Mathematical Biosciences 293 (2017) 11–20

Contents lists available at ScienceDirect

Mathematical Biosciences

journal homepage: www.elsevier.com/locate/mbs

Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment

Regina Padmanabhan[a], Nader Meskin[a,*], Wassim M. Haddad[b]

[a] The Department of Electrical Engineering, Qatar University, Qatar
[b] The School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, 30332-0150, USA

ARTICLE INFO

Article history:
Received 28 June 2016
Revised 8 August 2017
Accepted 9 August 2017
Available online 16 August 2017

Keywords:
Active drug dosing
Chemotherapy control
Reinforcement learning

ABSTRACT

The increasing threat of cancer to human life and the improvement in survival rate of this disease due to effective treatment has promoted research in various related fields. This research has shaped clinical trials and emphasized the necessity to properly schedule cancer chemotherapy to ensure effective and safe treatment. Most of the control methodologies proposed for cancer chemotherapy scheduling treatment are model-based. In this paper, a reinforcement learning (RL)-based, model-free method is proposed for the closed-loop control of cancer chemotherapy drug dosing. Specifically, the *Q*-learning algorithm is used to develop an optimal controller for cancer chemotherapy drug dosing. Numerical examples are presented using simulated patients to illustrate the performance of the proposed RL-based controller.

1. INTRODUCTION

Cancer is the common name that is given to a group of diseases that involve the repeated and uncontrolled division and spreading of abnormal cells. These abnormal tissues are called tumors. According to the Cancer Facts and Figures-2015 report published by the American Cancer Society, the five-year relative survival rate for several types of cancer diagnosed from the years 2004 to 2010 has been improved significantly [1]. The report highlights the competency in early diagnosis and enhancement in treatment methods as two main factors that contribute to the reduced morbidity and mortality rate.

The treatment schedule and drug dose vary according to the stage of the tumor, the weight of the patient, the white blood cell levels (immunity), concurrent illness, and age of the patient. Accordingly, the clinicians follow certain established standards in order to deduce the type of therapy and drug dose for each patient. However, limitations of this approach have been identified by the scientific and clinical communities [2,3]. This has motivated the investigation of novel methodologies to derive optimal drug dosing for cancer chemotherapy.

Since cancer is a dreadful disease, any process that enhances the therapeutic benefits of the treatment, and thus reducing organ damage and morbidity, is greatly desired. It is also important to evaluate the effectiveness of the chemotherapy plan and its feasibility [3]. Although clinical trials are more reliable to evaluate efficient chemotherapy treatment plans, they are limited by long trial times, high costs, and difficulty in conducting such trials. All these procedures exacerbate cost; and for this reason, it is desirable to devise cost effective chemotherapy treatment planning.

Research in cancer pharmacology is driven by the development of more effective and safe chemotherapeutic drugs, and improvement in drug delivery by gathering specific pharmacokinetic and pharmacodynamic details of the drug based on available clinical data. Engineering science has complemented this area of research by developing mathematical models that represent the distribution and effect of chemotherapeutic drugs. Such models have been widely used to devise and test various drug control methodologies. These *in silico* trials are cost effective and help clinicians and engineers to analyze the reliability of novel control methodologies for drug dosing in clinical pharmacology.

A mathematical model for cancer dynamics should address tumor growth, the reaction of the human immune system to the tumor growth, and the effects of chemotherapy treatment on immune cells, normal (host) cells, and tumor growth [3,4]. As in several clinical contexts, it is important to minimize, or rather, optimize the amount of drug(s) used in order to regulate the potentially lethal side effects of chemotherapy in cancer treatment [5,6]. Often, as a side effect of chemotherapy, the patient's immune mechanism weakens and the patient becomes prone to life-threatening infections. This, in turn, diminishes the capability of the immune system to eradicate the cancer.

* Corresponding author.
E-mail addresses: regina.ajith@qu.edu.qa (R. Padmanabhan), nader.meskin@qu.edu.qa (N. Meskin), wm.haddad@aerospace.gatech.edu (W.M. Haddad).

http://dx.doi.org/10.1016/j.mbs.2017.08.004
0025-5564/

In [2,7,8], and [9] cancer chemotherapy control algorithms are proposed using optimization methods. Specifically, in [2] the authors present a model predictive control (MPC) framework for cancer chemotherapy treatment, where an MPC-based optimized drug dosing schedule is applied for a given sampling period and the corresponding state transitions are measured. Then, based on the new state measurements, the model is adjusted and the optimal control problem is resolved. In [7], the efficacy of a combined immunochemotherapy plan is studied, where a multiobjective optimization strategy to treat cancer is used by incorporating immuno therapy and optimizing chemotherapy. MPC with adaptive parameter estimation is used in [8], and in [9] the problem of optimal control of cancer dynamics using four different mathematical models for cancer chemotherapy was investigated and the feasibility of various objective functions were compared to an optimization problem in cancer chemotherapy treatment.

An important challenge of the cancer chemotherapy control problem is the nonlinear relationship between the dynamic system states. Using the design flexibility properties of the state-dependent Riccati equation-based controller design for nonlinear systems [10], the problem of cancer chemotherapy is addressed in [11] and [12]. Specifically, the authors design a state-feedback-based patient specific controller. Specific disease conditions of a patient are accounted for by choosing appropriate state and control weights in the cost function. In [11], the authors additionally use a state estimator to predict the unavailable states of the system.

Computer modeling, evolutionary algorithms, and genetic algorithms based approaches are also used to optimize and automate chemotherapy [13,14]. These methods follow principles of natural selection to find global optimal solutions. Such methods depict competitive performance with respect to the other existing chemotherapy optimization methods. However, difficulty in the selection of initial population, fixing parameter values of the initial population, and considerable computation effort due to the inherent parallelism of these algorithms are the main pitfalls of such methods [3].

Due to system complexity, nonlinearity, and uncertainty in the mechanism of action for cancer, first principle mathematical models may not be able to account for all the variations in the patient dynamics. In general, model-based, open-loop control methods do not account for the discrepancy between a mathematical model of the patient and a specific patient. Furthermore, based on the clinical response of the patient during treatment, a closed-loop control approach can determine the appropriate changes required in the drug administration to account for the discrepancy between the system response and desired response.

Since optimal administration of a therapeutic drug is essential in increasing the chance of survival in cancer treatment [15], we propose to formulate the control problem as an optimization problem and solve the problem using reinforcement learning (RL)-based methods. An important aspect of the proposed RL-based method is that it is a model-free method. The RL-based methods require less computational effort and are less complex as compared to genetic algorithm-based methods. However, for the simulations in this paper, we use a nonlinear pharmacological model of cancer chemotherapy to represent a simulated patient. This is used to perform *in silico* trials to show the efficacy of the proposed RL-based cancer chemotherapy controller.

Reinforcement learning is a promising learning technique that initially emerged in the area of machine learning [16,17]. However, due to the efficacy of RL-based control methods in handling system uncertainties and nonlinearities, it is currently being used in many fields of engineering such as robot control, wind turbine speed control, image evaluation, clinical pharmacology, and autonomous helicopter control [18–20]. RL methods explore the response of a system for every possible action and then learn the optimal action by evaluating how close the last action drives the system towards a desired state. The controller then exploits the learned optimal policies. RL can be used for the control of drug disposition as it does not require a mathematical model for the system dynamics for designing a controller. In our context, the system refers to the dynamics of the cancer patient subjected to the chemotherapy drug. RL can learn the optimal control policy using the response of the patient to the applied control actions (drug infusion) [21].

In clinical pharmacology, reinforcement learning has been used for optimizing the continuous infusion of hormones and drugs. In [22], RL-based control was used for the long-term clinical planning tasks of erythropoietin dosage. Erythropoetin is a hormone used for the treatment of anemia associated with acute renal failure. In [23], the authors investigated the use of a RL-based controller for optimizing the infusion of anesthetic drugs for surgical patients. In subsequent research, they conducted the first closed-loop clinical trial for evaluating the use of reinforcement learning-based control for regulating propofol infusion in humans [20]. In [21], a RL-based method was proposed for the control of anesthesia administration for intensive care unit patients who require sedation. In [20] and [21], the RL-based controllers used demonstrated robust control of propofol infusion and very good performance with respect to control accuracy.

The main focus of this paper is to develop a RL-based control strategy for cancer chemotherapy treatment. We use a nonlinear model that captures the cancer drug dynamics to test the RL-based controller. The proposed approach follows the general framework presented in [21], and implements a Q-learning algorithm for the control of cancer chemotherapy drug dosing. The contents of the paper are as follows. In Section 2, we present a mathematical model for cancer chemotherapy. In addition, we present the development of a reinforcement learning-based controller for cancer chemotherapy treatment. Simulation results for three case studies and a discussion on the robustness of the proposed controller are provided in Section 3. Finally, in Section 4, conclusions and future research directions are presented.

2. METHODS

In this section, we first present a pharmacological model for cancer chemotherapy treatment. In addition, a RL-based control agent is developed for the control of cancer chemotherapy.

2.1. Mathematical model of cancer chemotherapy

There exists several mathematical models that capture tumor growth dynamics with and without an external curing agent [3,24]. It should be noted that the growth rate of a tumor varies according to the type of the tumor, the organ which is affected or site of the tumor, the capability of body's immune system to resist the tumor growth, and whether the tumor stage is avascular (without blood vessels), vascular (with blood vessels), or metastatic. Metastatic cancer refers to the spread of the cancer from the part of the body where it initially started to other healthy parts of the body [1]. Clinicians often recommend to immediately remove any identified abnormally grown tissue in order to avoid possible metastases.

In this paper, we use the nonlinear four-state model given in [11,24] to demonstrate the implementation of the proposed RL-based control agent for cancer chemotherapy. The model involves four states representing the number of immune cells $I(t)$, $t \geq 0$, the number of normal cells $N(t)$, $t \geq 0$, the number of tumor cells $T(t)$, $t \geq 0$, and the drug concentration $C(t)$, $t \geq 0$, and captures the logistic growth of the tumor while accounting for the response of the body's immune system to chemotherapy. The site of the tumor involves the host cells (normal cells) and tumor cells.

The model additionally involves terms that account for the proliferation and death of cells. As any other cells in the body, immune cells proliferate to create new cells and die after their lifetime. The per capita cell death rate is denoted by d_1 and it is assumed that the growth of the tumor cells and normal cells follow a logistic growth law [4,24]. With the state variables $x_1(t) = N(t)$, $x_2(t) = T(t)$, $x_3(t) = I(t)$, and $x_4(t) = C(t)$, the cancer chemotherapy model is given by

$$\dot{x}_1(t) = r_2 x_1(t)[1 - b_2 x_1(t)] - c_4 x_1(t)x_2(t) - a_3 x_1(t)x_4(t), \quad x_1(0) = x_{10}, \quad t \geq 0, \tag{1}$$

$$\dot{x}_2(t) = r_1 x_2(t)[1 - b_1 x_2(t)] - c_2 x_3(t)x_2(t) - c_3 x_2(t)x_1(t) - a_2 x_2(t)x_4(t), \quad x_2(0) = x_{20}, \tag{2}$$

$$\dot{x}_3(t) = s + \frac{\rho x_3(t)x_2(t)}{\alpha + x_2(t)} - c_1 x_3(t)x_2(t) - d_1 x_3(t) - a_1 x_3(t)x_4(t), \quad x_3(0) = x_{30}, \tag{3}$$

$$\dot{x}_4(t) = -d_2 x_4(t) + u(t), \quad x_4(0) = x_{40}, \tag{4}$$

where $u(t)$, $t \geq 0$, is the drug infusion rate, s denotes the (constant) influx rate of immune cells to the site of the tumor, r_1 and r_2 represent the per capita growth rate of the tumor cells and normal cells, respectively, b_1 and b_2 represent the reciprocal carrying capacities of both the cells, d_2 denotes the per capita decay rate of the injected drug, and a_1, a_2, and a_3 represent the fractional cell kill rates of the immune cells, tumor cells, and normal cells, respectively [11,24].

The number of the immune cells are increased if a tumor is present in the body and the immune cells try to eradicate the tumor cells [4]. In most cases, even though the immune system is capable of destroying the tumor cells, often times the immune mechanism is not strong enough to combat the rapid growth rate of the tumor. Once the body's immune system cannot control the growth of the abnormal cells, a detectable tumor appears in the body. As a response to the development of tumor cells, the immune system will increase production of the immune cells. This positive nonlinear growth is incorporated into the model via the term $\frac{\rho x_3(t)x_2(t)}{\alpha + x_2(t)}$ in (3), where α and ρ are positive constants that denote the immune threshold rate and immune response rate, respectively [4,24].

Tumor cells and host cells compete for available nourishment in the blood stream for their survival. When the immune cells are large in number, the existence of the tumor cells is low and vice versa [4]. An encounter of the immune cells with the tumor cells ends up either in the death of the tumor cells or the inaction of the immune cells. This phenomenon is captured using the terms $-c_2 x_3(t)x_2(t)$ and $-c_1 x_3(t)x_2(t)$, respectively, in the model equations. Note that the survival of the normal cells, tumor cells, and immune cells are mutually dependent. The competition terms c_i, $i = 1, 2, \ldots, 4$, in (1)–(3) are used to model the interdependency in the survival rate between the normal cells, tumor cells, and immune cells [24].

In general, the immune cells may either succeed in destroying the tumor cells or may get inactivated. Likewise, the injected drug can effect the normal cells, tumor cells, and immune cells. These competing relations between the system states are accounted for by using the parameters c_i, $i = 1, \ldots, 4$, in the model. The effect of the chemotherapy drug is reflected in (1)–(3) through the different response coefficients a_1, a_2, and a_3.

It should be noted that in addition to the desired effect, the drugs used for chemotherapy can also annihilate normal cells and immune cells. The control objective is thus to design an optimal control input $u(t)$, $t \geq 0$, for chemotherapy drug dosing that maximizes the desired drug effect and minimizes the drug induced side effects. The desired drug effect is the eradication of the cancer cells and the reduction of some of the common drug induced side effects, which include nausea, hair loss, recurrent infections due to destruction of white blood cells, anemia, neuropathy, and damage to vital organs such as the heart, kidneys, and lungs [1].

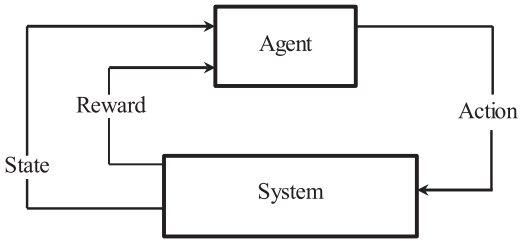

Fig. 1. Reinforcement learning schematic.

2.2. RL-based optimal control for chemotherapic drug dosing

In general, the problem of designing optimal controllers for nonlinear systems is challenging [25]. If the system dynamics are known, the optimal control law for a linear system is given by the algebraic Riccati equation using standard linear-quadratic optimal control. However, in the case of nonlinear systems this requires the solution of the Hamilton–Jacobi–Bellman partial differential equation [26]. In this section, we develop a methodology for RL-based control for cancer chemotherapy drug dosing. Our framework uses the nonlinear four-state model for cancer chemotherapy treatment given by (1)–(4). The nonlinear model represents the dynamics of the tumor cells and comprises a system of four coupled ordinary differential equations characterizing the normal cells, tumor cells, immune cells, and drug concentration.

Watkin's Q-learning is a RL-based approach that has gained considerable attention in recent years as a learning method that does not require an accurate system model and can be used online while the system dynamics change during the learning process [21,27]. In a learning-based approach, the agent or controller applies an action on the system and observes the corresponding reward to learn a useful control policy or action plan; see Fig. 1.

The problem of deriving control laws for regulating the number of tumor cells $x_2(t)$, $t \geq 0$, involves sequential decision making based on the response of the patient to drug administration. Reinforcement learning-based approaches make use of a finite Markov decision process (MDP) framework for developing algorithms that can learn optimal decisions iteratively [16,17]. In the case of cancer chemotherapy treatment, the aim is to transition from a nonzero initial state $x_2(t) \geq 0$, $t \geq 0$, to the desired final state $x_2(t) = 0$ at some time t. This can be achieved by identifying the best sequence of chemotherapic drug infusion that will transition the cancer patient from $x_2(t) \geq 0$, $t \geq 0$, to the terminal state $x_2(t) = 0$. The nonlinear system given by (1)–(4) can be cast in the state space form

$$\dot{x}(t) = f(x(t)) + G(x(t))u(t), \quad x(0) = x_0, \quad t \geq 0, \tag{5}$$

$$y(t) = h(x(t)), \tag{6}$$

where $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $G : \mathbb{R}^n \to \mathbb{R}^m$, $h : \mathbb{R}^n \to \mathbb{R}^l$, $x(t) \in \mathbb{R}^n$, $t \geq 0$, is the state vector, $u(t) \in \mathbb{R}$, $t \geq 0$, is the control input, and $y(t) \in \mathbb{R}^l$, $t \geq 0$, is the output of the system.

Analogous to the role of a mathematical model for a dynamical system in control theory, in a finite MDP framework the system dynamics are captured by the four finite sequences $\mathcal{S}$, $\mathcal{A}$, $\mathcal{R}$, and $\mathcal{P}$,

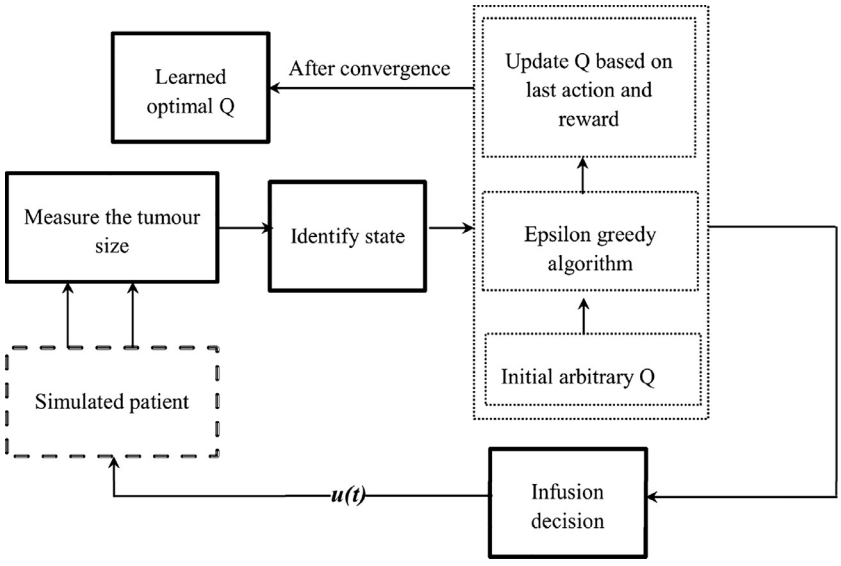

Fig. 2. Schematic of training sequence to obtain optimal Q table.

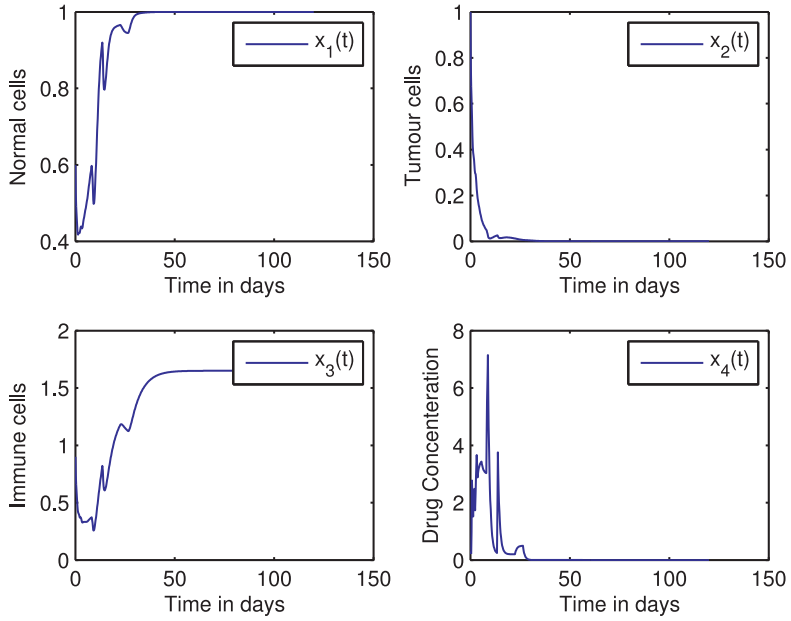

Fig. 3. Response of young patient with cancer (Case 1), $u_{\max} = 10$ mg l^{-1} day^{-1}.

where $\mathcal{S}$ is a finite set of states, $\mathcal{A}$ a finite set of actions defined for the states $s_k \in \mathcal{S}$, $\mathcal{R}$ represents the reward function that guides the agent in accordance to the desirability of an action $a_k \in \mathcal{A}$, and $\mathcal{P}$ is a state transition probability matrix. The state transition probability matrix $\mathcal{P}_{a_k}(s_k, s_{k+1})$ gives the probability that an action $a_k \in \mathcal{A}$ takes the state $s_k \in \mathcal{S}$ to the state s_{k+1} in a finite time step. Furthermore, the discrete states in the finite sequence $\mathcal{S}$ are represented as $(\mathcal{S}_i)_{i\in\mathbb{I}^+}$, where $\mathbb{I}^+ \triangleq \{1, 2, \ldots, q\}$ and q denotes the total number of states. Likewise, the discrete actions in the finite sequence $\mathcal{A}$ are represented as $(\mathcal{A}_j)_{j\in\mathbb{J}^+}$, where $\mathbb{J}^+ \triangleq \{1, 2, \ldots, p\}$ and p denotes the total number of actions. The transition probability matrix $\mathcal{P}$ can be formulated based on the system dynamics. Note that, since the Q-learning framework does not require $\mathcal{P}$ for deriving the optimal control policy, we assume $\mathcal{P}$ is unknown [21].

The reinforcement learning method starts with an initial arbitrary policy and learns the optimal policy by interacting with the system. In RL frameworks, a policy can be a path plan to transition from an initial position to the target position; it can be a rule base or a look-up-table such as "if in this state, then do this," and in general is a mapping from states to (control) actions [17]. The algorithm progresses iteratively by interacting with the system. Accordingly, as the agent receives more information in terms of state, action, and reward, the agent's decision set approaches the optimal decision set or optimal control policy. In the case of the Q-learning algorithm [27], each tuple of information involving the state, action, and reward are used to update an entry in the table Q. The entry $Q_k(s_k, a_k)$ in the Q table represents the desirability of each action in the finite sequence $(\mathcal{A}_j)_{j\in\mathbb{J}^+}$ with respect to each discrete state of the finite sequence $(\mathcal{S}_i)_{i\in\mathbb{I}^+}$.

As shown in Fig. 1, the main elements of the reinforcement learning framework include an agent and a system. At each time step k, the agent first observes the current state s_k of the system and then imparts an action a_k from the sequence of actions in $\mathcal{A}$. Accordingly, the system stochastically transitions from the current

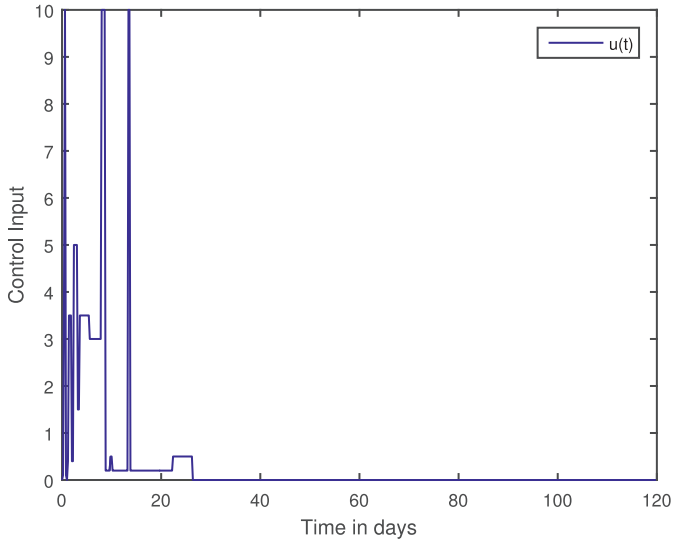

Fig. 4. Amount of drug administrated (Case 1), $u_{\max} = 10$ mg l^{-1} day^{-1}.

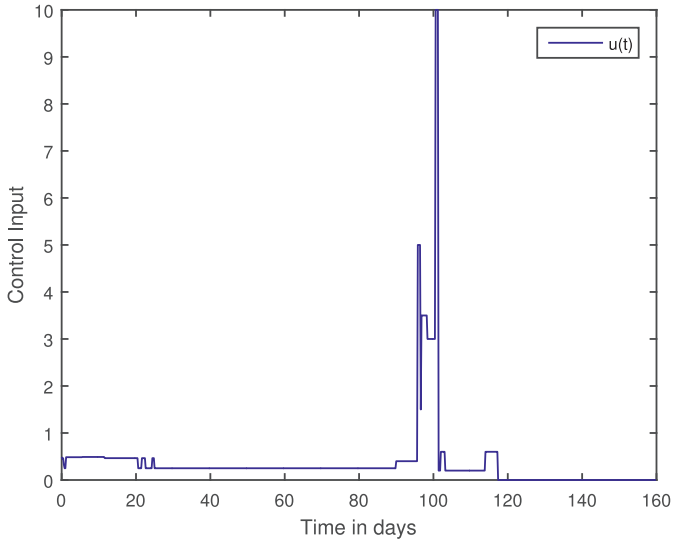

Fig. 6. Amount of drug administrated (Case 2), $u_{\max} = 0.5$ mg l^{-1} day^{-1} until delivery (90 days) and then $u_{\max} = 10$ mg l^{-1} day^{-1}.

system state s_k to the new state s_{k+1}. The desirability of the selected action a_k, at time step k, can be captured by using an appropriate reward $r_{k+1} \in \mathbb{R}$, which assigns a numerical value to a state action pair. The value of the reward r_{k+1} received gives the agent information on whether the last action chosen was "good" or "bad." The agent utilizes the Q-learning algorithm to find an optimal policy that maximizes the expected value $\mathbb{E}[\,\cdot\,]$ of the discounted reward it receives over an infinite horizon given by

$$J(r_k) = \mathbb{E}\left[\sum_{k=1}^{\infty} \theta^{(k-1)} r_k\right], \tag{7}$$

where the discount rate parameter θ represents the importance of immediate and future rewards. The parameter θ can take values $\theta \in [0, 1]$, where $\theta = 0$ constrains the agent to consider only the current reward, whereas for θ approaching 1 the agent considers current as well as future rewards.

Thus, once the agent receives a reward r_{k+1}, with respect to the state transition $s_k \to s_{k+1}$ and the action a_k, the Q table is updated by using the Q-learning algorithm given by

$$Q_k(s_k, a_k) \leftarrow Q_{k-1}(s_k, a_k) + \eta_k(s_k, a_k) \times [r_{k+1} + \theta \max_{a_{k+1}} Q_{k-1}(s_{k+1}, a_{k+1}) - Q_{k-1}(s_k, a_k)], \tag{8}$$

where $\eta_k(s_k, a_k) \in [0, 1)$, $k = 1, 2, \ldots$, denote the learning rates that effect the size of the correction after each iteration. It should be noted that the Q-learning algorithm starts with an initial arbitrary $Q_1(s_1, a_1)$. Then, with each observation, the Q table is updated until convergence is reached. We use a tolerance parameter δ with condition $\Delta Q_k \triangleq |Q_k - Q_{k-1}| \le \delta$ to assign the minimum threshold required for convergence. For further details on the proofs and conditions required for the convergence of the Q-learning algorithm; see [17,27,28].

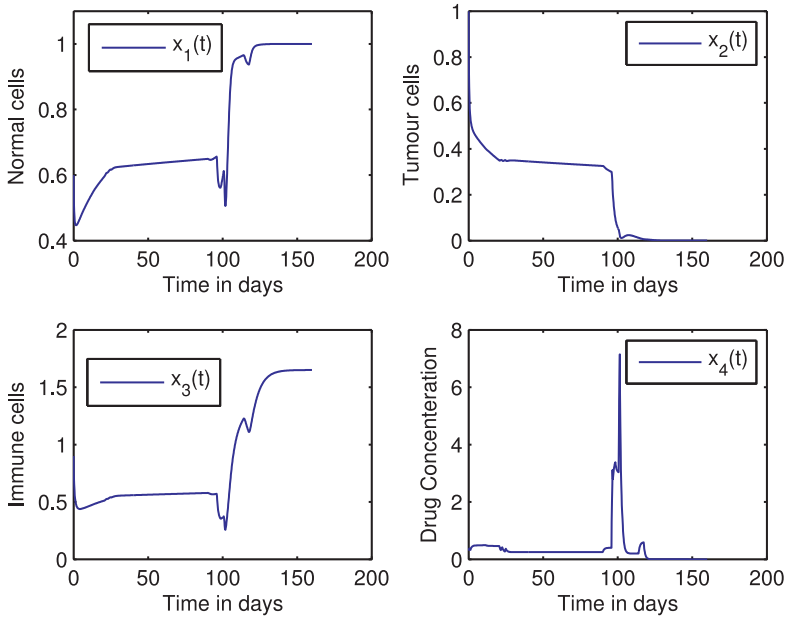

Fig. 5. Response of young pregnant woman with cancer (Case 2), $u_{\max} = 0.5$ mg l^{-1} day^{-1} until delivery (90 days) and then $u_{\max} = 10$ mg l^{-1} day^{-1}.

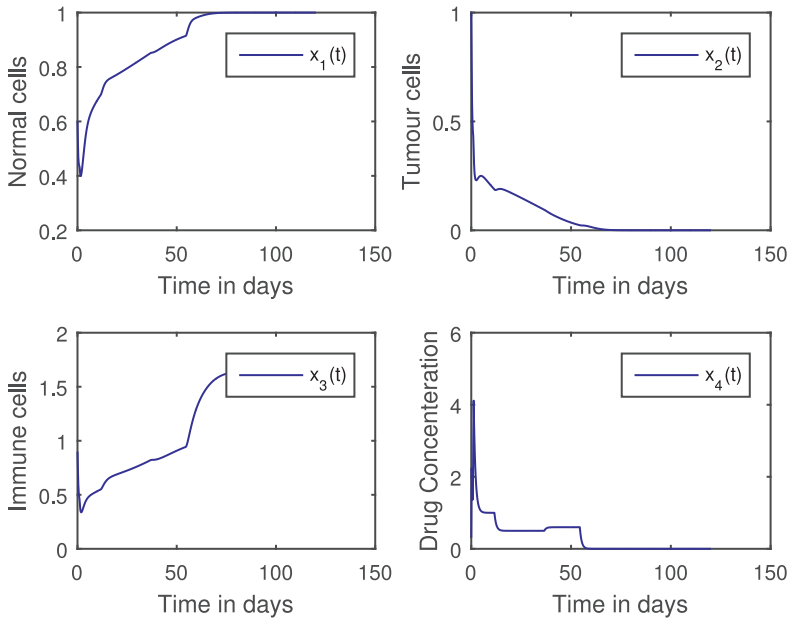

Fig. 7. Response of an elderly patient who has cancer along with other critical illnesses (Case 3), $u_{\max} = 10$ mg l^{-1} day^{-1}.

The methods adopted for measuring tumor size vary according to the type and site of the tumor. The size of a peripheral tumor can be assessed manually by using a caliper. However, if the tumor is in the brain or an other internal organ, then imaging techniques such as ultrasound imaging, magnetic resonance imaging, or computer tomographic imaging are required to assess the tumor volume [29]. In some situations measuring the number of normal cells is easier than measuring the number of tumor cells. In such cases, the number of tumor cells can be estimated using the available measurement of normal cells [11,30,31].

The schematic of the training sequence that we followed to obtain the optimal Q table for cancer chemotherapy treatment is shown Fig. 2. To learn the optimal Q table, the discrete states $s_k \in \mathcal{S}$ representing the status quo of the system should be available. In this paper, we define the states s_k of the cancer patient in terms of an available output $y(t)$, $t \geq 0$, as $s_k = g(y(t))$, $kT \leq t < (k+1)T$, where $g : \mathbb{R}^l \to \mathcal{S} \subset \mathbb{R}$. For the problem of drug dosing for cancer treatment, the aim is to derive the optimal sequence of actions in terms of drug infusion rates that result in a minimum tumor size, ideally $x_2(t) = 0$. Thus, we assume that the number of tumor cells is available and define the system state s_k based on the value of $x_2(t)$, $t \geq 0$, [4,32].

Recall that the reward function is used to guide the agent in assessing whether the action chosen at the last time step was desirable or not. This information is used to reinforce the agent's decision making. Note that at every time step k and state s_k, the controller or agent chooses the action a_k as

$$a_k = (\mathcal{A}_j)_{j \in \mathbb{J}^+}, \quad j = \arg\max Q_k(s_k, \cdot). \tag{9}$$

The reward r_{k+1} is computed by using the error $e(t)$ as

$$r_{k+1} = \begin{cases} \frac{e(kT) - e((k+1)T)}{e(kT)}, & e((k+1)T) < e(kT), \\ 0, & e((k+1)T) \geq e(kT), \end{cases} \tag{10}$$

where $e(t)$, $t \geq 0$, involves a particular combination of the system states; see Section 3. As shown in Fig. 2, an ϵ-greedy policy is implemented to derive the optimal policy in which the agent imparts random actions to the system with probability ϵ, where ϵ is a small positive number [17]. According to the information on the current state, action, and new state gathered during each interaction, the agent assesses the reward acquired to update the Q table.

The further the agent explores the system, the more it learns. Ideally, with exploration $k \to \infty$, the algorithm can converge to the optimal Q table starting from an arbitrary Q table. However, in most cases, convergence is achieved with an acceptable tolerance δ satisfying $\Delta Q_k \leq \delta$ for some finite k, well before the exploration approaches infinity. One of the conditions to ensure convergence is to reduce the learning rate $\eta_k(s_k, a_k)$ as the algorithm progress over time [17]. Fig. 2 shows the schematic of the training sequence to obtain the optimal Q table. See [21] for further details on implementing a Q-learning algorithm.

3. Results and discussion

In this section, we present numerical examples that illustrate the efficacy of the proposed RL approach for the closed-loop control of cancer chemotherapy drug dosing. There are several factors that oncologists consider when deciding on the drug dose for a cancer patient. For example, age and gender of the patient, whether the patient is suffering from any other disease, whether the patient is pregnant, etc. In general, the growth rate of normal cells and immune cells are age-dependent and the growth rate in a young patient will be larger than that of an elderly adult [11]. Therefore, in the case of a young patient, an oncologist prefers to immediately minimize the number of cancerous cells with less regard to normal cell and immune cell damage. This is mainly to prevent cancer metastasis. The reduced number of normal cells, which results as a side effect of chemotherapy, will be regenerated by the body if the patient is young.

Alternatively, if the cancer patient has other diseases, or if the site of the cancer is in a vital organ such as the brain, then destroying the normal cells is not recommended. All these conditions can be accounted for by choosing an appropriate reward function (10). Moreover, in the case of specific patient groups such as infants, children, and pregnant women, the oncologist must restrict the upper limits of the drug dose. This can be achieved by appropriately choosing the maximum value of the drug infusion rate $u_{\max}$ while training the RL agent.

In this paper, we illustrate the use of RL-based control for chemotherapic drug dosing using a simulated patient represented by (1)–(4) with the parameters given in Table 1 [11]. For our simulation, we iterated on 50,000 (arbitrarily high) scenarios, where a

Table 1
Parameter values used to generate simulated patient [11,24].

Parameter	Parameter description	Value	Unit
a_1	Fractional immune cell kill rate	0.2	$mg^{-1}l\ day^{-1}$
a_2	Fractional tumor cell kill rate	0.3	$mg^{-1}l\ day^{-1}$
a_3	Fractional normal cell kill rate	0.1	$mg^{-1}l\ day^{-1}$
b_1	Reciprocal carrying capacity of tumor cells	1	$cell^{-1}$
b_2	Reciprocal carrying capacity of normal cells	1	$cell^{-1}$
c_1	Immune cell competition term (competition between tumor cells and immune cells)	1	$cell^{-1}day^{-1}$
c_2	Tumor cell competition term (competition between tumor cells and immune cells)	0.5	$cell^{-1}day^{-1}$
c_3	Tumor cell competition term (competition between normal cells and tumor cells)	1	$cell^{-1}day^{-1}$
c_4	Normal cell competition term (competition between normal cells and tumor cells)	1	$cell^{-1}day^{-1}$
d_1	Immune cell death rate	0.2	day^{-1}
d_2	Decay rate of injected drug	1	day^{-1}
r_1	Per unit growth rate of tumor cells	1.5	day^{-1}
r_2	Per unit growth rate of normal cells	1	day^{-1}
s	Immune cell influx rate	0.33	cell day^{-1}
α	Immune threshold rate	0.3	cell
ρ	Immune response rate	0.01	day^{-1}

scenario represents the series of transitions from an arbitrary initial state to the required terminal state s_k. The action a_k at the kth time step is represented by $(\mathcal{A}_j)_{j\in\mathbb{J}^+}$, where $\mathbb{J}^+ = \{1, 2, \ldots, 20\}$. Furthermore, we initially assigned $\eta_k(s_k, a_k) = 0.2$ for the first 499 scenarios and then the value of $\eta_k(s_k, a_k)$ is subsequently halved after every 500th scenario. After convergence of the Q table to the optimal Q function, for every state s_k, the agent chooses an action $a_k = (\mathcal{A}_j)_{j\in\mathbb{J}^+}$, where $j = \arg\max Q_k(s_k, \cdot)$. For our simulation, with a chemotherapic agent, we consider three cases. Namely, (1) an adult with cancer, (2) a pregnant woman with cancer, and (3) an elderly patient who has cancer along with other critical illnesses. Note that we train different RL agents for each of the aforementioned cases. We used MATLAB® for our simulations.

Case 1: First, we consider the case of a young patient with cancer. In this case, since the patient has good growth ability, the patient's body can more easily compensate for the loss of normal cells and immune cells as the side effect of chemotherapy. In such a situation, the oncologist will generally try to annihilate the cancer cells $x_2(t)$, $t \geq 0$, completely. Thus, the aim is to eradicate the tumor cells and achieve the desired state $x_{2d} = 0$. Therefore, the error $e(t)$, $t \geq 0$, can be defined as $e(t) = x_2(t) - x_{2d}$. Table 2 shows the criteria used for the state assignment based on the error $e(t)$, $kT \leq t < (k+1)T$. The reward r_{k+1} is computed by using $e(t) = x_2(t)$. For this case, we use a RL agent trained with $u_{max} = 10$ mg l^{-1} day^{-1}.

Fig. 3 shows the response of the patient when a chemotherapeutic drug is administrated using a RL-based controller and includes the plots of the number of normal cells, the number of tumor cells, the number of immune cells, and the concentration of chemotherapeutic drug in blood. It can be seen that with chemotherapy, the number of tumor cells have decreased and the normal cells have increased. However, note that initially the number of immune cells decrease due to chemotherapy, whereas later their number improves. The amount of drug administrated for Case 1 is shown in Fig. 4.

Case 2: For our second case, we consider a young pregnant woman with cancer. In this case, the oncologist tries to keep the chemotherapy drug dose to a minimum level so as to ensure the safety of the fetus. Subsequently, after child birth, the oncologist can increase the drug dose. For our simulations, we assume that the patient is in her seventh month of pregnancy. Hence, the oncologist schedules the treatment in two stages. During the first stage, we restrict the maximum drug dose by choosing $u_{max} = 0.5$ mg l^{-1} day^{-1}. After child birth, we use a maximum drug dose of $u_{max} = 10$ mg l^{-1} day^{-1} [11].

Table 2
State assignment for cases $1-3$ based on $e(t)$. Case 1: young cancer patient, case 2: pregnant woman with cancer, and case 3: an elderly patient who has cancer along with other critical illnesses.

Cases 1, 2		Case 3	
State s_k	$e(kT)$	State s_k	$e(kT)$
1	[0, 0.0063]	1	[0, 0.03]
2	(0.0063, 0.0125]	2	(0.03, 0.1]
3	(0.0125, 0.025]	3	(0.1, 0.2]
4	(0.025, 0.01]	4	(0.2, 0.3]
5	(0.01, 0.05]	5	(0.3, 0.4]
6	(0.05, 0.1]	6	(0.4, 0.5]
7	(0.1, 0.2]	7	(0.5, 0.6]
8	(0.2, 0.25]	8	(0.6, 0.7]
9	(0.25, 0.3]	9	(0.7, 0.8]
10	(0.3, 0.35]	10	(0.8, 0.9]
11	(0.35, 0.4]	11	(0.9, 1]
12	(0.4, 0.45]	12	(1, 1.2]
13	(0.45, 0.5]	13	(1.2, 1.4]
14	(0.5, 0.55]	14	(1.4, 1.6]
15	(0.55, 0.6]	15	(1.6, 1.8]
16	(0.6, 0.65]	16	(1.8, 2]
17	(0.65, 0.7]	17	(2, 2.2]
18	(0.7, 0.8]	18	(2.2, 2.5]
19	(0.8, 0.9]	19	(2.5, 3]
20	(0.9, ∞]	20	(3, ∞]

For this case, two RL agents are trained; one for the first stage with $u_{max} = 0.5$ mg l^{-1} day^{-1} and the other for the second stage with $u_{max} = 10$ mg l^{-1} day^{-1}. Figs. 5 and 6 show the simulation results for the two-stage chemotherapy for the young pregnant woman using RL-based controllers. It can be seen that over the initial period of 90 days, the drug concentration in the plasma is restricted to 0.5 mg l^{-1}, whereas after child birth the drug dose is increased to eradicate the tumor completely. *Case 3:* Finally, we consider the case of an elderly patient who has cancer along with other critical illnesses. For such a scenario, it becomes essential that a greater number of normal cells be preserved while eradicating the tumor cells. Hence, we use a weighing factor β to trade off between the emphasis on eradicating the tumor cells and preserving the number of tumor cells. The aim here is to achieve $x_{1d} = 1$ and $x_{2d} = 0$, where x_{1d} and x_{2d} denote the desired values of $x_1(t)$, $t \geq 0$, and $x_2(t)$, $t \geq 0$, respectively. For this case, we train an RL agent using a reward function defined based on the deviation of the number of normal cells and tumor cells from the respective desired values. Specifically, we define the state s_k, $kT \leq t < (k+1)T$,

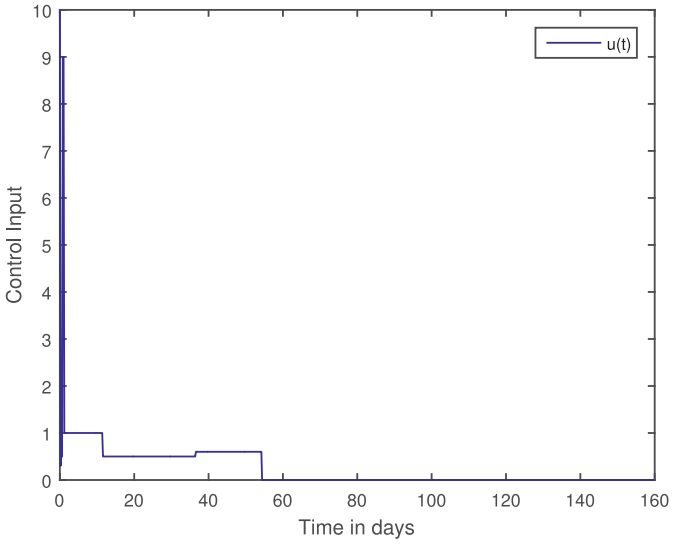

Fig. 8. Amount of drug administrated (Case 3), $u_{\max} = 10$ mg l^{-1} day^{-1}.

in terms of the error

$$e(t) = \beta x_2(t) + (1-\beta)[1 - x_1(t)]. \tag{11}$$

The reward (10) is calculated with respect to the error $e(t)$, $kT \le t < (k+1)T$. For our simulation, we set the values of discount factor $\theta = 0.7$, initial learning rate $\eta = 0.2$, and weighing factor $\beta = 0.9$.

Figs. 7 and 8 show the response of the patient when a chemotherapeutic drug is administrated using an RL-based controller which was trained with respect to the error (11). It can be seen that, as compared to Case 1, the amount of drug administrated is less so as to reduce the damage to the normal cells; see Figs. 3, 4, 7, and 8. However, for Case 3 the tumor cells are eradicated slowly as compared to Case 1. This trade off accounts for the necessity in preserving the normal cells and the necessity in ruling out possible metastasis. As discussed earlier, if the site of cancer is in a vital organ such as the brain, then destroying the normal cells is not recommended.

Table 2 shows the criteria used for the state assignment for Cases $1-3$. For Cases 1, 3, and the second stage of Case 2, we use the finite action sequence $\mathcal{A} = (0, 0.01, 0.02, 0.03, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1)$. However, for the first stage of Case 2, with $u_{\max} = 0.5$ mg l^{-1} day^{-1}, we use the finite action sequence $\mathcal{A} = (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.78, 0.80, 0.82, 0.85, 0.87, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.97, 0.98, 1)$. Moreover, for Cases 1, 3, and the second stage of Case 2 during the training of RL agent, we set the goal state as $s_k = 1$ to eradicate the tumor completely. However, for the first stage of Case 2, during the training of RL agent, we set the goal state as $s_k = 7$, which represents a limited tumor size. In order to demonstrate the robustness of the proposed controller, we use the trained optimal RL-based controller for the drug dosing of three different simulated patients. In case (i), we consider the simulated patient with a nominal model generated using the parameters given in Table 1. In cases (ii) and (iii), we use simulated patients with -10% and $+15\%$ parameter variations with respect to the values given in Table 1. Figs. 9 and 10 show the corresponding simulation results. It can be seen that the controller is able to impart patient specific infusion rates in accordance with the parameter variations. This is mainly due to the fact that the drug dosing decision is made using the optimal Q table with respect to the state s_k. Recall that the state s_k is defined based on the error $e(t)$, $t \ge 0$, which reflects the patient specific response to drug intake. Thus, the value of the error $e(t)$, $t \ge 0$, varies according to the patient characteristics. As we use the error value to decide the state s_k, and hence, the optimal action a_k, the controller exhibits robust performance.

Table 3
Statistical analysis for 15 simulated patients.

Parameter		N_{dev}	T_{per}
Number of days to achieve the target value.	Min	13	6
	Max	50	52
	Mean	28	27
Percent value; before chemotherapy.	Min	40	100
	Max	40	100
	Mean	40	100
Percent value; after 1 week of chemotherapy.	Min	10.17	19.34
	Max	87.75	0.0096
	Mean	45.05	2.50
Percent value; after 4 weeks of chemotherapy.	Min	0	0.5324
	Max	3.47	0
	Mean	0.4271	0.1708
Percent value; after 7 weeks of chemotherapy.	Min	0	0.0634
	Max	0.0560	0
	Mean	0.0059	0.0064

Table 3 shows the statistical results of the simulations performed on 15 simulated patients using the RL agent trained for Case 1. We generated 15 simulated patients with the parameter ranges of: fraction cell kill a_i, $i = 1, 2, 3$, $0 < a_i \le 0.5$, $a_3 \le a_1 \le a_2$, carrying capacities $b_1^{-1} \le b_2^{-1} = 1$, competition terms $0.3 \le c_i \le 1$, $i = 1, \ldots, 4$, death rates $0.15 \le d_1 \le 0.3$, $d_2 = 1$, per unit growth rates, $1.2 \le r_1 \le 1.6$, $r_2 = 1$, immune source rate $0.3 \le s \le 0.5$, immune threshold rate $0.3 \le \alpha \le 0.5$, and immune response rate $0.01 \le \alpha \le 0.05$. See [24] for further details on the parameter ranges of the cancer chemotherapy model.

Table 3 shows the minimum value, the maximum value, and the mean values of the number of normal cells, as well as the number of tumor cells at various weeks of chemotherapy treatment for the 15 simulated patients. The table also shows the minimum, maximum, and mean number of days for achieving the target values of $x_1(t)$, $t \ge 0$, and $x_2(t)$, $t \ge 0$, for the 15 simulated patients. The percent deviation of the number of normal cells from the target value $(x_{1\text{d}} = 1)$ is calculated as

$$\begin{aligned} N_{\text{dev}} &= \frac{|\text{Measured value} - \text{Target value}|}{\text{Target value}} \times 100 \\ &= |x_1(t^*) - 1| \times 100, \end{aligned}$$

where $t^* = 0, 1, 4$, or 7 weeks. The percent value of the number of tumor cells with respect to the initial value is calculated as

$$T_{\text{per}} = \frac{\text{Measured value}}{\text{Initial value}} \times 100 = \frac{x_2(t^*)}{x_2(0)} \times 100.$$

It can be seen from Table 3 that by week 7, the percent deviation of the number of normal cells from the target value is 0.0059 and the percent value of the number of tumor cells with respect to the initial value is 0.0064 for the 15 simulated patients. Comparing our simulation results with those in [11], it can be seen that both methods result in very similar responses. In both cases the tumor is eradicated using optimal chemotherapy drug dosing and the controllers are robust to parameter variations. However, the advantage of the proposed RL-based method is that it does not require a model of the system in order to develop a controller.

4. Conclusions and future work

In this paper, we investigated the efficacy of the proposed RL-based method for different cases of cancer treatment. The method results in an optimal as well as robust controller. In order to preserve normal cells while eradicating tumor cells, we proposed the use of a scaled value of the error in the reward function. The

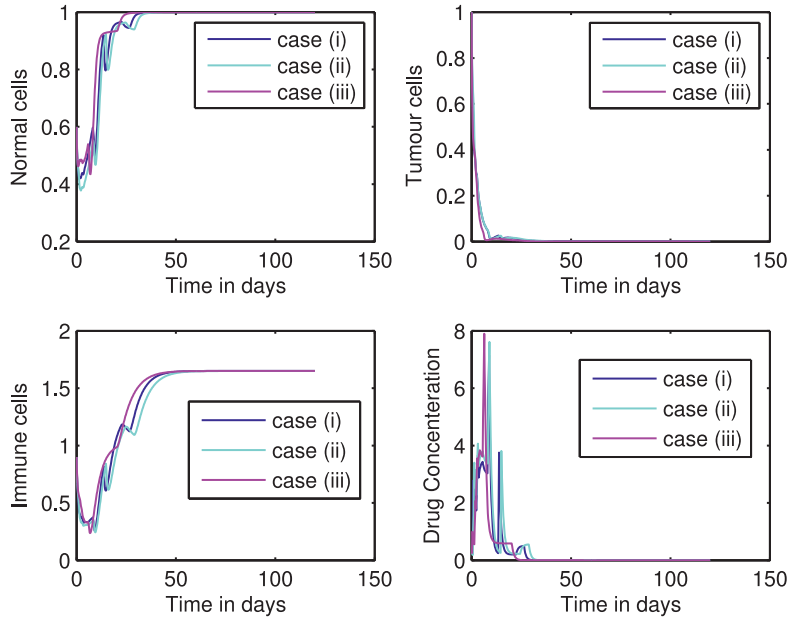

Fig. 9. Response for three different patient models; case (i) with nominal model, case (ii) with −10% parameter variation, case (iii) with +15% parameter variation.

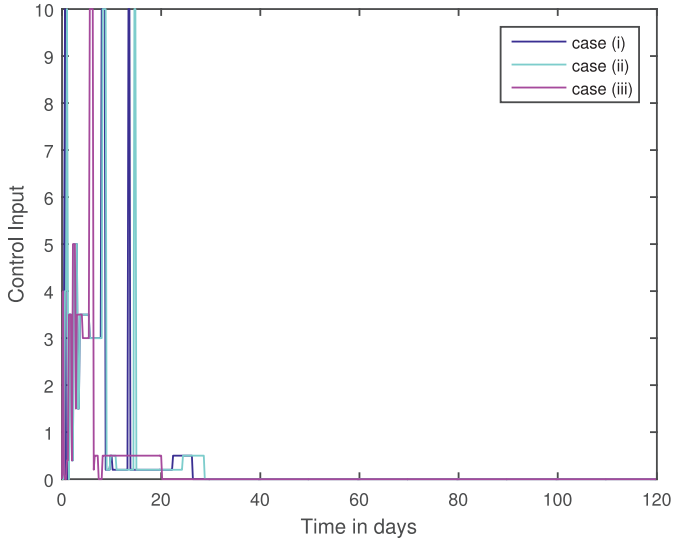

Fig. 10. Control input for three different patient models; case (i) with nominal model, case (ii) with −10% parameter variation, case (iii) with +15% parameter variation.

proposed controller using the RL method can be extended to account for different constraints in cancer treatment by appropriately choosing the reward function. The main advantage of the proposed RL-based control method is that the algorithm does not require knowledge of the system dynamics. However, different RL-agents need to be trained to account for the patient characteristics of different patient groups. This difficulty can be overcome by using adaptive reinforcement learning algorithms and will be addressed in future research.

Acknowledgment

This publication was made possible by the GSRA grant No. GSRA1-1-1128-13016 from the Qatar National Research Fund (a member of Qatar Foundation). The findings achieved herein are solely the responsibility of the authors.

References

[1] ACS, Cancer Facts and Figures 2015., Technical Report, American Cancer Society, Atlanta, Georgia. http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf.
[2] T. Chen, N.F. Kirkby, R. Jena, Optimal dosing of cancer chemotherapy using model predictive control and moving horizon state/parameter estimation, Comput. Methods Progr. Biomed. 108 (3) (2012) 973–983.
[3] H. Sbeity, R. Younes, Review of optimization methods for cancer chemotherapy treatment planning, J. Comput. Sci. Syst. Biol. 8 (2015) 074–095.
[4] L.G.D. Pillis, A. Radunskaya, A mathematical tumor model with immune resistance and drug therapy: an optimal control approach, Comput. Math. Methods Med. 3 (2) (2001) 79–100.
[5] J.C. Doloff, D.J. Waxman, Transcriptional profiling provides insights into metronomic cyclophosphamide-activated, innate immune-dependent regression of brain tumor xenografts, BMC Cancer 15 (1) (2015) 375.
[6] C.-S. Chen, J.C. Doloff, D.J. Waxman, Intermittent metronomic drug schedule is essential for activating antitumor innate immunity and tumor xenograft regression, Neoplasia 16 (1) (2014) 84W22–96W27.
[7] K.L. Kiran, D. Jayachandran, S. Lakshminarayanan, Multi-objective optimization of cancer immuno-chemotherapy, in: Proceedings of the 13th International Conference on Biomedical Engineering, 2009, pp. 1337–1340.
[8] S.L. Noble, E. Sherer, R.E. Hannemann, D. Ramkrishna, T. Vik, A.E. Rundell, Using adaptive model predictive control to customize maintenance therapy chemotherapeutic dosing for childhood acute lymphoblastic leukemia, J. Theor. Biol. 264 (3) (2010) 990–1002.
[9] M. Engelhart, D. Lebiedz, S. Sager, Optimal control for selected cancer chemotherapy ODE models: a view on the potential of optimal schedules and choice of objective function, Math. Biosci. 229 (1) (2011) 123–134.
[10] J.R. Cloutier, State-dependent Riccati equation techniques: an overview, in: Proceedings of the American Control Conference(1997) 932–936.
[11] Y. Batmani, H. Khaloozadeh, Optimal chemotherapy in cancer treatment: state dependent Riccati equation control and extended Kalman filter, Optim. Control Appl. Methods 34 (5) (2013) 562–577.
[12] T. Çimen, Systematic and effective design of nonlinear feedback controllers via the state-dependent Riccati equation (SDRE) method, Annu. Rev. Control 34 (1) (2010) 32–51.
[13] K.C. Tan, E.F. Khor, J. Cai, C. Heng, T.H. Lee, Automating the drug scheduling of cancer chemotherapy via evolutionary computation, Artif. Intell. Med. 25 (2) (2002) 169–185.
[14] S.-M. Tse, Y. Liang, K.-S. Leung, K.-H. Lee, T.S.-K. Mok, A memetic algorithm for multiple-drug cancer chemotherapy schedule optimization, IEEE Trans. Syst. Man Cybern Part B (Cybern) 37 (1) (2007) 84–91.
[15] L.G.D. Pillis, W. Gu, K.R. Fister, T.A. Head, K. Maples, A. Murugan, T. Neal, K. Yoshida, Chemotherapy for tumors: an analysis of the dynamics and a study of quadratic and linear optimal controls, Math. Biosci. 209 (1) (2007) 292–315.
[16] D. Vrabie, K.G. Vamvoudakis, F.L. Lewis, Optimal Adaptive Control and Differential Games by Reinforcement Learning Principle, Institution of Engineering and Technology, London, UK, 2013.
[17] R.S. Sutton, A.G. Barto, Reinforcement Learning: An Introduction, MIT Press, Cambridge, MA, 1998.

[18] M. Sedighizadeh, A. Rezazadeh, Adaptive PID controller based on reinforcement learning for wind turbine control, World Acad. Sci. Eng. Technol. 2 (2008) 01–23.
[19] P. Abbeel, A. Coates, M. Quigley, A.Y. Ng, An application of reinforcement learning to aerobatic helicopter flight, Neural Inf. Process. Syst 19 (2007) 1–8.
[20] B.L. Moore, L.D. Pyeatt, V. Kulkarni, P. Panousis, Kevin, A.G. Doufas, Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers, J. Mach. Learn. Res. 15 (2014) 655–696.
[21] R. Padmanabhan, N. Meskin, W.M. Haddad, Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning, Biomed. Signal Process. Control 22 (2015) 54–64.
[22] J. Martin-Guerrero, F. Gomez, E. Soria-Olivas, J. Schmidhuber, M. Climente–Marti, N. Jemenez-Torres, A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients, Expert Syst. Appl. 36 (2009) 9737–9742.
[23] B.L. Moore, P. Panousis, V. Kulkarni, L.D. Pyeatt, A.G. Doufas, Reinforcement learning for closed-loop propofol anesthesia, in: Proceedings of 22th Annual Conference on Innovative Applications of Artificial Intelligence, Atlanta. Georgia, USA, July, 2010, pp. 1807–1813.
[24] L.G.D. Pillis, A. Radunskaya, The dynamics of an optimally controlled tumor model: a case study, Math. Comput. Model. 37 (11) (2003) 1221–1244.
[25] N.V. Balashevich, R. Gabasov, A.I. Kalinin, F.M. Kirillova, Optimal control of nonlinear systems, Comput. Math. Math. Phys. 42 (7) (2002) 931–956.
[26] W.M. Haddad, V. Chellaboina, Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach, Princeton University Press, Princeton, New Jersey, United States, 2008.
[27] C.J.C.H. Watkins, P. Dayan, Q-learning, Mach. Learn. J. 8 (3) (1992) 279–292.
[28] D.P. Bertsekas, J.N. Tsitsiklis, Neuro-Dynamic Programming, Athena Scientific, MA, 1996.
[29] C. Suzuki, H. Jacobsson, T. Hatschek, M.R. Torkzad, K. Boden, Y. Eriksson-Alm, E. Berg, H. Fujii, A. Kubo, L. Blomqvist, Radiologic measurements of tumor response to treatment: Practical approaches and limitations, Radiographics 28 (2) (2008) 329–344, doi:10.1148/rg.282075068.
[30] B. Gholami, N.Y. Agar, F.A. Jolesz, W.M. Haddad, A.R. Tannenbaum, A compressive sensing approach for glioma margin delineation using mass spectrometry, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society (2011) 5682–5685.
[31] J. Huang, B. Gholami, N.Y. Agar, I. Norton, W.M. Haddad, A.R. Tannenbaum, Classification of astrocytomas and oligodendrogliomas from mass spectrometry data using sparse kernel machines, in: Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, 2011, pp. 7965–7968.
[32] K. Pachmann, P. Heiß, U. Demel, G. Tilz, Detection and quantification of small numbers of circulating tumour cells in peripheral blood using laser scanning cytometer (lsc®), Clin. Chem. Lab. Med. 39 (9) (2001) 811–817.