Mathematical Biosciences 309 (2019) 131–142

Contents lists available at ScienceDirect

Mathematical Biosciences

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

Optimal adaptive control of drug dosing using integral reinforcement learning

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

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

ARTICLE INFO

Keywords:
Drug dosing
Optimal adaptive control
Reinforcement learning
Anesthesia administration

ABSTRACT

In this paper, a reinforcement learning (RL)-based optimal adaptive control approach is proposed for the continuous infusion of a sedative drug to maintain a required level of sedation. To illustrate the proposed method, we use the common anesthetic drug propofol used in intensive care units (ICUs). The proposed online integral reinforcement learning (IRL) algorithm is designed to provide optimal drug dosing for a given performance measure that iteratively updates the control solution with respect to the pharmacology of the patient while guaranteeing convergence to the optimal solution. Numerical results are presented using 10 simulated patients that demonstrate the efficacy of the proposed IRL-based controller.

1. Introduction

Personalized medicine and precision medicine are two emerging initiatives in modern health care that focus on creating awareness in these interdisciplinary areas [1]. The necessity for patient-specific drug administration in these areas has led to new research vistas [2–4]. The primary motivation that has fostered such initiatives is the fact that different patients respond differently to the same drug and its dosage due to genetic and molecular variabilities between patients and within the same patient. Personalized medicine aims to deliver personalized drug doses and drug types for each patient according to current and predicted responses of the patient collected from experimental data and statistical analysis [4]. In this paper, we focus on developing an online controller design method that can deliver an optimal and patient-specific drug dose based on the patient's current response state to the drug. Specifically, we address the "right dose" problem of personalized medicine.

Critically ill patients in intensive care units often require sedation to facilitate various clinical procedures and to comfort patients during treatment [5,6]. The task of anesthesia administration for patients in intensive care units is quite challenging as oversedation or undersedation can result in detrimental physiological, psychological, and economical impacts to patients. Several clinical and in silico trials carried out in this area have recommended closed-loop control of anesthesia administration to enhance the safety of patients and to facilitate the effective use of clinician expertise [5,7,8].

Any drug that is introduced intravenously to the human body is dispersed to various internal organs by the blood, which is then metabolized in the liver and later eliminated through the kidneys. The mechanism involved in drug dispersal can be captured using mathematical models that are generally based on clinical trials conducted using healthy volunteers or patient data available on drug response to certain diseases [9–11]. However, given that the internal organs, such as the heart, liver, and kidneys, play a key role in distributing and eliminating any drug induced into the human body, there are significant differences in the drug pharmacology between the healthy volunteers and patients with respiratory, cardiac, hepatic, or renal illness. Thus, it is difficult to account for all such variabilities in a mathematical model, calling into question the reliability of model-based optimal controllers and leading to the necessity for developing controller design strategies that provide optimal and adaptive control solutions.

Several closed-loop control strategies, such as model predictive control, optimal control, and adaptive disturbance rejection control, have been suggested for the control of anesthesia administration [11–16]. The control strategies that are currently in use for regulating patient drug dosing have focused on optimal drug infusion with respect to given performance measures or adaptive drug infusion that addresses patient parameter uncertainty. The main advantage of adaptive controllers is that they can derive patient-specific infusion profiles even in the absence of an accurate patient model. However, such controllers may not account for certain desired performance constraints. On the

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

https://doi.org/10.1016/j.mbs.2019.01.012
Received 20 October 2018; Received in revised form 24 January 2019; Accepted 31 January 2019
Available online 05 February 2019
0025-5564/

other hand, optimal controllers are predicated on nominal patient models leading to suboptimal performance or even instability of the closed-loop system in the face of drug titration for actual patients.

The challenge here is to design an optimal drug infusion profile that accounts for gender, age, weight, pharmacokinetic and pharmacodynamic intrapatient and interpatient variability, as well as health conditions of the patient under treatment. In contrast to standard controller design methods, reinforcement learning (RL)-based approaches allow the development of control algorithms that can be used in real-time to affect optimal and adaptive drug dosing in the presence of pharmacokinetic and pharmacodynamic patient variability. The method presented in this paper can be used to derive patient-specific drug infusion profiles for generating a desired drug response of a patient without requiring an accurate patient model. Specifically, we use a learning-based controller design strategy that can be used to facilitate patient-specific and optimal drug titration.

Learning-based control strategies have been used in drug dosing control to optimize the dosing of erythropoietin during hemodialysis [17], develop dynamic treatment regimens for patients with lung cancer [18], assist insulin regulation in diabetic patients [19], infuse cytotoxins during chemotherapy [20], and administer anesthetic drugs to maintain required levels of sedation [21]. Both clinical and in silico trials using reinforcement learning methods for improving control accuracy of anesthetic drug infusion have been recently reported in [21,22]. Compared to [17–22], the advantage of the proposed method is that apart from being optimal as well as adaptive, the controller design is presented in the continuous-time domain using integral reinforcement learning [23]. Moreover, while Q-learning-based approaches involve an off-line training phase to train the controller, the proposed IRL-based approach employs an online algorithm, and hence, the controller can adapt its gains with respect to the actual patient parameters.

Integral reinforcement learning is a RL-based method in which the controller (RL agent) can learn the unknown and time-varying dynamics of the system by interacting with the system [23]. The actor-critic structure of the algorithm evaluates the current control policy and iteratively updates it to meet a given performance measure. The control policy update is carried out by observing the response of the system predicated on the current control policy. Therefore, the IRL-based controller can learn optimal actions in the presence of system parameter uncertainty and in the absence of the complete knowledge of the system dynamics. Thus, when the IRL-based controller is used for real-time drug administration, iterative tuning of the infusion profile is executed with respect to the drug pharmacology of the patient in order to derive the optimal control policy.

In [24], an online integral reinforcement learning-based algorithm is developed for the tracking control of partially unknown linear systems. Specifically, the solution to an algebraic Riccati equation associated with the linear-quadratic tracking (LQT) problem for partially unknown continuous-time systems with the knowledge of an initial stabilizing control policy is derived online. The convergence and stability properties of the IRL algorithm are also addressed in [24]. In this paper, we use the IRL approach to develop a reliable closed-loop controller to maintain the required level of sedation quantified in terms of the well-known bispectral (BIS) index [16].

The remainder of the paper is organized as follows. Section 2 presents an overview to the pharmacokinetics and pharmacodynamics of the drug propofol and the design of the proposed IRL-based controller for the closed-loop control of anesthesia administration. Simulation results for two different patient age groups are given in Section 3, followed by a detailed discussion of these results in Section 4. Finally, in Section 5, we present conclusions and future research directions.

2. Methods

In this section, we first introduce the mathematical formulation of the pharmacokinetics and pharmacodynamics of propofol, and then present the IRL-based controller design in conjunction with a hybrid extended Kalman filter (EKF) used to reconstruct the system states.

2.1. Drug disposition model

Even though the IRL algorithm implementation does not require complete system knowledge, in this subsection we introduce a mathematical model of the pharmacokinetics and pharmacodynamics of the drug propofol in the human body for the following reasons. First, the model is used for the in silico simulations provided in the paper. Second, the proposed IRL-based iterative algorithm requires an initial stabilizing control policy to generate the patient response so that the controller can observe the response of the patient and learn the pharmacological characteristics of the patient. Furthermore, instead of using an arbitrary initial control policy, we use a feasible (i.e., stabilizing) control policy predicated on a nominal patient model for addressing patient safety. And finally, a nominal patient model is required to construct a state estimator.

As shown in Fig. 1, we use a four-compartment model to represent the pharmacokinetics and pharamcodynamics of propofol in the human body. Specifically, Compartment 1 models the intravascular blood to which the drug is administrated through one of the veins, Compartment 2 models muscle tissue, Compartment 3 models fat, and the effect-site compartment models the time-lag in the drug dynamics at the locus of the drug effect [25].

Drug types, such as anesthetics, analgesics, and neuromuscular blockades, hormones, such as insulin, and chemical agents, such as cytotoxin, colloids, and crystalloids, are some of the substances that are infused intravenously into the human body. In this paper, we use the common anesthetic drug propofol to illustrate the design of the proposed IRL-based controller. The drug dynamics of a patient varies according to the physiology of the patient. Hence, we use the following drug disposition model that is dependent on the patient parameters such as age, weight, etc., [26–29]

$$\begin{aligned}\dot{x}_1(t) &= -(k_{10}+k_{12}+k_{13})x_1(t)+k_{21}\frac{v_2}{v_1}x_2(t)+k_{31}\frac{v_3}{v_1}x_3(t)+u(t),\\ &\quad x_1(0)=x_{10},\quad t\geq 0,\end{aligned} \tag{1}$$

$$\dot{x}_2(t) = k_{12}\frac{v_1}{v_2}x_1(t)-k_{21}x_2(t),\quad x_2(0)=x_{20}, \tag{2}$$

$$\dot{x}_3(t) = k_{13}\frac{v_1}{v_3}x_1(t)-k_{31}x_3(t),\quad x_3(0)=x_{30}, \tag{3}$$

$$\dot{c}_{\text{eff}}(t) = k_{e0}x_1(t)-k_{e0}c_{\text{eff}}(t),\quad c_{\text{eff}}(0)=c_{\text{eff0}}, \tag{4}$$

where $x_i(t)$, $t\geq 0$, $i=1$, 2, and 3, denotes the mass of the drug in the first, second, and third compartments, respectively, $c_{\text{eff}}(t)$, $t\geq 0$, is the effect-site concentration of the drug, k_{ji}, $i\neq j$, represents the rate of mass transfer between the jth and ith compartments, v_i, $i=1$, 2, and 3, denotes the volumes of the three compartments, and $u(t)$, $t\geq 0$, is the infusion rate of the drug. For our model, the state vector is given by $x(t)=[x_1(t), x_2(t), x_3(t), c_{\text{eff}}(t)]^{\text{T}}$.

The values of k_{ji}, $i,j=1$, 2, and 3, in the pharmacokinetic and pharmacodynamic model given by (1)–(4) depend on the patient features such as age, weight, height, and gender, and are given in Table 1. In Table 1, *lbm* denotes the lean body mass of the patient and is given by $lbm=1.07weight-148(weight^2/height^2)$, C_1 is the rate at which the drug is removed by excretion, C_2 and C_3 are the rates of drug clearances between the central compartment and Compartments 2 and 3, respectively, and k_{e0} represents the effect-site elimination rate constant.

The drug effect in terms of the BIS is linear for lower drug doses; however, higher drug dosing and prolonged drug titration result in a nonlinear saturation (i.e., sigmoidal) effect described by the Hill equation given by ([25])

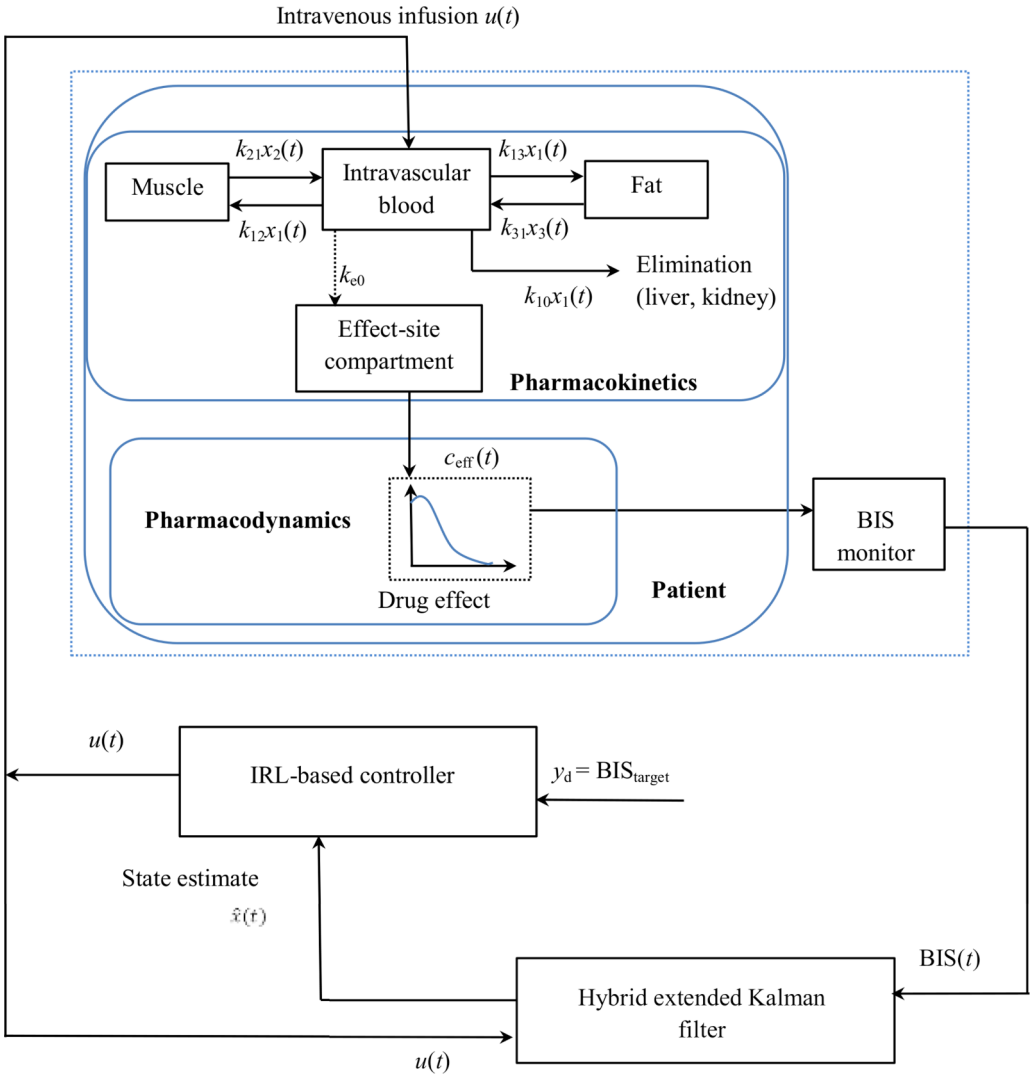

Fig. 1. IRL-based closed-loop control of drug administration.

$$\mathrm{BIS}(c_{\mathrm{eff}}(t)) = \mathrm{BIS}_0\left(1 - \frac{(c_{\mathrm{eff}}(t))^{\gamma}}{(c_{\mathrm{eff}}(t))^{\gamma} + (\mathrm{C}_{50})^{\gamma}}\right), \tag{5}$$

where BIS_0 is the base line value that represents an awake state, C_{50} is the drug concentration that causes 50% drug effect, and γ denotes the steepness of the drug concentration versus drug response relation. BIS ($c_{\mathrm{eff}}(t)$) is the measured value of the BIS index with a value in the range 0 to 100, where and BIS = 100 indicate an isoelectric electroencephalogram (EEG) signal and an EEG signal of a fully conscious patient, respectively. Note that (1)–(5) can be written as

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad t \geq 0, \tag{6}$$

$$y(t) = h(x(t)), \tag{7}$$

Table 1
Patient model parameters and parameter relations for the drug propofol [27,29].

Parameter	Model	Unit
v_1	4.27	l
v_2	$18.9 - 0.391(age - 53)$	l
v_3	2.38	l
C_1	$1.89 + 0.0456(weight - 77) - 0.681(lbm - 59) + 0.0264(height - 177)$	l min^{-1}
C_2	$1.29 - 0.024(age - 53)$	l min^{-1}
C_3	0.836	l min^{-1}
k_{e0}	0.456	min^{-1}
k_{10}	C_1/v_1	min^{-1}
k_{12}	C_2/v_1	min^{-1}
k_{13}	C_3/v_1	min^{-1}
k_{21}	C_2/v_2	min^{-1}
k_{31}	C_3/v_3	min^{-1}

where $A \in \mathbb{R}^{4\times4}$ is the system matrix, $B \in \mathbb{R}^{4\times1}$ is an input matrix, $x(t)$, $t \geq 0$, is the state vector, $y(t) = \mathrm{BIS}(t)$, $t \geq 0$, is the system measurement, and $u(t)$, $t \geq 0$, is the control input. Here, we assume that the pair (A, B) is stabilizable.

The system measurement as given by (5) is a nonlinear function of $c_{\mathrm{eff}}(t)$, $t \geq 0$. However, a linear approximation of the system measurement is required to design an IRL-based tracking controller. Hence, using a linear regression model in the region of the required target value of BIS(t), $t \geq 0$, the nonlinear measurement (5) can be approximated as [30]

$$y(t) = mc_{\mathrm{eff}}(t) + d, \tag{8}$$

where the constants m and d can be determined by multiple linear regression using a least-squares method on randomly selected patient data relating the patients pharmacokinetic and pharmacodynamic parameters and measured responses. Thus, using (8), (6) and (7) can be written as

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(0) = x_0, \quad t \geq 0, \tag{9}$$

$$y(t) = Cx(t) + d, \tag{10}$$

where, for $t \geq 0$, $x(t) \in \mathbb{R}^n$, $u(t) \in \mathbb{R}$, and $y(t) \in \mathbb{R}$.

2.2. Integral reinforcement learning-based controller design

In this subsection, we develop an integral reinforcement learning-based controller design method for the control of anesthesia administration. The objective is to design an online optimal adaptive tracking controller using an integral control action to account for system parameter uncertainties. Specifically, the integral tracking error is given by

$$e(t) = \int_0^t [y_{\mathrm{d}} - y(\tau)]\mathrm{d}\tau, \tag{11}$$

where $y(t)$, $t \geq 0$, and y_{d} are the measured response and the desired constant reference signal, respectively. Using (11), we obtain

$$\dot{e}(t) = y_{\mathrm{d}} - y(t) = \tilde{y}_{\mathrm{d}} - Cx(t), \quad e(0) = 0, \quad t \geq 0, \tag{12}$$

where $\tilde{y}_{\mathrm{d}} \triangleq y_{\mathrm{d}} - d$.

Using (9), (10), and (12) the augmented system (9) and (12) can be written as

$$\dot{x}_{\mathrm{a}}(t) = A_{\mathrm{a}}x_{\mathrm{a}}(t) + B_{\mathrm{a}}u(t) + G\tilde{y}_{\mathrm{d}}, \quad x_{\mathrm{a}}(0) = x_{\mathrm{a}0}, \quad t \geq 0, \tag{13}$$

where $x_{\mathrm{a}}(t) = [x^{\mathrm{T}}(t), e(t)]^{\mathrm{T}} \in \mathbb{R}^{\hat{n}}$, $\hat{n} = n + 1$,

$$A_{\mathrm{a}} = \begin{bmatrix} A & 0 \\ -C & 0 \end{bmatrix}, \quad B_{\mathrm{a}} = \begin{bmatrix} B \\ 0 \end{bmatrix}, \quad G = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Now, using the feedback control law

$$u(t) = k_1x_{\mathrm{a}}(t) + k_2\tilde{y}_{\mathrm{d}}, \tag{14}$$

where $k_1 \in \mathbb{R}^{1\times\hat{n}}$ and $k_2 \in \mathbb{R}$, the closed-loop system is given by

$$\dot{x}_{\mathrm{a}}(t) = \tilde{A}_{\mathrm{a}}x_{\mathrm{a}}(t) + \tilde{B}_{\mathrm{a}}\tilde{y}_{\mathrm{d}}, \quad x_{\mathrm{a}}(0) = x_{\mathrm{a}0}, \quad t \geq 0, \tag{15}$$

where $\tilde{A}_{\mathrm{a}} = A_{\mathrm{a}} + B_{\mathrm{a}}k_1$ and $\tilde{B}_{\mathrm{a}} = B_{\mathrm{a}}k_2 + G$, and $\tilde{A}_{\mathrm{a}}$ is Hurwitz.

Next, in order to track a desired constant reference signal, we consider the discounted cost function

$$V(x_{\mathrm{a}}(t), u(t)) = \frac{1}{2}\int_t^\infty e^{-\gamma_{\mathrm{d}}(\tau-t)}[x_{\mathrm{a}}^{\mathrm{T}}(\tau)Qx_{\mathrm{a}}(\tau) + u^{\mathrm{T}}(\tau)Ru(\tau)]\mathrm{d}\tau, \tag{16}$$

where γ_{d} is the discount factor, $Q \geq 0$, and $R > 0$. Here, we assume that the pair (A_{a}, Q) is observable. Note that since we are tracking a constant reference signal, the discount factor γ_{d} is introduced in the cost function to ensure (16) is finite over the infinite horizon. See Remarks 1 and 2 in [24] for further details.

An integral reinforcement learning algorithm is an iteration-based policy wherein the iteration starts with an initial arbitrary control policy that is stabilizing. Then, the control policy is progressively updated based on certain design criteria and until it achieves certain prespecified performance requirements. When we adopt any control algorithm for drug dosing, it is imperative to ensure patient safety. Hence, instead of initializing our algorithm with an arbitrary initial control policy, we assume that a nominal model of the patient is available and design an initial control policy based on the nominal model. This is a pragmatic assumption as there exist several models that depict the drug disposition mechanism in the human body and it is common to use such models to facilitate anesthesia administration [8,10,31,32]. However, it should be noted that the IRL algorithm does not use the knowledge of the system dynamics in designing an optimal control solution, rather it uses the input-output data of the system for tuning the controller.

Next, to derive an optimal control input using the IRL method, we show that the cost (16) can be written in terms of a LQT Bellman equation [24]. First, however, the following proposition is needed.

Proposition 1. *Consider the dynamical system (9) and (10) with reference dynamics (12) and stabilizing feedback control law (14). Then, the value function (16) can be written in a quadratic form*

$$V(X(t)) = \frac{1}{2}X^{\mathrm{T}}(t)PX(t), \tag{17}$$

where $X(t) = [x_{\mathrm{a}}^{\mathrm{T}}(t) \;\; \tilde{y}_{\mathrm{d}}]^{\mathrm{T}}$ *and some* $P = P^{\mathrm{T}} > 0$.

Proof. Substituting (14) into (16) and rearranging terms yields

$$V(X(t)) = \frac{1}{2}\int_t^\infty e^{-\gamma_{\mathrm{d}}(\tau-t)}[x_{\mathrm{a}}^{\mathrm{T}}(\tau)M_1x_{\mathrm{a}}(\tau) + 2x_{\mathrm{a}}^{\mathrm{T}}(\tau)M_2\tilde{y}_{\mathrm{d}} + \tilde{y}_{\mathrm{d}}M_3\tilde{y}_{\mathrm{d}}]\mathrm{d}\tau, \tag{18}$$

where $M_1 \triangleq Q + k_1^{\mathrm{T}}Rk_1$, $M_2 \triangleq k_1^{\mathrm{T}}Rk_2$, and $M_3 \triangleq k_2Rk_2$. Now, setting $\tau - t = \tau'$, (18) can be written as

$$V(X(t)) = \frac{1}{2}\int_0^\infty e^{-\gamma_{\mathrm{d}}\tau'}[x_{\mathrm{a}}^{\mathrm{T}}(\tau' + t)M_1x_{\mathrm{a}}(\tau' + t) + 2x_{\mathrm{a}}^{\mathrm{T}}(\tau' + t)M_2\tilde{y}_{\mathrm{d}} + \tilde{y}_{\mathrm{d}}M_3\tilde{y}_{\mathrm{d}}]\mathrm{d}\tau'. \tag{19}$$

Next, setting $\tau = \tau'$ in (19) and using

$$x_{\mathrm{a}}(\tau + t) = e^{\tilde{A}_{\mathrm{a}}\tau}x_{\mathrm{a}}(t) + \int_0^\tau e^{\tilde{A}_{\mathrm{a}}\tau'}\tilde{B}_{\mathrm{a}}\mathrm{d}\tau'\tilde{y}_{\mathrm{d}}, \tag{20}$$

it follows that

$$V(X(t)) = \frac{1}{2}[x_{\mathrm{a}}^{\mathrm{T}}(t) \;\; \tilde{y}_{\mathrm{d}}]P\begin{bmatrix} x_{\mathrm{a}}^{\mathrm{T}}(t) \\ \tilde{y}_{\mathrm{d}} \end{bmatrix}^{\mathrm{T}}, \tag{21}$$

where $P = \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$,

$$p_{11} = \int_0^\infty e^{-\gamma_{\mathrm{d}}\tau}\left[(e^{\tilde{A}_{\mathrm{a}}\tau})^{\mathrm{T}}M_1e^{\tilde{A}_{\mathrm{a}}\tau}\right]\mathrm{d}\tau, \tag{22}$$

$$p_{12} = \int_0^\infty e^{-\gamma_{\mathrm{d}}\tau}\left[(e^{\tilde{A}_{\mathrm{a}}\tau})^{\mathrm{T}}M_1L + (e^{\tilde{A}_{\mathrm{a}}\tau})^{\mathrm{T}}M_2\right]\mathrm{d}\tau, \tag{23}$$

$$p_{21} = \int_0^\infty e^{-\gamma_{\mathrm{d}}\tau}\left[L^{\mathrm{T}}M_1e^{\tilde{A}_{\mathrm{a}}\tau} + M_2e^{\tilde{A}_{\mathrm{a}}\tau}\right]\mathrm{d}\tau, \tag{24}$$

$$p_{22} = \int_0^\infty e^{-\gamma_{\mathrm{d}}\tau}\left[L^{\mathrm{T}}M_1L + 2L^{\mathrm{T}}M_2\right]\mathrm{d}\tau + M_3, \tag{25}$$

and $L \triangleq \int_0^\tau e^{\tilde{A}_{\mathrm{a}}\tau'}\tilde{B}_{\mathrm{a}}\mathrm{d}\tau'$, which proves (17). Finally, $P = P^{\mathrm{T}} > 0$ follows from the observability of (A_{a}, Q) and (22)–(25). □

Next, we obtain the Bellman equation for the closed-loop system (15) and the quadratic cost function (17). Specifically, consider the cost function (16), which can be equivalently written as

$$V(X(t),u(t)) = \frac{1}{2}\int_t^{\infty} e^{-\gamma_d(\tau-t)}[X^{\mathrm{T}}(\tau)S^{\mathrm{T}}QSX(\tau)+u^{\mathrm{T}}(\tau)Ru(\tau)]\mathrm{d}\tau, \tag{26}$$

where $S=\begin{bmatrix} I_{\hat{n}\times\hat{n}} & 0_{\hat{n}\times p} \\ 0_{1\times\hat{n}} & 0_{1\times p}\end{bmatrix}^{\mathrm{T}}$, and note that

$$\dot{V}(X(t),u(t)) = -\frac{1}{2}[X^{\mathrm{T}}(t)S^{\mathrm{T}}QSX(t)+u^{\mathrm{T}}(t)Ru(t)] + \gamma_d\frac{1}{2}X^{\mathrm{T}}(t)PX(t). \tag{27}$$

Next, differentiating (17), we obtain

$$\dot{V}(X(t)) = \frac{1}{2}X^{\mathrm{T}}(t)P\dot{X}(t)+\frac{1}{2}\dot{X}(t)^{\mathrm{T}}PX(t), \tag{28}$$

where

$$\dot{X}(t) = A_1X(t)+B_1u(t), \quad X(0)=X_0, \quad t\geq 0, \tag{29}$$

and

$$A_1 \triangleq \begin{bmatrix} A_a & G \\ 0_{1\times\hat{n}} & 0_{1\times 1}\end{bmatrix}, \quad B_1 \triangleq \begin{bmatrix} B_a \\ 0_{1\times 1}\end{bmatrix}.$$

Here, we assume that the pair (A_1, B_1) is stabilizable and the pair $(A_1, S^{\mathrm{T}}QS)$ is observable. Now, equating (27) and (28) yields the LQT Bellman equation

$$\begin{aligned} 0 &= [A_1X(t)+B_1u(t)]^{\mathrm{T}}PX(t)+X^{\mathrm{T}}(t)P[A_1X(t)+B_1u(t)] \\ &\quad -\gamma_dX^{\mathrm{T}}(t)PX(t) \\ &\quad +X^{\mathrm{T}}(t)S^{\mathrm{T}}QSX(t)+u^{\mathrm{T}}(t)Ru(t). \end{aligned} \tag{30}$$

Finally, to derive the optimal control for the infinite horizon LQT problem, define the Hamiltonian

$$\begin{aligned} H(X,u,P) &= (A_1X+B_1u)^{\mathrm{T}}PX+X^{\mathrm{T}}P(A_1X+B_1u)-\gamma_dX^{\mathrm{T}}PX \\ &\quad +X^{\mathrm{T}}S^{\mathrm{T}}QSX+u^{\mathrm{T}}Ru, \end{aligned} \tag{31}$$

where PX is the Fréchet derivative of the value function (17). Now, the necessary conditions for optimality yield

$$\frac{\partial H}{\partial u} = B_1^{\mathrm{T}}PX+Ru=0, \tag{32}$$

and hence,

$$u^* = K^*X, \tag{33}$$

where $K^* = -R^{-1}B_1^{\mathrm{T}}P$. Next, substituting (17) and (33) into (30) yields the algebraic Riccati equation

$$A_1^{\mathrm{T}}P+PA_1+S^{\mathrm{T}}QS-\gamma_dP-PB_1R^{-1}B_1^{\mathrm{T}}P=0. \tag{34}$$

Stabilizability of the pair (A_1, B_1) and observability of the pair $(A_1, S^{\mathrm{T}}QS)$ ensures that there exists a unique positive-definite solution P satisfying (34).

In order to compute the optimal gain K^*, one needs to solve (34), which depends on the system matrix A. Next, we show how one can iteratively find the solution of the algebraic Riccati Eq. (34) using the IRL Bellman equation when the system dynamics are unknown. Note that integrating (27) over the time interval $[t, t+T]$ we obtain

$$\begin{aligned} V(X(t)) &= \frac{1}{2}\int_t^{t+T} e^{-\gamma_d(\tau-t)}[X^{\mathrm{T}}(\tau)S^{\mathrm{T}}QSX(\tau)+u^{\mathrm{T}}(\tau)Ru(\tau)]\mathrm{d}\tau \\ &\quad +e^{-\gamma_dT}V(X(t+T)). \end{aligned} \tag{35}$$

Now, using (17), (35) becomes

$$\begin{aligned} X^{\mathrm{T}}(t)\ PX(t) &= \\ &\frac{1}{2}\int_t^{t+T} e^{-\gamma_d(\tau-t)}[X^{\mathrm{T}}(\tau)S^{\mathrm{T}}QSX(\tau)+u^{\mathrm{T}}(\tau)Ru(\tau)]\mathrm{d}\tau \\ &\quad +e^{-\gamma_dT}X^{\mathrm{T}}(t+T)PX(t+T). \end{aligned} \tag{36}$$

In order to implement a data-based integral reinforcement learning policy converging to the optimal control policy, we solve (36) by constructing two approximators consisting of a critic and an actor as outlined in Algorithm 1.

Algorithm 1. Online integral reinforcement learning policy iteration algorithm for solving the linear-quadratic tracking problem [24].

- Initialization: Initialize the control input $u_0(t) = K_0X(t)$.
- Policy evaluation: Using $u_k(t)$, $k=0,1,\ldots,$ $t\in[kT,(k+1)T]$, find P_k by solving

$$\begin{aligned} X^{\mathrm{T}}(kT)P_kX(kT) &= \frac{1}{2}\int_{kT}^{(k+1)T} e^{-\gamma_d(\tau-t)}[X^{\mathrm{T}}(\tau)S^{\mathrm{T}}QSX(\tau) \\ &\quad +u_k^{\mathrm{T}}(\tau)Ru_k(\tau)]\mathrm{d}\tau \\ &\quad +e^{-\gamma_dT}X^{\mathrm{T}}((k+1)T)P_kX((k+1)T). \end{aligned} \tag{37}$$

- Policy improvement: Iteratively update the control input $u_{k+1}(t)$ using

$$K_{k+1} = -R^{-1}B^{\mathrm{T}}P_k, \tag{38}$$

until $\|K_{k+1}-K_k\|_{\mathrm{F}} \leq \epsilon$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius matrix norm and ϵ is a preassigned tolerance.

The iterative IRL algorithm is similar to the IRL algorithms discussed in [23] and [24], which is shown to be equivalent to the Newton's method discussed in [33] and is quadratically convergent to the solution of an associated algebraic Riccati equation. Assuming the stabilizability of the pair (A_1, B_1) and the observability of the pair $(A_1, S^{\mathrm{T}}QS)$, and using a stabilizing initial controller K_0, the policy iteration given by Algorithm 1 converges to the optimal solution given by (33), where P satisfies the algebraic Ricatti equation (34) [23]. For the proof of asymptotic stability of LQT ARE solution see Theorem 2 in [24].

Algorithm 1 has an actor-critic structure in which (38) and (37) represent the actor and critic, respectively. Using Algorithm 1, the controller (38) can evaluate how a patient responds to the drug infusion $u_k(t)$, $t\geq 0$, in order to calculate P_k at each iteration k, and thus, obtain $u_{k+1}(t)$, $t\geq 0$, such that the cost (16) is minimized. The initial control input $u_0(t)$ is calculated using the nominal model of a patient. Once the IRL algorithm converges, the controller gives the *optimal* and patient-specific drug input.

2.3. Adaptive online implementation of IRL algorithm

In order to implement the iterative algorithm given by Algorithm 1, rewrite (37) as

$$\begin{aligned} X^{\mathrm{T}}(kT)P_kX(kT)-e^{-\gamma_dT}\ X^{\mathrm{T}}((k+1)T)P_kX((k+1)T) &= \\ \frac{1}{2}\int_{kT}^{(k+1)T} e^{-\gamma_d(\tau-t)}[X^{\mathrm{T}}(\tau)S^{\mathrm{T}}QSX(\tau) & \\ +u_k^{\mathrm{T}}(\tau)Ru_k(\tau)]\mathrm{d}\tau, & \end{aligned} \tag{39}$$

or, equivalently,

$$\vec{P}_k^{\mathrm{T}}Z_k = d_k, \tag{40}$$

where $d_k \triangleq V((k+1)T)-V(kT)$ is the integral reinforcement on the time interval $[kT,(k+1)T]$, $Z_k \triangleq \vec{X}(kT)-e^{-\gamma_dT}\vec{X}((k+1)T)$, and the $(\tilde{n}\times 1)$, $\left(\frac{\tilde{n}(\tilde{n}+1)}{2}\times 1\right)$, and $\left(\frac{\tilde{n}(\tilde{n}+1)}{2}\times 1\right)$ vectors

$$X(kT) = \begin{bmatrix} x_1(kT) \\ \vdots \\ x_{\tilde{n}}(kT) \end{bmatrix},$$

$$\vec{p}_k = \begin{bmatrix} p_{11}^k \\ \vdots \\ p_{ij}^k \\ 2p_{12}^k \\ \vdots \\ 2p_{1j}^k \\ 2p_{23}^k \\ \vdots \\ 2p_{2j}^k \\ \vdots \\ 2p_{(\tilde{n}-1)\tilde{n}}^k \end{bmatrix},$$

$$\vec{X}(kT) = \begin{bmatrix} x_1^2(kT) \\ \vdots \\ x_j^2(kT) \\ x_1(kT)x_2(kT) \\ \vdots \\ x_1(kT)x_j(kT) \\ x_2(kT)x_3(kT) \\ \vdots \\ x_2(kT)x_j(kT) \\ \vdots \\ x_{(\tilde{n}-1)}(kT)x_{\tilde{n}}(kT) \end{bmatrix},$$

$i, j = 1 \ldots \tilde{n}$, $\tilde{n} = \hat{n} + 1$, are the vectors derived using the entries of P_k and $X(t)$. Here, k denotes the iteration number, i, j denote the matrix indices, and $\vec{p}$ is obtained by stacking the diagonal entries followed by the upper triangular part of P_k into a column vector, where the off-diagonal entries are denoted as $2p_{ij}$.

The *desired value function* d_k can be computed by using

$$d_k = V((k+1)T) - V(kT), \tag{41}$$

where $\dot{V}(X(t), u(t)) = e^{-\gamma_d(\tau - t)}[X^{\mathrm{T}}(t)S^{\mathrm{T}}QSX(t) + u_k^{\mathrm{T}}(t)Ru_k(t)]$. Using (40) and (41) yields

$$\vec{p}_k = (Z_k Z_k^{\mathrm{T}})^{-1} Z_k d_k^{\mathrm{T}}. \tag{42}$$

During the time interval $[kT, (k+1)T]$, the matrix P_k is calculated after collecting a sufficient number of data points from the system trajectory, which is generated by applying the current control policy $u_k(t)$, $t \in [kT, (k+1)T]$, to the system. The vector $\vec{p}_k$ can be calculated by minimizing, in the least squares sense, the error between the target value function and the parameterized left-hand side of (42). The value of the vector $\vec{p}_k$, and thus the matrix P_k, is estimated by using N data points of the variables $X(t)$ and $V(X(t), u(t))$ collected during the time interval $[kT, (k+1)T]$ in the least squares Eq. (42).

Since the vector $\vec{p}_k$ has $\frac{\tilde{n}(\tilde{n}+1)}{2}$ independent components, at least $N \geq \frac{\tilde{n}(\tilde{n}+1)}{2}$ data points must be used to compute $\vec{p}_k$ [23]. Using the calculated value of P_k, the actor (38) calculates K_{k+1} to obtain the control policy $u_{k+1}(t) = K_{k+1}X(t)$, $t \in [kT, (k+1)T]$. This is repeated until the algorithm converges to the optimal control gain K^*.

Note that Algorithm 1 requires the knowledge of the system state $x(t)$, $t \geq 0$. However, in the case of anesthesia administration, it is impossible to measure the full state $x(t)$, $t \geq 0$. Hence, we use the measurable output BIS(t), $t \geq 0$, which is a nonlinear function of $x(t)$, $t \geq 0$, in conjunction with a hybrid EKF to reconstruct the system states for state feedback [34,35]. The discrete-time samples of the measured outputs BIS(t), $t \geq 0$, at the k'th time step are given by

$$y_{k'} = \mathrm{BIS}(c_{\mathrm{eff}}(t)), \quad t = k'T_{\mathrm{s}} \quad k' = 1, 2, \ldots, \tag{43}$$

where T_{s} is the sampling time.

2.4. Hybrid extended Kalman filter [35]

Using the continuous-time dynamics (6) and discrete-time measurement (43) we obtain

$$\dot{x}(t) = Ax(t) + Bu(t) + w(t), \quad x(0) = x_0, \quad t \geq 0, \tag{44}$$

$$y_{k'} = h(x_{k'}) + v_{k'}, \quad k' = 1, 2, \ldots, \tag{45}$$

where $x_{k'} = x(k'T_{\mathrm{s}})$, $w(t)$, $t \geq 0$, denotes a white process noise with intensity $\mathcal{N}(0, Q_{\mathrm{n}})$, and $v_{k'}$, $k' = 1, 2, \ldots$, denotes discrete-time white observation noise with covariance $\mathcal{N}(0, R_{\mathrm{n}})$. The hybrid extended Kalman filter for (44) and (45) is given as follows:

(1) Initialize the filter so that

$$\hat{x}_0^+ = \mathbb{E}[x_0], \tag{46}$$

$$Q_{\mathrm{e}0}^+ = \mathbb{E}[(x_0 - \hat{x}_0^+)(x_0 - \hat{x}_0^+)^{\mathrm{T}}], \tag{47}$$

where $\mathbb{E}[\cdot]$ denotes the expectation operator.

(2) For $k' = 1, 2, \ldots,$ perform the following steps.

(a) Integrate the continuous-time model for the state estimate $\hat{x}(t)$, $t \geq 0$, and covariance $Q_{\mathrm{e}}(t)$, $t \geq 0$, as

$$\begin{aligned} \dot{\hat{x}}(t) &= A\hat{x}(t) + Bu(t), \quad (k'-1)T_{\mathrm{s}} \leq t \leq k'T_{\mathrm{s}}, \\ \hat{x}((k'-1)T_{\mathrm{s}}) &= \hat{x}_{k'-1}^+, \\ \dot{Q}_{\mathrm{e}}(t) &= AQ_{\mathrm{e}}(t) + Q_{\mathrm{e}}(t)A^{\mathrm{T}} + Q_{\mathrm{n}}, \quad (k'-1)T_{\mathrm{s}} \leq t \leq k'T_{\mathrm{s}}, \\ Q_{\mathrm{e}}((k'-1)T_{\mathrm{s}}) &= (Q_{\mathrm{e}})_{k'-1}^+, \end{aligned} \tag{48}$$

where $\hat{x}_{(k'-1)}^+$ and $(Q_{\mathrm{e}})_{(k'-1)}^+$ are the initial conditions at the beginning of the integration process, and at the end of the integration the terminal condition satisfies $\hat{x}_{k'}^- = \hat{x}(k'T_{\mathrm{s}})$ and $(Q_{\mathrm{e}})_{k'}^- = Q_{\mathrm{e}}(k'T_{\mathrm{s}})$.

(b) At time instant k', incorporate the measurement $y_{k'}$ into the state estimate and error covariance as

$$F_{k'} = (Q_{\mathrm{e}})_{k'}^- J_{k'}^{\mathrm{T}} \left(J_{k'}(Q_{\mathrm{e}})_{k'}^- J_{k'}^{\mathrm{T}} + R_{\mathrm{n}} \right)^{-1}, \tag{49}$$

$$\hat{x}_{k'}^+ = \hat{x}_{k'}^- + F_{k'}(y_{k'} - h(\hat{x}_{k'}^-)), \tag{50}$$

$$(Q_{\mathrm{e}})_{k'}^+ = (I - F_{k'}J_{k'})(Q_{\mathrm{e}})_{k'}^-, \tag{51}$$

where $J_{k'}$ is the partial derivative of $h(x_{k'})$ with respect to $x_{k'}$ evaluated at $\hat{x}_{k'}^-$.

3. Simulation results

In this section, we present simulation results to illustrate the efficacy of the proposed IRL-based control approach for the closed-loop optimal adaptive control of drug dosing. The simulations were carried out using MATLAB®. In [9] and [36], it is shown that the value of C_{50}, which indicates the drug concentration that causes 50% drug effect, is

different for different age groups and it decreases as age increases. Given the significant effect of age on the pharmacodynamics of a patient, two different age groups are used in our simulations. Namely, Group-I is composed of elderly patients and Group-II involves young patients.

For both groups, a constant reference trajectory of $y_{\rm d}(t) = 50$, $t \geq 0$, is used and the discount factor is selected at $\gamma_{\rm d} = 0.9$. Note that the IRL algorithm does not use the system matrix A for learning the optimal P matrix. Instead, we use input-output data to demonstrate the efficacy of the proposed IRL-based controller design method. As discussed in Section 2, we use a hybrid EKF to reconstruct an estimate of the system states for feedback. At every k'th time step, the estimator gain $F_{k'}$, $k' = 1, 2, \ldots,$ is updated using the measured value of BIS(t), $t = k'T_{\rm s}$, where $T_{\rm s} = 0.2$ min. As noted earlier, the condition on the number of data points required for the least squares estimation problem is $N \geq \frac{\tilde{n}(\tilde{n}+1)}{2}$.

For $X(t)$, $t \geq 0$, we have $\tilde{n} = 6$, and hence, in each iteration we collect $N = 40$ data points. The time duration of integration in (37) is set to $T = 0.2$ min. Thus, the time duration of an iteration, denoted by $T_{\rm I} \triangleq [kT, (k+1)T]$, is $T \times N = 8$ min. Setting the time duration T to a very small value results in redundant information in the matrix Z_k. Alternatively, if the time duration T is set to a large value, then the controller may fail to detect certain drug response characteristics of the patient. As noted earlier, the parameter value m and the constant d in (8) can be determined by linear regression using the least-squares method on randomly selected patient data relating the patient's pharmacokinetic and pharmacodynamic parameters and measured response. For $\text{BIS}_{\text{target}} = 50$, we set $c_{\rm eff}(t) = C_{50}$, $t \geq 0$, and write the linearized form of (5) as

$$\text{BIS}(c_{\rm eff}(t)) \ = \text{BIS}(C_{50}) + \left.\frac{\partial \text{BIS}(c_{\rm eff}(t))}{\partial c_{\rm eff}(t)}\right|_{C_{50}} (c_{\rm eff}(t) - C_{50}) + \text{HOT}. \tag{52}$$

Group-I: In this group, we consider elderly patients of *age* = 58 ± 2 years, *height* = 156 ± 6 cm, and *weight* = 82 ± 8 kg. Table 2 shows the pharmacological parameters of the 5 simulated patients in Group-I. For the hybrid EKF, we set $R_{\rm n} = 100$, $Q_{\rm n} = I_{4\times4} \times 0.1$, and $Q_{\rm e0} = I_{4\times4}$. We use the model of Patient 1 to derive the estimator gain $F_{k'}$, $k' = 1, 2, \ldots,$ for all the 5 simulated patients in Group-I. It is a common practice among clinicians to use a nominal model derived using averaged patient parameters to facilitate target controlled infusion (TCI) [31,32]. To derive the values of m and d, we use the pharmacodynamic values of Patient 1 with $C_{50} = 3$ μg/ml and $\gamma = 2$ in (52) to obtain

$$\text{BIS}(c_{\rm eff}(t)) \approx 99.78 - 1.6566 \times 10^4 c_{\rm eff}(t). \tag{53}$$

We denote the optimal value of the state feedback gain obtained by solving the Riccati equation (34) by $K_{\rm R}^*$ and that obtained using Algorithm 1 by $K_{\rm A}^*$. The value of $K_{\rm R}^*$ for each patient is calculated using the respective pharmacokinetic model of the patient obtained using the model (1)–(4) with pharmacokinetic parameters and patient features given in Tables 1 and 2. Table 3 shows the initial feedback gain K_0 and the optimal feedback gains $K_{\rm A}^*$ and $K_{\rm R}^*$ for 3 out of the 5 patients in Group-I. We use the same initial stabilizing gain K_0 to derive the optimal value of the state feedback gain $K_{\rm A}^*$ for all of the 5 patients in Group-I. Starting with the initial feedback gain K_0, the algorithm converges iteratively to the optimal gain $K_{\rm A}^*$ by learning from the interactions with the patient and the response obtained. Note that $K_{\rm R}^*$ is calculated using (34), which does not involve the knowledge of the pharmacodynamic parameters of the patient. However, the algorithm relies on the patient's response to a drug to derive the optimal gain required for maintaining a certain level of drug response in the patient's body. Hence, the value of $K_{\rm A}^*$ reflects both the pharmacokinetics and pharmacodynamics of the patient.

Table 2
Patient parameters used to generate simulated patients in age Group-I.

Patient no.	*Age* [years]	*Height* [cm]	*Weight* [kg]	C_{50} [μg/ml]	γ
1	56	160	88	3.0	2.0
2	57	160	90	3.0	2.0
3	60	150	87	2.9	2.1
4	60	162	75	3.0	2.4
5	56	162	75	3.1	2.0

Figs. 2and 4 show the simulation results when the proposed IRL-based controller is used for the tracking control of the target BIS value in elderly patients. Note that the controller is able to achieve tracking performance with a deviation of ± 5 units from the desired set point. Fig. 3 shows the control input for the 5 elderly patients given in Table 2. Fig. 4 shows the convergence of the gain matrix K. In this figure, we have plotted $\|K_{\rm R}^* - K_k\|_{\rm F}$ versus the number of iterations for all the 5 patients in Group-I.

Group-II: In this group, we consider 5 young patients of *age* = 23 ± 2 years, *height* = 162 ± 3 cm, and *weight* = 55 ± 5 kg. Table 4 shows the pharmacological parameters of the 5 simulated patients in Group-II. For the hybrid EKF, we set $R_{\rm n} = 100$, $Q_{\rm n} = I_{4\times4} \times 0.1$, and $Q_{\rm e0} = I_{4\times4} \times 10$. We use the model of Patient 6 to derive the estimator gain $F_{k'}$, $k' = 1, 2, \ldots,$ for all of the 5 simulated patients in Group-II. Since the patient's sensitivity to the anesthetic drug propofol increases with increase in age [9,36], younger patients require more drug as compared to older patients to achieve the same level of sedation. We use the pharmacodynamic parameter values of Patient 6 with $C_{50} = 5$ μg/ml and $\gamma = 3$ in (52) to obtain

$$\text{BIS}(c_{\rm eff}(t)) \ \approx 124.75 - 1.4938 \times 10^4 c_{\rm eff}(t). \tag{54}$$

For all of the 5 patients in Group-II, we use the same initial stabilizing gain K_0 in Algorithm 1 to derive the optimal value of the state feedback gain $K_{\rm A}^*$; see Table 5. For each patient, the algorithm iteratively converges to the optimal value of the feedback gain $K_{\rm A}^*$ by accounting for the interactions with the patient and the response obtained. See Table 5 for the initial feedback gain K_0 and the optimal feedback gains $K_{\rm A}^*$ and $K_{\rm R}^*$ for three patients in Group-II.

Figs. 5and 6 show the simulation results when the proposed IRL-based controller is used for tracking control of the target BIS value in 5 young patients. Note that the controller is able to achieve tracking performance with a deviation of ± 10 units from the desired set point. Moreover, the value of $u(t)$, $t \geq 0$, as shown in Fig. 6 is within the acceptable clinical range of control inputs [6]. Fig. 7 shows the convergence of the gain matrix K. In this figure, we have plotted $\|K_{\rm R}^* - K_k\|_{\rm F}$ versus the number of iterations for all the 5 patients in Group-II.

4. Discussion

In this section, we discuss the performance along with some of the limitations of the proposed IRL-based controller design method based on the simulation results presented in Section 3. Here, the patient features of Patients 1 and 6 are obtained from [27] and [37], respectively. In order to show the performance of the proposed IRL-based controller when used for patients with varying patient features, we choose random values in the range *age* = 58 ± 2 years and *age* = 23 ± 2 years for Patients 2 to 5 in Group-I, and *age* = 23 ± 2 years, *height* = 162 ± 3 cm, and *weight* = 55 ± 5 kg for Patients 7 to 10 in Group-II. For Group-I, we used the pharmacodynamic parameter values of Patient 1 to derive the regression model parameters in (53) and obtain the initial stabilizing controller gain K_0. However, in order to show that the proposed controller can achieve robustness to system parameter uncertainities, we use the nominal values of the initial stabilizing controller gain and regression model parameters in Algorithm 1 to derive the optimal values of the state feedback gain $K_{\rm A}^*$ for all of the 5 patients in Group-I. Similarly, for Group-II, we used the

Table 3
Optimal feedback gains for Group-I.

Patient no.	Gain	K_{11}	K_{12}	K_{13}	K_{14}	K_{15}	K_{16}
All	K_0	− 2.2499	− 0.1602	− 0.1573	− 0.0000	-0.0006	-2.2440×10^{-05}
	K_A^*	− 1.7477	− 0.0000	− 0.0000	− 0.0000	− 0.0002	− 0.0001
	K_R^*	− 1.7321	− 0.2156	− 0.1356	− 0.0000	− 0.0002	− 0.0001
	K_A^*	− 1.7137	− 0.0000	− 0.0000	− 0.0000	− 0.0002	− 0.0001
	K_R^*	− 1.7935	− 0.2019	− 0.1382	− 0.0000	− 0.0002	− 0.0001
	K_A^*	− 1.6145	− 0.0000	− 0.0000	− 0.0000	− 0.0002	− 0.0001
	K_R^*	− 1.7796	− 0.2184	− 0.1376	− 0.0000	− 0.0002	− 0.0001

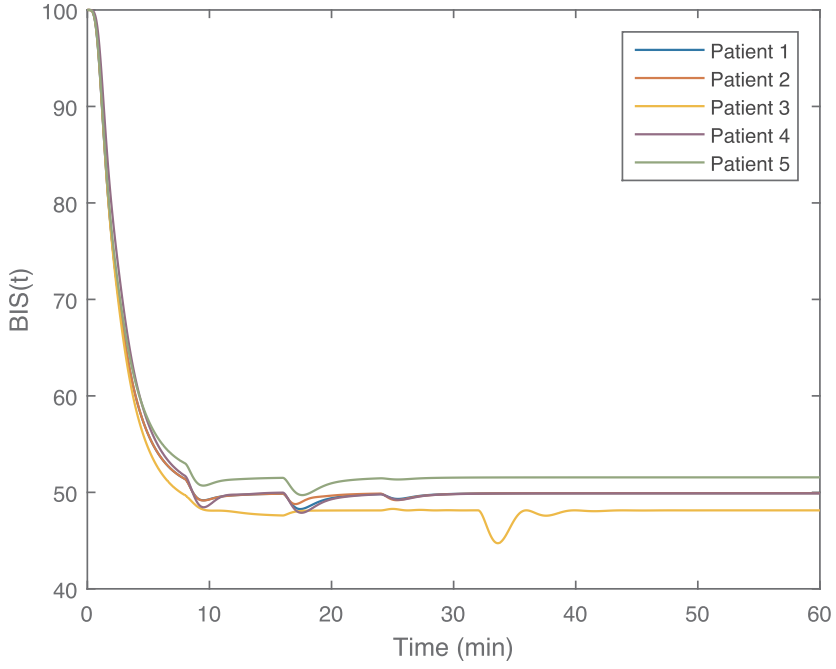

Fig. 2. BIS(t) versus time for the 5 patients in Group-I with $\text{BIS}_{\text{target}} = 50$.

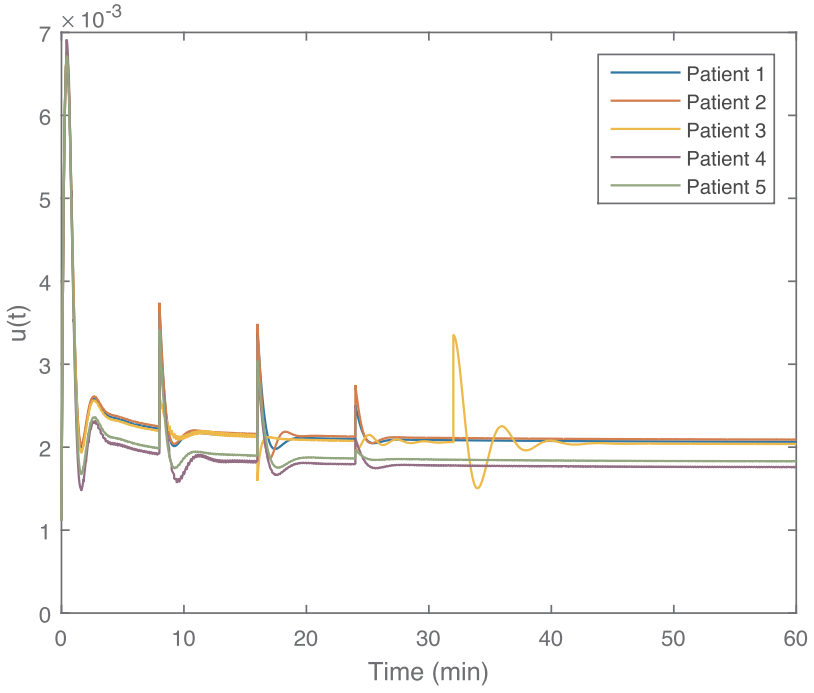

Fig. 3. Control inputs versus time for the 5 patients in Group-I.

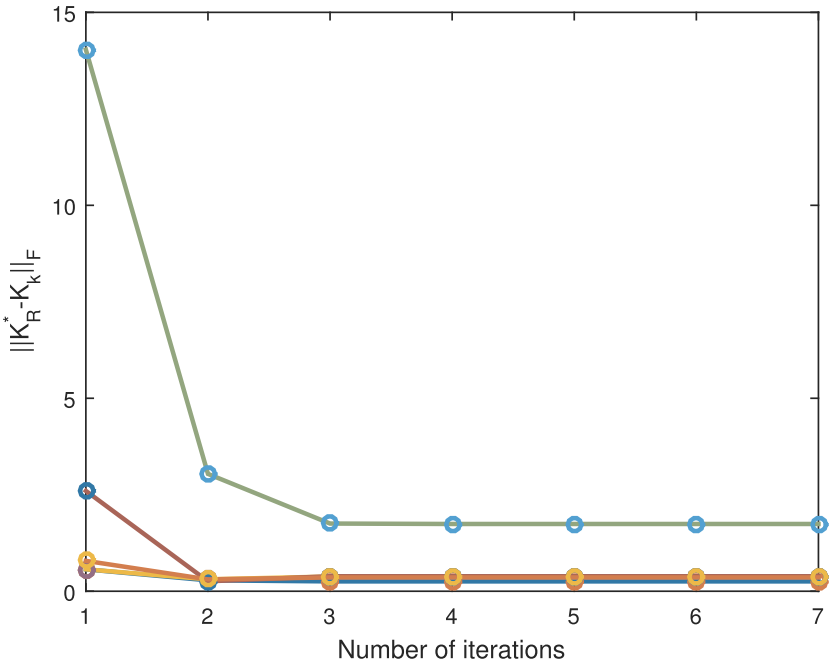

Fig. 4. Convergence of gain matrix K: $\|K_R^* - K_k\|_F$ versus the number of iterations for the 5 patients in Group-I.

Table 4
Patient parameters used to generate simulated patients in Group II.

Patient No.	*Age* [years]	*Height* [cm]	*Weight* [kg]	C_{50} [μg/ml]	γ
6	22	164	50	5.0	3.0
7	25	160	60	5.0	3.2
8	24	159	59	5.1	3.0
9	23	162	50	5.0	3.0
10	25	159	60	5.1	3.2

pharmacodynamic parameter values of Patient 6 to derive the regression model parameters in (54) and to obtain the initial stabilizing controller gain K_0. However, we use the nominal values of the initial stabilizing gain and regression model parameters in Algorithm 1 to derive the optimal values of the state feedback gain K_A^* for all of the 5 patients in Group-II.

During the 60 min drug infusion period presented in Section 3, the range of values of the induction phase duration for all the 5 simulated patients in Groups-I and -II are 3.95 ± 0.22 and 6.04 ± 0.24, respectively. The induction phase duration is the initial time from when the drug is administrated to the time when the drug effect reaches and remains within the range of BIS_{target} ± 10 for 30 seconds [38]. The minimum and maximum values of the BIS variable after reaching $BIS_{target} = 50$ for the first time is in the range 44.81 − 51.42 and 40.27 − 51.40 for all the 5 simulated patients in Groups-I and -II, respectively. All these performance metrics are within the acceptable range given in [39].

However, it can be seen from Figs. 2 and 5 that there is a small tracking error in the simulation results for both Groups-I and -II. The offset in tracking in the steady state region of Figs. 2 and 5 for some patients is due to the discrepancy between the linearized BIS model (8) that is used for the controller design and the actual nonlinear BIS output (5). In fact the tracking error is calculated using (8) instead of (5). In order to show this, we plotted $y(t)$, $t \geq 0$, given by (8) for all of the 5 elderly patients in Group-I. It can be seen from Fig. 8 that, in contrast to Fig. 2, the IRL-based controller is able to track $y(t) = 50$, $t \geq 0$, without any offset for all of the 5 simulated patients in Group-I. Note that Fig. 2 shows the measured BIS value given by (5), which is nonlinear. Similar comments hold for all of the 5 young patients in Group-II.

Another important point to note is the persistence of excitation (PE) condition on the system input that is required for the convergence of the IRL-algorithm [23,24]. In [23] and [24] persistence of excitation is ensured by injecting a probing noise along with the control input. Since we are dealing with patients, we do not add any probing noise during our simulations. It has been shown in [40] that the classical persistency of excitation-type conditions on the regression vectors of past inputs, outputs, and noise terms can be translated into corresponding conditions involving the inputs alone. Instead of using a probing noise, we assume that the regular persistence of excitation condition is satisfied with the feasible (i.e., stabilizing) initial control input that we used. Since the proposed IRL algorithm converges to the optimal control input, this assumption seems reasonable. However, we also note that the algorithm diverges whenever the Z_k matrix in (42) has an ill condition number. To avoid this situation, we need to formulate sufficient

Table 5
Optimal feedback gains for Group-II.

Patient no.	Gain	K_{11}	K_{12}	K_{13}	K_{14}	K_{15}	K_{16}
All	K_0	− 5.4593	− 0.0000	− 5.0191	− 0.2280	− 0.0008	-1.3600×10^{-05}
	K_A^*	− 1.5487	− 0.0000	− 0.0000	− 0.0000	− 0.0002	− 0.0001
	K_R^*	− 1.5581	− 0.3474	− 0.1297	− 0.0000	− 0.0002	− 0.0002
	K_A^*	− 1.5609	− 0.0000	− 0.0000	− 0.2623	− 0.0002	− 0.0002
	K_R^*	− 1.5656	− 0.3441	− 0.1300	− 0.0000	− 0.0002	− 0.0002
	K_A^*	− 1.5245	− 0.0000	− 0.0000	− 0.0000	− 0.0002	− 0.0001
	K_R^*	− 1.5669	− 0.3360	− 0.1301	− 0.0000	− 0.0002	− 0.0002

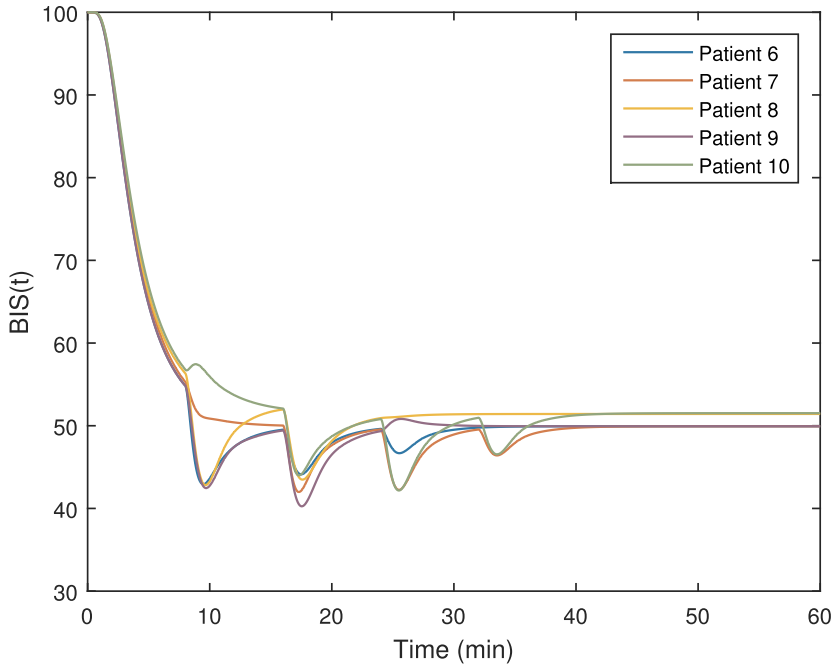

Fig. 5. BIS(t) versus time for the 5 patients in Group-II with $\text{BIS}_{\text{target}} = 50$.

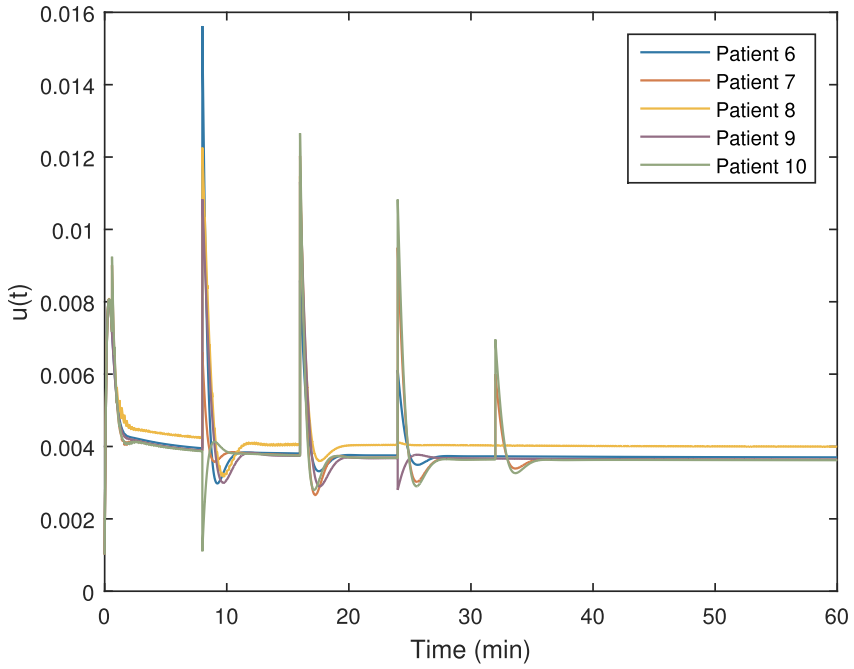

Fig. 6. Control inputs versus time for the 5 patients in group-II.

conditions on the initial control input with regard to the PE conditions; this will be considered in future research.

Finally, we note that even though the reinforcement learning framework requires the stabilizability and controllability of certain system matrix pairs, these assumptions are only needed to make sure that the Riccati equation (34) has a positive-definite solution. In the case of linear compartmental systems characterizing pharmacokinetic and pharmacodynamic drug dynamics with drug elimination, these systems are asymptotically stable [25], and hence, these geometric properties are automatically satisfied without requiring knowledge of the system matrices. Alternatively, assuming the availability of a nominal model of the patient along with a stabilizing nominal controller, it can be shown that the required minimality properties for the reinforcement learning framework are also satisfied.

5. Conclusions and future research directions

In this paper, an integral reinforcement learning-based controller design for the continuous infusion of a sedative drug to maintain a desired level of sedation in the human body is proposed. Simulation results using 10 patients with different pharmacological parameters show that the proposed IRL-based controller can achieve robustness to system parameter uncertainties and provide an optimal control solution. Further investigation of the performance of such controllers in the

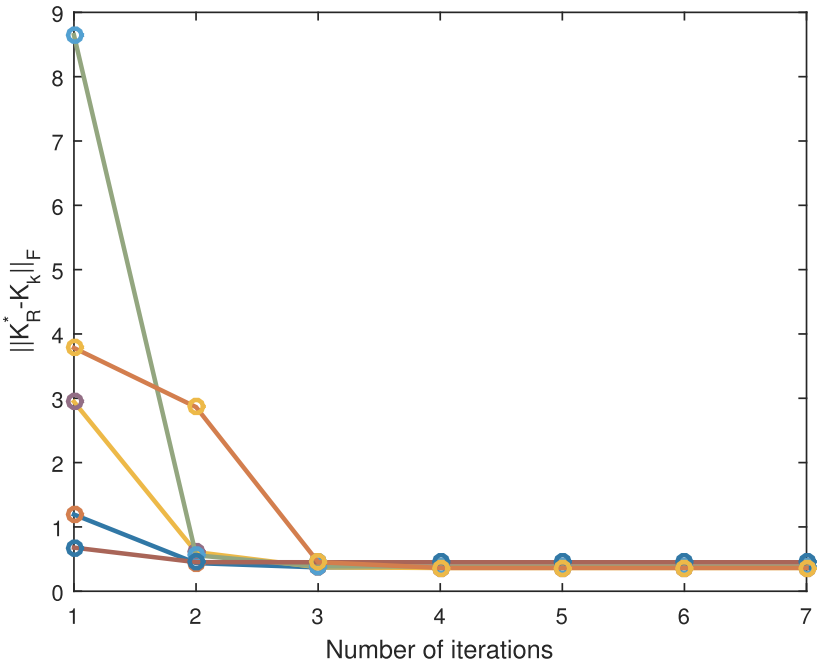

Fig. 7. Convergence of gain matrix K: $\|K_R^* - K_k\|_F$ versus the number of iterations for the 5 patients in Group-II.

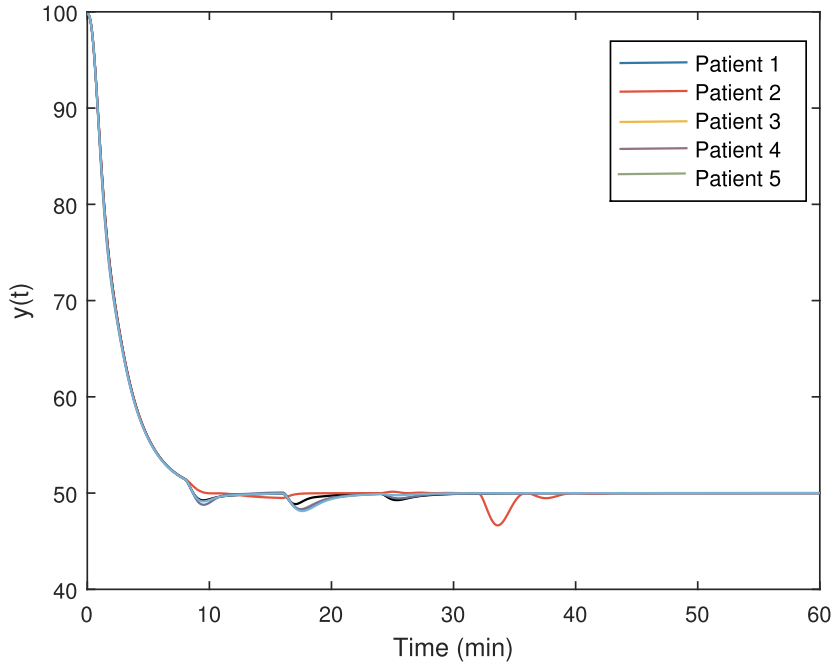

Fig. 8. $y(t)$ versus time for the 5 patients in Group-I with $\text{BIS}_{\text{target}} = 50$.

face of time delays, nonlinearities, and nonnegative constraints on the system inputs, states, and outputs will be considered in future research.

Acknowledgement

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

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.mbs.2019.01.012.

References

[1] B. Gholami, W.M. Haddad, J.M. Bailey, AI in the ICU, IEEE Spectr. 55 (10) (2018) 31–35.
[2] R.W. Peck, Precision medicine is not just genomics: the right dose for every patient, Annu. Rev. Pharmacol. Toxicol. 58 (1) (2018) 105–122.
[3] R.J. Gordon, Standardized care versus precision medicine: do we really need to wait for point-of-care testing? Anesth. Analg. 125 (6) (2017) 2161.
[4] S.J. Bielinski, et al., Preemptive genotyping for personalized medicine: design of the right drug, right dose, right time-using genomic data to individualize treatment protocol, Mayo Clin. Proc. 89 (1) (2014) 25–33.
[5] W.M. Haddad, J.M. Bailey, B. Gholami, A.R. Tannenbaum, Clinical decision support and closed-loop control for intensive care unit sedation, Asian J. Control 15 (2) (2013) 317–339.

[6] S. Mehta, L. Burry, S. Fischer, J.C.M. Motta, D. Hallet, D. Bowman, C. Wong, M.O. Meade, T.E. Stewart, D.J. Cook, Canadian survey of the use of sedatives, analgesics, and neuromuscular blocking agents in critically ill patients, Crit. Care Med. 34 (2) (2006) 374–380.
[7] A.R. Absalom, R.D. Keyser, M.M.R.F. Struys, Closed-loop anesthesia: are we getting close to finding the holygrail? Anesth. Analg. 112 (3) (2011) 516–518.
[8] J.P. Van Den Berg, H.E.M. Vereecke, J.H. Proost, D.J. Eleveld, J.K.G. Wietasch, A.R. Absalom, M.M.R.F. Struys, 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) (2017) 44.
[9] J. Barr, K. Zomorodi, E.J. Bertaccini, S.L. Shafer, E. Geller, A double blind randomised comparison of IV lorazepam versus midazolam for sedation of ICU patients via a pharmacologic model, Anesthesiology 95 (2001) 286–298.
[10] T.W. Schnider, C.F. Minto, P.L. Gambus, C. Andresen, D.B. Goodale, S.L. Shafer, E.J. Youngs, The influence of method of administration and covariates on the pharmacokinetics of propofol in adult volunteers, Anesthesiology 88 (5) (1998) 1170–1182.
[11] B. Gholami, W.M. Haddad, J.M. Bailey, A.R. Tannenbaum, Optimal drug dosing control for intensive care unit sedation using a hybrid deterministic-stochastic pharmacokinetic and pharmacodynamic model, Optim. Control Appl. Methods 34 (2013) 547–561.
[12] E. Furutani, K. Tsuruoka, S. Kusudo, A hypnosis and analgesia control system using a model predictive controller in total intravenous anesthesia during day-case surgery, Proceedings of the SICE Annual conference, Taipei, Taiwan, (August 2010), pp. 223–226.
[13] W.M. Haddad, T. Hayakawa, J.M. Bailey, Adaptive control for nonnegative and compartmental dynamical systems with applications to general anesthesia, Int. J. Adapt Control Signal Process. 17 (2003) 209–235.
[14] K. Soltesz, J.O. Hahn, T. Hagglund, G.A. Dumont, J.M. Ansermino, Individualized closed-loop control of propofol anesthesia: a preliminary study, Biomed Signal Process. Control 8 (6) (2013) 500–508.
[15] J.O. Hahn, G.A. Dumont, J.M. Ansermino, Robust closed-loop control of hypnosis with propofol using WAVcns index as the controlled variable, Biomed. Signal Process. Control 7 (5) (2012) 517–524.
[16] J.M. Bailey, W.M. Haddad, Drug dosing control in clinical pharmacology, IEEE Control Syst. Mag. 23 (2) (2005) 35–51.
[17] J. Martin-Guerrero, F. Gomez, E. Soria-Olivas, J. Schmidhuber, M. Climente-Marti, N. Jemenez-Torres, A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients, Expert Syst. Appl. 36 (2009) 9737–9742.
[18] Y. Zhao, D. Zeng, M.A. Socinski, M.R. Kosorok, Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer, Biometrics 67 (4) (2011) 1422–1433.
[19] E. Daskalaki, P. Diem, S.G. Mougiakakou, Personalized tuning of a reinforcement learning control algorithm for glucose regulation, Proceedings of the 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (2013) 3487–3490.
[20] R. Padmanabhan, N. Meskin, W.M. Haddad, Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment, Math. Biosci. 293 (2017) 11–20.
[21] B.L. Moore, L.D. Pyeatt, V. Kulkarni, P. Panousis, Kevin, A.G. Doufas, Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers, J. Mach. Learn. Res. 15 (2014) 655–696.
[22] R. Padmanabhan, N. Meskin, W.M. Haddad, Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning, Biomed. Signal Process. Control 22 (2015) 54–64.
[23] D. Vrabie, K.G. Vamvoudakis, F.L. Lewis, Optimal Adaptive Control and Differential Games by Reinforcement Learning Principle, Institution of Engineering and Technology, London, UK, 2013.
[24] H. Modares, F.L. Lewis, Linear quadratic tracking control of partially-unknown continuous-time systems using reinforcement learning, Proceedings of the IEEE Transactions on Automatic Control, 59 (2014), pp. 3051–3058.
[25] W.M. Haddad, V. Chellaboina, Q. Hui, Nonnegative and Compartmental Dynamical Systems, Princeton University Press, Princeton NJ, 2010.
[26] C.M. Ionescu, R. De Keyser, M.M. Struys, Evaluation of a propofol and remifentanil interaction model for predictive control of anesthesia induction, Proceedings of the 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC) (2011) 7374–7379.
[27] F. Nogueira, T. Mendonca, P. Rocha, Positive state observer for the automatic control of the depth of anesthesia-clinical results, Comput. Methods Programs Biomed. (2016), https://doi.org/10.1016/j.cmpb.2016.08.019.
[28] T. Mendonca, H. Alonso, M.M.D. Silva, S. Esteves, M. Seabra, Comparing different identification approaches for the depth of anesthesia using BIS measurements, IFAC Proceedings 45 (16) (2012) 781–785.
[29] C.M. Ionescu, I. Nascu, R. De Keyser, Lessons learned from closed loops in engineering: Towards a multivariable approach regulating depth of anaesthesia, J. Clin. Monit. Comput. 28 (6) (2014) 537–546.
[30] I. Nascu, C.M. Ionescu, I. Nascu, R. De Keyser, Evaluation of three protocols for automatic doa regulation using propofol and remifentanil, 9th IEEE International Conference on Control and Automation (ICCA) (2011) 573–578.
[31] B. Marsh, M. White, N. Morton, G.N. Kenny, Pharmacokinetic model driven infusion of propofol in children, Br. J. Anesth. 67 (1991) 41–48.
[32] A.R. Absalom, V. Mani, T. Smet, M.M. Struys, Pharmacokinetic models for propofol defining and illuminating the devil in the detail, Br. J. Anaesth. 103 (1) (2009) 26–37.
[33] D. Kleinman, On an iterative technique for Riccati equation computations, Proceedings of the IEEE Transactions on Automatic Control, 13(1) (1968), pp. 114–115.
[34] R.E. Kalman, A new approach to linear filtering and prediction problems, J. Basic Eng. 82 (1) (1960) 35–45.
[35] D. Simon, Optimal State Estimation: Kalman, H Infinity, and Nonlinear Approaches, Wiley-Interscience, Hoboken NJ, 2006.
[36] T.W. Schnider, C.F. Minto, S.L. Shafer, P.L. Gambus, C. Andresen, D.B. Goodale, E.J. Youngs, The influence of age on propofol pharmacodynamics, Anesthesiology 90 (6) (1999) 1502–1516.
[37] T. Kazama, K. Ikeda, K. Morita, M. Kikura, M. Doi, T. Ikeda, T. Kurita, Y. Nakajima, Comparison of the effect-site keos of propofol for blood pressure and eeg bispectral index in elderly and younger patients, Anesthesiology 90 (6) (1999) 1517–1527.
[38] K. Soltesz, G.A. Dumont, J.M. Ansermino, Assessing control performance in closed-loop anesthesia, Proceedings of the 21st Mediterranean Conference on Control and Automation (2013) 191–196.
[39] A.R. Absalom, K.P. Mason, Total Intravenous Anesthesia and Target Controlled Infusions: A Comprehensive Global Anthology, Springer, Switzerland AG, 2017.
[40] T.L. Lai, C.Z. Wei, On the concept of excitation in least squares identification and adaptive control, Stochastics 16 (3–4) (1986) 227–254.