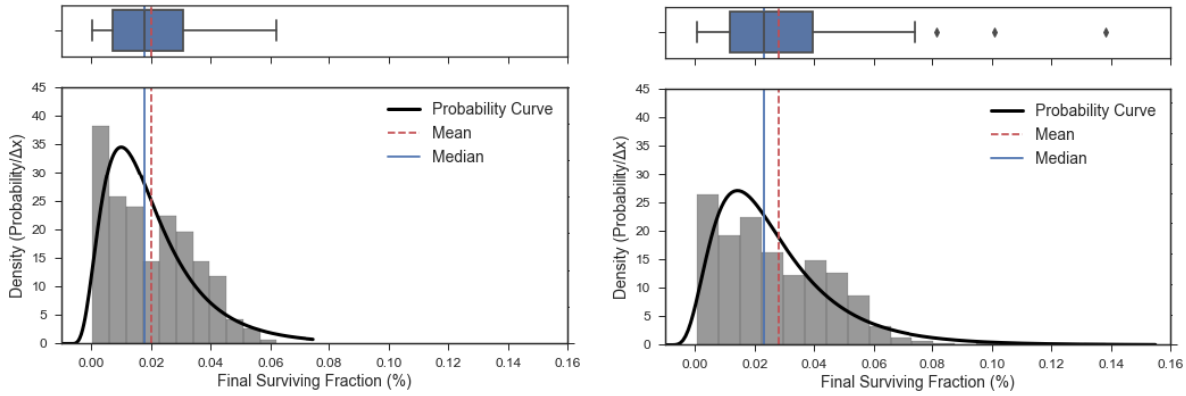
We further evaluate the variability in the solution for each plan using commonly used variability metrics in statistics, including median, range, and interquartile range (IQR). In Figure 14, mean and median values are marked with a red dash line and a blue line, respectively. We observed that the ART plan resulted in a smaller variability compared to the reference plan. First, the BED distribution of the ART plan has a narrower spread. Second, both the mean and median values of the ART plan are higher than those of the reference plan. Furthermore, the ART plan resulted in a smaller difference between the mean and median

values compared to the reference plan. Third, the IQR can be visualized using the box width in the box plot. We can see that the ART plan exhibited a 25% narrower IQR than the reference plan. Also, the range of tumor BED values is reduced by 21%. Overall, all variability measures of the optimal ART plan were lower than those of the reference plan. Hence, the ART plan outperformed the reference plan with respect to improving the final BED^T and reducing its variability in uncertain biological parameters.



(a) ART Plan (b) Reference Plan

Figure 14: Histogram, estimated probability density function (PDF), and box plot of the final tumor BED for the 500 generated scenarios based on (a) the ART plan, and (b) the Reference plan.



(a) ART plan (b) Reference plan

Figure 15: Histogram, estimated probability density function (PDF), and box plot of the final tumor SF for the 500 generated scenarios based on (a) the ART plan, and (b) the Reference plan.

We used the normal distribution for estimating the PDF for the BED values because it has the best fitting results compared with other well-known distributions (i.e., Lognormal, Gamma, and Beta). This can be explained using the central limit theorem (Montgomery and Runger, 2014) and the randomness of biological parameters. However, a lognormal distribution was used to estimate PDF of SF values because the relationship between the SF and the BED of the tumor is logarithmic. Figure 15 shows the histogram, the estimated probability density function, and the box plot of the final tumor SF from each plan.

The ultimate goal of treatment planning is to produce a plan whose final SF is minimized. As shown in the figure, the final SF distribution is skewed right. The long tail on the right side corresponds to less undesired cases having a large SF (i.e., worst cases). Comparing the two treatment plans in Figure 15, the ART plan reduced the right tail distribution of the final tumor SF as well as the upper quartile of the box plot. Hence, the ART plan outperformed the reference plan in terms of controlling the worst-case tumor SF. Moreover, the final SF values of the ART plan showed a tighter distribution and higher density around lower values compared to the reference plan. The ART plan also resulted in reducing the IQR and value ranges by 11.95% and 46.79%, respectively. Therefore, we claim that the ART plan has an advantage over the reference plan in developing treatment plans under the biological parameter uncertainty.

3.4.2. Effect of different decision making

We further investigate the effect of changing a planner's preference and assumptions on the final optimal fractionation schedule as well as the quality of treatment in terms of biological and dosimetric measures. Based on the patient's characteristics and the type of cancer, the treatment planner can set different priorities and goals in the RT treatment planning. For instance, one may want to develop a plan to ensure that the target is receiving the required dose by controlling the surviving fraction ($SF < \epsilon$) while reducing the radiation damage to OARs in terms of the BED. This can be done by assigning a higher weight on the OAR BED part of the reward function than the tumor BED while satisfying the surviving fraction. We expect that the OAR toxicity will improve as a result, and this leads to better patient recovery from radiotherapy. In this regard, we also explored the effect of changing OARs sparing factor (θ) in the BED formulation by finding the optimal plan based on two values of $\theta = 0.7$ and $\theta = 0.4$. Table 8 shows the optimal fractionation schedule, total dose, and OAR BED reduction percentage (ΔBED^{Φ}) based on this plan along with different values of θ. Compared to the conventional plan with a prescription dose of 70 Gy, the ART plan reduced the total dose by 2.14% and 0.71% based on the new preference (or priority) with $\theta = 0.7$ and $\theta = 0.4$, respectively. This reduction in total radiation dose may not seem to be significant, but this will result in better OAR sparing with a lower amount of radiation. In both plans, the total treatment dose is decreased because the priority was made to reduce the OAR BED while having the same or better surviving fraction of the tumor cells and limiting the BED in target volume to a clinically desired level. As we can see from Table 8, the OAR BED is improved by 9.26% and 16.12% using $theta = 0.7$ and $theta = 0.4$, respectively, compared to the reference plan on the tumor.

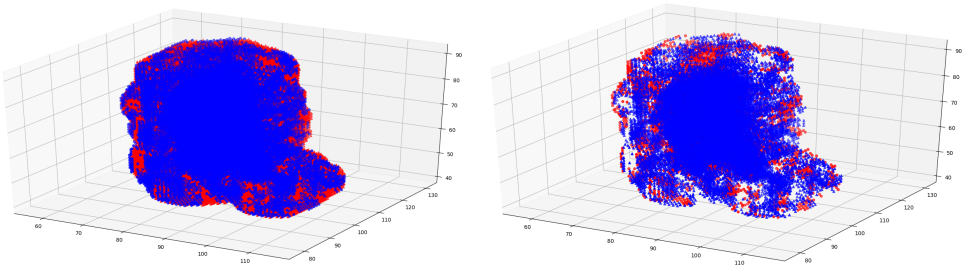
Furthermore, the ART with $\theta = 0.7$ resulted in a more aggressive plan than the one with $\theta = 0.4$. This is because the OAR will likely reduce the radiation exposure by lowering the value of θ. Thus, a lower penalty value was assigned to the OAR BED in the reward function and we can see a higher BED^{Φ} depletion even with a smaller amount of dose reduction. The results from this experiment show that the proposed treatment planning framework is effective in developing a plan that preserves more healthy cells. Therefore, the planner can develop the best plan according to the patient characteristics and the physician's preference.

3.4.3. Plan evaluation: dose-volume results

Dose-volume metrics are commonly used to evaluate the treatment plan quality. For the purpose of simulating the ART procedure in this section, treatment plans were adapted to the patient volumetric changes

Table 8: Weekly dose per fraction and OAR BED based on the new decision making preference and two values of θ

	Weekly dose per fraction (Gy)								
	W1	W2	W3	W4	W5	W6	W7	Total Dose	ΔBED^{Φ}
$\theta = 0.7$	2.00	2.00	1.90	1.90	1.90	2.00	2.00	68.50	-9.26%
$\theta = 0.4$	2.00	2.00	2.00	2.00	1.95	1.95	2.00	69.50	-16.12 %



(a) First adaptation point (b) Second adaptation point

Figure 16: The simulated residual tumor volume (blue points) and the removed voxels (red points) at (a) the first adaptation point (beginning of week 4), and (b) the second adaptation point (beginning of week 5).

at each adaptation point. Planning target volumes (PTVs) were generated analogously on the basis of the average estimated residual tumor volume over all scenarios using the proposed tumor response model. We removed the tumor cells receiving a dose higher than the tolerance threshold from k outer layers (i.e., *2.5mm* per layer) of the tumor volume until achieving the estimated residual tumor volume at each adaptation point. Figure 16 shows the residual tumor volume and the removed voxels at each adaptation point after the first iteration.

To evaluate the OAR toxicity of the generated plans, the dose-volume metrics were calculated for the heart and total lung for three plans: the reference plan, ART(I) with $\theta = 0.7$ (see Section 3.4), and ART(II) with $\theta = 0.7$ (see Section 3.4.2). For the heart, V_{45} and D_2 were measured to examine the level of a high dose of radiation, which is critical for a serial organ. Also, V_{20} and mean dose were measured to account for the average spread of radiation dose in the total lung, which is a parallel organ. Table 9 summarizes these dose-volume metrics for the proposed plan and the reference plan. Note that all dose-volume values were normalized to have 95% of PTV receiving at least 70 Gy for the comparison purpose. We make the following observations regarding the ART plans in comparison to the reference plan based on the table. First, all metrics of the ART plans for the heart and total lung were lower compared to the reference plan. Second, ART(I) reduced V_{45} and D_2 of the heart by 3.78% and 0.10%, respectively. Third, ART(II) decreased V_{45} and D_2 of the heart by 6.01% and 2.48%, respectively. Finally, both ART(I) and ART(II) reduced V_{20} of total lung by 3.56% and 4.28%, and mean lung dose by 2.14% and 5.63%, respectively.

Overall, the ART plans outperformed the reference plan by reducing OAR toxicity. ART(II) showed slightly lower values on all measured dose-volume metrics, which is a direct result of the planner's preference to keep the OAR toxicity at the desired level.

Table 9: Comparison of dose-volume metrics for the optimal ART plans and the reference plan

	Optimal ART plan		
	ART(I)	ART(II)	Reference plan
Heart			
$V_{45}(\%)$	11.69	11.42	12.15
D_2(Gy)	71.84	70.13	71.91
Total Lung			
$V_{20}(\%)$	30.85	30.62	31.99
$Mean$(Gy)	18.77	18.10	19.18

4. Conclusion

Multiple studies demonstrate the benefits of ART in terms of healthy tissue sparing and tumor cell reduction. Considering the biological features of the tumor and healthy organs in treatment planning and adapting the plan to biological changes during the course of treatment is the key motivation for ART. In this paper, we developed a novel biological response model that incorporates important biological factors for tumors and healthy organs to predict the tumor volume regressions during the treatment. Then, we proposed an automated framework using Reinforcement Learning and an optimization method to find the optimal adaptation points for ART and dynamically adapt the plan considering the tumor's uncertain biological response over time. We aimed to achieve a plan with a maximum final tumor control while minimizing or maintaining the OARs toxicity levels by finding the actions to maximize the RL reward function. After finding the adaptation points, an optimization model was solved to find the optimal beamlet intensities satisfying clinical dose-volume requirements for the patient based on the predicted tumor volume and the proposed fractionation dose determined by the RL approach at each adaptation point.

We evaluated the performance of our proposed RT treatment planning framework using a clinical non-small cell lung cancer (NSCLC) case. We also analyzed the proposed approach under various assumptions and decision priorities to see the trade-off in terms of tumor coverage and OARs toxicity. The proposed ART plans were assessed and compared with the reference plan (i.e, equal dose fractionation) based on biological and dose-volume metrics. The results showed that the proposed approach can help the treatment planner to achieve a robust solution under high levels of uncertainty in the biological parameters. Using the proposed method, it is not only possible to control the biological aspect of the treatment and tumor biological response uncertainty, but it also helps satisfy dose-volume requirements and clinical limits of the treatment. Furthermore, the proposed reinforcement learning framework can help achieve a robust solution under uncertainty in the biological parameters, while reducing the variability in the solution and improving the control on the worst cases. The proposed approach enables the physicians to find appropriate personalized ART plan in terms of fractionation dose and the timing of the adaptations. Two major benefits of this approach are to reduce the time and effort to collect large-scale datasets and avoid the need for taking expensive CT images at each visit. The proposed RL approach can be easily applied to various types of cancer, ART methods, and different treatment planning preferences.

For future work, the predicted tumor response to the radiation should be validated or corrected by obtaining actual imaging data for every visit during the course of RT treatment (e.g., at the determined adaptation

points). This will help further enhance the proposed approach. Furthermore, the proposed framework can be extended by adjusting the reinforcement learning environment to account for other radiation therapy treatment planning problems, such as radiation beam angle optimization.

References

Ashrafi, H. and Thiele, A. C. (2021). A study of robust portfolio optimization with european options using polyhedral uncertainty sets. *Operations Research Perspectives*, 8:100178.

Bai, X., Lim, G., Grosshans, D., R., M., and Cao, W. (2020). A biological effect-guided optimization approach using beam distal-edge avoidance for intensity-modulated proton therapy. *Medical Physics*, 49(9):3816–3825.

Bazaraa, M. S., Sherali, H. D., and Shetty, C. M. (2013). *Nonlinear programming: theory and algorithms*. John Wiley & Sons.

Belfatto, A., Riboldi, M., Ciardo, D., Cecconi, A., Lazzari, R., Jereczek, B. A., Orecchia, R., Baroni, G., and Cerveri, P. (2016). Adaptive mathematical model of tumor response to radiotherapy based on CBCT data. *IEEE Journal of Biomedical and Health Informatics*, 20(3):802–809.

Berkovic, P., Paelinck, L., Lievens, Y., Gulyban, A., Goddeeris, B., Derie, C., Surmont, V., Neve, W. D., and Vandecasteele, K. (2015). Adaptive radiotherapy for locally advanced non-small cell lung cancer, can we predict when and for whom? *Acta Oncologica*, 54(9):1438–1444.

Bibault, J.-E., Fumagalli, I., Ferté, C., Chargari, C., Soria, J.-C., and Deutsch, E. (2013). Personalized radiation therapy and biomarker-driven treatment strategies: a systematic review. *Cancer Metastasis Reviews*, 32(3-4):479–492.

Bortfeld, T., Ramakrishnan, J., Tsitsiklis, J. N., and Unkelbach, J. (2015). Optimization of radiation therapy fractionation schedules in the presence of tumor repopulation. *INFORMS Journal on Computing*, 27(4):788–803.

Brenner, D. J., Hlatky, L. R., Hahnfeldt, P. J., Hall, E. J., and Sachs, R. K. (1995). A convenient extension of the linear-quadratic model to include redistribution and reoxygenation. *International Journal of Radiation Oncology, Biology, Physics*, 32(2):379–390.

Carlson, D. J., Stewart, R. D., and Semenenko, V. A. (2006). Effects of oxygen on intrinsic radiation sensitivity: A test of the relationship between aerobic and hypoxic linear-quadratic (LQ) model parameters a. *Medical Physics*, 33(9):3105–3115.

Coronato, A., Naeem, M., De Pietro, G., and Paragliola, G. (2020). Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964.

Dial, C., Weiss, E., Siebers, J. V., and Hugo, G. D. (2016). Benefits of adaptive radiation therapy in lung cancer as a function of replanning frequency. *Medical Physics*, 43(4):1787.

Douglas, B. and Fowler, J. (2012). The effect of multiple small doses of x rays on skin reactions in the mouse and a basic interpretation. *Radiation Research*, 178(2):AV125–AV138.

El Sharouni, S. Y., Kal, H., and Battermann, J. (2003). Accelerated regrowth of non-small-cell lung tumours after induction chemotherapy. *British Journal of Cancer*, 89(12):2184–2189.

Fowler, J. F. (1989). The linear-quadratic formula and progress in fractionated radiotherapy. *The British Journal of Radiology*, 62(740):679–694. PMID: 2670032.

Fowler, J. F. (2001). Biological factors influencing optimum fractionation in radiation therapy. *Acta Oncologica*, 40(6):712–717.

Fox, I. and Wiens, J. (2019). Reinforcement learning for blood glucose control: Challenges and opportunities.

Futoma, J., Lin, A., Sendak, M., Bedoya, A., Clement, M., O'Brien, C., and Heller, K. (2018). Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks.

Ghate, A. (2011). Dynamic Optimization in Radiotherapy. In *Transforming Research into Action*, INFORMS TutORials in Operations Research, pages 60–74. INFORMS.

Glavic, M., Fonteneau, R., and Ernst, D. (2017). Reinforcement learning for electric power system decision and control: Past considerations and perspectives. *IFAC-PapersOnLine*, 50(1):6918–6927.

Guckenberger, M., Richter, A., Wilbert, J., Flentje, M., and Partridge, M. (2011). Adaptive radiotherapy for locally advanced non–small-cell lung cancer does not underdose the microscopic disease and has the potential to increase tumor control. *International Journal of Radiation Oncology* Biology* Physics*, 81(4):e275–e282.

Guidi, G., Maffei, N., Meduri, B., D'Angelo, E., Mistretta, G., Ceroni, P., Ciarmatori, A., Bernabei, A., Maggi, S., Cardinali, M., et al. (2016). A machine learning tool for re-planning and adaptive rt: a multicenter cohort investigation. *Physica Medica*, 32(12):1659–1666.

Hall, E. J. and Giaccia, A. J. (2006). *Radiobiology for the Radiologist*, volume 6. Lippincott Williams & Wilkins.

Jeong, J., Oh, J. H., Sonke, J.-J., Belderbos, J., Bradley, J. D., Fontanella, A. N., Rao, S. S., and Deasy, J. O. (2017). Modeling the cellular response of lung cancer to radiation therapy for a broad range of fractionation schedules. *Clinical Cancer Research*, 23(18):5469–5479.

Jeong, J., Shoghi, K., and Deasy, J. (2013). Modelling the interplay between hypoxia and proliferation in radiotherapy tumour response. *Physics in Medicine & Biology*, 58(14):4897.

Kawata, Y., Arimura, H., Ikushima, K., Jin, Z., Morita, K., Tokunaga, C., Yabu-uchi, H., Shioyama, Y., Sasaki, T., Honda, H., et al. (2017). Impact of pixel-based machine-learning techniques on automated frameworks for delineation of gross tumor volume regions for stereotactic body radiation therapy. *Physica Medica*, 42:141–149.

Khaled, S. and Held, K. D. (2012). Radiation biology: a handbook for teachers and students.

Kim, M., Ghate, A., and Phillips, M. H. (2009). A Markov decision process approach to temporal modulation of dose fractions in radiation therapy planning. *Physics in Medicine and Biology*, 54(14):4455–4476.

Kim, M., Ghate, A., and Phillips, M. H. (2012). A stochastic control formalism for dynamic biologically conformal radiation therapy. *European Journal of Operational Research*, 219(3):541–556.

Kober, J., Bagnell, J. A., and Peters, J. (2013). Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274.

Lawrence, Y. R., Werner-Wasik, M., and Dicker, A. P. (2008). Biologically conformal treatment: biomarkers and functional imaging in radiation oncology. *Future Oncology*.

Lee, H., Ahn, Y. C., Oh, D., Nam, H., Kim, Y. I., and Park, S. Y. (2014). Tumor volume reduction rate measured during adaptive definitive radiation therapy as a potential prognosticator of locoregional control in patients with oropharyngeal cancer. *Head & Neck*, 36(4):499–504.

Li, Y., Zhang, W., Wang, C.-X., Sun, J., and Liu, Y. (2020). Deep reinforcement learning for dynamic spectrum sensing and aggregation in multi-channel wireless networks. *IEEE Transactions on Cognitive Communications and Networking*, 6(2):464–475.

Lim, G. J., Kardar, L., Ebrahimi, S., and Cao, W. (2020). A risk-based modeling approach for radiation therapy treatment planning under tumor shrinkage uncertainty. *European Journal of Operational Research*, 280(1):266–278.

Ling, Y., Hasan, S. A., Datla, V., Qadir, A., Lee, K., Liu, J., and Farri, O. (2017). Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study. In *Machine Learning for Healthcare Conference*, pages 271–285.

Liu, X., Yu, W., Liang, F., Griffith, D., and Golmie, N. (2021). On deep reinforcement learning security for industrial internet of things. *Computer Communications*, 168:20–32.

Liu, Z., Yao, C., Yu, H., and Wu, T. (2019). Deep reinforcement learning with its application for lung cancer detection in medical internet of things. *Future Generation Computer Systems*, 97:1–9.

McMahon, S. J. (2018). The linear quadratic model: usage, interpretation and challenges. *Physics in Medicine & Biology*, 64(1):01TR01.

Mnih, V., Kavukcuoglu, K., Silver, D., Graves, A., Antonoglou, I., Wierstra, D., and Riedmiller, M. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

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

Montgomery, D. C. and Runger, G. C. (2014). *Applied statistics and probability for engineers*. Wiley.

Naeem, M., Rizvi, S. T. H., and Coronato, A. (2020). A gentle introduction to reinforcement learning and its application in different fields. *IEEE Access*.

Nahum, A. and Kutcher, G. (2007). Biological evaluation of treatment plans. In Mayles, P., Nahum, A., and Rosenwald, J. C., editors, *Handbook of radiotherapy physics theory and practice*, chapter 36, pages 731–771. Taylor & Francis, London, UK.

Nahum, A. E. and Uzan, J. (2012). (radio) biological optimization of external-beam radiotherapy. *Computational and Mathematical Methods in Medicine*, 2012.

Paul-Gilloteaux, P., Potiron, V., Delpon, G., Supiot, S., Chiavassa, S., Paris, F., and Costes, S. V. (2017). Optimizing radiotherapy protocols using computer automata to model tumour cell death as a function of oxygen diffusion processes. *Scientific Reports*, 7(1):1–14.

Petersen, B. K., Yang, J., Grathwohl, W. S., Cockrell, C., Santiago, C., An, G., and Faissol, D. M. (2018). Precision medicine as a control problem: Using simulation and deep reinforcement learning to discover adaptive, personalized multi-cytokine therapy for sepsis. *arXiv preprint arXiv:1802.10440*.

Ramella, S., Fiore, M., Silipigni, S., Zappa, M. C., Jaus, M., Alberti, A. M., Matteucci, P., Molfese, E., Cornacchione, P., Greco, C., Trodella, L., Ippolito, E., and D'Angelillo, R. M. (2017). Local control and toxicity of adaptive radiotherapy using weekly CT imaging: Results from the LARTIA trial in stage III NSCLC. *Journal of Thoracic Oncology*, 12(7):1122–1130.

Roach, M. C., Bradley, J. D., and Robinson, C. G. (2018). Optimizing radiation dose and fractionation for the definitive treatment of locally advanced non-small cell lung cancer. *Journal of Thoracic Disease*, 10(Suppl 21):S2465.

Saberian, F., Ghate, A., and Kim, M. (2016). Optimal fractionation in radiotherapy with multiple normal tissues. *Mathematical Medicine and Biology: a Journal of the IMA*, 33(2):211–252.

Saka, B., Rardin, R. L., Langer, M. P., and Dink, D. (2011). Adaptive intensity modulated radiation therapy planning optimization with changing tumor geometry and fraction size limits. *IIE Transactions on Healthcare Systems Engineering*, 1(4):247–263.

Santiago, A., Barczyk, S., Jelen, U., Engenhart-Cabillic, R., and Wittig, A. (2016). Challenges in radiobiological modeling: can we decide between lq and lq-l models based on reviewed clinical nsclc treatment outcome data? *Radiation Oncology*, 11(1):67.

Scott, J. G., Berglund, A., Schell, M. J., Mihaylov, I., Fulp, W. J., Yue, B., Welsh, E., Caudell, J. J., Ahmed, K., Strom, T. S., et al. (2017). A genome-based model for adjusting radiotherapy dose (gard): a retrospective, cohort-based study. *The Lancet Oncology*, 18(2):202–211.

Seppenwoolde, Y., Lebesque, J. V., De Jaeger, K., Belderbos, J. S., Boersma, L. J., Schilstra, C., Henning, G. T., Hayman, J. A., Martel, M. K., and Ten Haken, R. K. (2003). Comparing different ntcp models that predict the incidence of radiation pneumonitis. *International Journal of Radiation Oncology* Biology* Physics*, 55(3):724–735.

Siegel, R. L., Miller, K. D., and Jemal, A. (2019). Cancer statistics, 2019. *CA: A Cancer Journal for Clinicians*, 69(1):7–34.

Sonke, J.-J., Aznar, M., and Rasch, C. (2019). Adaptive radiotherapy for anatomical changes. *Seminars in Radiation Oncology*, 29(3):245–257.

South, C., Partridge, M., and Evans, P. (2008). A theoretical framework for prescribing radiotherapy dose distributions using patient-specific biological information. *Medical physics*, 35(10):4599–4611.

Steel, G. G., McMillan, T. J., and Peacock, J. (1989). The 5Rs of radiobiology. *International Journal of Radiation Biology*, 56(6):1045–1048.

Stember, J. and Shalu, H. (2020). Deep reinforcement learning to detect brain lesions on mri: a proof-of-concept application of reinforcement learning to medical images. *arXiv preprint arXiv:2008.02708*.

Stuschke, M. and Pöttgen, C. (2010). Altered fractionation schemes in radiotherapy. In *Controversies in the Treatment of Lung Cancer*, volume 42, pages 150–156. Karger Publishers.

Surucu, M., Shah, K. K., Mescioglu, I., Roeske, J. C., Small Jr, W., Choi, M., and Emami, B. (2016). Decision trees predicting tumor shrinkage for head and neck cancer: Implications for adaptive radiotherapy. *Technology in Cancer Research & Treatment*, 15(1):139–145.

Tejedor, M., Woldaregay, A. Z., and Godtliebsen, F. (2020). Reinforcement learning application in diabetes blood glucose control: A systematic review. *Artificial Intelligence in Medicine*, 104:101836.

Tseng, H.-H., Luo, Y., Cui, S., Chien, J.-T., Ten Haken, R. K., and El Naqa, I. (2017). Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical Physics*, 44(12):6690–6705.

Unkelbach, J., Craft, D., Hong, T., Papp, D., Ramakrishnan, J., Salari, E., Wolfgang, J., and Bortfeld, T. (2014). Exploiting tumor shrinkage through temporal optimization of radiotherapy. *Physics in Medicine and Biology*, 59(12):3059–3079.

Uzan, J. and Nahum, A. (2012). Radiobiologically guided optimisation of the prescription dose and fractionation scheme in radiotherapy using biosuite. *The British Journal of Radiology*, 85(1017):1279–1286.

van de Schoot, A. J., de Boer, P., Visser, J., Stalpers, L. J. A., Rasch, C. R. N., and Bel, A. (2017). Dosimetric advantages of a clinical daily adaptive plan selection strategy compared with a non-adaptive strategy in cervical cancer radiation therapy. *Acta Oncologica*, 56 5:667–674.

Van Hasselt, H., Guez, A., and Silver, D. (2016). Deep reinforcement learning with double Q-learning. In *Thirtieth AAAI conference on artificial intelligence*.

Van Leeuwen, C., Oei, A., Crezee, J., Bel, A., Franken, N., Stalpers, L., and Kok, H. (2018). The alfa and beta of tumours: a review of parameters of the linear-quadratic model, derived from clinical radiotherapy studies. *Radiation Oncology*, 13(1):1–11.

Veresezan, O., Troussier, I., Lacout, A., Kreps, S., Maillard, S., Toulemonde, A., Marcy, P.-Y., Huguet, F., and Thariat, J. (2017). Adaptive radiation therapy in head and neck cancer for clinical practice: state of the art and practical challenges. *Japanese Journal of Radiology*, 35(2):43–52.

Watanabe, Y., Dahlman, E. L., Leder, K. Z., and Hui, S. K. (2016). A mathematical model of tumor growth and its response to single irradiation. *Theoretical Biology and Medical Modelling*, 13(1):6.

Wein, L. M., Cohen, J. E., and Wu, J. T. (2000). Dynamic optimization of a linear–quadratic model with incomplete repair and volume-dependent sensitivity and repopulation. *International Journal of Radiation Oncology • Biology • Physics*, 47(4):1073–1083.

Wen, Z., O'Neill, D., and Maei, H. (2015). Optimal demand response using device-based reinforcement learning. *IEEE Transactions on Smart Grid*, 6(5):2312–2324.

Withers, H. R. (1975). The four R's of radiotherapy. In *Advances in radiation biology*, volume 5, pages 241–271. Elsevier.

Wouters, B. G. (2009). Cell death after irradiation: how, when and why cells die. *Basic Clinical Radiobiology*, page 27.

Yang, Y. and Xing, L. (2005). Optimization of radiotherapy dose-time fractionation with consideration of tumor specific biology. *Medical Physics*, 32(12):3666–3677.

Yousefi, S., Amrollahi, F., Amgad, M., Dong, C., Lewis, J. E., Song, C., Gutman, D. A., Halani, S. H., Vega, J. E. V., Brat, D. J., et al. (2017). Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models. *Scientific Reports*, 7(1):1–11.

Yu, C., Dong, Y., Liu, J., and Ren, G. (2019a). Incorporating causal factors into reinforcement learning for dynamic treatment regimes in hiv. *BMC medical informatics and decision making*, 19(2):19–29.

Yu, C., Liu, J., and Nemati, S. (2019b). Reinforcement learning in healthcare: A survey. *arXiv preprint arXiv:1908.08796*.

Zaghian, M., Cao, W., Liu, W., Kardar, L., Randeniya, S., Mohan, R., and Lim, G. (2017). Comparison of linear and nonlinear programming approaches for "worst case dose" and "minmax" robust optimization of intensity-modulated proton therapy dose distributions. *Journal of Applied Clinical Medical Physics*, 18(2):15–25.

Zarepisheh, M., Long, T., Li, N., Tian, Z., Romeijn, H. E., Jia, X., and Jiang, S. B. (2014). A dvh-guided imrt optimization algorithm for automatic treatment planning and adaptive radiotherapy replanning. *Medical Physics*, 41(6Part1):061711.

Zhang, P., Rimner, A., Yorke, E., Hu, J., Ravindranath, B., Mageras, G., and Deasy, J. (2015). Predicting spatial distribution of residual tumor post radiation therapy based on pretreatment pet/ct for locally advanced non-small cell lung cancer. *International Journal of Radiation Oncology• Biology• Physics*, 93(3):E557.

Zheng, Y., Singh, H., Zhao, L., Ramirez, E. V., Rana, S., Prabhu, K., Doh, L. S., and Larson, G. L. (2015). Adaptive radiation therapy for lung cancer using uniform scanning proton beams: Adaptation strategies, practical considerations, and clinical outcomes. *International Journal of Radiation Oncology • Biology • Physics*, 93(3):S29.

Zhu, X., Ge, Y., Li, T., Thongphiew, D., Yin, F.-F., and Wu, Q. J. (2011). A planning quality evaluation tool for prostate adaptive IMRT based on machine learning. *Medical Physics*, 38(2):719–726.

Appendix A. Action set construction algorithm

Algorithm 2 is developed to determine the appropriate action set of the reinforcement learning method for any type of cancer.

Algorithm 2 Action set construction method

Input: d, $\underline{d}$, $\bar{d}$ and Δ

if $0 < \bar{d} - d \leq \Delta \;\; and \;\; 0 < d - \underline{d} \leq \Delta$ **then**

$a \in \left\{\underline{d}, d, \bar{d}\right\}$

else if $\bar{d} - d > \Delta \;\; and \;\; 0 \leq d - \underline{d} \leq \Delta$ **then**

$a \in \left\{\underline{d}, d, d + \frac{|\bar{d}-d|}{p}, ..., d + i(\frac{|\bar{d}-d|}{p})\right\}, \forall i = 1, 2, ..., p$ and $p = \left\lceil \frac{|\bar{d}-d|}{\Delta} \right\rceil$

else if $d - \underline{d} > \Delta \;\; and \;\; 0 \leq \bar{d} - d \leq \Delta$ **then**

$a \in \left\{d - i(\frac{|\underline{d}-d|}{p}), ..., d - (\frac{|\underline{d}-d|}{p}), d, \bar{d}\right\}, \forall i = 1, 2, ..., p$ and $p = \left\lceil \frac{|\underline{d}-d|}{\Delta} \right\rceil$

else if $\underline{d} - d > \Delta \;\; and \;\; \bar{d} - d > \Delta$ **then**

$a \in \left\{d - i(\frac{|\underline{d}-d|}{\underline{p}}), ..., d - (\frac{|\underline{d}-d|}{\underline{p}}), d, d + \frac{|\bar{d}-d|}{\bar{p}}, ..., d + j(\frac{|\bar{d}-d|}{\bar{p}})\right\}$,

$\forall i = 1, 2, ..., \underline{p}, \; \forall j = 1, 2, ..., \bar{p}, \; \underline{p} = \left\lceil \frac{|\underline{d}-d|}{\Delta} \right\rceil$ and $\bar{p} = \left\lceil \frac{|\bar{d}-d|}{\Delta} \right\rceil$

else

$a \in \{d - \Delta, d, d + \Delta\}$

end

Appendix B. Proof of Theorem 2

The BED function in (16) is a quadratic function of d_i. It is continuous and twice differentiable. The first derivative of $BEDi^T(d_i)$ with respect to d_i is strictly positive for all $d_i \geq 0$:

$$\frac{\partial BEDi^T(d_i)}{\partial d_i} = \frac{OER\, u_{i-1} + m_{i-1}}{OER\, v_{i-1}} + 2\left(\frac{OER^2\, u_{i-1} + m_{i-1}}{OER^2\, \alpha/\beta\, v_{i-1}}\right) d_i > 0.$$

Also, the second derivative of $BEDi^T(d_i)$ with respect to d_i is strictly positive for all d_i:

$$\frac{\partial^2 BEDi^T(d_i)}{\partial d_i^2} = 2\left(\frac{OER^2\, u_{i-1} + m_{i-1}}{OER^2\, \alpha/\beta\, v_{i-1}}\right) > 0.$$

Hence, $BEDi^T(d_i)$ is strictly convex and it is an increasing function for $d_i \geq 0$.

Next, we show the existence of a lower bound (d_l) on dose d_i. For the notational convenience, we use the following abstract form of $BEDi^T(d_i)$ function, $ad_i^2 + bd_i + c$, where a, b, and c are defined as follows:

$$a = \frac{OER^2\, u_{i-1} + m_{i-1}}{OER^2\, \alpha/\beta\, v_{i-1}},$$

$$b = \frac{OER\, u_{i-1} + m_{i-1}}{OER\, v_{i-1}},$$

$$c = -\left(\frac{u_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha\, \tau_g} + \left(\frac{w_{i-1}}{v_{i-1}}\right)\frac{\Delta t_i}{\alpha\, \tau_d}.$$

The lower bound d_l can be found by examining the roots of $ad_i^2 + bd_i + c = 0$. The roots of a quadratic function are given by $d_i = \frac{-b \pm \sqrt{\delta}}{2a}$, where $\delta = b^2 - 4ac$. If $\delta \leq 0$, then the $BED_i^T(d_i)$ is non-negative for any $d_i \geq 0$. If $\delta > 0$, then there are two roots for $BED_i^T(d_i) = 0$. Since the first derivative of the function is equal to zero ($\frac{\partial BED_i^T(d_i)}{\partial d_i} = 0$) for a negative d_i, at least one of the roots must be negative. For a positive δ, there are two possibilities: (1) no positive root if $BED_i^T(0) \geq 0$ or (2) a positive root if $BED_i^T(0) < 0$. In the former case, the c value is positive and we have $\frac{u_{i-1}}{w_{i-1}} \leq \frac{\tau_g}{\tau_d}$. If we consider two negative roots as d_1 and d_2 where $d_1 > d_2$, the $BED_i^T(d_i)$ is non-negative for any $d_i \geq d_1$; since $d_1 < 0$ we conclude that $BED_i^T(d_i)$ is non-negative for any $d_i \geq 0$. In case (2), the c value is negative and we have $\frac{u_{i-1}}{w_{i-1}} > \frac{\tau_g}{\tau_d}$. For the positive root as in d_1, $BED_i^T(d_i)$ is non-negative for any $d_i \geq d_1$. Therefore, the lower bound is $d_l = \frac{-b+\sqrt{\delta}}{2a}$ for this case.