Reinforcement learning-based control of drug dosing with applications to anesthesia and cancer therapy

9

Regina Padmanabhan[a], *Nader Meskin*[a], *Wassim M. Haddad*[b]
[a]Department of Electrical Engineering, Qatar University, Doha, Qatar, [b]School of Aerospace Engineering, Georgia Institute of Technology, Atlanta, GA, United States

1 Introduction

This chapter presents a general framework that utilizes reinforcement learning (RL)-based method to regulate multiple parameters during intravenous drug administration. First, the Q-learning algorithm which is an RL-based method is used to fine tune continuous infusion of the drug propofol for patients in the ICU. We control the infusion of propofol so as to keep the bispectral index (BIS) and mean arterial pressure (MAP) of the patient at a desired range. Next, the use of a similar Q-learning-based controller is discussed to regulate the drug titration while different drugs with synergistic interactive effects are administered simultaneously. Finally, an RL-based controller design strategy for cancer chemotherapy treatment is also presented.

1.1 Motivation

During the last few decades, the critical and complex task of anesthesia administration has been widely studied and discussed in the literature using clinical as well as in silico trials. Consequently, many recent reviews on the currently adopted strategies highlight several aspects of the problem that need further research attention (Ionescu et al., 2014; Van Den Berg et al., 2017). Moreover, common anesthetics, such as propofol and midazolam that are necessary for various medical procedures, have side effects that include cough, nausea, skin irritations, numbness, delirium, seizures, muscle pain, weak or shallow breathing, and hemodynamic instability. The overdosing of some anesthetic and analgesic drugs is known to cause death (Jacobi et al., 2002; Mehta et al., 2006). Typically, patients admitted to the intensive care units suffer from multiple illnesses which necessitate the use of many drugs for life support and treatment. When it comes to the combined administration of several drugs such as anesthetic, analgesic, neuromuscular blockades, and cardiac drugs, the fact that the mechanisms of action are complex, interlaced, and not yet completely understood makes the problem very challenging. In the case of the continuous and simultaneous

Control Applications for Biomedical Engineering Systems. https://doi.org/10.1016/B978-0-12-817461-6.00009-3

infusion of these drugs for long periods, it is evident that an appropriate closed-loop control approach can be used to improve patient safety (Absalom et al., 2011; Ionescu et al., 2014; Jacobi et al., 2002).

Cancer chemotherapy treatment is another important therapeutic approach that involves continuous infusion of intravenous drugs. Surveys conducted in the area of cancer diagnosis and treatment highlight that the relative survival rate for many types of cancer have improved significantly over the years (ACS, 2015; WHO, 2018). These reports suggest that proficiency in early diagnosis and improvement in treatment methods are the important factors that contributed to reducing the morbidity rate and mortality rate associated with cancer. Even though there has been obvious improvement in the overall prognosis, diagnosis, and treatment of cancer, a steady increase in the incidence of this disease is a matter of concern. Like any other drug-dosing application, there are several factors that determine the drug dose required for a patient to culminate certain desired response. In the case of cancer chemotherapy, these factors include the type and stage of cancer, age, and weight of the patient, immune response of the patient, and presence any other illness. Accordingly, the clinician chooses the type and amount of the drug to be given to a patient by following certain established treatment protocols and guidelines.

However, several clinical trials and scientific studies point out the limitations of this approach and highlight the need for optimal and patient-specific dosing of chemotherapeutic drugs (Chen et al., 2012; Sbeity and Younes, 2015). These literatures highlight the importance to conduct clinical and in silico trials to study the effectiveness and feasibility of novel chemotherapy, plan to improve the therapeutic benefits of the treatment (Sbeity and Younes, 2015). However, clinical trials are often tedious to conduct, require long trial time, and are expensive. On the other hand in silico trials are cost effective and provide flexible techniques to evaluate novel treatment plans.

Even though several control methodologies have been suggested in the literature for the closed-loop control of intravenous drug administration, very few findings have attracted the attention of clinicians. This is mainly due to the discrepancy between the actual clinical requirement and the one that is considered for study. An ideal closed-loop controller that can effectively facilitate the complex task of drug delivery should account for multiple clinical phenomena such as drug interaction, drug overdosing and underdosing, significant variabilities in the drug response(s) of different patients, nonlinearities and disturbances in the system, and major drug-induced side effects such as immunosuppression or hemodynamic instability.

1.2 Literature review

1.2.1 Drug-dosing control for anesthesia administration

Anesthesia is mainly used to facilitate invasive and painful clinical procedures such as endotracheal intubation, ventilation, suction, and hemodialysis. Too much or too little anesthetic can cause increased morbidity. Hence, the rate of infusion of anesthetic drugs is critical, requiring continuous monitoring and repeated adjustments (Haddad et al., 2010). Typically, open-loop drug infusion is facilitated by a medical

practitioner or via target controlled infusion (TCI) pump (Absalom and Mason, 2017; Absalom et al., 2011; Masui et al., 2010). TCI pumps are programmed to derive the required drug dose for a patient by using a nominal model of the patient. However, recent investigations in the area of anesthetic and analgesic drug dosing have documented several positive outcomes of the closed-loop control approaches compared to open-loop ones (Absalom et al., 2011; Brogi et al., 2017; Kuizenga et al., 2016; Soltesz et al., 2013). Specific advantages of the closed-loop control approaches include improved patient safety, early recovery time, and reduced treatment cost. Moreover, closed-loop control relieves the clinicians from doing frequent mechanical adjustments which in turn allow them to indulge in more critical aspects of therapy to improve overall well being of the patient (Haddad et al., 2010).

Patients admitted to ICU often suffer from multiple illnesses or even organ system failure. Hence, it is necessary to evaluate the health of these patients using various physiological monitors and provide required assistance using life-supporting devices. Some of the life-supporting procedures such as mechanical ventilation involve invasive endotracheal tube insertion which leaves the patient in physical as well as mental distress. Moreover, due to anxiety and discomfort related to these procedures the patients are often restless and in an incoherent state of mind. Hence, in order to comfort the patients and to perform painful clinical procedures in a cooperative and safe manner, often these patients are kept in a state of moderate sedation for a long period of time. Apart from the complications in the normal physiological functioning of the body which arise due to an inherent illness, side effects of the drugs used for treatment can also have an adverse effect on the overall health of these patients. For instance, most of the sedatives and analgesics used these days are identified to impair cardiac and respiratory functions (Absalom et al., 2011; Jacobi et al., 2002; Minto et al., 2000; Robinson et al., 1997). Thus, the critically ill patients in the ICUs who are treated using multiple intravenous drugs for long periods also demand the regulation of multiple physiological variables such as MAP, heart rate, respiratory rate, level of unconsciousness and pain sensation, and other vital parameters within acceptable safe limits (Heusden et al., 2018; Jacobi et al., 2002).

Analyzing drug anesthetic effects requires pharmacokinetic (PK) models to account for the drug disposition and pharmacodynamic (PD) models to capture drug concentration effects. In order to formulate the mathematical equivalent of a human drug disposition system with a time-dependent drug dose as an input signal, several physiological and nonphysiological models have been proposed (Absalom et al., 2009; Haddad et al., 2010). Among these, deterministic PK models, represented by compartmental models, which involve single or multiple compartments to capture the drug distribution and metabolism have gained wide acceptance (Absalom et al., 2009; Masui et al., 2010). In the case of intravenous infusion of anesthetic drugs, the mechanism of drug disposition can be effectively represented using a three-compartmental model with an additional effect-site compartment to model the time-lag in the drug dynamics at the locus of the drug effect (Masui et al., 2010). It should be noted that underlying illness, drug interaction, and other clinical disturbances alter the drug requirements (Absalom et al., 2011; Jacobi et al., 2002; Minto et al., 2000; Robinson et al., 1997).

Advancements in the area of automation and control engineering have fostered human health care in many ways. There exist many control methods that have been successfully used to design controllers for applications that require tracking a certain desired response. However, the requirement for an accurate mathematical model that depicts human physiology and difficulty in measuring certain system parameters that are required for feedback are the two main hurdles that limit the utilization of such control methods in the area of drug dosing. Several clinical and in silico trials conducted to evaluate the efficacy of the fixed-gain and linear controllers for the closed-loop control of anesthesia administration have proved inadequate (Absalom et al., 2011; Bailey and Haddad, 2005; Haddad et al., 2013; Hahn et al., 2012; Soltesz et al., 2013). This set back is mainly due to the complexity and uncertainty involved in the intricate task of anesthesia administration.

Furutani et al. (2010) reported 79 clinical trials conducted to evaluate the performance of model predictive controllers (MPC) for the closed-loop control of anesthesia administration. This study marks improved performance of the closed-loop control approach over manual control in terms of the amount of drug used and tracking error in reference output (BIS). However, the performance of the MPC-based controller was not so good compared to the that reported by Morley et al. (2000), Absalom and Kenny (2003), Liu et al. (2006), and Struys et al. (2001). Even though optimal control methods can account for system state constraints and control constraints, as pointed out by Furutani et al. (2010) such methods demand more accurate mathematical model to improve the tracking ability and robustness of the closed-loop control system. Haddad et al. (2003) documented the improved performance of adaptive disturbance rejection controller in addressing the system uncertainties and system disturbances associated with anesthesia administration. However, adaptive controllers cannot embody optimality requirements of the system optimality. Thus, it is necessary to develop novel methods that are capable of addressing problems that arise due to the system disturbances and system uncertainties, while deriving at optimal control laws to enhance the applicability and safety of automated anesthesia administration.

1.2.2 Drug-dosing control for cancer chemotherapy

Most of the cancer chemotherapy control algorithms reported in the literature are implemented using optimization methods (Chen et al., 2012, 2014; Doloff and Waxman, 2015; Engelhart et al., 2011; Kiran et al., 2009; Noble et al., 2010; Swierniak et al., 2003). Chen et al. (2012) and Noble et al. (2010) discussed an MPC-based controller which uses a new state measurement at the end of each sampling period to update the model used for solving the optimization problem. Kiran et al. (2009) used a multiobjective optimization approach to regulate the use of therapeutic agents and derive optimal treatment schedule for immunotherapy and chemotherapy. Similarly, Engelhart et al. (2011) investigated the problem of deriving optimal treatment plan for immunotherapy, chemotherapy, or/and antiangiogenic therapy with respect to various objective functions.

Batmani and Khaloozadeh (2013) and Çimen (2010) resorted to state-dependent Riccati equation (SDRE)-based controller design approach for deriving treatment

schedule for cancer chemotherapy. Specifically, Batmani and Khaloozadeh (2013) used a state observer to estimate the unavailable system states. In Babaei and Salamci (2015), a hybrid method that compounds SDRE and model reference adaptive controller design method is used to determine a personalized drug dose for cancer treatment. As mentioned earlier, the efficacy of the optimal control approaches depends on the accuracy of the mathematical used. However, it is often impossible to derive an ideal mathematical model which can accommodate all the complex dynamics involved in the tumor microenvironment (Pillis and Radunskaya, 2001; Sbeity and Younes, 2015; Swan, 1990). Typically, these dynamics include the tumor growth, immune response to tumor growth, changes in the vascular network that supply nutrients to the tumor, and the effect of the drug on various cell types in the tumor microenvironment to name some.

Evolutionary algorithms (EA)-based approaches have also been used to derive optimal drug-dosing schedules for chemotherapy (Tan et al., 2002; Tse et al., 2007). Even though the EA-based approaches exhibit competitive performance compared to the other existing chemotherapy optimization approaches, difficulty in the selection of the initial population and significant computation effort involved limits the acceptance of these methods (Sbeity and Younes, 2015).

1.2.3 RL-based algorithms

Even though several control methodologies have been suggested in the available literature for the closed-loop control of intravenous drug administration, very few findings have attracted the attention of clinicians. This is mainly due to the discrepancy between the actual clinical situation and the one that is considered for study. An ideal closed-loop controller that can effectively facilitate the complex task of drug delivery should account for multiple clinical phenomena such as drug interaction, drug overdosing and underdosing, significant variabilities in the drug response(s) of different patients, nonlinearities and disturbances in the system, and major drug-induced side effects such as immunosuppression or hemodynamic instability. RL-based control is a novel promising approaches for the control of intravenous drug administration to achieve multiple clinical objectives simultaneously.

RL-based methods have been used for many years to derive optimal control inputs in the presence of system disturbances and in the absence of knowledge of complete system dynamics (Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998; Vrabie et al., 2013). RL algorithms arrive at an optimal solution by performing control policy updates based on a reward or performance index defined with respect to the controlled system. Such algorithms are based on dynamic programming and give optimal solutions when the iterations converge (Barto et al., 1983; Sutton, 1988; Sutton and Barto, 1998). Moreover, these are interactive algorithms which can account for time-varying system dynamics and performance requirements (Vrabie et al., 2013). RL methods rely on the speed, efficiency, and computational advantages of the digital computers to assess the impact of each possible control action on the system and derive the best control action in an uncertain noisy environment.

RL-based control strategies have exhibited satisfactory performance in the areas of aeronautics, robotics, and clinical pharmacology when used for the control, automation, motion planning, signal processing, and networking (Abbeel et al., 2007; Dadhich et al., 2016; Hong et al., 2016; Sedighizadeh and Rezazadeh, 2008). RL methods can derive an optimal controller by exploring the advantage of each possible action in driving the system to a target (goal) state. After training, the controller uses the learned optimal control policies to regulate the transience of the system under control from an arbitrary initial state to the goal state. RL is suitable for deriving optimal drug-dosing schedules mainly because this method does not require the model of the system and it can learn the best sequence of actions or the optimal control law using the response of the patient to the control input (drug infusion). In the context of RL, the term agent is used as a synonym of the term controller in the field of control theory (Sutton and Barto, 1998). Here, a policy can be either a function of system states, or a path or a plan to transition the system from an arbitrary initial state to the goal state, or it can even be rule-based such as "if in this state, then do this." A reward function is used to assess the advantage of an action with respect to system states.

Recently, RL-based control strategies have been used in the drug-dosing control scenario to optimize the dosing of erythropoietin during hemodialysis, develop dynamic treatment regime for patients with lung cancer, assist insulin regulation in diabetic patients, regulate heparin dosing, and administer anesthetic drugs to induce and maintain the desired sedation level (Daskalaki et al., 2013; Martin-Guerrero et al., 2009; Moore et al., 2014; Nemati et al., 2016; Zhao et al., 2011). Moore et al. (2010) discussed RL-based optimal controller for the regulation of hypnosis during surgery. Specifically, the authors derived optimal control solutions by penalizing the control actions that correspond to an increase or decrease in BIS value from the target BIS value and rewarding the control actions that maintain the BIS output of the patient at the target value of BIS. The authors used three control actions (drug dose) u such as 0, 20, or 40 mg to train the RL-based controller. Currently, bedside monitors that can provide a measure of the depth of anesthesia in terms of BIS index value are available (Johansen et al., 2000). Extending the RL-based controller presented by Moore et al. (2010, 2014) and conducted the first clinical trial to evaluate the closed-loop control of hypnosis using human volunteers. This RL-based controller showed a patient-specific control of hypnosis with an enhanced control accuracy with respect to other similar investigations in the literature.

The remaining chapter is organized as follows. In Section 2, a general framework is presented to formulate the intravenous drug-dosing control problem in a finite Markov decision process (MDP) framework and develop a Q-learning based controller to design a multiobjective controller to regulate anesthesia administration by accounting for important physiological parameters of the patients. In Section 3, an optimal drug-dosing profile is derived by accounting for PK and PD disturbances such as drug interaction in the human body under treatment. Finally, in Section 4, the Q-learning-based controller design approach presented in Section 2 is adapted to address specific cases in cancer chemotherapy treatment.

2 Control of BIS by accounting for MAP

The aim of this section is to develop a multiobjective controller that can regulate sedation by simultaneously accounting for drug-induced hemodynamic instability in a patient. Administration of sedative drug such as propofol can have adverse effects on the hemodynamic stability of the patient. Specifically, propofol causes vasodilation leading to the decrease in MAP, drug overdose, and even cardiovascular collapse (Fan et al., 2012). Consequently, during propofol infusion, along with regulating the desired drug response (BIS index) it is important to maintain hemodynamics parameters (e.g., MAP) of the patient in an clinically acceptable and safe range (Haddad et al., 2010; Rao and Bequette, 2000). Toward this end, first, a general framework for the development of an RL-based controller for the control of nonlinear dynamical systems is presented. Next, the PK and PD models of propofol related to patient responses such as the BIS index and MAP are discussed. Here, this model serves as a patient model which is then used to generate input-output data required to train the RL agent or the RL-based controller.

2.1 Problem formulation

The problem of obtaining control solutions to follow a desired system trajectory often requires sequential decision making and can be solved by representing it in a finite MDP framework (Vrabie et al., 2013). During anesthesia administration, the aim is to reach a defined goal state (desired BIS value) with decisions predicated on the best sequence of control actions (propofol infusion) required to transition the system from a given arbitrary initial condition to the desired goal state. Toward this end, consider the nonlinear dynamical system given by

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0, \quad t \geq 0, \tag{1}$$

$$y(t) = h(x(t)), \tag{2}$$

where $x(t) \in \mathbb{R}^n$, $t \geq 0$, is a vector with n states of the system as the elements, $u(t) \in \mathbb{R}$, $t \geq 0$, denotes the control input, $y(t) \in \mathbb{R}^l$, $t \geq 0$, represents l number of outputs or responses of the system, $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ is locally Lipschitz continuous and $h : \mathbb{R}^n \to \mathbb{R}^l$ is continuous. RL-based control approaches are suitable for problems that require a goal-oriented decision making (Sutton and Barto, 1998).

A finite MDP can be defined using the four-tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$, where $\mathcal{S}$ is a finite set of states of the system or environment, $\mathcal{A}$ is a finite set of possible actions when in the states $s_k \in \mathcal{S}$, $\mathcal{P}$ represents a state transition probability matrix, $\mathcal{P}_{a_k}(s_k, s_{k+1})$ is the probability that the state $s_k \in \mathcal{S}$ at k transits to the state s_{k+1} at $k+1$ with an action $a_k \in \mathcal{A}$ at k, and $\mathcal{R}$ represents a reward that assesses the advantage of an action $a_k \in \mathcal{A}$ for all $s_k \in \mathcal{S}$. Note that the transition probability matrix denoted as $\mathcal{P}$ represents the system dynamics and is equivalent to the function $f(\cdot, \cdot)$ which is assumed to be unknown. The discrete states in the finite set $\mathcal{S}$ are denoted as $(\mathcal{S}_i)_{i \in \mathbb{I}^+}$, where $\mathbb{I}^+ \triangleq \{1, 2, \ldots, q\}$ and q

represents the total number of states. Similarly, the discrete actions in the finite sequence $\mathcal{A}$ are denoted as $(\mathcal{A}_j)_{j\in\mathbb{J}^+}$, where $\mathbb{J}^+ \triangleq \{1,2,\ldots,p\}$ and p represents the total number of actions.

RL-based methods, such as Q-learning (Watkins and Dayan, 1992), have gained significant attention in recent years. The main reason for the increased acceptance of RL-based methods is the fact that it does not rely on a model of the system for the design of the controller. Moreover, RL-based methods can account for changes in the system that happens during the learning phase. Fig. 1 shows the schematic diagram of an RL-based approach in which the agent or controller learn a useful policy or an action plan using the information on the action taken, reward observed, and a new state to which the system reached due to the current action. In other words, the Q-learning-based controller design method can train an RL agent to learn the best sequence of control actions to regulate the states s_k of the system without using the system state $x(t)$, $t \geq 0$. Instead it uses the information gathered at time steps $k \in \{1, 2, \ldots\}$, $kT \leq t < (k+1)T$ along the system trajectories. At every time step k, the RL agent identifies the current state s_k from the set $\mathcal{S}$ and then it chooses an action a_k from the defined action set $\mathcal{A}$. Consequently, the system stochastically transitions from the current state s_k to a new state s_{k+1} incurring a numerical reward $r_{k+1} \in \mathbb{R}$.

Since learning is predicated on the knowledge of the discrete states $s_k \in \mathcal{S}$ and which should be measurable at time step k. Hence, the states s_k of the RL environment are defined with respect to the system response given by $y(t)$, $t \geq 0$, as

$$s_k = g(y(t)), \quad kT \leq t < (k+1)T, \tag{3}$$

where $g: \mathbb{R}^l \to \mathcal{S} \subset \mathbb{R}$ is a mapping between the system response $y(t)$ and the state representation s_k, $k = 1, 2, \ldots$. Here, $T > 0$ is the sampling time.

The agent aims to maximize the reward it earns over an infinite horizon. This can be achieved by using different strategies (Watkins and Dayan, 1992). A straightforward approach is to choose each action a_k such that it maximizes the expected value of the discounted return (Moore et al., 2014; Watkins and Dayan, 1992). In this case, the objective function is given by

$$J(R_k) = \mathbb{E}\left[\sum_{i=1}^{\infty} \theta^{(i-1)} r_{i+k}\right], \tag{4}$$

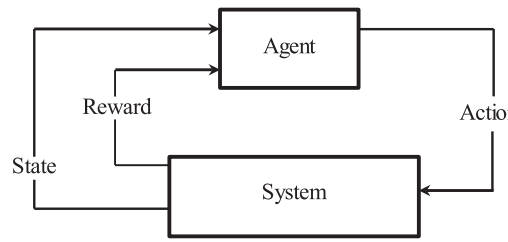

Fig. 1 Reinforcement learning schematic (Padmanabhan et al. (2015)).

where $\mathbb{E}[\cdot]$ denotes expectation, R_k denotes the total discounted return, and $\theta \in [0, 1]$ is a *discount rate parameter* which represents the horizon of interest to the agent. For $\theta = 0, J(R_k) = r_k$, that is, for learning, the agent considers only the current reward. Alternatively, for θ approaching 1, the weight of the costs incurred in the future is increased.

2.2 Learning an optimal policy

RL-based control relies on learning an optimal control policy while interacting with the system. Information obtained while interacting with the system is used to enhance the agent's decision-making policy over time. Thus, the agent interacts with the system to learns the optimal policy starting from an initial arbitrary policy. In the case of linear systems, optimal control law pertaining to the certain defined objective function and system constraints can be derived by solving associated algebraic Riccati equation. However, deriving optimal control law for nonlinear systems is tedious and requires the solution of complex Hamilton-Jacobi-Bellman partial differential equation (Balashevich et al., 2002; Haddad and Chellaboina, 2008).

Watkin's Q-learning is an RL-based approach which uses each state transition to update each entry of a table Q which forms the control policy. The policy is stored in a table so that appropriate responses can be retrieved quickly with respect to the state of the system. The entry $Q(s_k, a_k)$ of the Q table for each pair of state s_k and action a_k, $k \in \{1, 2, \ldots\}$ shows the value of the state s_k when associated with action a_k. The controller or RL agent assess the measured variables, and implement control actions according to the learned optimal policy given by $Q(s_k, a_k): \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ (see Fig. 1).

For every k and state s_k, the controller or agent selects the control action a_k as

$$a_k = \arg\max_{a \in \mathcal{A}} Q(s_k, a). \tag{5}$$

The numerical reward $r_k \in \mathbb{R}$ guides the agent to whether the action chosen at the time step k was "good" or "bad." After the transition $s_k \to s_{k+1}$, having taken an action a_k and received a reward r_{k+1}, the Q table is updated by

$$Q_k(s_k, a_k) \leftarrow Q_{k-1}(s_k, a_k) + \eta_k(s_k, a_k)[r_{k+1} + \theta \max_{a_{k+1}} Q_{k-1}(s_{k+1}, a_{k+1}) - Q_{k-1}(s_k, a_k)], \tag{6}$$

where $\eta_k(s_k, a_k) \in [0, 1)$ is the learning rate or step size parameter that is related to the size of adjustment after each experiment and θ denotes the discount rate parameter.

It has been shown by Bertsekas and Tsitsiklis (1996), Sutton and Barto (1998), and Watkins and Dayan (1992) that the Q-learning algorithm (6) converges to the optimal Q-function while maximizing Eq. (4) with probability one as long as

$$\sum_{k=1}^{\infty} \eta_k(s_k, a_k) = \infty, \quad \sum_{k=1}^{\infty} \eta_k^2(s_k, a_k) < \infty, \quad (s_k, a_k) \in \mathcal{S} \times \mathcal{A}. \tag{7}$$

Note that $\sum_{k=1}^{\infty}\eta_k(s_k,a_k)=\infty$ requires that all state-action pairs (s_k, a_k) are visited infinitely often, whereas $\sum_{k=1}^{\infty}\eta_k^2(s_k,a_k)<\infty, (s_k,a_k)\in\mathcal{S}\times\mathcal{A}$ is the condition required to ensure convergence of the algorithm with probability one. As mentioned earlier, the Q-learning algorithm starts with an initial arbitrary estimate of the unknown $Q(s, a)$ and then the algorithm iteratively updates the Q table until convergence is achieved, that is, until $\Delta Q = 0$, where ΔQ is the change in the Q table, or when the updates satisfy a minimum threshold $\Delta Q \leq \delta$, where δ is a prespecified tolerance parameter. Note that the optimal value of the Q table depends on the parameter values that are used for each iteration.

The framework introduced in this section is used to develop a Q-learning-based controller for the closed-loop regulation of the BIS and MAP by controlling the continuous infusion of propofol. The controller is designed to regulate the system output $y(t)\in\mathbb{R}^l, t\geq 0$, using the control input $u(t)\in\mathbb{R}, t\geq 0$. For our simulation, we use the simulated patient model as introduced in the following section to train the RL agent. Thus, the RL system or the environment shown in Fig. 1 is replaced by the nominal PK and PD model of the patient as shown in Fig. 2.

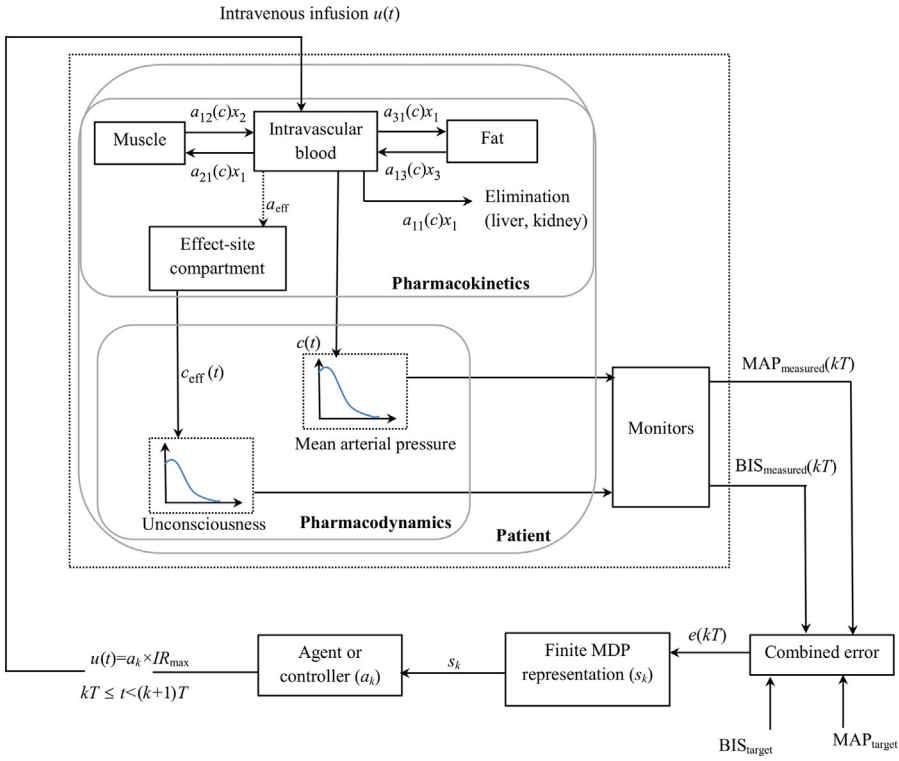

Fig. 2 Closed-loop control using RL showing the interaction between the agent and the simulated patient for simultaneous regulation of BIS and MAP management (Padmanabhan et al. (2015)).

2.3 Pharmacokinetic and pharmacodynamic patient model

As mentioned earlier, propofol has interlaced effects on the consciousness and hemodynamic stability of the patient. Specifically, most anesthetic drugs reduce sympathetic tone and alter arterial pressure of the patient by bringing about venodilation. Propofol infusion also decreases cardiac output and thereby reduces drug disposition. This reduction in disposition of the drug in body is compensated by titrating more drug which may lead to overdose. A nonlinear dynamical patient model given by the function $f(x(t), u(t))$ is used to represent the nonlinear PK and PD of the drug, where $u(t)$, $t \geq 0$, denotes the intravenous infusion of the drug propofol and $x(t)$, $t \geq 0$, is the system states.

As shown in Fig. 2, a nonlinear three-compartment model with an effect-site compartment is used to represent the patient dynamics. The control input is the continuous infusion of propofol to the central compartment. In this model, $x_1(t)$, $t \geq 0$, represents the amount of the drug in the arteries and veins (intravascular blood). In addition to the intravascular blood, $x_1(t)$, $t \geq 0$, also includes the mass of the drug in organs with very high blood supply such as the brain, heart, kidneys, and liver. The states $x_2(t)$, $t \geq 0$, and $x_3(t)$, $t \geq 0$, represent the rest of the drug in the body which is assumed to be in two peripheral compartments, comprised of muscle and fat, respectively. These two peripheral compartments receive less than 20% of the overall blood supply (cardiac output) in the body.

The three-compartment model is given by Haddad et al. (2010), Soltesz et al. (2013), and Absalom et al. (2009)

$$\dot{x}_1(t) = -[a_{11}(c(t)) + a_{21}(c(t)) + a_{31}(c(t))]x_1(t) + a_{12}(c(t))x_2(t) + u_1(t), \quad x_1(0) = x_{10}, \quad t \geq 0, \tag{8}$$

$$\dot{x}_2(t) = a_{21}(c(t))x_1(t) - a_{12}(c(t))x_2(t), \quad x_2(0) = x_{20}, \tag{9}$$

$$\dot{x}_3(t) = a_{31}(c(t))x_1(t) - a_{13}(c(t))x_3(t), \quad x_3(0) = x_{30}, \tag{10}$$

$$\dot{c}_{\mathrm{eff}}(t) = a_{\mathrm{eff}}(x_1(t)/V_{\mathrm{c}} - c_{\mathrm{eff}}(t)), \quad c_{\mathrm{eff}}(0) = c_{\mathrm{eff0}}, \tag{11}$$

where $c(t) = x_1(t)/V_{\mathrm{c}}$, $t \geq 0$, denotes the drug concentration in the central compartment denoted by $x_1(t)$, $t \geq 0$, and V_{c} represents the volume of the central compartment, $a_{ij}(c(t)) = A_{ij}\left(\mathrm{AC}_{50}^{\gamma_{\mathrm{a}}}/\left(\mathrm{AC}_{50}^{\gamma_{\mathrm{a}}} + (c(t))^{\gamma_{\mathrm{a}}}\right)\right)$, $i, j = 1, 2, 3$, denote the nonnegative mass transfer coefficients between the jth and ith compartment, A_{ij} are positive constants, γ_{a} is a parameter that determines the steepness of the concentration-effect relationship, and AC_{50} is the drug concentration associated with a 50% decrease in the transfer coefficient. The relation between the amount of the drug in the system and its effect on the output variables such as BIS and MAP follow a nonlinear sigmoidal dynamics and thus, the function $h(\cdot)$ in Eq. (2) can be modeled using the Hill equation given by Haddad et al. (2010)

$$h(x(t)) = [\mathrm{BIS}_{\mathrm{measured}}(c_{\mathrm{eff}}(t)), \mathrm{MAP}_{\mathrm{measured}}(c(t))]^T, \tag{12}$$

where $\mathrm{BIS}_{\mathrm{measured}}(c_{\mathrm{eff}}(t))$ and $\mathrm{MAP}_{\mathrm{measured}}(c(t))$ are the drug effects captured by

$$\mathrm{BIS}_{\mathrm{measured}}(c_{\mathrm{eff}}(t)) = \mathrm{BIS}_0\left(1-\frac{(c_{\mathrm{eff}}(t))^{\gamma}}{(c_{\mathrm{eff}}(t))^{\gamma}+(\mathrm{C}_{50})^{\gamma}}\right), \tag{13}$$

$$\mathrm{MAP}_{\mathrm{measured}}(c(t)) = \mathrm{MAP}_0\left(1-\frac{(c(t))^{\alpha}}{(c(t))^{\alpha}+(\mathrm{MC}_{50})^{\alpha}}\right), \tag{14}$$

where BIS_0 represents the baseline value, which is typically assigned a value of 100 to denote an awake state, C_{50} denotes the concentration of the drug related to the half-maximal effect of the BIS and models the patient's sensitivity to the drug, γ denotes the degree of nonlinearity, MAP_0 is the initial value of MAP of the patient before drug infusion, MC_{50} denotes the concentration of the drug related to the half-maximal effect of the MAP, and α denotes the degree of nonlinearity associated with MAP of the patient (Haddad et al., 2010).

2.4 Closed-loop control of BIS and MAP using RL

In this section, the Q-learning algorithm is used to develop a drug-dosing agent for the simultaneous regulation of anesthesia and hemodynamic status. The control variable $u(t)$, $t \geq 0$, in the dynamical system given by Eq. (1) is the continuous intravenous infusion of propofol. In the RL framework, since the agent interacts with the patient at discrete time steps the propofol infusion rate at each time step k is defined as

$$IR_k = a_k \times IR_{\max}, \tag{15}$$

where $k \in \{1, 2, \ldots\}$, $IR_{\max}$ is the maximum allowable infusion rate, and a_k is a particular action from the action set $\mathcal{A}$ selected at the kth time step. Thus, between any two time steps k and $k+1$, the infusion rate remains constant and is given by $u(t) = IR_k$, $kT \leq t < (k+1)T$, where T is the time duration between any two time steps. The action $a_k \in \mathcal{A}$ at the kth time step can vary from 0 (no infusion) to 1 (maximum rate of infusion) within the finite action set $\mathcal{A} = \{0, 0.01, 0.02, 0.03, 0.04, 0.05, 0.08, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$, where $\mathbb{J}^+ = \{1, 2, \ldots, 20\}$. Since $IR_{\max}$ is a configurable parameter, one of the benefits of the infusion rate scheme given by Eq. (15) is that it is easy to set its value according to the sedation requirements of the patient in the ICU.

In the RL framework, the controller or agent makes a decision about the action to be taken at each time step based on the current state of the system $s_k = g(y(t))$, $s_k \in \mathcal{S}$, $t \in [kT, (k+1)T)$. Hence, the state s_k of the system should be observable for decision making. Therefore, the states s_k of the RL system is defined based on the measurable parameters $\mathrm{BIS}_{\mathrm{measured}}(c_{\mathrm{eff}}(t))$ and $\mathrm{MAP}_{\mathrm{measured}}(c(t))$, $kT \leq t < (k+1)T$. The state s_k is defined based on the error $e(t)$, $kT \leq t < (k+1)T$, given by

$$e(t) = \sqrt{\beta_{\mathrm{w}}\mathrm{BIS}^2_{\mathrm{error}}(t) + \mathrm{MAP}^2_{\mathrm{error}}(t)}, \tag{16}$$

where $\beta_w > 0$ is a weighing factor, which can be used to weigh the importance of anesthesia control over hemodynamic control,

$$\mathrm{BIS}_{\mathrm{error}}(t) = \frac{\mathrm{BIS}_{\mathrm{measured}}(c_{\mathrm{eff}}(t)) - \mathrm{BIS}_{\mathrm{target}}}{\mathrm{BIS}_{\mathrm{target}}} \times 100, \tag{17}$$

and

$$\mathrm{MAP}_{\mathrm{error}}(t) = \frac{\mathrm{MAP}_{\mathrm{measured}}(c(t)) - \mathrm{MAP}_{\mathrm{target}}}{\mathrm{MAP}_{\mathrm{target}}} \times 100. \tag{18}$$

For ICU sedation, the agent aims to learn the best sequence of infusion rates which minimize $\mathrm{BIS}_{\mathrm{error}}$ and $\mathrm{MAP}_{\mathrm{error}}$. Hence, defining the system states denoted by s_k with respect to the error $e(t)$, $t \geq 0$, is reasonable. Moreover, using $e(t)$, $t \geq 0$, for training purposes has the advantage of involving single measurement given by Eq. (16) rather than two separate measurements of $\mathrm{BIS}_{\mathrm{error}}(t)$ and $\mathrm{MAP}_{\mathrm{error}}(t)$, $t \geq 0$. This decreases the complexity of the training algorithm. In this case, the action of the agent is predicated on the values of BIS and MAP. For our simulations, the parameters BIS and the MAP are calculated using the propofol concentration in the PD models (13), (14). In real time, both these variables can be measured in ICU using corresponding bedside monitors. To model possible measurement limitations in the BIS and MAP monitors, a sampling time $T = 6$ s is used. Thus, the agent interacts with the patient at every 6 s (Moore et al., 2014).

Here, the agent seeks to learn the best action sequence that will transition the system from given initial state to target states identified as $\mathrm{BIS}_{\mathrm{target}} = 65$ and $\mathrm{MAP}_{\mathrm{target}} = 80$. The range of output variables that are considered for our simulation are $\mathrm{BIS}_{\mathrm{measured}}(t) \in [0, 100]$ and $\mathrm{MAP}_{\mathrm{measured}}(t) \in [0, 120]$. Note that $\mathrm{BIS}_{\mathrm{error}}(t)$, $t \geq 0$, remains positive when $\mathrm{BIS}_{\mathrm{measured}}(t) \in (65, 100]$ and negative when $\mathrm{BIS}_{\mathrm{measured}}(t) \in [0, 65)$. However, this change in sign is not reflected in the value of $e(t)$ for $\mathrm{BIS}_{\mathrm{measured}}(t) \in (65, 100]$ and $\mathrm{BIS}_{\mathrm{measured}}(t) \in [30, 65)$. See that, as shown in Fig. 3, $e(t)$ when calculated using Eq. (16) gives the same value for $\mathrm{BIS}_{\mathrm{measured}}(t) \in (65, 100]$ and $\mathrm{BIS}_{\mathrm{measured}}(t) \in [30, 65)$. The agent should increase the infusion of the sedative drug when $\mathrm{BIS}_{\mathrm{error}}(t)$ is positive and decrease it when $\mathrm{BIS}_{\mathrm{error}}(t)$ is negative. In order to account for this, separate set of states is assigned for positive and negative values of $\mathrm{BIS}_{\mathrm{error}}(t)$, $t \geq 0$. Specifically, $s_k \in \{1, 2, \ldots, 13\}$ is assigned for $e(kT) \in [0, e_p(t)]$ and $s_k \in \{14, 15, \ldots, 20\}$ is assigned for $e(kT) \in [0, e_n(t)]$, where $e_p(t)$ and $e_n(t)$ denote the maximum error in the region of error $e(kT)$ where $\mathrm{BIS}_{\mathrm{error}}(t)$ is positive and negative, respectively. See Table 1 for the mapping between the error $e(kT)$ and state s_k. The entries in the Q table which corresponds to the states 1–13 for positive values of $\mathrm{BIS}_{\mathrm{error}}(t)$, $t \geq 0$, and 14–20 for negative values of $\mathrm{BIS}_{\mathrm{error}}(t)$, $t \geq 0$, are updated using Eq. (6).

It can be seen from Table 1 that there is a dense discretization of $e(kT)$ near the region where $e(kT) = 0$. Moreover, compared the case when $\mathrm{BIS}_{\mathrm{error}}(t)$ is negative, more number of states are assigned when $\mathrm{BIS}_{\mathrm{error}}(t)$ is positive. This is to account for the fact that when $\mathrm{BIS}_{\mathrm{error}}(t)$ is negative the patient is oversedated and hence the ideal infusion rate is zero as $e(kT)$ approaches the value 300.

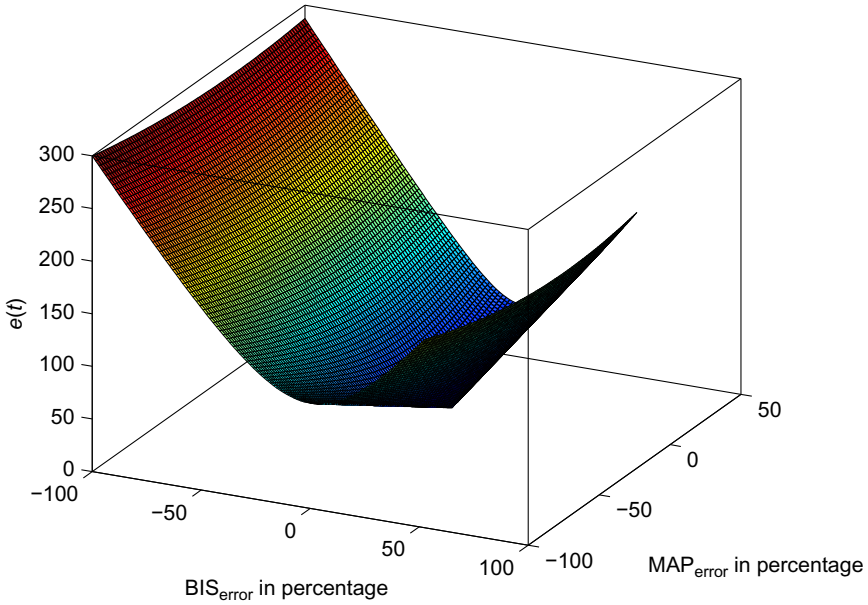

Fig. 3 Normalized percentage error of BIS and MAP versus combined error $e(t)$ (Padmanabhan et al. (2015)).

Table 1 State assignment based on $e(t)$ (Padmanabhan et al., 2015)

$\mathbf{BIS_{error} > 0}$		$\mathbf{BIS_{error} < 0}$	
State s_k	$e(kT)$	**State s_k**	$e(kT)$
1	[0, 2]	14	[0, 10]
2	[2, 4]	15	[10, 50]
3	[4, 10]	16	[50, 100]
4	[10, 15]	17	[100, 150]
5	[15, 25]	18	[150, 200]
6	[25, 35]	19	[200, 250]
7	[35, 45]	20	[250, 300]
8	[45, 60]		
9	[60, 80]		
10	[80, 100]		
11	[100, 120]		
12	[120, 140]		
13	[140, 165]		

On the other hand, when $\text{BIS}_{\text{error}}(t)$ is positive the patient is undersedated and the infusion rate should vary considerably according to how close is the patient from targeted BIS value.

Choosing an appropriate reward function is a very important step during the implementation of Q-learning-based algorithm. Note that the reward function is used to assess the advantage of each action in the action set. Reward function plays a key role in reinforcing the agent's decision-making policies and hence choosing reward function requires a careful consideration. For ICU sedation, it is apparent that the action that decreases the difference between the measured value of BIS and MAP denoted by $\text{BIS}_{\text{measured}}$ and $\text{MAP}_{\text{measured}}$ and the targeted value of BIS and MAP denoted by $\text{BIS}_{\text{target}}$ and $\text{MAP}_{\text{target}}$, respectively, must incur more reward. An appropriate reward has to steer the agent to learn the optimal policy for the regulation of BIS and MAP responses toward the required target values. Hence, the reward r_{k+1} corresponding to action a_k at k is computed by

$$r_{k+1} = \begin{cases} \dfrac{e(kT) - e((k+1)T)}{e(kT)}, & e((k+1)T) < e(kT), \\ 0, & e((k+1)T) \geq e(kT). \end{cases} \tag{19}$$

For an error $e((k+1)T) \geq e(kT)$, the algorithm assigns $r_{k+1} = 0$. This means that if certain action imparted to the system at k could not reduce the error at the time step $k+1$ then that action is given a zero reward. This assignment penalizes bad control actions. On the other hand, if certain action imparted to the system at the current time step reduces the error at the next time step, then that action is given a reward proportional to the difference in error $(e(kT) - e((k+1)T))$ between two time steps. Note that the Q table is updated using Eqs. (6), (19). Here, for each state s_k, the action in the set $\mathcal{A}$ that results in maximum value for $e(kT) - e((k+1)T)$ is assigned the highest value of reward.

RL-based algorithms exploit the computational power of computers to execute all possible actions from each state to asses which action will steer the system closer toward the desired target state. The aim is to drive the system from a given initial state $s_0 \in \mathcal{S}$, $\mathcal{S} = \{1, 2, \ldots, 20\}$, to the desired target state 1 as $k \to \infty$. A *policy* is defined as the sequences of state actions which can steer the system from an initial arbitrary state to the target state. Among the possible policies, the *optimal policy* is the one which earns a maximum reward. Thus, successful training is achieved when, for each state $s_k \in \mathcal{S}$, the agent identifies the best action a_k^* among all possible actions $a_k \in \mathcal{A}$ resulting in a maximum reward. Maximizing the reward in turn implies that the action a_k^* will drive the system closer to the desired state 1 as compared to all other possible actions in the given action set. The learned optimal policy is unique for a given set of states and action set (Sutton and Barto, 1998).

First, the RL agent learns by experimenting with the system using the permissible actions and assessing the response (output) of the simulated patient as shown in Fig. 4. For our simulations, the patient model is assigned an arbitrary initial condition.

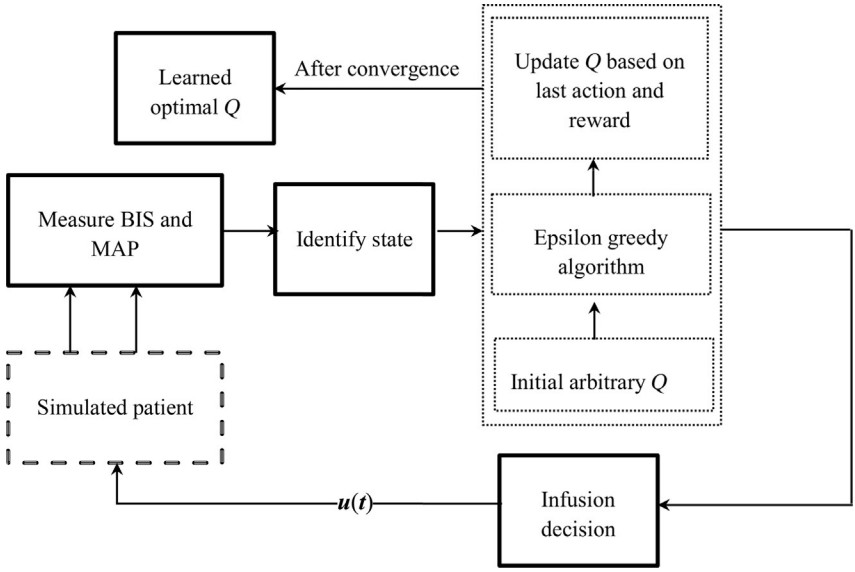

Fig. 4 Schematic representation of training sequence to obtain the optimal Q table (Padmanabhan et al. (2015)).

Note that, as shown in Fig. 2, the patient is replaced by a nominal population model which represents the PK and PD of the drug. The combined error $e(t)$, $t \geq 0$, is derived using the PD model for the response variables BIS and MAP given in Eqs. (13), (14), respectively. As shown in Fig. 4, using the value of $e(t)$, $t \geq 0$, the current state of the system is identified. Initially, the values of the Q table are set to zero. Using Eq. (5), in a Q table with all zero entries, the agent is always directed to choose the same action as a_k^*. In order to avoid this initial difficulty, and to facilitate learning, an ϵ-greedy policy can be used (Sutton and Barto, 1998). Using ϵ-greedy policy, the agent executes random actions with probability ϵ, where ϵ is a small positive number. These random actions help the agent to gather information according to the pharmacology of a patient. Toward this end, the agent infuses propofol at different rates defined in the action set and observes the response of the patient. After each experiment, the agent calculates the reward incurred and updates the corresponding state-action entry in the Q table using Eqs. (6), (19), respectively. This is done to associate each state with the best action in the action set $\mathcal{A}$.

Given the current state $s_k \in \mathcal{S}$, a "good" control action $a_k \in \mathcal{A}$ imparted by the controller results in a positive reward ($r_{k+1} > 0$) for $e((k+1)T) < e(kT)$. Similarly, a "bad" control action $a_k \in \mathcal{A}$ results in zero reward ($r_{k+1} = 0$), for $e((k+1)T) \geq e(kT)$. Note that r_{k+1} is used in Eq. (6) to update the Q table. To facilitate learning, the agent tries every possible action $a_k \in \mathcal{A}$ for all possible states $s_k \in \mathcal{S}$ and observes the utility of each action in earning a positive reward (Sutton and Barto, 1998). According to Eq. (7), to arrive at an optimal policy with respect to a defined set of states and actions,

the controller or the RL agent should explore all states and actions and utilize the information pertaining to the previous trails that were useful or effective in incurring more reward (Sutton and Barto, 1998). With $k \rightarrow \infty$, all the defined states and actions in the Q table will be executed recurrently which enables the Q table to converge the optimal Q table. Another condition required to ensure convergence of the Q table and to learn the optimal policy is to reduce the learning rate $\eta_k(s_k, a_k)$ defined in Eq. (7) over time (Sutton and Barto, 1998).

2.5 Details of the simulation

In this section, simulation results are presented to illustrate the use of RL-based controller for the closed-loop control of BIS and MAP. Simulations are conducted by setting iteration number to 50,000 (arbitrarily high) scenarios, where a scenario represents the series of transitions from an arbitrary initial state to the required final state 1. Furthermore, initially $\eta_k(s_k, a_k) = 0.2$ (for scenarios 1–499) is assigned and subsequently halved $\eta_k(s_k, a_k)$ every 500th scenario. For each scenario, a new set of randomized initial states $x_1(0) \in [0, 0.084]$ g, $x_2(0) \in [0, 0.067]$ g, $x_3(0) \in [0, 0.039]$ g, and $c_{\text{eff}}(0) \in [0, 0.005]$ g L^{-1} of propofol was assigned to the simulated patient model and then the learning phase was repeated until convergence and the performance goals were met; that is, keeping the BIS and the MAP values within the desired ranges. For our simulation, the Q table converged before reaching the maximum iteration. After convergence, for every state s_k, the agent chose an action $a_k = \arg\max_{a \in \mathcal{A}} Q(s_k, a)$.

After the training phase, that is, once the agent learned the optimal sequence of infusion rates required for each state $s_k \in \mathcal{S}$ to reach the desired goal state, the efficacy of the learned agent in a sequence of scenarios is evaluated over individual patients to check how well the agent can perform drug administration based on its optimal control policy during practical situations.

During anesthesia administration oversedation and undersedation is not acceptable. Hence, after training exploration or random actions are avoided to update the Q table, but used the optimal $Q(s_k, a_k)$ discussed in the previous section for making drug infusion decisions for the 30 simulated patients.

The value of $\text{BIS}_{\text{error}}(t)$, $t \geq 0$, and $\text{MAP}_{\text{error}}(t)$, $t \geq 0$, is in range the range of $0\%-100\%$. Next, in order to prioritize the control of BIS over MAP a positive-weighing parameter β_{w} is used. A high value for β_{w} decreases the regulation of MAP, on the other hand choosing a small value for β_{w} will reduce the control of BIS. For our simulation $\beta_{\text{w}} = 8$ is used which is set by trial and error. The recommended dose of propofol given in ASHP guidelines (Jacobi et al., 2002) is an initial bolus (20 mg) followed by continuous infusion ($5-80$ $\mu\text{gkg}^{-1}\text{min}^{-1}$). The maximum amount of drug required for a 100-kg patient during the maintenance phase of anesthesia administration is 8 mg min^{-1}. Hence, for training the RL agent $IR_{\max} = 20$ mgmin^{-1} is used.

The evaluation of the performance of the Q-learning algorithm-based controller is conducted in 30 simulated patient models using hypnosis scenarios that lasted for 2 h.

Table 2 Perturbation values (Padmanabhan et al., 2015)

Parameter	Perturbation range
Concentration at half maximal effect of BIS, C_{50}	0.004 ± 0.001 g L^{-1}
Concentration at half maximal effect of MAP, MC_{50}	0.004 ± 0.001 g L^{-1}
Concentration at half maximal effect of a_{ij}, AC_{50}	0.004 ± 0.001 g L^{-1}
Degree of nonlinearity of BIS(c_{eff}), γ	3 ± 1
Degree of nonlinearity of MAP(c), α	3 ± 1
Degree of nonlinearity of a_{ij}, γ_a	3 ± 1
Time lag between $c_{eff}(t)$ and $c(t)$, a_{eff}	$\in [0.17, 1]$ (min^{-1})
Volume of central compartment, V_c	16 ± 1 L
Transfer coefficients, a_{ij}	$\pm 0.5\%$ (min^{-1})

The pharmacological parameter values of the 30 simulated patients are taken randomly from a predefined parameter range as given in Table 2. In addition, $A_{11} = 0.119\text{min}^{-1}$, $A_{12} = 0.0550\text{min}^{-1}$, $A_{21} = 0.112\text{min}^{-1}$, $A_{31} = 0.0419\text{min}^{-1}$, and $A_{13} = 0.0033\text{min}^{-1}$ are used (Bailey and Haddad, 2005). Unlike surgery which requires deep sedation, often the procedures in ICU can be carried out with moderate sedation. Thus, for our simulation, the target values of output variables are set to $\text{BIS}_{\text{target}} = 65$ and $\text{MAP}_{\text{target}} = 80$. Simulation results showing the steady-state performance of the Q-learning-based anesthesia control approach for two case studies are presented. Statistical analysis pertaining to these simulation studies is also conducted.

2.6 Results and discussion

Fig. 5 shows the implementation of the closed-loop control strategy using the learned optimal policy. At each time step K, the RL agent chooses an infusion rate based on the learned optimal Q table. Note that, even though the training of the agent is done using a simulated patient model, performance evaluation is conducted on a population of 30 simulated patients. To study the efficacy of the trained RL agent in the closed-loop regulation of anesthesia, common performance matrices such as the root mean square error (RMSE), median performance error (MDPE), and median absolute performance error (MDAPE) are used (Moore et al., 2014). The performance error (PE) is defined as

$$\text{PE}_i(j) \triangleq \frac{\text{Measured value}_i(j) - \text{Target value}}{\text{Target value}} \times 100, \quad j = 1, \ldots, N, \tag{20}$$

where $i \in \{1, \ldots, 30\}$ represents the ith patient, j represents the set of PE measurements for an individual, N is the number of measurements for each patient, and Measured value and Target value in Eq. (20) refer to BIS and MAP, respectively. Note that for the controlled variables BIS and MAP, the PE is the same as the $\text{BIS}_{\text{error}}(t)$ and

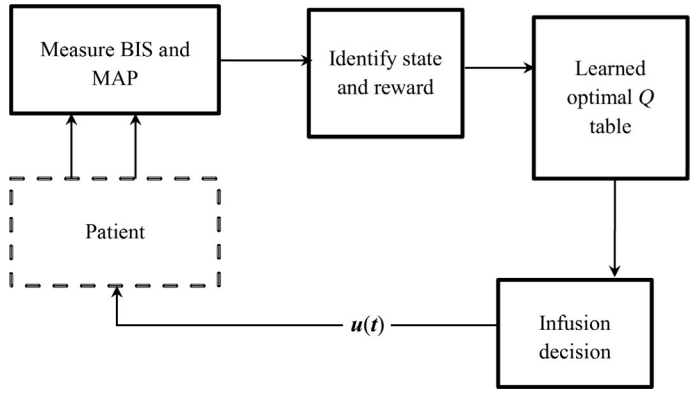

Fig. 5 RL-based optimal and robust closed-loop control of BIS and MAP (Padmanabhan et al. (2015)).

$\text{MAP}_{\text{error}}(t)$, $t \geq 0$, given by Eqs. (17), (18), respectively. The MDPE gives the control bias observed and is computed by

$$\text{MDPE}_i = \text{median}(\text{PE}_i(j)), \quad j = 1, \ldots, N, \tag{21}$$

whereas

$$\text{MDAPE}_i = \text{median}(|\text{PE}_i(j)|), \quad j = 1, \ldots, N, \tag{22}$$

and

$$\text{RMSE}_i = \sqrt{\frac{\sum_{j=1}^{N} (\text{Measured value}_i(j) - \text{Target value})^2}{N}}, \tag{23}$$

where MDAPE_i denotes the median of the absolute value of PE and it reflects the accuracy of the trained RL agent in keeping the targeted values of the control variables BIS and MAP for each of the 30 simulated patients (Moore et al., 2014). RMSE_i is the RMSE for each patient. Table 3 shows the performance metrics for the Q-learning-based agent for 30 simulated patients during the 2 h of hypnosis scenario considered. The amount of inaccuracy reflected in the values of the MDAPE metrics listed in Table 3 is in the acceptable clinical performance range (Moore et al., 2014).

To further elucidate the performance of the RL agent, the central tendency as well as the range of measured variables is evaluated for all of 30 simulated patients. Specifically, the amount of time that the outputs are within a desired band of the targeted values, that is, ± 5, and the percentage of all of the patients for which the outputs are within a predefined band is calculated. For the 2-h drug infusion period considered in our simulation, the measured value of the output variable BIS is within ± 5 of $\text{BIS}_{\text{target}}$ for 90.41% of the time for all 30 simulated patients. Similarly, the measured value of

Table 3 Performance metrics for control variables BIS and MAP (Padmanabhan et al., 2015)

Performance metrics (for 30 patients)	Controlled variables	
	BIS	MAP
MDPE (%)	3.97 ± 2.32	4.05 ± 2.50
MDAPE (%)	4.19 ± 6.43	5.31 ± 5.30
Min−max	66.43−68.25	75.52−89.46
Interquartile range	0.55	7.16
RMSE	2.12−3.30	2.30−9.50

the output variable MAP is within ±5 of MAP_{target} for 76.65% of the time for 60% of 30 simulated patients.

Table 3 shows the minimum and maximum values of BIS and MAP, respectively, during the maintenance period of drug administration. The time range $t \in [0, 10)$ and $t \in [10, 120]$ are considered as the induction period and maintenance period of anesthesia administration, respectively. This table also lists the variability or mid-spread of the controlled variables determined in terms of the interquartile range (IQR). IQR is the value of the middle of a data set arranged in ascending order. In order to obtain the IQR, the average value of the controlled variables BIS and MAP during $t \in [10, 120]$ for all of the simulated patients is used. The IQR of BIS variable is 0.55 and that of MAP is 7.16. Note that BIS has comparatively lesser variability than MAP.

Fig. 6 shows the closed-loop anesthesia control scenario for three randomly selected simulated patients from 30 simulated patients. These plots further elucidate the variations of the controlled variables BIS and MAP around the target MAP_{target} and BIS_{target} values with respect to RL-based control. Often postsurgical patients are kept in ICU under moderate sedation to facilitate treatment procedures. Patient 1 is assumed to be a postsurgical patient and hence a nonzero initial condition is assigned to indicate the presence of propofol in the patient's body. This is to model the residual quantity of anesthetic drugs in the patient's body that has been administered during surgery. The initial conditions of Patients 2 and 3 are set to zero. The RL-based controller is able to regulate the output BIS and MAP value close to the target values. The trained RL-based controller demonstrates acceptable performance with respect to the simultaneous control of BIS and MAP (Moore et al., 2014). Fig. 6 along with the performance evaluation metrics given in Table 3 demonstrates the significance of the β_w parameter in Eq. (16) for prioritizing the control of BIS relative to MAP.

Note that our simulations show similar performance compared to the clinical trial conducted by Moore et al. (2014) for the evaluation of RL-based closed-loop control of intraoperative hypnosis. With respect to this clinical trial, the authors report that the range of the percentage values of MDPE and MDAPE is −2.8 to 8.8 and 3.4–9.6, respectively. For the 15 patients considered for the experiment, the range of the value of RMSE is 3.3–6.5. These figures are comparable with our results given in Table 2 for

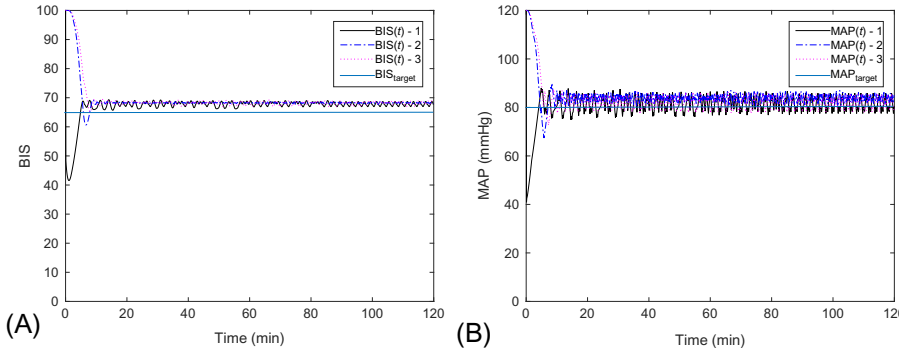

Fig. 6 Simulation results for three patients chosen randomly from the test set of 30 patients. (A) BIS index versus time for $\text{BIS}_{\text{target}} = 65$. (B) MAP versus time for $\text{MAP}_{\text{target}} = 80$ (Padmanabhan et al. (2015)).

the 30 simulated patients. Apart from the closed-loop regulation of the BIS, a methodology for the control of MAP is developed. In addition to that the RL-based controller does not rely on a system model and it demonstrates optimal and robust performance (see Fig. 6 and Table 3). This is an added advantage when it comes to the control of uncertain biological systems wherein developing accurate system models are very challenging.

In the context of simultaneous control of output variables such as BIS and MAP, one could debate that including an additional parameter (secondary variable(s)) can adversely affect the regulation of BIS (primary variable). However, instead of regulating sedation alone, simultaneous and balanced maintenance of the level of sedation along with other important requirements such as hemodynamic and respiratory system stability, pain management, muscle relaxation, etc. should be considered for improving patient safety. This is important as many of the sedative drugs (e.g., propofol) are known to induce significant changes in the heart rate, cardiac output, MAP, and respiratory rate of the patient. For our simulations, we consider MAP as the secondary control variable as propofol infusion reduces the sympathetic tone of the patient and induces venodilation. As a consequence, significant changes in the cardiac output and MAP is reported in the literature (Robinson et al., 1997).

Next, two case studies are presented to further elucidate the effect of simultaneous control of the BIS and MAP of a patient. First, a hemodynamic disturbance is simulated to account for the effect of hemorrhage on MAP by altering the MAP values by d units. For the second case study, irrespective of propofol infusion, the value of the secondary controlled variable (MAP) is held constant throughout the simulation period. This is to model the case of the intubated patients in the ICU who suffer from complications due to postaortic aneurysm repair. Another similar clinical situation in which the MAP becomes dangerously low is in the case of the septic patients.

First, to test the efficacy of the RL agent due to exogenous hemodynamic disturbance, a random patient from the population of 30 generated patients is simulated for: (i) $\text{MAP}(t)$, $t \geq 0$; (ii) $\text{MAP}(t) + d$, $t > 20$; and (iii) $\text{MAP}(t) - d$, $t > 40$. Here, the value of the

exogenous disturbance on MAP is set to $d = 10$ (see Fig. 7). Simulation results reflect the effect of prioritizing the control of BIS over MAP by using the parameter $\beta_w = 8$. For scenarios (i)–(iii), the average values of BIS and MAP are obtained as BIS = 68.18, MAP = 81.82, BIS = 66.84, MAP = 90.17, and BIS = 66.86, MAP = 69.85, for the interval $t \in [10, 120]$, $t \in [20, 120]$, and $t \in [40, 120]$, respectively. It should be noted that, because of the exogenous disturbance quantified by the parameter d, the value of $\text{MAP}_{\text{error}}$ is more and so the error signal is $e(t)$, $t \geq 0$. This explains the reason why the patient is sedated slightly more in the case of scenarios (ii) and (iii) compared to that of scenario (i). The control of BIS is affected by the increase in $e(t)$, $t \geq 0$, contributed mainly by the disturbance in the MAP value. However, the RL agent is able to keep the variation in BIS value within the acceptable range given by ± 5 units of the $\text{BIS}_{\text{target}}$ (Moore et al., 2014).

During the second case study, irrespective of the propofol infusion, the value of MAP(t), $t \geq 0$, is kept constant at the values 120, 100, 60, and 40 for all of the 30 simulated patients. The efficacy of the RL agent in regulating the value of BIS for these constant values of MAP is analyzed. Note that for all these simulation studies, the effect of propofol infusion on MAP is not considered. Instead, the value of the MAP is held constant. As shown in Table 4, for the cases with MAP(t), $t \geq 0$, kept at values 100 and 60, the RL agent is able to keep the variation in BIS value within the acceptable range given by ± 5 units of the $\text{BIS}_{\text{target}}$ (Moore et al., 2014). However, for the cases in which MAP(t), $t \geq 0$, is kept constant at values 120 and 40, the variations in the value of BIS are in the range of $\text{BIS}_{\text{target}} \pm 10$. For such extreme scenarios, it is recommended to use an RL agent which is trained by setting a large value for the parameter β_w to improve the regulation of BIS.

Efficacy of RL methods is demonstrated in several real-time applications, however, for clinical scenarios, decision making predicated on online identification requires a careful consideration. This work is a preliminary study toward the

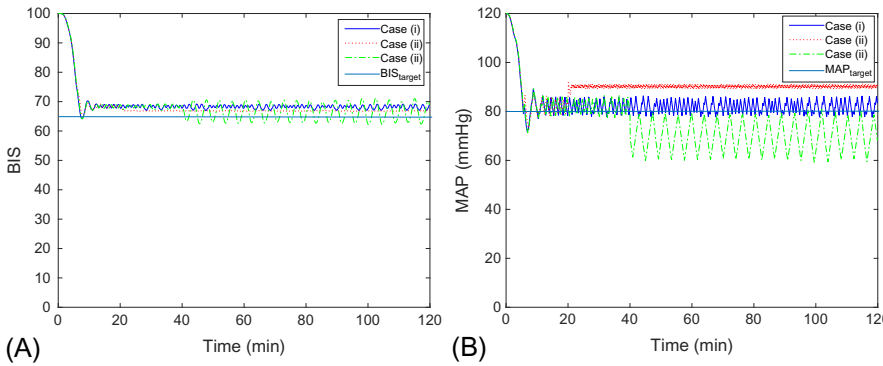

Fig. 7 Simulation results for a patient chosen randomly from the test set of 30 patients; Case (i) MAP(t), $t \geq 0$; Case (ii) MAP(t) + d, $t > 20$; and Case (iii) MAP(t) − d, $t > 40$, where $d = 10$ units represents a disturbance in the hemodynamic system of patient. (A) BIS index versus time for $\text{BIS}_{\text{target}} = 65$. (B) MAP versus time for $\text{MAP}_{\text{target}} = 80$ (Padmanabhan et al. (2015)).

Table 4 Performance metrics for the control variable BIS by keeping MAP constant; for 30 simulated patients (Padmanabhan et al., 2015)

Performance metrics	MAP(t) = 120	MAP(t) = 100	MAP(t) = 60	MAP(t) = 40
MDPE (%)	9.43 ± 0.63	-0.84 ± 0.45	-0.87 ± 0.49	9.44 ± 0.62
MDAPE (%)	9.43 ± 0.95	1.56 ± 0.43	1.58 ± 0.45	9.44 ± 0.56
Min−max	69.18−74.28	62.51−66.50	62.48−66.64	69.14−74.31
Interquartile range	2.6	2.15	2.16	2.67

implementation of RL-based closed-loop control of anesthesia. Some of the factors that contribute to the interindividual variations in the pharmacological parameters within a patient population are the patient physiological features, age, and concurrent illness. Drug habituation due to the frequent use of certain drugs also affect the response of a patient to the drug. Note that compared to the nominal model used for training, the variation in the pharmacologic parameters of the patient under treatment will be reflected in the response of the patient. Consequently, the error signal $e(t)$, $t \geq 0$, varies accordingly and thus the state s_k. This implies that as the RL agent executes control actions with respect to the state s_k, it can indirectly address pharmacological variations in patient to a certain extent. However, if the drug habituation or any other clinical situation results in significant and nonlinear changes in the patient pharmacology, then adaptive decision making is essential. As mentioned, the amount of propofol required to result in certain desired sedation level changes with the gender, age, weight, and height of the patient. These patient features are reflected in the pharmacological model parameters such as a_{ij}, a_{eff}, C_{50}, MC_{50}, V_c, γ, and α. The RL agent is trained by setting $IR_k = a_k \times IR_{\max}$, where $IR_{\max} = 20\,\text{mg}\,\text{min}^{-1}$. Table 2 shows the range of the patient pharmacological features used to obtain the 30 simulated patients.

Figs. 6 and 7 and Tables 3 and 4 show that the RL-based controller demonstrates acceptable performance for the 30 simulated patients with a wide range of pharmacological features. However, patient pharmacological features will considerably vary between different patient populations such as elderly, adults, children, and infants. Accordingly, the value of $IR_{\max}$ should be fixed and the RL agent needs to be trained with the new value of $IR_{\max}$. Similarly, to address the drug-dosing requirements of each patient population with vivid pharmacological features, it is recommended to use a bank of RL agents in which each agent is trained by using an appropriate $IR_{\max}$.

Finally, even though the RL-based controller demonstrates good performance, one of the limitations of this approach is the use of discrete state space and action space. Continuous-time state space and action space can enhance the robust adaptation of the RL-based controller and thereby derive more patient-specific and optimal control solution. However, this improved performance comes at the cost of increased

computational cost. The performance of the RL agent can be further enhanced by adjusting the value of the discount factor θ, learning rate $\eta(s_k, a_k)$, and by choosing a more appropriate reward function (Matignon et al., 2006).

3 Control of BIS by accounting for synergistic drug interaction

In this section, the use of an RL-based controller to fine tune the drug titration while different drugs with interactive effects are administered simultaneously is discussed. It is important to consider the interactive effects of the drugs to restrict the drug usage to the optimal level required to achieve certain therapeutic effects. In Section 2, a general framework is presented to formulate the problem of closed-loop control of intravenous drug administration using a finite MDP framework and the development of the Q-learning-based controller (Padmanabhan et al., 2015). In view of automated drug delivery for ICU sedation, another relevant factor that needs attention is the interactive effects of the drugs that are administered together.

In the following section, a Q-learning-based controller is developed to account for the synergistic effect during the combined administration of sedatives and analgesics. Then, the simulated patients who are used to train the RL agent and to conduct in silico trials are explained. The aim is to develop a controller to derive an optimal drug-dosing profile by accounting for the PK and PD disturbances in the human body under treatment.

3.1 Training the RL agent

The real-time system in this context is a dynamical system that represents the PK and PD of the multiple drugs that are administered together. The system description that follows is required to comprehend the input-output information needed to train the RL agent.

Consider the nonlinear dynamical system given by

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0, \quad t \geq 0, \tag{24}$$

$$y(t) = h(x(t)), \tag{25}$$

where for $t \geq 0$, $x(t) \in \mathbb{R}^{(n+p)}$ is the state vector, n and p are the number of states used to represent the PK of the sedative agent and analgesic agent, respectively, $u(t) \in \mathbb{R}^{(m+r)}$ is the control input, m and r are the number of sedative agents and analgesic agent infused, $y(t) \in \mathbb{R}^l$ is the output (controlled variable) of the system, $f: \mathbb{R}^{(n+p)} \times \mathbb{R}^{(m+r)} \rightarrow \mathbb{R}^{(n+p)}$ is locally Lipschitz continuous, and $h: \mathbb{R}^{(n+p)} \rightarrow \mathbb{R}^l$ is continuous. Here, the controlled variable of interest is the sedation level of the patient.

The aim is to develop an RL-based agent for the closed-loop control of the primary drug during their combined administration with any other drugs with a synergistic interactive effect.

Toward this end, the equivalent finite MDP representation of the system presented in Section 2.1, which involves a finite set of states $\mathcal{S}$ of the system, a finite set of action $\mathcal{A}$ that is available for each state $s_k \in \mathcal{S}$, a scalar reward $r_k \in \mathbb{R}$, and the transition probability matrix $\mathcal{P}$ that depends on the function $f(\cdot, \cdot)$ defined in Eq. (24) which is assumed to be unknown is used. With respect to the infusion of the sedative agent, the finite action set with p number of discrete actions defined as $(\mathcal{A}_j)_{j\in\mathbb{J}^+}$, $\mathbb{J}^+ \triangleq \{1,2,\ldots,p\}$ is considered. As explained in Section 2.2, a Q-function is progressively updated as per Eq. (6) using the available information with respect to system (24), which involve current state, action taken, new state reached, and reward received for the state transition.

3.2 Simulated patient

In this section, the patient models used for our simulations are presented. A superscript S or A denote that the parameter is associated with a sedative or an analgesic drug, respectively. First, consider the dynamical system

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad t \geq 0, \tag{26}$$

where $A \in \mathbb{R}^{(n+p)\times(n+p)}$ is a compartmental matrix, $B \in \mathbb{R}^{(n+p)\times(m+r)}$ is an input matrix, $x(t) \in \mathbb{R}^{(n+p)}$, $t \geq 0$, is the state vector, and $u(t) \in \mathbb{R}^{(m+r)}$, $t \geq 0$, is given by $u(t) = [(u^{\mathrm{S}}(t))^T, (u^{\mathrm{A}}(t))^T]^T$, where $u^{\mathrm{S}}(t) \in \mathbb{R}^m$, $t \geq 0$, and $u^{\mathrm{A}}(t) \in \mathbb{R}^r$, $t \geq 0$, represent the sedative and analgesic drug infusion, respectively. Next, rewrite Eq. (26) as

$$\dot{x}(t) = Ax(t) + B\bar{u}^{\mathrm{S}}(t) + d(t), \quad x(0) = x_0, \quad t \geq 0, \tag{27}$$

where $\bar{u}^{\mathrm{S}}(t) = [(u^{\mathrm{S}}(t))^T, 0]^T$, $d(t) = B\bar{u}^{\mathrm{A}}(t)$, and $\bar{u}^{\mathrm{A}}(t) = [0, (u_{\mathrm{A}}(t))^T]^T$. For each drug, a three-compartment model with an effect-site compartment is used to represent the drug disposition in the human body. While infusing several drugs simultaneously, the mass distribution of each drug in these three compartments and the effect site can be represented using the respective system states for each drug.

For the simultaneous infusion of a sedative and an analgesic drug, we consider the state vector $x(t) = [x_1(t), x_2(t), x_3(t), c^{\mathrm{S}}_{\mathrm{eff}}(t), x_5(t), x_6(t), x_7(t), c^{\mathrm{A}}_{\mathrm{eff}}(t)]^T$, where $x_i(t)$, $t \geq 0$, $i = 1, 2, 3$, and $x_i(t)$, $t \geq 0$, $i = 5, 6, 7$, denote the masses of the sedative and analgesic in the ith compartment, respectively, and $c^{\mathrm{S}}_{\mathrm{eff}}(t)$, $t \geq 0$, and $c^{\mathrm{A}}_{\mathrm{eff}}(t)$, $t \geq 0$, are the effect-site concentrations of the sedative and analgesic, respectively. In particular,

$$\dot{x}_1(t) = -(a^{\mathrm{S}}_{11} + a^{\mathrm{S}}_{21} + a^{\mathrm{S}}_{31})x_1(t) + a^{\mathrm{S}}_{12}x_2(t) + u^{\mathrm{S}}(t), \quad x_1(0) = x_{10}, \quad t \geq 0, \tag{28}$$

$$\dot{x}_2(t) = a^{\mathrm{S}}_{21}x_1(t) - a^{\mathrm{S}}_{12}x_2(t), \quad x_2(0) = x_{20}, \tag{29}$$

$$\dot{x}_3(t) = a^{\mathrm{S}}_{31}x_1(t) - a^{\mathrm{S}}_{13}x_3(t), \quad x_3(0) = x_{30}, \tag{30}$$

$$\dot{c}^{\mathrm{S}}_{\mathrm{eff}}(t) = a^{\mathrm{S}}_{\mathrm{eff}}(x_1(t)/V_{\mathrm{c}} - c^{\mathrm{S}}_{\mathrm{eff}}(t)), \quad c^{\mathrm{S}}_{\mathrm{eff}}(0) = c^{\mathrm{S}}_{\mathrm{eff0}}, \tag{31}$$

and

$$\dot{x}_5(t) = -(a^{\mathrm{A}}_{11} + a^{\mathrm{A}}_{21} + a^{\mathrm{A}}_{31})x_5(t) + a^{\mathrm{A}}_{12}x_6(t) + u^{\mathrm{A}}(t), \quad x_5(0) = x_{50}, \quad t \geq 0, \tag{32}$$

$$\dot{x}_6(t) = a^{\mathrm{A}}_{21}x_5(t) - a^{\mathrm{A}}_{12}x_6(t), \quad x_6(0) = x_{60}, \tag{33}$$

$$\dot{x}_7(t) = a^{\mathrm{A}}_{31}x_5(t) - a^{\mathrm{A}}_{13}x_7(t), \quad x_7(0) = x_{70}, \tag{34}$$

$$\dot{c}^{\mathrm{A}}_{\mathrm{eff}}(t) = a^{\mathrm{A}}_{\mathrm{eff}}(x_5(t)/V_{\mathrm{c}} - c^{\mathrm{A}}_{\mathrm{eff}}(t)), \quad c^{\mathrm{A}}_{\mathrm{eff}}(0) = c^{\mathrm{A}}_{\mathrm{eff0}}, \tag{35}$$

where a^{S}_{ij} and a^{A}_{ij} denote the rate of mass transfer between the jth and ith compartment for the sedative and analgesic drug, respectively, and V_{c} is the volume of the central compartment (blood).

When two drugs with interactive effects are administered simultaneously, their drug effect varies according to the ratio of the two drugs denoted as ϕ and their normalized drug concentration U. We use the common sedation assessment measure given by the BIS (Johansen et al., 2000) to assess the sedation level of the patient. The net sedative effect of an anesthetic drug when administered along with an analgesic drug which has synergistic interactive effect is given by

$$\mathrm{BIS}_{\mathrm{measured}}(t) = \mathrm{BIS}_0\left(1 - \frac{\left(\dfrac{U^{\mathrm{S}}(t) + U^{\mathrm{A}}(t)}{U_{50}(\phi)}\right)^{\gamma(\phi(t))}}{1 + \left(\dfrac{U^{\mathrm{S}}(t) + U^{\mathrm{A}}(t)}{U_{50}(\phi(t))}\right)^{\gamma(\phi)}}\right), \tag{36}$$

where $\phi(t) \triangleq \frac{U^{\mathrm{S}}(t)}{U^{\mathrm{S}}(t)+U^{\mathrm{A}}(t)}$, $\gamma(\phi(t))$, $t \geq 0$, is the steepness of the concentration-response relation at ratio $\phi(t)$, and $U_{50}(\phi(t))$ is the number of units associated with 50% of maximum effect at ratio $\phi(t)$ (Minto et al., 2000). Furthermore, $U^{\mathrm{S}}(t)$, $t \geq 0$, and $U^{\mathrm{A}}(t)$, $t \geq 0$, are the normalized drug concentrations of the sedative and analgesic drugs and are given by $U^{\mathrm{S}}(t) = \frac{c^{\mathrm{S}}_{\mathrm{eff}}(t)}{C^{\mathrm{S}}_{50}}$ and $U^{\mathrm{A}}(t) = \frac{c^{\mathrm{A}}_{\mathrm{eff}}(t)}{C^{\mathrm{A}}_{50}}$, where C^{S}_{50} and C^{A}_{50} are the drug concentrations of the sedative and analgesic that cause 50% drug effects, respectively. The BIS value corresponding to fully conscious patient is denoted by BIS_0. For training the RL agent, $e(t) = \mathrm{BIS}_{\mathrm{error}}(t)$ is assigned, where $\mathrm{BIS}_{\mathrm{error}}(t)$ is given by Eq. (17).

3.3 Results and discussion

In this section, the efficacy of the RL-based controller in deriving optimal infusion rates of an anesthetic drug so as to achieve certain desired sedation level by simultaneously accounting for the infusion of an synergistic analgesic is discussed. For our simulation, the most widely used sedative and analgesic drugs, propofol and remifentanil, respectively, are used. These drugs have synergistic interactive effects (Mehta et al., 2006).

For our simulations, 25 simulated patients using clinically relevant patient parameters are used. The pain experienced by a patient during the clinical procedures such as surgery, tracheal tube insertion, or physiotherapy treatment varies considerably. In the case of analgesic drugs, the drug concentration that causes half-maximal effect (pain relief) denoted by C_{50}^{A} varies with the intensity of the pain associated. For instance, the C_{50}^{S} and C_{50}^{A} of patients with or without liver disorders varies considerably (Mehta et al., 2006). Hence, to account for such variations in the pharmacological parameter C_{50}^{A} with respect to different pain stimulus, the values in the range 0.025 ± 0.007 mg L^{-1} are used (Mehta et al., 2006). Table 5 summarizes the range of PK and PD parameters of the drugs propofol and remifentanil that are used to generate 25 simulated patients. For training the RL agent using a simulated patient, the PK parameter values $C_{50}^{S} = 5.6$ μg L^{-1} for propofol, and $C_{50}^{A} = 30$ ng L^{-1} for remifentanil are used. The response of the patient given by Eq. (36) are calculated using the relation $U_{50}(\phi) = 1 - \theta_B\phi + \theta_B\phi^2$, where $\theta_B = 0.22$ and $\gamma(\phi) = 0.85$ (Padmanabhan et al., 2014).

At each time step k, the Q-learning algorithm (6) requires the values of s_k, a_k, s_{k+1}, and r_{k+1} to progressively derive the optimal action set. Toward this end, the states s_k are defined based on the error $e(kT)$. The values $s = 10$ when $\text{BIS}_{\text{error}} < 0$ and $s_k \in \{1, 2, \ldots, 9\}$ when $\text{BIS}_{\text{error}} > 0$ are used. The range of values of the error $e(kT)$ for each $s_k \in \{1, 2, \ldots, 9\}$ is ([0, 1], (1, 3], (3, 8], (8, 12], (12, 18], (18, 25], (25, 35], (35, 45], (45, 54]), respectively. We use a finite action set $\mathcal{A} = \{0, 0.02, 0.04, 0.1, 0.25, 0.5, 0.7, 0.8, 0.9, 1\}$ for the RL agent. At each time step k, the agent imparts an infusion rate $u(t) = IR_k$, $IR_k = a_k \times IR_{\max}$, where $IR_{\max}$ is the maximum allowable infusion rate for the sedative drug propofol. For our simulations, $IR_{\max} = 25\,\text{mg}\,\text{min}^{-1}$ and $\text{BIS}_{\text{target}} = 65$ are used. The action a_k at the kth time step is chosen from the finite action set $\mathcal{A}$.

Table 5 Range of values used to generate 25 simulated patients (Padmanabhan et al., April, 2017a)

Parameter	Propofol	Remifentanil
C_{50}	0.004 ± 0.001 g L^{-1}	0.025 ± 0.007 mg L^{-1}
V_c (L)	16 ± 1	16 ± 1
a_{ij} (min^{-1})	$\pm 0.5\%$	$\pm 0.5\%$
a_{eff} (min^{-1})	$\pm 0.5\%$	$\pm 0.5\%$

Another factor to consider is that the patient PK and PD vary significantly according to the health condition of the patient. The recommended propofol infusion rate for a patient treated for ailments in renal, hepatic, or cardiac function is 2.8 ± 1.1 mg kg^{-1} h^{-1} of propofol. Likewise, for a patient with respiratory ailments, the recommended drug dose titration rate of propofol is 1.25 ± 0.87 mg kg^{-1} h^{-1}. The recommended remifentanil infusion rate for the combined administration of propofol and remifentanil is $0.6-15$ μg kg^{-1} h^{-1} (Mehta et al., 2006). For an 80-kg patient, this range is equivalent to $0.008-0.02$ mg min^{-1}. Hence, the efficacy of the Q-learning-based controller with respect to two different drug infusion rates; 0.05 and 0.1 mg min^{-1} are tested. Figs. 8 and 9 show the controlled variable (BIS) for the two different infusion rates of remifentanil. For both cases, first the drug propofol alone during the time interval $t \in [0, 60)$ min is administered and then

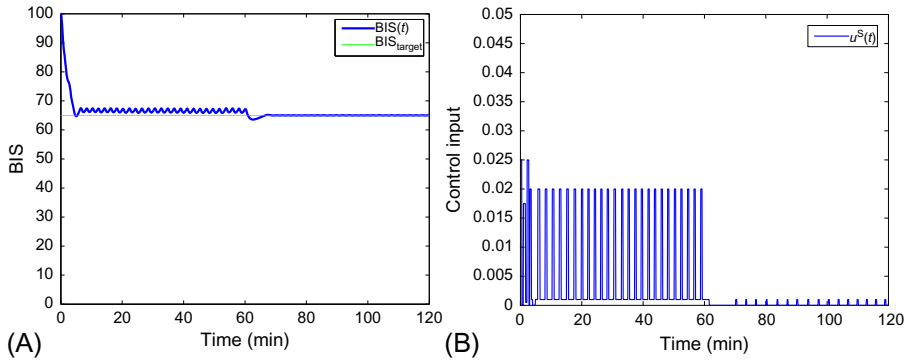

Fig. 8 Simulation results with remifentanil infusion rate of $u^{A}(t) = 0.1$ mg min^{-1} during $t =$ [60, 120]. (A) BIS index versus time for $\text{BIS}_{\text{target}} = 65$. (B) Control input $u^{S}(t)$ versus time (Padmanabhan et al., April, 2017a).

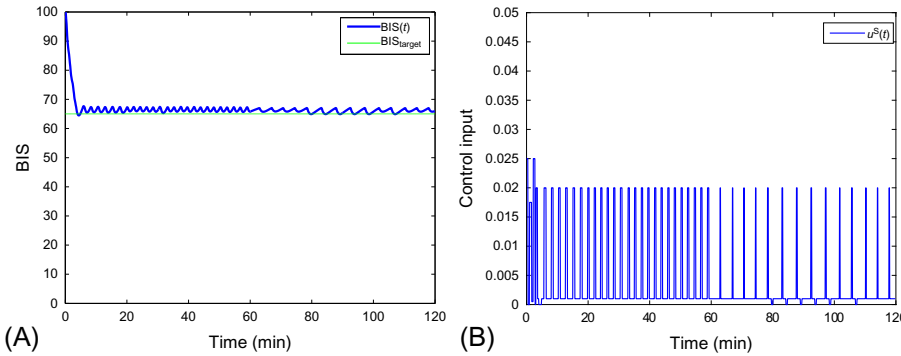

Fig. 9 Simulation results with remifentanil infusion rate of $u^{A}(t) = 0.05$ mg min^{-1} during $t =$ [60, 120]. (A) BIS index versus time for $\text{BIS}_{\text{target}} = 65$. (B) Control input $u^{S}(t)$ versus time (Padmanabhan et al., April, 2017a).

Table 6 Performance metrics for 25 patients for the controlled variable BIS for $u^{A}(t) = 0.1\text{mg}\,\text{min}^{-1}$ (Padmanabhan et al., April, 2017a)

Performance metrics	Drugs	
	Propofol	Propofol and remifentanil
MPE (%)	1.4657 ± 0.3110	0.499 ± 4.865
MDPE (%)	1.4349 ± 0.3127	− 1.678 to 0
MDAPE (%)	1.4349 ± 0.9479	0−1.678

Table 7 Performance metrics for 25 patients for the controlled variable BIS for $u^{A}(t) = 0.05\text{mg}\,\text{min}^{-1}$ (Padmanabhan et al., April, 2017a)

Performance metrics	Drugs	
	Propofol	Propofol and remifentanil
MPE (%)	1.2675 ± 0.3535	0.6111 ± 0.4006
MDPE (%)	1.4349 ± 0.3125	0.1182 ± 1.5518
MDAPE (%)	1.4349 ± 0.9479	0.2698 ± 1.4079

during $t \in [60, 120]$ remifentanil along with propofol is infused. Since the two drugs have synergistic interactive effects, the desired drug effect of $\text{BIS}_{\text{target}} = 65$ can be achieved using lower doses of propofol when administered along with remifentanil. It can be seen from Figs. 8 and 9 that the target anesthetic effect is achieved and maintained using a lower dose of propofol when both drugs are administered together.

As mentioned earlier, for our simulations two different drug infusion rates of the analgesic drug remifentanil are considered. Tables 6 and 7 show the statistical performance indices such as mean performance error (MPE), MDPE, and MDAPE used to evaluate the RL-based controller for the two different values of u^{A} used (Moore et al., 2014; Padmanabhan et al., 2015). It can be seen from Figs. 8 and 9, and Tables 6 and 7 that the performance of the RL-based controller is within the acceptable clinical ranges (Moore et al., 2014).

Simulations are conducted to generate the 25 simulated patients using the pharmacological parameters given in Table 5. Our simulation results (see Tables 6 and 7) show comparable performance with the recent in silico trial conducted on 24 virtually generated patients using model-based predictive control algorithm for automatic induction and regulation of depth of anesthesia (Nascu et al., 2011). In this in silico trial, the range of the percentage value of PE is -2.12 ± 15.13, value of MDPE is 0.8664, and the value of MDAPE is 1.114 for the 24 patients.

4 Control of cancer chemotherapy treatment

In this section, an RL-based controller design approach for the closed-loop control of cancer chemotherapy is developed. Specifically, a multiobjective RL-based controller to eradicate tumor cells by simultaneously accounting for the damage of normal cells and immune cells in a patient is discussed. The Q-learning-based approach presented in this section follows the general framework discussed in Section 2.1 to implement a similar controller for the control drug dosing pertaining to chemotherapy. The efficacy of the RL-based controller is evaluated using a nonlinear model of cancer chemotherapy controller. In the following section, a pharmacological model for cancer chemotherapy treatment is presented.

4.1 Mathematical model of cancer chemotherapy

Recently, there have been considerable efforts to develop various mathematical models to depict cancer dynamics. This is mainly to support research activities associated with the prediction of cancer incidence, drug development and its validation, and evaluation of novel drug-dosing approaches (De Pillis and Radunskaya, 2003; Sbeity and Younes, 2015). A mathematical model of cancer chemotherapy essentially accounts for the growth, death, mutation, PK, and PD in the tumor microenvironment. Typically, tumor microenvironment involves many types of cells, extracellular matrix, proteins, blood vessels, lymph vessels, etc. However, for the studies related to cancer chemotherapy, the three main cell types identified are tumor cells, immune cells, and host (normal) cells. All these cell types share common habitat (tumor microenvironment) and resources (nutrition and oxygen), resulting in nonlinear and interdependent cell dynamics. In general, a tumor without blood vessels and with blood vessels are referred as benign cancer and malignant cancer, respectively. Malignant cancers are capable of spreading (metastasize) from a tumor microenvironment to a healthy new site (ACS, 2015). Metastatic cancer is potentially lethal and hence it is often recommended to eradicate the tumor in the initial stage itself to avoid metastases.

Similar to the previous section concerning the RL-based control of anesthesia administration, in this section, a nonlinear model representing the cancer dynamics as given by Batmani and Khaloozadeh (2013) and De Pillis and Radunskaya (2003) is used to train the RL agent. In this model, the cell dynamics involved in the tumor microenvironment is explained by using the number of tumor cells, immune cells, normal cells, and drug concentration denoted by $T(t), t \geq 0, I(t), t \geq 0, N(t), t \geq 0$, $C(t)$, $t \geq 0$, respectively. The model is given by

$$\dot{x}_1(t) = r_2 x_1(t)[1 - b_2 x_1(t)] - c_4 x_1(t) x_2(t) - a_3 x_1(t) x_4(t), \quad x_1(0) = x_{10}, \quad t \geq 0, \tag{37}$$

$$\begin{aligned} \dot{x}_2(t) &= r_1 x_2(t)[1 - b_1 x_2(t)] - c_2 x_3(t) x_2(t) - c_3 x_2(t) x_1(t) - a_2 x_2(t) x_4(t), \\ x_2(0) &= x_{20}, \end{aligned} \tag{38}$$

$$\dot{x}_3(t) = s + \frac{\rho x_3(t)x_2(t)}{\alpha_c + x_2(t)} - c_1 x_3(t)x_2(t) - d_1 x_3(t) - a_1 x_3(t)x_4(t), \quad x_3(0) = x_{30}, \tag{39}$$

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

where $x_1(t) = N(t)$, $t \geq 0$, $x_2(t) = T(t)$, $t \geq 0$, $x_3(t) = I(t)$, $t \geq 0$, and $x_4(t) = C(t)$, $t \geq 0$, $u(t)$, $t \geq 0$, 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_1 is the death rate of immune cells, d_2 denotes the per capita decay rate of the injected drug, and a_1, a_2, and a_3 denote the fractional cell kill rates of the immune cells, tumor cells, and normal cells, respectively (Batmani and Khaloozadeh, 2013; De Pillis and Radunskaya, 2003; Pillis and Radunskaya, 2001).

The common response of the immune system of our body toward any identified harmful infection or disease is to increase the number of immune cells. This happens whenever the body's immunosurveillance identifies a tumor cell and is modeled in Eq. (39) using the term $\frac{\rho x_3(t)x_2(t)}{\alpha_c + x_2(t)}$, where ρ and α_c represent the immune response rate and immune threshold rate, respectively (De Pillis and Radunskaya, 2003; Pillis and Radunskaya, 2001). As mentioned earlier, since all the three main cell types share common habitat and resources the increase in the survival rate of one cell type adversely effects the existence of the other type of cell. These interaction between the cell types are modeled in Eqs. (37)–(39) using the terms $-c_1 x_3(t)x_2(t)$, $-c_2 x_3(t)x_2(t)$, $c_3 x_2(t)x_1(t)$, and $c_4 x_1(t)x_2(t)$, where c_i, $i = 1, 2, \ldots, 4$, represent the competition terms (De Pillis and Radunskaya, 2003). Similarly, the effect of the chemotherapeutic drug on all the three cell types is modeled in Eqs. (37)–(39) using the terms $a_i x_4(t)$, $i = 1, 2, \ldots, 3$.

Note that, apart from annihilating the tumor cells, the chemotherapeutic agent can also adversely affect the proliferation and survival of the normal cells and immune cells. Other typical side effects of chemotherapeutic drugs include hair loss, nausea, frequent infections due to the reduction in immune cell number, neuropathy, anemia, and organ damage (ACS, 2015). Hence, the aim is to derive optimal control input $u(t)$, $t \geq 0$, for the control of drug infusion during chemotherapy so that the desired drug effect is maximized and the drug-induced side effects are minimized.

4.2 RL-based optimal control for chemotherapic drug dosing

The model (37)−(40) is used along with the general framework discussed earlier in this chapter to develop an RL-based control approach for the closed-loop control of cancer chemotherapy. Similar to anesthesia administration discussed earlier, the problem of obtaining control solution for eradicating tumor cells using chemotherapeutic agents requires sequential decision making and can be solved using RL-based approaches. Here, the objective is to drive the system given by Eqs. (37)–(40) from a nonzero initial condition to the final goal state such as $x_2(t) = 0$ at some time t. This requires deriving the best sequence of actions in terms of the drug infusion rates

so as to steer the cancer patient from the state $x_2(t) \geq 0, t \geq 0$, to the desired state $x_2(t) = 0$. Toward this end, first rewrite Eqs. (37)–(40) in the state space form given by

$$\dot{x}(t) = f(x(t), u(t)), \quad x(0) = x_0, \quad t \geq 0, \tag{41}$$

$$y(t) = h(x(t)), \tag{42}$$

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

In this control problem, the number of tumor cells, given by $x_2(t), t \geq 0$, is the feedback parameter used to derive the optimal amount of drug to be titrated to the patient. In case of a peripheral tumor, the number of tumor cells can be measured employing a caliper. However, for externally inaccessible tumors, such as that in the brain or lungs, imaging techniques can be used to assess the volume of the tumor (Batmani and Khaloozadeh, 2013; Gholami et al., 2011; Huang et al., 2011; Suzuki et al., 2008).

Fig. 10 shows the schematic diagram of the training steps that is used to obtain the optimal Q function for cancer chemotherapy. Here, the objective is to derive the optimal control sequence that results in a minimum tumor volume, ideally $x_2(t) = 0$. As the training of the RL agent is predicated on the availability of the discrete states $s_k \in \mathcal{S}$, it is straightforward to define the state s_k of the patient in terms of measurable output $y(t)$, $t \geq 0$. Thus, $s_k = g(y(t))$, $kT \leq t < (k+1)T$, is assigned, where $g: \mathbb{R}^l \to \mathcal{S} \subset \mathbb{R}$ (Pachmann et al., 2001; Pillis and Radunskaya, 2001).

Sections 2.1 and 2.2 explain the use of RL-based approaches for the development of algorithms for optimal decisions making (Padmanabhan et al., 2017; Sutton and

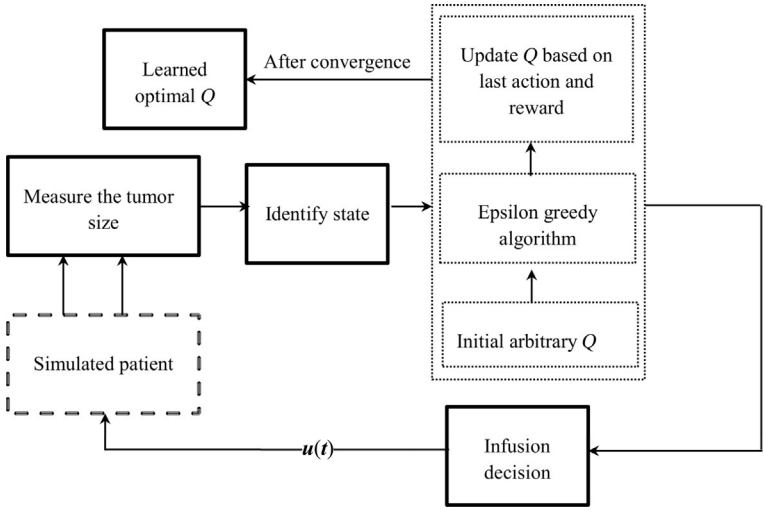

Fig. 10 Schematic diagram of training sequence to obtain optimal Q table (Padmanabhan et al., April, 2017b).

Barto, 1998; Vrabie et al., 2013). In this section, the RL framework discussed in Sections 2.1 and 2.2 is used to develop a closed-loop controller for regulating cancer chemotherapy treatment. As shown in Fig. 1, the main elements include an agent and a system. Note that here also, as the Q-learning algorithm does not use transition probability matrix denoted by $\mathcal{P}$ in order to derive the optimal control policy, it is assumed to be unknown. A controller or an agent is developed to maximize the reward it receives over an infinite horizon defined by Eq. (4). As explained in Section 2.2, a Q function is progressively updated as per Eq. (6) using the available information with respect to system (41) involving the current state, action taken, new state reached, and reward received for the state transition.

4.3 Results and discussion

This section details the numerical examples that demonstrate the performance of the Q-learning-based approach for the closed-loop control of drug dosing related to chemotherapy. To account for real-time situations, three different clinical scenarios were used to train the RL-based controller. Specifically, the case of an adult patient with cancer, a pregnant woman with cancer, and a critically ill elderly patient with cancer is considered. Here, different RL agents are developed to address the drug-dosing control in each of these cases. It is apparent that the ability of the human body to grow, repair, and defend disease is different for different age groups (Batmani and Khaloozadeh, 2013). The reason behind choosing these three case studies is to demonstrate the changes required in the RL algorithm to implement clinically relevant treatment strategies.

For instance, in case of a young cancer patient, the first preference of an oncologist will be to eradicate tumor cells to prevent metastasis. As young patients have a good growth ability, even if some of the normal cells are damaged as a side effect of chemotherapy that will be easily compensated by the body. However, this is not the case with an elderly patient who suffers from cancer as well as other diseases. For an elderly patient with cancer, the oncologist will try to eradicate cancer while preserving normal cells as well. Similarly, if the patient is suffering from brain cancer or cancer in any of vital organ, then also it is important to restrict damage of normal cells. These conditions are accounted for by selecting appropriate reward function (19). Moreover, for specific patient population like infants, children, and pregnant women, the oncologist needs to restrict the upper limits of the drug dose. This can be achieved by appropriately choosing the maximum value of the drug infusion rate $u_{\max}$ in $IR_k = a_k \times u_{\max}$ while training the RL agent.

The parameters listed in Table 8 are used in Eqs. (37)−(40) to generate simulated patients for the training and testing of the RL-based control algorithm. In the simulation, the maximum number of iteration is assigned as 50,000 scenarios. Here, a scenario is the series of state transitions from a random initial state to the desired final state s_k. The value of $\eta_k(s_k, a_k) = 0.2$ is assigned initially 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

Table 8 Parameter values used to generate simulated patient (Batmani and Khaloozadeh, 2013; De Pillis and Radunskaya, 2003) (Padmanabhan et al., April, 2017b)

Parameter	Parameter description	Value	Unit
a_1	Fractional immune cell kill rate	0.2	$\text{mg}^{-1}\text{L}\,\text{day}^{-1}$
a_2	Fractional tumor cell kill rate	0.3	$\text{mg}^{-1}\text{L}\,\text{day}^{-1}$
a_3	Fractional normal cell kill rate	0.1	$\text{mg}^{-1}\text{L}\,\text{day}^{-1}$
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 (between T and I cells)	1	$\text{cell}^{-1}\,\text{day}^{-1}$
c_2	Tumor cell competition term (between T and I cells)	0.5	$\text{cell}^{-1}\,\text{day}^{-1}$
c_3	Tumor cell competition term (between N and T cells)	1	$\text{cell}^{-1}\,\text{day}^{-1}$
c_4	Normal cell competition term (between N and T cells)	1	$\text{cell}^{-1}\,\text{day}^{-1}$
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	$\text{cell}\,\text{day}^{-1}$
α_c	Immune threshold rate	0.3	cell
ρ	Immune response rate	0.01	day^{-1}

chooses an action $a_k = \arg\max_{a\in A} Q(s_k, a)$. Next, the changes required in the development of the RL-based agent for three clinical situations are discussed.

Case 1. First, the case of a young patient with cancer is considered. In this case, since the patient has a 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 typically tries to eradicate the cancer cells $x_2(t)$, $t \geq 0$, completely. Thus, here the objective is to annihilate the tumor cells to attain the desired state $x_{2\text{d}} = 0$. Therefore, the error $e(t)$, $t \geq 0$, is defined as $e(t) = x_2(t) - x_{2\text{d}}$. The criteria used for the state assignment based on the error $e(t)$, $kT \leq t < (k+1)T$ is shown in Table 9. In this case, the reward r_{k+1} is calculated by setting $e(t) = x_2(t)$. For this case, an RL agent trained with $u_{\max} = 10\,\text{mg}\,\text{L}^{-1}\text{day}^{-1}$ is used.

Fig. 11 shows the response of the patient when a chemotherapeutic drug is administered using an 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. The number of normal cells and tumor cells given in Fig. 11 are normalized values. Note that with treatment, the number of tumor cells have reduced and the normal cells have increased. However, see that initially the

Table 9 State assignment for Cases 1–3 based on $e(t)$ (Padmanabhan et al., April, 2017b)

Cases 1 and 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, ∞]

Notes: 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.

number of immune cells decreases as an adverse effect of chemotherapy, and later their number improves. The amount of drug administered for Case 1 is shown in Fig. 12.

Case 2. For this case, a young pregnant woman with cancer is considered. Here, the aim is to keep the amount of the chemotherapeutic drug used to a minimum level and thus not harm the fetus. After child birth, the use of the chemotherapeutic drug can be increased to the required level to eradicate tumor. In similar situations, the oncologist often resorts to a two-stage chemotherapy. Here, for our simulations, it is assumed that the patient 7 months pregnant. In the first stage, the maximum amount of the drug infused is restricted by setting $u_{\max} = 0.5\text{mgL}^{-1}\text{day}^{-1}$. However, after child birth, the maximum amount of the drug infused is increased to $u_{\max} = 10\text{mgL}^{-1}\text{day}^{-1}$ (Batmani and Khaloozadeh, 2013).

For training the RL agent to derive drug infusion rates during the first stage, the value of $u_{\max}$ was set to $0.5\ \text{mgL}^{-1}\text{day}^{-1}$. Similarly, for the second state, the value of $u_{\max}$ was set to $10\ \text{mgL}^{-1}\text{day}^{-1}$. Figs. 13 and 14 show the simulation results for the two-stage chemotherapy for the young pregnant woman using RL-based controllers. Note that during the initial 90 days, the drug concentration in the plasma is

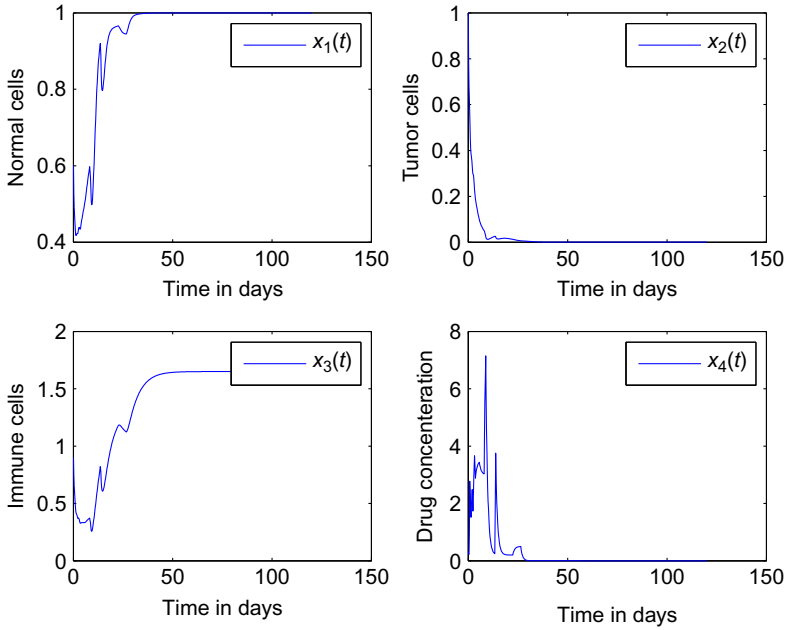

Fig. 11 Response of young patient with cancer (Case 1), $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

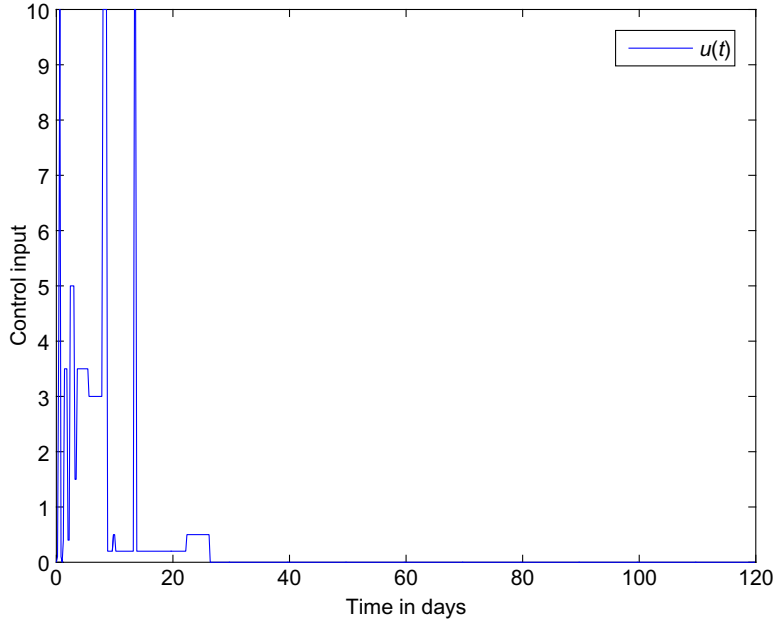

Fig. 12 Amount of drug administered (Case 1), $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

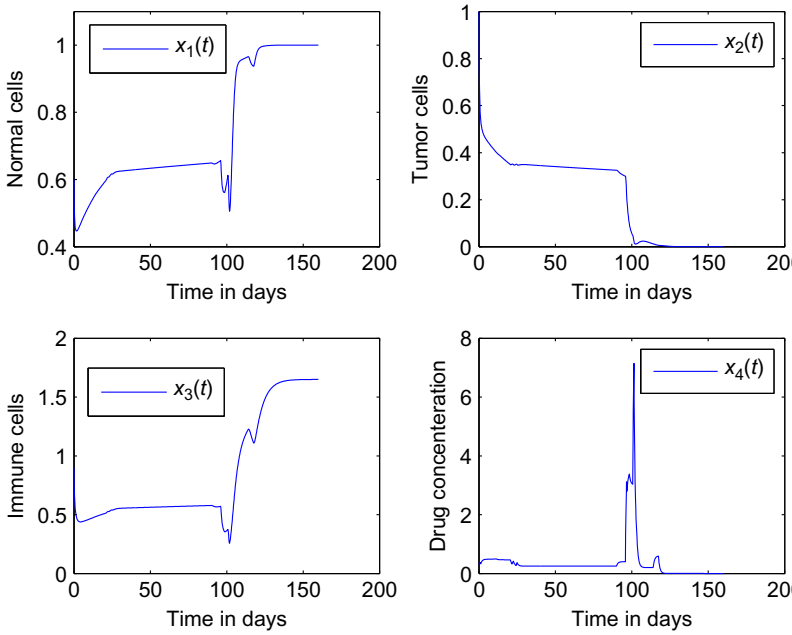

Fig. 13 Response of young pregnant woman with cancer (Case 2), $u_{\max} = 0.5\text{mg L}^{-1}\text{day}^{-1}$ until delivery (90 days) and then $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

restricted to 0.5 mgL^{-1}, however, after child birth the amount of drug used to eradicate the tumor completely has increased.

Case 3. For this case, an elderly patient with cancer and other illnesses is considered. This scenario represents the clinical situations wherein it is essential to minimize the damage to the normal cells while annihilating maximum number of tumor cells. In order to account for this requirement while training the RL agent, the parameter β_w which denotes a weighing factor is used to prioritize between the normal cells and cancer cells. The objective is to attain $x_{1d} = 1$ and $x_{2d} = 0$, where x_{1d} and x_{2d} represent the target values of $x_1(t)$, $t \geq 0$, and $x_2(t)$, $t \geq 0$, respectively. Here, an RL agent is trained using a reward function defined based on the deviation of the number of normal cells and tumor cells from the respective desired values. Specifically, the state s_k, $kT \leq t < (k+1)T$, is defined in terms of the error

$$e(t) = \beta_w x_2(t) + (1 - \beta_w)[1 - x_1(t)]. \tag{43}$$

The reward is calculated using the value of error $e(t)$, $kT \leq t < (k+1)T$ in Eq. (19).

For our simulation, the parameter values used are $\theta = 0.7$, $\eta = 0.2$, and $\beta_w = 0.9$. Simulation results showing the response of the closed-loop control of the chemotherapeutic drug for Case 3 is shown in Figs. 15 and 16. Here the RL agent is trained with respect to the error defined by Eq. (43). Note that, compared to Case 1 that is shown in

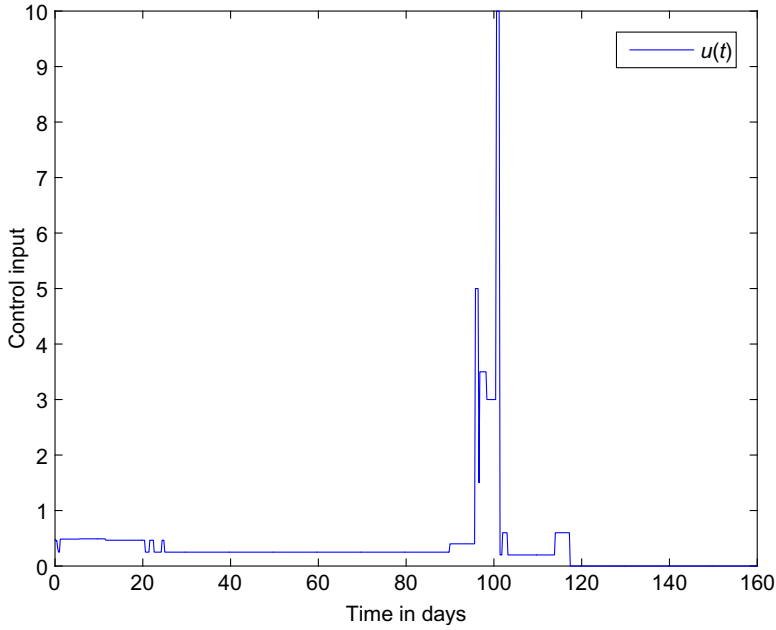

Fig. 14 Amount of drug administered (Case 2), $u_{\max} = 0.5\text{mg L}^{-1}\text{day}^{-1}$ until delivery (90 days) and then $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

Figs. 11 and 12, the amount of drug used in this case is lesser (see Figs. 15 and 16). This is to minimize the damage to the healthy normal cells.

Table 9 shows the criteria used for the state assignment for Cases 1–3. For Cases 1 and 3, and the second stage of Case 2, the finite action set $\mathcal{A} = \{0, 0.01, 0.02, 0.03, 0.04, 0.06, 0.08, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ is used. However, for the first stage of Case 2, with $u_{\max} = 0.5\text{mgL}^{-1}\text{day}^{-1}$, the finite action set $\mathcal{A} = \{0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.78, 0.80, 0.82, 0.85, 0.87, 0.9, 0.91, 0.92, 0.93, 0.94, 0.95, 0.97, 0.98, 1\}$ is used. Moreover, for Cases 1 and 3, and the second stage of Case 2 during the training of RL agent, the goal state as $s = 1$ is used to eradicate the tumor completely. However, for the first stage of Case 2, during the training of RL agent, the goal state is set as $s = 7$, which represents a limited tumor size.

In order to demonstrate the robustness of the controller, the trained optimal RL-based controller is used 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 8. In Cases (ii) and (iii), simulated patients with -10% and $+15\%$ parameter variations with respect to the values given in Table 8 are used. Figs. 17 and 18 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

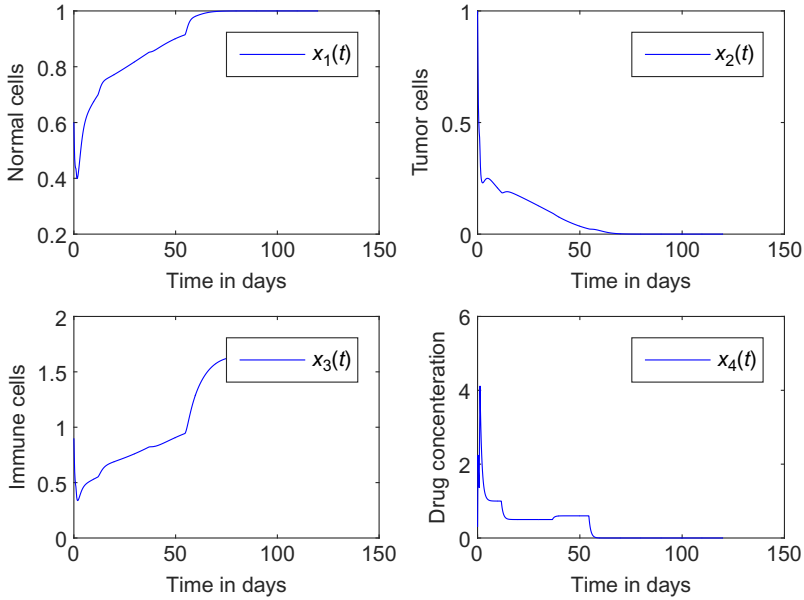

Fig. 15 Response of an elderly patient who has cancer along with other critical illnesses (Case 2), $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

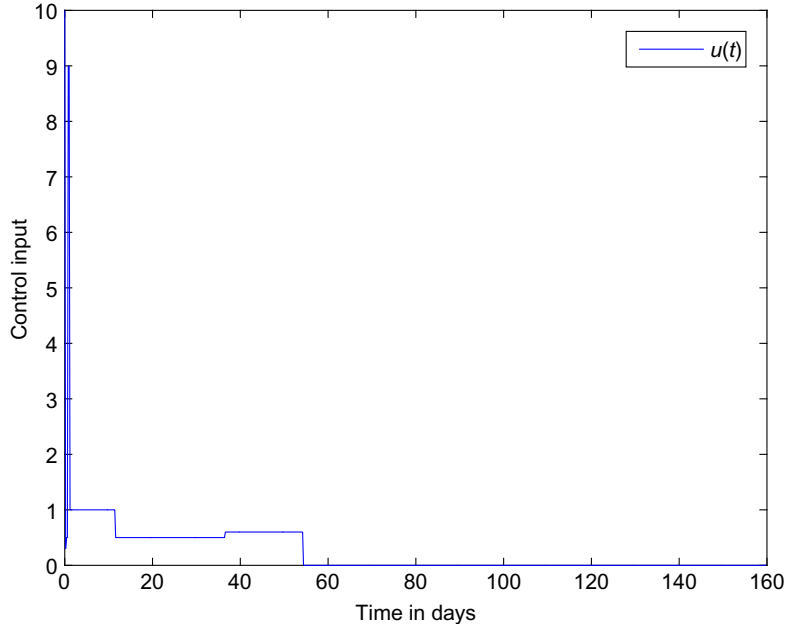

Fig. 16 Amount of drug administered (Case 2), $u_{\max} = 10\text{mg L}^{-1}\text{day}^{-1}$ (Padmanabhan et al., 2017b).

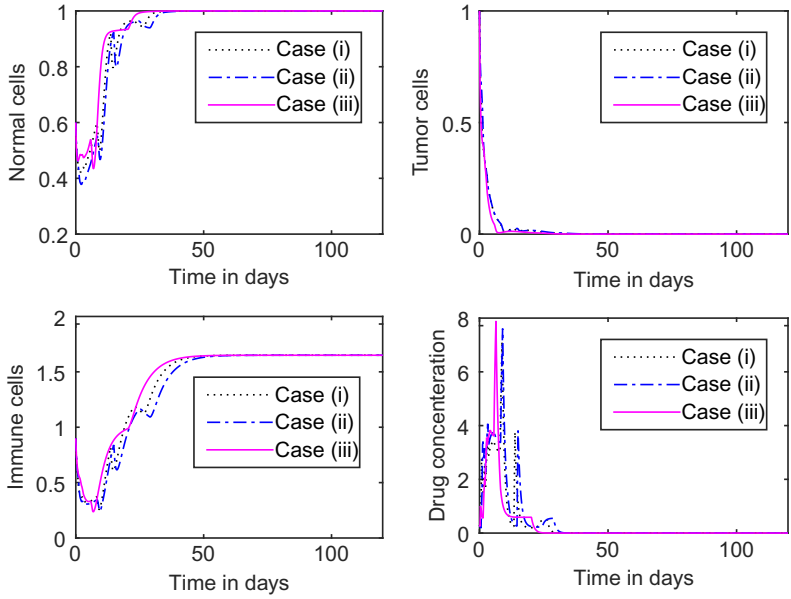

Fig. 17 Response for three different patient models: Case (i) with nominal model, Case (ii) with −10% parameter variation, and Case (iii) with +15% parameter variation (Padmanabhan et al., 2017b).

Table 10 Statistical analysis for 15 simulated patients (Padmanabhan et al., April, 2017b)

Parameter		N_{dev}	T_{per}
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

based on the error $e(t)$, $t \geq 0$, which reflects the patient-specific response to drug intake. Thus, the value of the error $e(t)$, $t \geq 0$, varies according to the patient characteristics.

Table 10 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

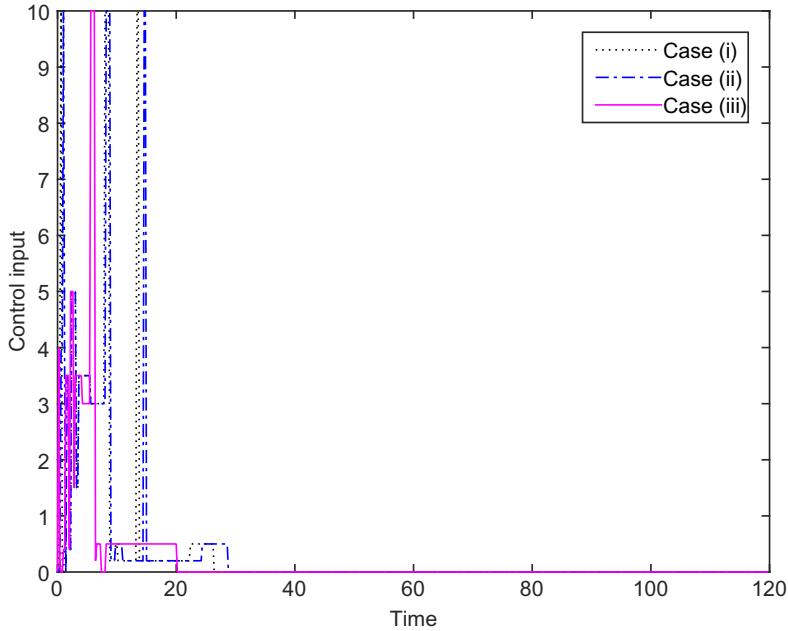

Fig. 18 Control input for three different patient models: Case (i) with nominal model, Case (ii) with −10% parameter variation, and Case (iii) with +15% parameter variation (Padmanabhan et al., April, 2017b).

with the parameter ranges of fraction cell kill a_i, $i = 1, 2, 3, 0 < a_i \leq 0.5$, $a_3 \leq a_1 \leq a_2$, carrying capacities $b_1^{-1} \leq b_2^{-1} = 1$, competition terms $0.3 \leq c_i \leq 1$, $i = 1, \ldots, 4$, death rates $0.15 \leq d_1 \leq 0.3$, $d_2 = 1$, per unit growth rates, $1.2 \leq r_1 \leq 1.6$, $r_2 = 1$, immune source rate $0.3 \leq s \leq 0.5$, immune threshold rate $0.3 \leq \alpha_c \leq 0.5$, and immune response rate $0.01 \leq \alpha_c \leq 0.05$. See De Pillis and Radunskaya (2003) for further details on the parameter ranges of the cancer chemotherapy model.

The percent deviation of the number of normal cells from the target value ($x_{1d} = 1$) given in Table 10 is calculated as

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

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 10 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. The minimum, maximum, and mean number of days for achieving the target values of $x_1(t)$, $t \geq 0$, and $x_2(t)$, $t \geq 0$, are 13, 50, 28, and 6, 52, 27 days, respectively, for the 15 simulated patients. Comparing our simulation results with those by Batmani and Khaloozadeh (2013), 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 RL-based method is that it does not require a model of the system in order to develop a controller.

5 Summary

First, in Section 2, an RL-based controller design approach for the simultaneous regulation of sedation and MAP using the controlled titration of the sedative drug propofol is detailed. Simulation studies conducted using 30 simulated patients with varying pharmacological parameters show that the RL-based controller design approach is promising in developing closed-loop controllers for ICU sedation while regulating multiple vital physiological parameters simultaneously.

Next, in Section 3, an RL-based controller that can account for the simultaneous administration of drugs with synergistic interactive effect is presented. The RL agent is trained using Q-learning algorithm with respect to the states defined in terms of the error associated with the desired output. A similar method can be used for the case of the drugs with inhibitive drug interactive effect. Moreover, our simulations demonstrate that the RL-based methods can be used to implement closed-loop control systems which are robust to system uncertainties. Further experiments are warranted to refine the optimal control agent and to extend it by including the simultaneous control of additional vital physiological parameters such as heart rate, cardiac output, and respiratory rate.

Finally, in Section 4, the efficacy of the RL-based method is investigated for different cases of cancer treatment. The method results in an optimal and robust controller. In order to preserve normal cells while eradicating tumor cells, a scaled value of the error is used in the reward function. The 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 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.

It is apparent that the credibility of feedback information that is used to guide the controller is one of the main factors that determines the closed-loop performance of the controller. In case of biomedical applications, most of the monitoring systems has to account for various errors in the measured signal which may arise due to the movement of cables, electrodes, and patient, and also the noise and interferences from other devices. Hence, modern biomedical monitors include intricate filtering algorithms to improve signal-to-noise ratio of the measurement. This in turn increases the

computation time and hence introduces time delay in measurement. Hence, studying the scope for improvement in performance of the proposed Q-learning-based controller design methodology by accounting for possible time delays during intravenous drug administration is desirable.

Acknowledgments

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

Abbeel, P., Coates, A., Quigley, M., Ng, A.Y., 2007. An application of reinforcement learning to aerobatic helicopter flight. In: Neural Information Processing Systems, vol. 19, pp. 1–8.

Absalom, A.R., Kenny, G.N.C., 2003. Closed-loop control of propofol anaesthesia using bispectral index: performance assessment in patients receiving computer-controlled propofol and manually controlled remifentanil infusions for minor surgery. Br. J. Anesth. 90 (6), 737–741.

Absalom, A.R., Mason, K.P., 2017. Total Intravenous Anesthesia and Target Controlled Infusions: A Comprehensive Global Anthology. Springer, Switzerland.

Absalom, A.R., Mani, V., De Smet, T., Struys, M.M., 2009. Pharmacokinetic models for propofol defining and illuminating the devil in the detail. Br. J. Anaesth. 103 (1), 26–37.

Absalom, A.R., De Keyser, R., Struys, M.M.R.F., 2011. Closed-loop anesthesia: are we getting close to finding the holy grail? Anesth. Analg. 112 (3), 516–518.

ACS, 2015. Cancer facts and figures 2015. American Cancer Society, Atlanta, Georgia. http://www.cancer.org/acs/groups/content/@editorial/documents/document/acspc-044552.pdf.

Babaei, N., Salamci, M.U., 2015. Personalized drug administration for cancer treatment using model reference adaptive control. J. Theor. Biol. 371, 24–44.

Bailey, J.M., Haddad, W.M., 2005. Drug dosing control in clinical pharmacology. IEEE Control Syst. Mag. 23 (2), 35–51.

Balashevich, N.V., Gabasov, R., Kalinin, A.I., Kirillova, F.M., 2002. Optimal control of nonlinear systems. Comput. Math. Math. Phys. 42 (7), 931–956.

Barto, A.G., Sutton, R.S., Anderson, C.W., 1983. Neuron like adaptive elements that can solve difficult learning control problems. IEEE Trans. Syst. Man Cybernet. 13, 834–846.

Batmani, Y., Khaloozadeh, H., 2013. Optimal chemotherapy in cancer treatment: state dependent Riccati equation control and extended Kalman filter. Optimal Control Appl. Methods 34 (5), 562–577.

Bertsekas, D.P., Tsitsiklis, J.N., 1996. Neuro-Dynamic Programming. Athena Scientific, Belmont, MA.

Brogi, E., Cyr, S., Kazan, R., Giunta, F., Hemmerling, T.M., 2017. Clinical performance and safety of closed-loop systems: a systematic review and meta-analysis of randomized controlled trials. Anesth. Analg. 124 (2), 446–455.

Chen, T., Kirkby, N.F., Jena, R., 2012. Optimal dosing of cancer chemotherapy using model predictive control and moving horizon state/parameter estimation. Comput. Methods Programs Biomed. 108 (3), 973–983.

Chen, C.S., Doloff, J.C., Waxman, D.J., 2014. Intermittent metronomic drug schedule is essential for activating antitumor innate immunity and tumor xenograft regression. Neoplasia 16 (1), 84–96.

Çimen, T., 2010. Systematic and effective design of nonlinear feedback controllers via the state-dependent Riccati equation (SDRE) method. Annu. Rev. Control 34 (1), 32–51.

Dadhich, S., Bodin, U., Sandin, F., Andersson, U., 2016. Machine learning approach to automatic bucket loading. In: 24th Mediterranean Conference on Control and Automation, pp. 1260–1265.

Daskalaki, E., Diem, P., Mougiakakou, S.G., 2013. Personalized tuning of a reinforcement learning control algorithm for glucose regulation. In: 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 3487–3490.

De Pillis, L.G., Radunskaya, A., 2003. The dynamics of an optimally controlled tumor model: a case study. Math. Comput. Model. 37 (11), 1221–1244.

Doloff, J.C., Waxman, D.J., 2015. Transcriptional profiling provides insights into metronomic cyclophosphamide-activated, innate immune-dependent regression of brain tumor xenografts. BMC Cancer 15 (1), 375.

Engelhart, M., Lebiedz, D., Sager, S., 2011. 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), 123–134.

Fan, S.Z., Wei, Q., Shi, P.F., Chen, Y.J., Liu, Q., Shieh, J.S., 2012. A comparison of patient's heart rate variability and blood flow variability during surgery based on the Hilbert Huang transform. Biomed. Signal Process. Control 7 (5), 465–473.

Furutani, E., Tsuruoka, K., Kusudo, S., 2010. A hypnosis and analgesia control system using a model predictive controller in total intravenous anesthesia during day-case surgery. In: SICE Annual conference, Taipei, Taiwan, pp. 223–226.

Gholami, B., Agar, N.Y.R., Jolesz, F.A., Haddad, W.M., Tannenbaum, A.R., 2011. A compressive sensing approach for glioma margin delineation using mass spectrometry. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE, pp. 5682–5685.

Haddad, W.M., Chellaboina, V., 2008. Nonlinear Dynamical Systems and Control: A Lyapunov-Based Approach. Princeton University Press, Princeton, NJ.

Haddad, W.M., Hayakawa, T., Bailey, J.M., 2003. Adaptive control for nonnegative and compartmental dynamical systems with applications to general anesthesia. Int. J. Adapt Control Signal Process. 17, 209–235.

Haddad, W.M., Chellaboina, V., Hui, Q., 2010. Nonnegative and Compartmental Dynamical Systems. Princeton University Press, Princeton, NJ.

Haddad, W.M., Bailey, J.M., Gholami, B., Tannenbaum, A.R., 2013. Clinical decision support and closed-loop control for intensive care unit sedation. Asian J. Control 15 (2), 317–339.

Hahn, J.O., Dumont, G.A., Ansermino, J.M., 2012. Robust closed-loop control of hypnosis with propofol using WAVCNS index as the controlled variable. Biomed. Signal Process. Control 7 (5), 517–524.

Heusden, K.V., Ansermino, J.M., Dumont, G.A., 2018. Robust MISO control of propofol-remifentanil anesthesia guided by the NeuroSENSE monitor. IEEE Trans. Control Syst. Technol. 26 (5), 1758–1770.

Hong, M., Razaviyayn, M., Luo, Z.Q., Pang, J.S., 2016. A unified algorithmic framework for block-structured optimization involving big data: with applications in machine learning and signal processing. IEEE Signal Process. Mag. 33 (1), 57–77.

Huang, J., Gholami, B., Agar, N.Y.R., Norton, I., Haddad, W.M., Tannenbaum, A.R., 2011. Classification of astrocytomas and oligodendrogliomas from mass spectrometry data using

sparse kernel machines. In: Annual International Conference of the IEEE Engineering in Medicine and Biology Society, IEEE, pp. 7965–7968.

Ionescu, C.M., Nascu, I., De Keyser, R., 2014. Lessons learned from closed loops in engineering: towards a multivariable approach regulating depth of anaesthesia. Int. J. Clin. Monit. Comput. 28 (6), 537–546.

Jacobi, J., Fraser, G.L., Coursin, D.B., Riker, R.R., Fontaine, D., Wittbrodt, E.T., Chalfin, D.B., Masica, M.F., Bjerke, H.S., Coplin, W.M., Crippen, D.W., Fuchs, B.D., Kelleher, R.M., Marik, P.E., Nasraway, S.A., Murray, M.J., Peruzzi, W.T., Lumb, P.D., 2002. Clinical practice guidelines for the sustained use of sedatives and analgesics in the critically ill adult. Am. J. Health Syst. Pharm. 59, 150–178.

Johansen, J.W., Sebel, P.S., Smet, T.D., Struys, M.M., 2000. Development and clinical application of electroencephalographic bispectrum monitoring. Anesthesiology 93, 1336–1344.

Kiran, K.L., Jayachandran, D., Lakshminarayanan, S., 2009. Multi-objective optimization of cancer immuno-chemotherapy. In: 13th International Conference on Biomedical Engineering, pp. 1337–1340.

Kuizenga, M.H., Vereecke, H.E., Struys, M.M., 2016. Model-based drug administration: current status of target-controlled infusion and closed-loop control. Curr. Opin. Anesthesiol. 29 (4), 475–481.

Liu, N., Chazot, T., Genty, A., Landais, A., Restoux, A., McGee, K., Laloë, P.A., Trillat, B., Barvais, L., Fischler, M., 2006. Titration of propofol for anesthetic induction and maintenance guided by the bispectral index: closed-loop versus manual control: a prospective, randomized, multicenter study. J. Am. Soc. Anesthesiol. 104 (4), 686–695.

Martin-Guerrero, J.D., Gomez, F., Soria-Olivas, E., Schmidhuber, J., Climente-Marti, M., Jemenez-Torres, N.V., 2009. A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients. Expert Syst. Appl. 36, 9737–9742.

Masui, K., Upton, R.N., Doufas, A.G., Coetzee, J.F., Kazama, T., Mortier, E.P., Struys, M.M., 2010. The performance of compartmental and physiologically based recirculatory pharmacokinetic models for propofol: a comparison using bolus, continuous, and target-controlled infusion data. Anesth. Analg. 111, 368–379.

Matignon, L., Laurent, G.J., Fort-Piat, N.L., 2006. Reward function and initial values: better choices for accelerated goal-directed reinforcement learning. In: 16th International Conference on Artificial Neural Networks, Athens, Greece, pp. 840–849.

Mehta, S., Burry, L., Fischer, S., Motta, J.C.M., Hallet, D., Bowman, D., Wong, C., Meade, M.O., Stewart, T.E., Cook, D.J., 2006. Canadian survey of the use of sedatives, analgesics, and neuromuscular blocking agents in critically ill patients. Crit. Care Med. 34 (2), 374–380.

Minto, C., Schnider, T., Short, T., Gregg, K., Gentilini, A., Shafer, S., 2000. Response surface model for anesthetic drug interactions. Anesthesiology 92, 1603–1616.

Moore, B.L., Panousis, P., Kulkarni, V., Pyeatt, L.D., Doufas, A.G., 2010. Reinforcement learning for closed-loop propofol anesthesia. In: Proceedings of 22th Annual Conference on Innovative Applications of Artificial Intelligence, Atlanta, Georgia, USA, pp. 1807–1813.

Moore, B.L., Pyeatt, L.D., Kulkarni, V., Panousis, P., Kevin, Doufas, A.G., 2014. Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers. J. Mach. Learn. Res. 15, 655–696.

Morley, A., Derrick, J., Mainland, P., Lee, B.B., Short, T.G., 2000. Closed loop control of anaesthesia: an assessment of the bispectral index as the target of control. Anaesthesia 55 (10), 953–959.

Nascu, I., Ionescu, C.M., Nascu, I., De Keyser, R., 2011. Evaluation of three protocols for automatic DOA regulation using propofol and remifentanil. In: 9th IEEE International Conference on Control and Automation, pp. 573–578.

Nemati, S., Ghassemi, M.M., Clifford, G.D., 2016. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. In: 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, pp. 2978–2981.

Noble, S.L., Sherer, E., Hannemann, R.E., Ramkrishna, D., Vik, T., Rundell, A.E., 2010. Using adaptive model predictive control to customize maintenance therapy chemotherapeutic dosing for childhood acute lymphoblastic leukemia. J. Theor. Biol. 264 (3), 990–1002.

Pachmann, K., Heiß, P., Demel, U., Tilz, G., 2001. Detection and quantification of small numbers of circulating tumour cells in peripheral blood using laser scanning cytometer (LSC®). Clin. Chem. Lab. Med. 39 (9), 811–817.

Padmanabhan, R., Meskin, N., Haddad, W.M., 2014. Direct adaptive disturbance rejection control for sedation and analgesia. In: Middle East Conference on Biomedical Engineering, Doha, Qatar, pp. 175–179.

Padmanabhan, R., Meskin, N., Haddad, W.M., 2015. Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning. Biomed. Signal Process. Control 22, 54–64.

Padmanabhan, R., Meskin, N., Haddad, W.M., 2017. Reinforcement learning-based control for combined infusion of sedatives and analgesics. In: 4th International Conference on Control, Decision and Information Technologies, Barcelona, Spain, pp. 505–509.

Padmanabhan, R., Meskin, N., Haddad, W.M., 2017b. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment, Mathematical Biosciences 293, 11–20.

Pillis, L.G.D., Radunskaya, A., 2001. A mathematical tumor model with immune resistance and drug therapy: an optimal control approach. Comput. Math. Methods Med. 3 (2), 79–100.

Rao, R.R., Bequette, B.W., 2000. Simultaneous regulation of hemodynamic and anesthetic states: a simulation study. Ann. Biomed. Eng. 28 (1), 71–84.

Robinson, B.J., Ebert, T.J., Brien, T.J.O., Colinco, M.D., Muzi, M., 1997. Mechanisms whereby propofol mediates peripheral vasodilation in humans: sympathoinhibition or direct vascular relaxation? Anesthesiology 86, 64–72.

Sbeity, H., Younes, R., 2015. Review of optimization methods for cancer chemotherapy treatment planning. J. Comput. Sci. Syst. Biol. 8, 74–95.

Sedighizadeh, M., Rezazadeh, A., 2008. Adaptive PID controller based on reinforcement learning for wind turbine control. World Acad. Sci. Eng. Technol. 2, 1–23.

Soltesz, K., Hahn, J.O., Hagglund, T., Dumont, G.A., Ansermino, J.M., 2013. Individualized closed-loop control of propofol anesthesia: a preliminary study. Biomed. Signal Process. Control 8 (6), 500–508.

Struys, M.M., De Smet, T., Versichelen, L.F., Van de Velde, S., Van den Broecke, R., Mortier, E.P., 2001. Comparison of closed-loop controlled administration of propofol using bispectral index as the controlled variable versus "standard practice" controlled administration. Anesthesiology 95 (1), 6–17.

Sutton, R.S., 1988. Learning to predict by the methods of temporal difference. Mach. Learn. 3, 9–44.

Sutton, R.S., Barto, A.G., 1998. Reinforcement Learning: An Introduction. MIT Press, Cambridge, MA.

Suzuki, C., Jacobsson, H., Hatschek, T., Torkzad, M.R., Boden, K., Eriksson-Alm, Y., Berg, E., Fujii, H., Kubo, A., Blomqvist, L., 2008. Radiologic measurements of tumor response to treatment: practical approaches and limitations. Radiographics 28 (2), 329–344. https://doi.org/10.1148/rg.282075068.

Swan, G.W., 1990. Role of optimal control theory in cancer chemotherapy. Math. Biosci. 101 (2), 237–284.

Swierniak, A., Lezewicz, U., Schattler, H., 2003. Optimal control for a class of compartmental models in cancer chemotherapy. Int. J. Appl. Math. Comput. Sci. 13 (3), 357–368.

Tan, K.C., Khor, E.F., Cai, J., Heng, C.M., Lee, T.H., 2002. Automating the drug scheduling of cancer chemotherapy via evolutionary computation. Artif. Intell. Med. 25 (2), 169–185.

Tse, S.-M., Liang, Y., Leung, K.-S., Lee, K.-H., Mok, T.S.-K., 2007. A memetic algorithm for multiple-drug cancer chemotherapy schedule optimization. IEEE Trans. Syst. Man Cybern. B Cybern. 37 (1), 84–91.

Van Den Berg, J.P., Vereecke, H.E.M., Proost, J.H., Eleveld, D.J., Wietasch, J.K.G., Absalom, A.R., Struys, M.M., 2017. Pharmacokinetic and pharmacodynamic interactions in anaesthesia. A review of current knowledge and how it can be used to optimize anaesthetic drug administration. Br. J. Anaesth. 118 (1), 44.

Vrabie, D., Vamvoudakis, K.G., Lewis, F.L., 2013. Optimal Adaptive Control and Differential Games by Reinforcement Learning Principle. Institution of Engineering and Technology, London.

Watkins, C.J.C.H., Dayan, P., 1992. Q-learning. Mach. Learn. J. 8 (3), 279–292.

WHO, 2018. Fact Sheets. Available from: http://www.who.int/mediacentre/factsheets/fs297/en/.

Zhao, Y., Zeng, D., Socinski, M.A., Kosorok, M.R., 2011. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. Biometrics 67 (4), 1422–1433.