scientific reports

Check for updates

OPEN

Delay-aware chemotherapy dosing via online critic learning

Farshad Rahimi[1,3]✉ & Mahdieh Samadi[2,3]

This paper proposes a delay-aware adaptive control framework for individualized chemotherapy dosing based on online critic learning. The approach explicitly compensates for pharmacokinetic and pharmacodynamic delays while adapting to patient-specific uncertainties. An online critic network estimates the value function to guide real-time dose adjustments. Simulation results on diverse patient profiles demonstrate effective tumor suppression and toxicity control, highlighting the robustness of the proposed scheme to variations in delay and patient dynamics.

Cancer remains one of the leading causes of mortality worldwide. Global research indicates that air pollution is directly linked to lung and bladder cancers and contributes to an increase in breast and pancreatic cancers[1,2]. Treatment approaches for cancer vary depending on the specific type of malignancy and the patient's individual condition[3].

Cancer treatment involves a range of approaches, including radiation, chemotherapy, and immunotherapy. Among these, chemotherapy continues to play a central role in managing the disease by targeting and destroying cancerous cells. It works through two main mechanisms: cytotoxic effects, which directly kill cancer cells by triggering apoptosis or necrosis, and cytostatic effects, where drugs like cytarabine slow down tumor growth by inhibiting DNA replication[4]. The effectiveness of these treatments depends on various factors such as the type of drug used, dosage, and specific characteristics of the cancer. This highlights the importance of accurate dosing strategies–like those explored in this study–to enhance treatment benefits while reducing harmful side effects. One of the major challenges with chemotherapy, however, is its impact on healthy cells, often leading to significant side effects[5]. These adverse effects limit how much of the drug can be safely given, making it crucial to strike a careful balance between patient safety and effective tumor reduction[6].

Assessing the effectiveness and feasibility of chemotherapy plans is critical for optimizing patient outcomes. While clinical trials provide robust evaluations, they are hindered by prolonged durations, high costs, and implementation challenges, leading to increased expenses[7,8]. Consequently, developing cost-effective chemotherapy strategies is a priority. Understanding the fundamental dynamics of tumor growth is essential before applying control methods to manage cancer. Significant research has advanced this field[9]. In this study, we adopt the model detailed in[10,11], selected for its ability to incorporate memory effects in tumor response and facilitate analysis of a stable equilibrium point, which supports the goal of reducing tumor cell populations.

The model introduced in[10] provides a foundational framework for developing control strategies to suppress tumor cell proliferation. Optimal control theory is instrumental in designing efficient drug administration protocols by accommodating various constraints and assumptions. Several studies[12,13] have built upon the framework in[10] to develop optimal control-based solutions. Notably, the study in[14] proposes a nominal-plus-neighboring optimal control method for cancer treatment through adoptive cellular immunotherapy, aiming to reduce tumor cell density and treatment costs while enhancing immune responses. Additionally, the work in[13] explores an integral reinforcement learning-based control strategy, which, while applied to drug infusion in other contexts, offers insights applicable to optimizing chemotherapy dosing.

Related works

This subsection reviews works most closely related to the proposed method.

In[15], an optimal control strategy is proposed for managing tumor growth. The authors linearize the nonlinear dynamics of tumor growth using time-varying approximations and apply a linear quadratic regulator to control tumor proliferation. Building on[16] and[15], the study in[11] develops drug regimens for cancer patients by integrating a state-dependent Riccati equation approach with an extended Kalman filter. Other control methods, including fuzzy control and model predictive control, have been explored to address chemotherapy challenges[17]. For instance,[18] introduces a model predictive control method for scheduling cancer therapy, effective even with incomplete measurements. This approach highlights the importance of estimating states and parameters to account for patient-specific variations in tumor growth and drug response, which may deviate significantly from

[1]Faculty of Electrical and Computer Engineering, Sahand University of Technology, Tabriz, Iran. [2]Faculty of Electrical and Computer Engineering, University of Tabriz, Tabriz, Iran. [3]Farshad Rahimi and Mahdieh Samadi have contributed equally to this work. ✉email: fa_rahimi@sut.ac.ir

the model. Recent studies have also explored the performance of data-driven MPC schemes under imperfect or uncertain inputs. For instance, Liu *et al.*[19] analyzed the regret bounds of MPC in such scenarios, providing theoretical insights into input uncertainty effects. Our work complements this direction by addressing the case of delayed and uncertain inputs within a critic-learning-based adaptive dosing framework.

Learning-based control methods offer distinct advantages, particularly in adaptability and handling nonlinearities. These methods adjust to evolving conditions and environments, making them ideal for complex, dynamic systems. Their ability to manage nonlinearities and uncertainties enhances both accuracy and reliability[20,21].

In recent reinforcement learning applications, an innovative method has been introduced that integrates Bayesian data assimilation with reinforcement learning to optimize chemotherapy dosing for cancer patients. This approach has shown promising results, notably lowering the incidence of neutropenia when compared to conventional treatment strategies[22]. Furthermore, a model-based optimal control strategy has been designed specifically for consolidation therapy in acute myeloid leukemia. By incorporating pharmacokinetic and pharmacodynamic modeling, this method aims to improve white blood cell recovery (nadir levels) while minimizing the required dosage of cytarabine[23].

In[24], a reinforcement learning-based control approach is developed for chemotherapy, utilizing a Q-learning algorithm tested on a nonlinear model of chemotherapy drug dynamics. In[25], a reinforcement learning-based optimal control strategy for chemotherapy is proposed, using drug input and tumor cell output without requiring a full-state observer. This strategy employs an actor-critic architecture with fuzzy-rule networks and a discontinuous reward function, validated through numerical simulations. Similarly,[26] presents a model-free adaptive controller combining fuzzy-rule networks and reinforcement learning for optimal chemotherapy drug administration, also validated numerically. Additionally,[27] introduces a model-free control approach using normalized advantage function reinforcement learning for cancer treatment, enhancing immune responses against tumor cell proliferation. This method integrates chemotherapy and anti-angiogenic drugs, demonstrating efficacy in reducing tumor cell populations with minimal drug doses, without relying on complex mathematical models.

The aforementioned studies assume real-time availability of measured data. However, time delays are inherent in chemotherapy processes. Typically, a delay of 1 to 14 days occurs between tumor measurement (e.g., via imaging or laboratory tests) and therapy adjustment, due to multidisciplinary team reviews, scheduling, and result processing[28]. These delays significantly affect system stability and performance, hindering timely dosing adjustments critical for effective treatment. Traditional control approaches often overlook these delays, an assumption impractical for chemotherapy. Incorporating delays into controller design is vital for achieving robust, adaptive control that reflects real-world treatment dynamics. However, managing time-delay systems in learning-based approaches poses unique challenges, as these systems are infinite-dimensional and complex, particularly within adaptive dynamic programming frameworks. Moreover, modeling delays–often represented by integral terms–complicates stability proofs and implementation. In traditional control methods, handling delays involves finding an upper bound for integral delay terms, whereas learning-based approaches require bounding the integral terms of the delays themselves, not their derivatives. To our knowledge, no prior studies have investigated optimal drug dosing with time delays using critic-only structure learning.

This study addresses this gap by developing an online critic learning method tailored for time-delay systems, optimizing drug administration while accounting for delays and customizing treatment to individual patient conditions, thereby advancing cancer chemotherapy control.

The primary contributions of this paper are outlined below:

- A novel value function is proposed for an online critic learning method to optimize drug dosing in cancer chemotherapy. This value function explicitly incorporates time delays in the treatment process and adapts dosing strategies to each patient's unique conditions through appropriate weighting factors.
- Unlike ?[24–26], this approach accounts for time delays between tumor cell measurements and drug administration, improving treatment responsiveness and accuracy.
- A bilinear matrix inequality is formulated to evaluate the impact of time delays on achieving equilibrium, providing a framework to analyze the stability of the proposed optimal chemotherapy approach under constant delays.

The manuscript is organized as follows: Sect. 2 establishes the mathematical framework of the problem under study and elaborates on the core objectives guiding the proposed control strategy. Additionally, it covers the equilibrium analysis of the cancer model and the development of a performance index. Section 3 demonstrates the effectiveness of the proposed methodology through simulation results. Finally, Sect. 4 provides the concluding remarks.

Mathematical formulation

In this paper, we analyze a nonlinear mathematical model of cancer proposed by de Pillis and Radunskaya[10]. The model describes the dynamics of three cell populations: normal (healthy) cells ($\mathscr{N}$), tumor cells ($\mathscr{T}$), and immune cells ($\mathscr{I}$), using ordinary differential equations to capture their growth and interaction. The model employs parameters that are representative of typical biological values, as defined in[10], to describe the system dynamics (e.g., growth rates and carrying capacities). These parameters, referred to as normalized in the sense of being standardized for the model, facilitate analysis of cancer dynamics across different patients. The normal cell population refers specifically to healthy host cells in the tissue near the tumor site, not a statistically normalized quantity[10].

$$\dot{\mathscr{N}}(t) = r_2\mathscr{N}(t)(1 - b_2\mathscr{N}(t)) - c_4\mathscr{N}(t)\mathscr{T}(t), \tag{1a}$$

$$\dot{\mathscr{T}}(t) = r_1\mathscr{T}(t)(1 - b_1\mathscr{T}(t)) - c_2\mathscr{I}(t)\mathscr{T}(t) - c_3\mathscr{T}(t)\mathscr{N}(t), \tag{1b}$$

$$\dot{\mathscr{I}}(t) = s + \frac{\rho\mathscr{I}(t)\mathscr{T}(t)}{\alpha + \mathscr{T}(t)} - c_1\mathscr{I}(t)\mathscr{T}(t) - d_1\mathscr{I}(t). \tag{1c}$$

In this model, all variables and parameters are positive due to physiological reasons. The differential equation for normal cells shows logistic growth as $\mathscr{N}(t)(1 - b_2\mathscr{N}(t))$ with growth rate r_2, where b_2^{-1} is the carrying capacity, and $-c_4\mathscr{T}(t)\mathscr{N}(t)$ represents normal cell decline due to competition with tumor cells. In (1b), $\mathscr{T}(t)(1 - b_1\mathscr{T}(t))$ shows logistic growth with rate r_1 and carrying capacity b_1^{-1}. Terms $-c_2\mathscr{I}(t)\mathscr{T}(t)$ and $-c_3\mathscr{T}(t)\mathscr{N}(t)$ describe tumor cell death from immune and normal cell interactions. In (1c), tumor cells stimulate immune cell growth, modeled by $\rho\mathscr{I}(t)\mathscr{T}(t)/(\alpha + \mathscr{T}(t))$, while immune cells die at rate d_1 in the absence of tumor cells. $-c_1\mathscr{I}(t)\mathscr{T}(t)$ represents immune cell inactivation by tumor cells. This model does not represent any specific cancer type and does not account for chemotherapy effects[10].

Analysis of equilibrium points in a drug-free model

The model presented in (1) has three distinct types of equilibrium points, which will be discussed below.

- **Case 1: Tumor-Free State** The tumor-free equilibrium is characterized by the absence of tumor cells and is given by:

$$\mathscr{T}^* = \left(\frac{1}{b_2}, 0, \frac{s}{d_1}\right).$$

This equilibrium is asymptotically stable if the following condition is satisfied:

$$r_1 < c_3 + \frac{c_2 s}{d_1}.$$

- **Case 2: Dead State** The dead state, where normal cells are absent, has two equilibrium points:

$$D_1^* = \left(0, 0, \frac{s}{d_1}\right),$$
$$D_2^* = (0, z, f(z)),$$

where z is a non-negative solution of the equation

$$z + \left(\frac{c_2}{r_1 b_1}\right) f(z) - \frac{1}{b_1} = 0, \tag{2}$$

and $f(z)$ is defined as

$$f(z) = \frac{s(z + a)}{c_1 z(z + a) + d_1(z + a) - \rho z}. \tag{3}$$

- D_1^* is always unstable. D_2^* may be stable or unstable depending on the system parameters.
- **Case 3: Coexisting State** The coexisting equilibrium, where all cell types are present, is given by:

$$C^* = (g(x), x, f(x)),$$

where x is a non-negative solution of the equation

$$x + \left(\frac{c_2}{r_1 b_1}\right) f(x) + \left(\frac{c_3}{r_1 b_1}\right) g(x) - \frac{1}{b_1} = 0,$$

and $g(x)$ is defined as

$$g(x) = \frac{1}{b_2} - \left(\frac{c_4}{r_2}\right) x.$$

In this paper, we utilize the parameter sets and variation ranges suggested in[10], as shown in Table 1. These parameters are not specific to any particular type of cancer and can vary between different cancer types and individual patients. For example, these parameter values could approximate the dynamics of a generic solid tumor with moderate growth and immune response, as seen in some clinical cases[29], though they remain theoretical and adaptable to various cancers per[10,11]. The parameter set in Table 1 includes several equilibrium points. For instance, a coexisting stable equilibrium point at (0.435, 0.565, 0.435) is depicted in Fig. 1.

Parameter	Description	Value	Unit
a_1	Immune cell kill rate	0.2	$\mathrm{mg}^{-1}\,\mathrm{L}\,\mathrm{day}^{-1}$
a_2	Tumor cell kill rate	0.3	$\mathrm{mg}^{-1}\,\mathrm{L}\,\mathrm{day}^{-1}$
a_3	Normal cell kill rate	0.1	$\mathrm{mg}^{-1}\,\mathrm{L}\,\mathrm{day}^{-1}$
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	1	$\mathrm{cell}^{-1}\mathrm{day}^{-1}$
c_2	Tumor cell competition term	0.5	$\mathrm{cell}^{-1}\mathrm{day}^{-1}$
c_3	Tumor-normal cell competition term	1	$\mathrm{cell}^{-1}\mathrm{day}^{-1}$
c_4	Normal-tumor cell competition term	1	$\mathrm{cell}^{-1}\mathrm{day}^{-1}$
d_1	Immune cell death rate	0.2	day^{-1}
d_2	Decay rate of injected drug	1	day^{-1}
r_1	Tumor cell growth rate	1.5	day^{-1}
r_2	Normal cell growth rate	1	day^{-1}
s	Immune cell influx rate	0.33	$\mathrm{cell}\,\mathrm{day}^{-1}$
α	Immune threshold rate	0.3	cell
ρ	Immune response rate	0.01	day^{-1}

Table 1. The parameters used in this paper come from[10,11].

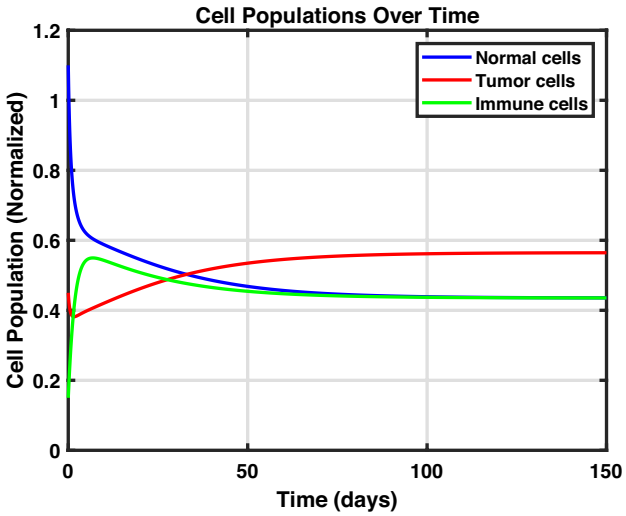

Fig. 1. Assessment of cell populations without chemotherapy.

The influence of chemotherapy on tumor progression can be analyzed from multiple perspectives. It is presumed that chemotherapy impacts all components at varying rates, specifically affecting normal cells ($\mathscr{N}$), tumor cells ($\mathscr{T}$), and immune cells ($\mathscr{I}$) through the drug concentration $\mathscr{M}(t)$, with the kill rates defined by parameters $a_3 = 0.1\,\mathrm{mg}^{-1}\mathrm{Lday}^{-1}$, $a_2 = 0.3\,\mathrm{mg}^{-1}\mathrm{Lday}^{-1}$, and $a_1 = 0.2\,\mathrm{mg}^{-1}\mathrm{Lday}^{-1}$, respectively, as listed in Table 1. In the model, the influence of chemotherapy is represented by an extra state $\mathscr{M}(t)$, which signifies the concentration of the drug in the bloodstream (mg/L). The pharmacokinetics (PK) of the chemotherapy drug follows a one-compartment model, described by $\dot{\mathscr{M}}(t) = -d_2\mathscr{M}(t) + u(t)$, where $d_2 = 1\,\mathrm{day}^{-1}$ is the drug decay rate (Table 1). The pharmacodynamics (PD) is modeled with a linear formulation, where the drug effect on each component–normal cells ($-a_3\mathscr{N}\mathscr{M}$), tumor cells ($-a_2\mathscr{T}\mathscr{M}$), and immune cells ($-a_1\mathscr{I}\mathscr{M}$)–is proportional to the drug concentration $\mathscr{M}(t)$, consistent with standard PK/PD modeling approaches in drug development[30]. The dynamics of each component under chemotherapy treatment are described by:

$$
\begin{aligned}
\dot{\mathscr{N}}(t) &= r_2\mathscr{N}(t)(1-b_2\mathscr{N}(t)) - c_4\mathscr{N}(t)\mathscr{T}(t) - a_3\mathscr{N}(t)\mathscr{M}(t),\\
\dot{\mathscr{T}}(t) &= r_1\mathscr{T}(t)(1-b_1\mathscr{T}(t)) - c_2\mathscr{I}(t)\mathscr{T}(t) - c_3\mathscr{T}(t)\mathscr{N}(t)\\
&\quad - a_2\mathscr{T}(t)\mathscr{M}(t),\\
\dot{\mathscr{I}}(t) &= s + \frac{\rho\mathscr{I}(t)\mathscr{T}(t)}{\alpha+\mathscr{T}(t)} - c_1\mathscr{I}(t)\mathscr{T}(t) - d_1\mathscr{I}(t) - a_1\mathscr{I}(t)\mathscr{M}(t),\\
\dot{\mathscr{M}}(t) &= -d_2\mathscr{M}(t) + u(t),
\end{aligned}
\tag{4}
$$

where d_2 represents the rate at which the chemotherapy drug degrades in the bloodstream, and $u(t) \in \mathbb{R}^m$ denotes the control input, or the externally administered drug dosage (mg/L/day), at time t. Delays are critical in many real-world systems, such as chemical reactions, where they impact efficiency, safety, and predictability by influencing reaction rates and enabling optimization of conditions like temperature and catalysts.

Modeling time delay in administering chemotherapy

The pharmacokinetic model is formulated as a one-compartment linear system to provide a tractable basis for developing and analyzing the proposed delay-aware control strategy. Although real chemotherapeutic drugs may exhibit multi-compartment kinetics and nonlinear pharmacodynamics (e.g., Hill-type effects), the adopted linear representation offers a locally valid approximation around the therapeutic operating point and is consistent with many control-oriented studies in the literature. Future extensions of this framework will consider nonlinear and drug-specific PK/PD models to further enhance physiological fidelity. It is important to account for the unavoidable delays between measuring tumor size and administering chemotherapy. Real-time measurement of tumor size and instant adjustment of drug dosage are impractical. Therefore, this time-delay should be incorporated into the model described in (4). Consequently, the model in (4) can be rewritten as follows:

$$
\begin{aligned}
\dot{\mathscr{N}}(t) &= r_2\mathscr{N}(t)(1-b_2\mathscr{N}(t)) - c_4\mathscr{N}(t)\mathscr{T}(t-d) - a_3\mathscr{N}(t)\mathscr{M}(t),\\
\dot{\mathscr{T}}(t) &= r_1\mathscr{T}(t-d)(1-b_1\mathscr{T}(t-d)) - c_2\mathscr{I}(t)\mathscr{T}(t-d) - c_3\mathscr{T}(t-d)\mathscr{N}(t) - a_2\mathscr{T}(t-d)\mathscr{M}(t),\\
\dot{\mathscr{I}}(t) &= s + \frac{\rho\mathscr{I}(t)\mathscr{T}(t-d)}{\alpha+\mathscr{T}(t-d)} - c_1\mathscr{I}(t)\mathscr{T}(t-d) - d_1\mathscr{I}(t) - a_1\mathscr{I}(t)\mathscr{M}(t),\\
\dot{\mathscr{M}}(t) &= -d_2\mathscr{M}(t) + u(t),
\end{aligned}
\tag{5}
$$

Here, d represents the time-delay between measuring the tumor size and administering chemotherapy. The model in (4) incorporates these delays into the cancer treatment process. In the revised cancer treatment model (5), we assume that the number of normal cells is updated based on the tumor cell count at time $t-d$. For instance, if $d=2$, it means that at the current time t, we only have information on the tumor cell count from two days ago $(t-2)$. This consideration is particularly important when updating the numbers of normal cells $\mathscr{N}$ and immune cells $\mathscr{I}$.

Remark 1 In model (5), we considered only the delays between measuring tumor size and administering chemotherapy. However, it is important to note that these delays can be extended to include normal cells $\mathscr{N}$ and immune cells $\mathscr{I}$, depending on the user's preference. In this study, we focus on the fact that real-time measurement of tumor size and immediate adjustment of drug dosage are impractical in clinical practice. Tumor assessments (e.g., via imaging or lab tests) occur periodically rather than continuously, with inherent delays of two weeks for processing and adjustment. Therefore, we have chosen to model delays primarily in tumor size to reflect these realistic, non-continuous monitoring constraints.

By defining a new variable as follows:

$$
\eta(t) = \left[\mathscr{N}(t), \mathscr{T}(t), \mathscr{I}(t), \mathscr{M}(t)\right],
$$

the model (5) can be rewritten as follows:

$$
\begin{aligned}
\dot{\eta}(t) &= f(\eta(t)) + f_d(\eta(t-d)) + Bu(t),\\
\eta(t) &= h(t), \quad t \in [-d, 0],
\end{aligned}
\tag{6}
$$

where $B = [0,0,0,1]^T$. The term $h(t)$ represents the history of the number of tumor cells. The functions $f(\eta(t)) : \mathbb{R}^n \to \mathbb{R}^n$ and $f(\eta(t-d)) : \mathbb{R}^n \to \mathbb{R}^n$ are known to be locally Lipschitz. It should be mentioned that the history of the number of normal and immune cells can be considered in $f(\eta(t-d))$. Considering the delay in other cell populations depends on the user or the injector of the drug. However, in this paper, we aim to consider only the effects of the history of tumor cells.

Control objective

The objective of chemotherapy is to guide the system into a region where either the tumor-free equilibrium is achieved or an equilibrium with minimal tumor presence is maintained. This study targets the tumor-free equilibrium, developing a closed-loop controller aimed at completely eliminating the tumor. Our approach centers on designing a controller using the adaptive dynamic programming algorithm for system (6), ensuring stability even with time delays in tumor cell population measurements. Also, the control protocols are tailored

to account for time delays, optimizing drug dosage for cancer treatment. In essence, the designed control law will integrate the time delay history of tumor cell numbers. We present a novel value function that incorporates these time delays, utilizing policy iteration to solve the Hamilton-Jacobi-Bellman (HJB) equation[31] with adaptive dynamic programming[32], approximated by a critic neural network[33]. Following the proposed control formulation, the definitions of the parameters provided in Table 2 help improve the overall understanding.

The goal is to derive an optimal feedback control strategy, $u(t)$, that effectively minimizes the infinite-horizon performance index linked to system (6). This cost function is formulated as:

$$V(\eta(t)) = \int_t^{\infty} \left(E(\eta(\tau), u(\tau)) + \int_{\tau-d}^{\tau} \eta^T(q) Q_d \eta(q) dq \right) d\tau, \tag{7}$$

here, the utility function $E(\eta(\tau), u(\tau))$ is given by $E(\eta(\tau), u(\tau)) = \eta^T(\tau) Q \eta(\tau) + u^T(\tau) R u(\tau)$.

Consider $U(\eta(\tau), u(\tau))$, defined as:

$$U(\eta(\tau), u(\tau)) = E(\eta(\tau), u(\tau)) + \int_{\tau-d}^{\tau} \eta^T(q) Q_d \eta(q) dq.$$

This function satisfies $U(0,0) = 0$ and is non-negative for all $\eta(t)$ and $u(t)$. Here, $Q, Q_d \in \mathbb{R}^{n \times n}$ are positive semi-definite weighting matrices, and $R \in \mathbb{R}^{m \times m}$ is a positive definite weighting matrix on the control input that penalizes the control effort (drug infusion rate) in the cost function. The expression (7) incorporates time delays. In the proposed control formulation, the weighting matrices Q, Q_d, and R can be viewed as clinical preference indicators. The matrices Q and Q_d penalize tumor growth and deviation from the desired therapeutic trajectory, while R penalizes excessive drug dosing. Thus, increasing R represents a stronger emphasis on toxicity management, whereas increasing Q or Q_d prioritizes tumor suppression. These parameters can be adjusted based on patient-specific characteristics–such as disease aggressiveness, drug tolerance, or comorbidities–allowing oncologists to align the control design with individualized treatment goals and established clinical protocols.

In what follows, we will show that employing the newly defined value function (7) for controller design ensures the stabilization of the closed-loop system, ultimately resulting in the complete elimination of tumors. Denote $V^*(\eta(t))$ as the optimal value function associated with $V(\eta(t))$, which is formally expressed as:

$$V^*(\eta(t)) = \min_{u(t) \in \varphi} V(\eta(t)). \tag{8}$$

The gradient associated with the optimal value function $V^*(\eta(t))$ is available thorough the Bellman optimality concept. This gradient, represented as $\nabla V^*(\eta(t)) = \frac{tial V^*(\eta(t))}{tial \eta}$, is governed by the following equation:

$$\min_{u(t) \in \varphi} H(\eta(t), \nabla V^*(\eta(t))) = 0, \tag{9}$$

where $H(\eta(t), \nabla V^*(\eta(t)))$ is known as the Hamiltonian function.

For ease of presentation, the variable t is excluded from the following equations.

The Hamiltonian function related to the cost function (7) is expressed as[34]:

$$\begin{aligned} &H(\eta, \nabla V^*(\eta), u) = U(\eta, u) + (\nabla V^*(\eta))^T \dot{\eta} = U(\eta, u) + (\nabla V^*(\eta))^T (f(\eta) + f_d(\eta)) \\ &+ (\nabla V^*(\eta))^T B u = E(\eta, u) + \int_{\tau-d}^{\tau} \eta^T(q) Q_d \eta(q) dq + (\nabla V^*(\eta))^T (f(\eta) + f_d(\eta)) + (\nabla V^*(\eta))^T B u. \end{aligned} \tag{10}$$

Our aim is to achieve minimization of the expression defined in Eq. (10) by incorporating Eqs. (10) and (9) to formulate the optimal control strategy, denoted as u^*. The optimal control input is derived by solving the condition $\frac{tial H(\eta, \nabla V^*(\eta), u)}{tial u} = 0$, as demonstrated below:

$$u^* = -\frac{1}{2} R^{-1} B^T \nabla V^*(\eta). \tag{11}$$

Applying a straightforward transformation to Eq. (11) results in:

Symbol	Description	Remarks / selection criteria
L_r	Learning rate matrix	Determines the adaptation speed of the critic/actor parameters. Typically chosen as a small positive-definite diagonal matrix.
R	Control weighting matrix	Penalizes the control effort in the cost function. A higher value reduces control aggressiveness.
Q	State weighting matrix	Balances state tracking performance versus control effort; usually positive-definite.
Q_d^c	Delay compensation matrix	Compensates for the effect of input/state delay; tuned to ensure BMI feasibility.
$\widehat{W}_c$	Estimated critic weights	Updated online using the adaptive law to approximate the value function.

Table 2. Design parameters used in the proposed control scheme.

$$(\nabla V^*(\eta))^T B = -2u^{*T} R. \tag{12}$$

Equation (12) will be needed later. By inserting Eq. (11) into Eq. (10), the HJB equation can be reformulated as:

$$\eta^T Q\eta - \frac{1}{4}(\nabla V^*(\eta))^T BR^{-1}B^T \nabla V^*(\eta) + \int_{\tau-d}^{\tau} \eta^T(q)Q_d\eta(q)dq + (\nabla V^*(\eta))^T(f(\eta)+f_d(\eta)) = 0. \tag{13}$$

It is important to demonstrate that the control protocol achieved from Eq. (11) can effectively stabilize the system expressed in Eq. (6). This assertion is confirmed by the following theorem.

Remark 2 In this study, the drug infusion rate is treated as a continuous control input for analysis and simulation purposes, representing an idealized continuous-infusion scenario. In clinical practice, however, chemotherapy is typically administered at discrete intervals (e.g., daily or per treatment cycle). The proposed control strategy can be readily implemented in a sampled form, where the computed control input is applied as a piecewise-constant dose between consecutive dosing instants. Since the underlying tumor dynamics evolve on a slower timescale, such discretization is not expected to significantly alter the system performance.

Theorem 1 *Consider the system described by Eq. (6) and the control protocol governed by Eq. (11). The control strategy in Eq. (11) ensures that the closed-loop nonlinear time-delay system (5) achieves uniform ultimate boundedness, given that there exist positive definite matrices Q and Q_d, as well as free-weighting matrices $\bar{M}$ and $\bar{N}$, which satisfy the following bilinear matrix inequality condition:*

$$\bar{\Upsilon} = \begin{bmatrix} R & 0 & 0 & 0 \\ * & Q-\bar{M}_d & -\bar{M}\bar{N}_d & -\bar{M} \\ * & * & -\bar{N}_d & -\bar{N} \\ * & * & * & 0 \end{bmatrix} \geq 0, \tag{14}$$

in which

$$\begin{aligned} \bar{M}_d &= d\bar{M}Q_d^{-1}\bar{M}^T, \\ \bar{M}\bar{N}_d &= d\bar{M}Q_d^{-1}\bar{N}^T, \\ \bar{N}_d &= d\bar{N}Q_d^{-1}\bar{N}^T. \end{aligned}$$

Proof We consider the following Lyapunov function:

$$L(\eta) = V^*(\eta). \tag{15}$$

Based on the definition of $V^*(\eta)$, it follows that $V^*(\eta) > 0$ for $z \neq 0$ and $V^*(\eta) = 0$ when $\eta = 0$. This confirms that $V^*(\eta)$ is a positive definite function, which further implies that $L(\eta)$ also possesses positive definiteness. Additionally, by evaluating the time derivative of the Lyapunov function (15) along the system trajectory $\dot{\eta} = f(\eta) + f_d(\eta) + Bu$, we obtain the following expression:

$$\dot{L}(\eta) = (\nabla V^*(\eta))^T\dot{\eta} = (\nabla V^*(\eta))^T(f(\eta)+f_d(\eta)+Bu). \tag{16}$$

Using (10), we obtain:

$$(\nabla V^*(\eta))^T(f(\eta)+f_d(\eta)) = -E(\eta,u) - \int_{t-d}^{t} \eta^T(q)Q_d\eta(q)dq - (\nabla V^*(\eta))^T Bu. \tag{17}$$

By substituting (17) and (11) into (16), Eq. (16) can be reformulated as follows:

$$\dot{L}(\eta) = -\eta^T Q\eta - u^{*T} Ru^* - \int_{t-d}^{t} \eta^T(q)Q_d\eta(q)dq. \tag{18}$$

Inspired by[35], we used the free-weighting matrices technique. This method allows us to bound the integral and convert it into a form where Lyapunov or stability conditions can be applied. Additionally, as mentioned in Proposition 3.11 of[36], we used Jensen's inequality, which plays a crucial role in handling the constant delay d in the stability analysis of the delay system. To bound the effects of the delay, free-weighting matrices[35] are incorporated. Then, by defining the free-weighting matrices $\bar{M}$ and $\bar{N}$, the term $-\int_{t-d}^{t} \eta^T(q)Q_d\eta(q)\,dq$ in (18) can be expressed as follows:

$$
\begin{aligned}
&-\int_{t-d}^{t}\eta^T(q)Q_d\eta(q)dq \le 2\eta^T\bar{M}\int_{t-d}^{t}\eta^T(q)dq + 2\eta^T(t-d)\bar{N}\int_{t-d}^{t}\eta^T(q)dq - \int_{t-d}^{t}\Big(\eta^T(q)\bar{M}\\
&+\eta^T(t-d)\bar{N}+\eta^T(q)Q_d\Big)Q_d^{-1}\Big(\eta^T(q)\bar{M}+\eta^T(t-d)\bar{N}+\eta^T(q)Q_d\Big)^T dq + \int_{t-d}^{t}\Big(\eta^T(q)\bar{M}+\eta^T(t-d)\bar{N}\Big) \qquad (19)\\
&Q_d^{-1}\Big(\eta^T(q)\bar{M}+\eta^T(t-d)\bar{N}\Big)^T dq \le \eta^T\Upsilon\eta,
\end{aligned}
$$

in which

$$
\eta^T = \left[\eta^T,\ \eta^T(t-d),\ \int_{t-d}^{t}\eta^T(q)dq\right],
$$
$$
\Upsilon = \begin{bmatrix} \bar{M}_d & \bar{M}\bar{N}_d & \bar{M} \\ * & \bar{N}_d & \bar{N} \\ * & * & 0 \end{bmatrix}.
$$

Referring to (18) and applying (19), we get:

$$
\dot{L}(\eta) = -\eta^T Q\eta - u^{*^T}Ru^* - \int_{t-d}^{t}\eta^T(q)Q_d\eta(q)dq \le -\eta^T Q\eta - u^{*^T}Ru^* + \eta^T\Upsilon\eta \le -(\bar{\eta}^T\bar{\Upsilon}\bar{\eta}), \qquad (20)
$$

where

$$
\bar{\eta}^T = \left[u^T,\ \eta^T,\ \eta^T(t-d),\ \int_{t-d}^{t}\eta^T(q)dq\right],
$$
$$
\bar{\Upsilon} = \begin{bmatrix} R & 0 & 0 & 0 \\ * & Q-\bar{M}_d & -\bar{M}\bar{N}_d & -\bar{M} \\ * & * & -\bar{N}_d & -\bar{N} \\ * & * & * & 0 \end{bmatrix}.
$$

According to (20), if $\bar{\Upsilon} \ge 0$, then $-(\bar{\eta}^T\bar{\Upsilon}\bar{\eta}) < 0$ holds, leading to $\dot{L}(\eta) < 0$. Thus, it completes the proof. □
As we discussed above, delays between measuring and applying chemotherapy are inevitable. We proposed an approach to account for these delays in treatment and drug dosing. To achieve this, we use the Lyapunov-Krasovskii function, $\int_{t-d}^{t}\eta^T(q)Q_d\eta(q)\,dq$, to analyze the stability of the optimal chemotherapy in cancer treatment with time-delay and incorporate these delays into the value function (7). The integration of this temporal element into our cost function enables us to model the effects of delayed responses inherent in cancer treatment protocols. This time-aware approach enhances our capacity to dynamically refine the control mechanism. Our analysis, as detailed in Theorem 1, demonstrates that incorporating a time-delay-sensitive Lyapunov-Krasovskii functional within the cost function contributes significantly to the stability and robustness of our proposed methodology. Leveraging the principles established in Theorem 1, we navigate the intricacies of the HJB equations to derive an effective control protocol. This process culminates in the development of a sophisticated drug administration system that not only accounts for but also adapts to these inherent temporal lags in treatment response.

It should be noted that in the presence of time delays, full state observability is lost, violating the Markov property and complicating reward assignment in reinforcement learning. To address this, we integrate a Lyapunov-Krasovskii functional into the value function, enabling delayed-state awareness. A critic-only neural network approximates the HJB equation, while stability is ensured via a bilinear matrix inequality (BMI) condition. Online weight updates preserve learnability, making the method suitable for real-time, delay-affected cancer treatment control.

Remark 3 Our proposed method addresses the challenges of time delays in a cancer treatment model, which disrupt the Markov property assumed in reinforcement learning (RL), where future states depend only on current states and actions. The delays in tumor size measurements (5) introduce historical state dependencies, making the system non-Markovian. We tackle this using a Lyapunov-Krasovskii functional in the value function (7) and a critic-only neural network to approximate the HJB equation, with a BMI (14) ensuring stability.

Remark 4 Our approach develops a cancer treatment model using a nonlinear cancer dynamics framework and a critic-only neural network to solve the HJB equation for optimal chemotherapy dosing, considering delays in tumor size measurements. It employs a BMI to ensure system stability under constant delays. In contrast, the method in[37] offers a general online actor-critic algorithm for nonlinear systems with state delays, using both actor and critic networks to approximate the HJB equation and control policy, relying on Lyapunov techniques for stability without using a BMI. However, our method is cancer-specific with a simpler critic-only design and BMI-based stability, while[37] provides a broader dual-network approach.

To implement the optimal control strategy, we employ computational methods to approximate solutions to the HJB equation for our time-delayed system. The high-dimensional and nonlinear dynamics, compounded by time delays, pose significant computational challenges. We address these by developing a streamlined neural

network architecture, which efficiently approximates the HJB solution for chemotherapy dosing optimization, building on established dynamic programming techniques[38].

Neural network implementation

It is well known that neural networks excel at approximating complex functions. Since the performance index function is generally complex and does not possess a straightforward analytical expression, we leverage a neural network to approximate its structure. In this work, a simple single-layer neural network is adopted as an effective means of capturing and estimating the underlying functional relationship. The function $V(\eta)$ is represented as:

$$V(\eta) = W_c^T S(\eta) + \varepsilon(\eta), \tag{21}$$

here, we interpret $S(\eta)$ as the neural activation map, with R^c representing the c-dimensional Euclidean space. W_c corresponds to the optimized parameter set, while c quantifies the neural units in the intermediate stratum. $\varepsilon(\eta)$ denotes the neural network's approximation discrepancy. The spatial derivative of Eq. (21) with respect to z can be articulated as:

$$\nabla V(\eta) = (\nabla S(\eta))^T W_c + \nabla \varepsilon(\eta), \tag{22}$$

where $\nabla S(\eta) = \frac{tial S(\eta)}{tial \eta} \in R^{c \times n}$ represents the gradient of the activation map, and $\nabla \varepsilon(\eta)$ denotes the gradient of the approximation error. Incorporating Eq. (22) into (9) results in:

$$\min_{u(t) \in \varphi} U(\eta, u) + ((\nabla S(\eta))^T W_c + \nabla \varepsilon(\eta))\dot{\eta} = 0. \tag{23}$$

Consequently, we can formulate the Hamiltonian as:

$$H(\eta, u, W_c) = U(\eta, u) + (W_c^T \nabla S(\eta))\dot{\eta} = -\nabla \varepsilon(\eta)\dot{\eta} \triangleq e_{rH}. \tag{24}$$

In this framework, e_{rH} encapsulates the residual discrepancy emanating from the neural approximation. Given that the ideal parameter set W_c remains undetermined, we utilize a critic neural architecture to estimate $V(\eta)$ as follows:

$$\hat{V}(\eta) = W_c^T S(\eta). \tag{25}$$

Consequently, the gradient of the approximated value function $\hat{V}(\eta)$ can be expressed as:

$$\nabla \hat{V}(\eta) = (\nabla S(\eta))^T \hat{W}_c, \tag{26}$$

Thus, the approximate Hamiltonian can be formulated as:

$$H(\eta, u, \hat{W}_c) = U(\eta, u) + (W_c^T \nabla S(\eta))\dot{\eta} \triangleq e_r. \tag{27}$$

The weight approximation error is defined as $\tilde{W}_c = W_c - \hat{W}_c$. By incorporating Eqs. (27) and (24), we derive:

$$e_r = e_{rH} - \tilde{W}_c^T \nabla S(\eta)\dot{\eta}. \tag{28}$$

The weight approximation error can be reformulated as:

$$\dot{\tilde{W}}_c = -\dot{\hat{W}}_c = L_r \left(e_{rH} - \tilde{W}_c^T \nabla S(\eta)\dot{\eta} \right) \nabla S(\eta)\dot{\eta}. \tag{29}$$

To optimize the parameter set $\hat{W}_c$ of the critic neural architecture, we minimize the cost function $E_c = \frac{1}{2} e_r^T e_r$ using a normalized gradient descent technique. The iterative refinement of $\hat{W}_c$ is governed by the following update rule:

$$\dot{\hat{W}}_c = -L_r e_r \nabla S(\eta)\dot{\eta}, \tag{30}$$

where $L_r > 0$ is the learning rate, controlling the speed of weight adjustments. This update occurs online, making the weights $\hat{W}_c$ time-dependent as they adapt to the system's dynamics and time delays during treatment. Consequently, by taking into account Eqs. (11) and (21), the optimal control policy can be formulated as:

$$u(\eta) = -\frac{1}{2} R^{-1} B^T \left((\nabla S(\eta))^T W_c + \nabla \varepsilon(\eta) \right), \tag{31}$$

and it can be estimated as:

$$\hat{u}(\eta) = -\frac{1}{2} R^{-1} B^T (\nabla S(\eta))^T \hat{W}_c. \tag{32}$$

The approximate control policy in Eq. (32) depends entirely on the critic neural network. By adjusting the weight vector of the critic neural network using Eq. (30), the necessity of training the action neural network is eliminated. This simplifies the overall process, making the method both practical and computationally efficient for implementation.

Theorem 2 *When the critic neural network weights are updated as per Eq. (29) for the chemotherapy treatment dynamics, as redefined in system (6), the approximation error in the weights remains uniformly ultimately bounded.*

Proof At the first step, we considered the following Lyapunov function:

$$\Gamma_2 = \frac{1}{2L_r}\tilde{W}_c^T\tilde{W}_c. \tag{33}$$

The time derivative of (33) is

$$\dot{\Gamma}_2 = \frac{1}{2L_r}\tilde{W}_c^T\dot{\tilde{W}}_c = \tilde{W}_c^T(e_{rH} - \tilde{W}_c^T\nabla S(\eta)\dot{\eta})\nabla S(\eta)\dot{\eta} = \tilde{W}_c^T e_{rH}\nabla S(\eta)\dot{\eta} - \left\|\tilde{W}_c^T\nabla S(\eta)\dot{\eta}\right\|^2 \leq \frac{1}{2}e_{rH}^2 - \frac{1}{2}\left\|\tilde{W}_c^T\nabla S(\eta)\dot{\eta}\right\|^2. \tag{34}$$

Consequently, the condition $\dot{\Gamma}_2 < 0$ is satisfied when $\tilde{W}_c$ falls within the compact domain characterized by $\|\tilde{W}_c\| \leq \|\frac{e_{rH}}{\theta_1}\|$, given the assumption $\|\nabla S(\eta)\dot{\eta}\| \leq \theta_1$, where θ_1 represents a positive scalar. Applying the principles of Lyapunov stability theory, we can deduce that the parameter estimation error exhibits uniform ultimate boundedness, thereby concluding the proof. □

Remark 5 The critic-only structure simplifies the computational framework by focusing solely on updating the critic weights to approximate the optimal value function, avoiding the dual complexity of simultaneously updating both the critic and the actor (policy) components. This reduction in computational burden is critical for real-time clinical applications, where rapid processing of delayed tumor size data is essential, and resources may be constrained. The critic-only method requires fewer parameters to tune and fewer iterative updates, leading to lower memory usage and faster convergence compared to the actor-critic approach.

The adaptive dynamic programming algorithm for online critic learning, which is designed to optimize drug administration in cancer therapy and is related to the method provided in this paper, is outlined in Algorithm 1.

Require: Design parameters L_r, R, Q, Q_d, c, and initial critic weights $\hat{W}_c$.

1: Initialize the system state $\eta(0)$ and control input $u(0)$.

2: Set iteration index $j = 0$ and define a small convergence threshold $\varepsilon > 0$.

3: **while** not converged **do**

4: **Policy Evaluation (Critic Update):**

5: Using the current control policy $u^{(j)}(\eta)$, solve for the updated value function $V^{(j+1)}(\eta)$ that satisfies the following Hamilton–Jacobi–Bellman (HJB) equation:

$$\begin{aligned} 0 =& E(\eta(t), u^{(j)}(t)) + \int_{\tau-d}^{\tau} \eta^\top(q) Q_d \eta(q)\, dq \\ &+ \left(\nabla V^{(j+1)}(\eta)\right)^\top \left(f(\eta) + f_d(\eta - d) + Bu^{(j)}(\eta)\right) \end{aligned}$$

6: Estimate or update critic weights $\hat{W}_c^{(j+1)}$ such that $V^{(j+1)}(\eta) \approx \hat{W}_c^{(j+1)\top}\phi(\eta)$, where $\phi(\eta)$ is the chosen basis or feature vector.

7: **Policy Improvement (Actor Update):**

8: Derive an improved control policy based on the updated critic using:

$$u^{(j+1)}(\eta) = -\frac{1}{2}R^{-1}B^\top \nabla V^{(j+1)}(\eta)$$

This step ensures that the control policy is updated toward minimizing the cost-to-go estimated by the critic.

9: **Convergence Check:**

10: If $\left\|V^{(j+1)}(\eta) - V^{(j)}(\eta)\right\| \leq \varepsilon$ and $\left\|u^{(j+1)}(\eta) - u^{(j)}(\eta)\right\| \leq \varepsilon$, stop and obtain the approximate optimal policy $u^*(\eta)$. Otherwise, set $j \leftarrow j + 1$ and continue iterations.

11: **end while**

12: **Output:** Approximate optimal value function $V^*(\eta)$ and optimal policy $u^*(\eta)$, ensuring convergence under standard ADP stability assumptions.

Algorithm 1. Adaptive dynamic programming algorithm for optimizing drug administration in cancer therapy.

Simulation numerical examples

In this section, we will conduct several simulations and analyze their results. All parameters used in the simulations are listed in Table 1.

Interested readers can access the computational scripts utilized in our simulations via this digital repository.

As previously stated, we assume the system's equilibrium point is located at the origin of the state space $\mathbb{R}^n$. By shifting the tumor-free equilibrium point T^* to the origin, we can rewrite Eq. (5) accordingly. we define the following new variables:

$$\begin{aligned} x_1(t) &= N(t) - \frac{1}{b_2}, \quad x_2(t) = T(t), \\ x_3(t) &= I(t) - \frac{s}{d_1}, \quad x_4(t) = M(t). \end{aligned} \tag{35}$$

Using these definitions, we can transform Eq. (5) into Eq. (36) as follows:

$$\begin{aligned} \dot{x}_1(t) &= -r_2 x_1(t)(1 + b_2 x_1(t)) - \left(\frac{c_4}{b_2} x_2(t-d) - \frac{a_3}{b_2}\right) x_2(t-d) - c_4 x_1(t) x_2(t-d) - a_3 x_1(t) x_4(t), \\ \dot{x}_2(t) &= r_1 x_2(t-d)(1 - b_1 x_2(t-d)) - \left(\frac{s c_2}{d_1} + \frac{c_3}{b_2}\right) x_2(t-d) - c_3 x_1(t) x_2(t-d) - c_2 x_2(t-d) \\ &\quad \times x_3(t) - a_2 x_2(t-d) x_4(t), \\ \dot{x}_3(t) &= -\frac{c_1 s}{d_1} x_2(t-d) - d_1 x_3(t) - \left(\frac{a_1 s}{d_1} x_4(t) + \frac{\rho s}{d_1(\alpha + x_2(t-d))} x_2(t-d)\right) + \frac{\rho}{\alpha + x_2(t-d)} \\ &\quad \times x_2(t-d) x_3(t) - c_1 x_2(t) x_3(t) - a_1 x_3(t) x_4(t), \\ \dot{x}_4(t) &= -d_2 x_4(t) + u(t). \end{aligned} \tag{36}$$

It is assumed that there is a $d = 1$ week delay between measuring the tumor population and administering the drug. We incorporated 1–2 week delays in our simulations to reflect robustness across short biomarker-based and long imaging-based monitoring, informed by discussions with oncology specialists at Emam Reza Hospital in Tabriz and their anonymized clinical records. Clinical studies, such as those on nadir neutrophil counts in breast cancer, AML consolidation therapy, and ctDNA-guided switching with weeks turnarounds[39], justify these delays. Our delay-aware online RL approach enhances dosing decisions in clinical settings with common two-week delays, outperforming offline or less frequently updated strategies. To verify the BMI condition presented in Theorem 1 for systems with delay, we check the feasibility of the proposed BMI (14). If the BMI is feasible, it indicates that convergence is guaranteed and the system can tolerate the corresponding delay. For one of the considered delay cases (1 week delay), the BMI parameters have been obtained as follows:

$$Q = \begin{bmatrix} 2.3 & 0.1 & 0.4 & 0.0 \\ 0.1 & 1.8 & 0.2 & 0.0 \\ 0.4 & 0.2 & 2.1 & 0.0 \\ 0.0 & 0.0 & 0.0 & 1.5 \end{bmatrix}, \quad Q_d = \mathrm{diag}(0.8,\ 0.9,\ 0.7,\ 1.2), \quad R = 0.4.$$

Then, the obtained BMI is feasible. Following Algorithm 1, model (36), (30) and control policy (32), the simulation results are obtained. The initial conditions for the simulations are as follows: $x_1(0) = -0.5$, $x_2(0) = 0.5$, $x_3(0) = -1.15$, $x_4(0) = 0$, $Q = \mathrm{diag}([q_{\mathscr{N}} = 1, q_{\mathscr{T}} = 12, q_{\mathscr{I}} = 1, q_{\mathscr{M}} = 0.02])$, $Q_d = 8Q$, $L_r = 0.02$. The number of neurons is chosen as $c = 10$. The activation function and initial weights of critic learning :

$$\begin{aligned} S(\eta) &= [\mathscr{N}^2; \mathscr{T}^2; \mathscr{I}^2; \mathscr{M}^2; \mathscr{N}\mathscr{T}; \mathscr{N}\mathscr{I}; \mathscr{N}\mathscr{M}; \mathscr{T}\mathscr{I}; \mathscr{T}\mathscr{M}; \mathscr{I}\mathscr{M}] \\ W_c(0) &= [3; 0.2; 2.2; 2.8; 5.4; 1; 3.4; 4.7; 4.5; 1.3]. \end{aligned} \tag{37}$$

The simulation results for this case are shown in Figs. 2 and 3. Figure 2 provides information about the cell populations in a patient. This figure indicates that the number of tumor cells converged to zero after

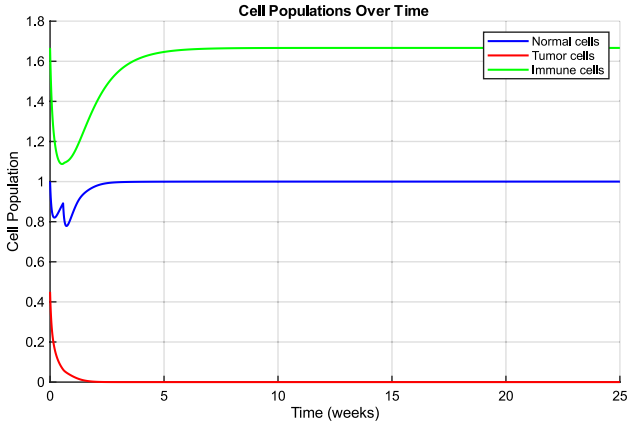

Fig. 2. Assessment of cell populations with control policy (32).

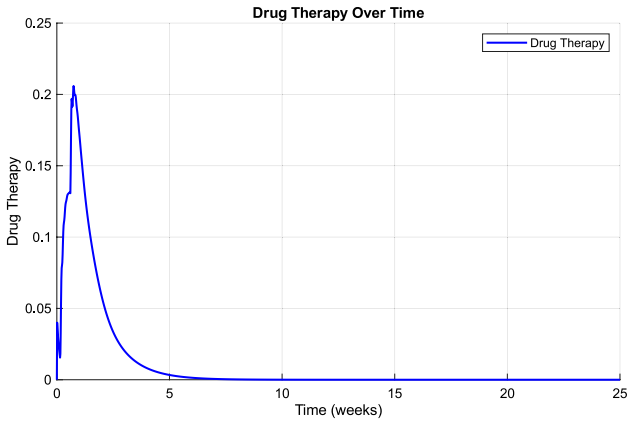

Fig. 3. The drug concentration.

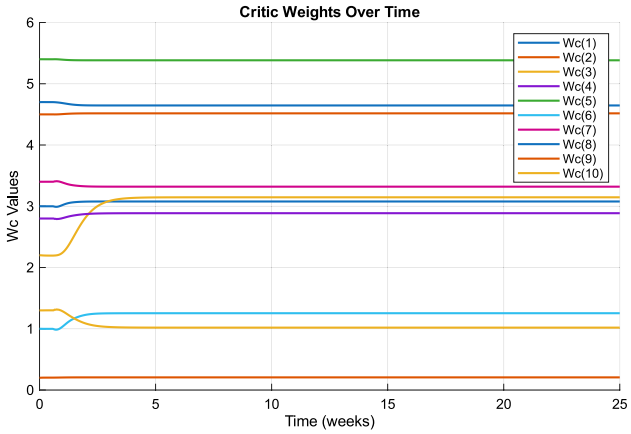

Fig. 4. The weights of critic learning structure.

approximately three weeks days. Additionally, the number of immune and normal cells reached the equilibrium point, signifying that the patient's treatment was successful under the control policy (32).

Figure 3 illustrates an intense initial drug regimen, with the concentration ($\mathscr{M}$) peaking at approximately 0.20 within the first two weeks to aggressively target the tumor burden ($\mathscr{T}(0) = 0.5$). This high initial dose is driven by the optimal control policy (Eq. (32)), which leverages the delayed tumor measurement ($x_2(t-1)$) and the initial critic weights ($W_c(0)$). A rapid decline follows the peak, reflecting the optimization's adjustment due to the 1-week delay in tumor measurement, as the online critic learning updates the weights to adapt to the decreasing tumor population (Fig. 2 shows $\mathscr{T}$ nearing zero by three weeks). The delay compensation ensures the dose is reduced to prevent overdosing as the tumor diminishes. By around five weeks, the critic weights converge (Fig. 4), and the drug dosage decreases to zero, indicating that the patient has regained health, with normal and immune cell populations stabilizing at their equilibrium points.

Remark 6 The values of $Q = \mathrm{diag}([q\mathscr{N}, q\mathscr{T}, q\mathscr{I}, q\mathscr{M}])$ have different implications in cancer chemotherapy depending on the patient. Younger patients typically have a higher growth capacity for normal and immune cells compared to older patients. Therefore, for younger patients, it is more important to reduce the number of cancerous cells than to preserve normal or immune cells. Consequently, an oncologist might assign a high value to $q\mathscr{T}$ and lower values to the other parameters. For pregnant patients, the oncologist might select higher values for $q\mathscr{N}$, $q\mathscr{I}$, and R until childbirth.

The simulation results for the case 2 where $Q = \mathrm{diag}([q\mathscr{N} = 14, q\mathscr{T} = 0, q\mathscr{I} = 12, q\mathscr{M} = 10])$ and $R = 20$ are displayed in Figs. 5, 6 and 7. Compared to Case 1 ($Q = [1, 10, 1, 0.01]$, $R = 0.4$), which prioritizes tumor reduction ($q\mathscr{T} = 10$) for a typical patient, Case 2 focuses on preserving normal and immune cells ($q\mathscr{N}, q\mathscr{I} = 10$) and minimizing drug concentration ($q\mathscr{M} = 10$) with a high control penalty ($R = 20$), without directly penalizing tumor cells ($q\mathscr{T} = 0$). This setup reflects a conservative treatment scenario, such as for a pregnant patient (Remark 6), where minimizing drug exposure and protecting healthy cells are critical. These settings were chosen to demonstrate the proposed method's adaptability to diverse clinical needs, validating the effectiveness across both aggressive and conservative treatment strategies.

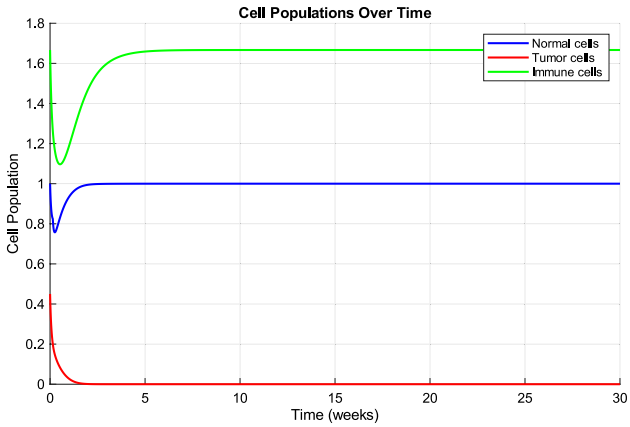

Fig. 5. Assessment of cell populations with control policy (32) for the case 2.

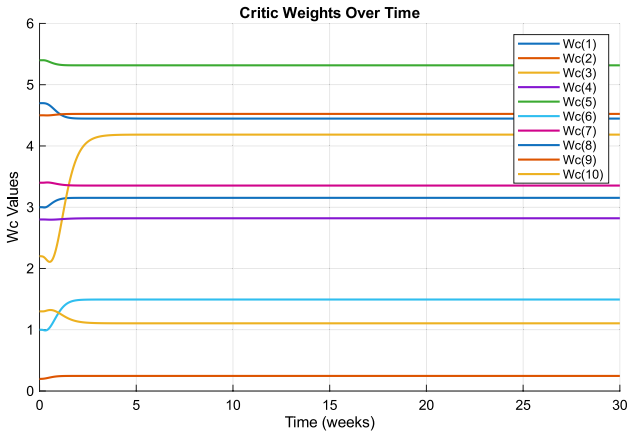

Fig. 6. The weights of critic learning structure for the case 2.

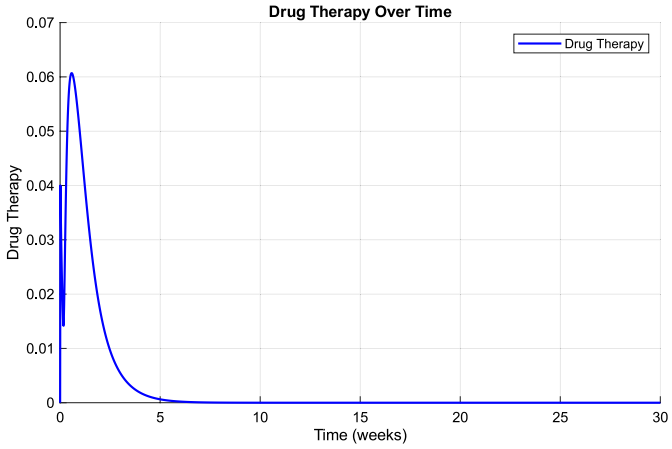

Fig. 7. The drug concentration for the case 2.

In the next step, we demonstrate that our proposed method–by incorporating an integral term into the Hamiltonian error–effectively compensates for time delays, leading to similar cell population trajectories and treatment outcomes for both 1-week and 2-week delays. This compensation significantly reduces the impact of delays, as illustrated in Figs. 8 and 9.

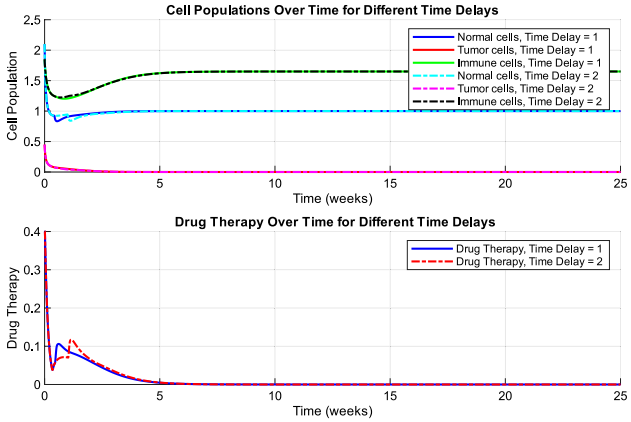

Fig. 8. Assessment of cell populations with control policy (32) for different delays.

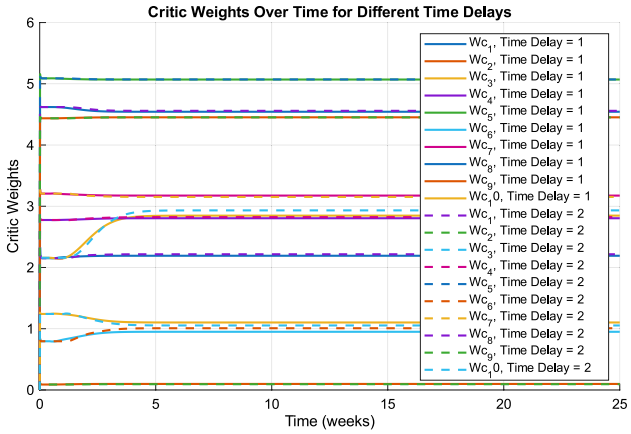

Fig. 9. The weights of critic learning structure for different delays.

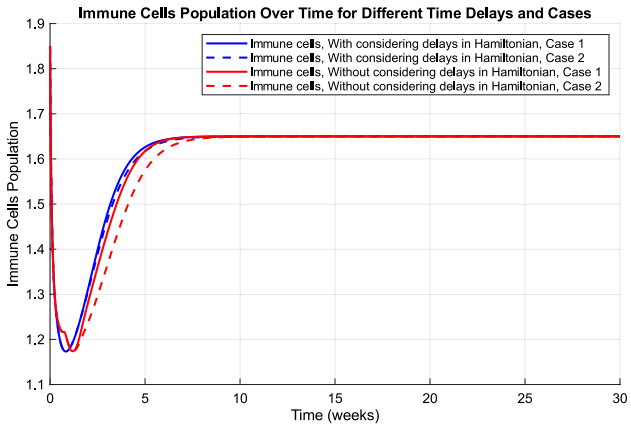

Fig. 10. Assessment of Immune cell population with control policy (32).

It is worth noting that with longer delays, drug dosage profiles may exhibit oscillations, as shown in the zoomed area of Fig. 8, indicating the influence of delay. Nonetheless, Figs. 8 and 9 confirm that acceptable treatment performance is maintained, validating the method's robustness against time delays in the therapy process.

In the next scenario, we illustrate the benefits of incorporating delay compensation into the control law. We use the same parameters as in Case 1 ($Q = [1, 10, 1, 0.01]$, $R = 0.4$) as a baseline. Figures 10 and 11 compare

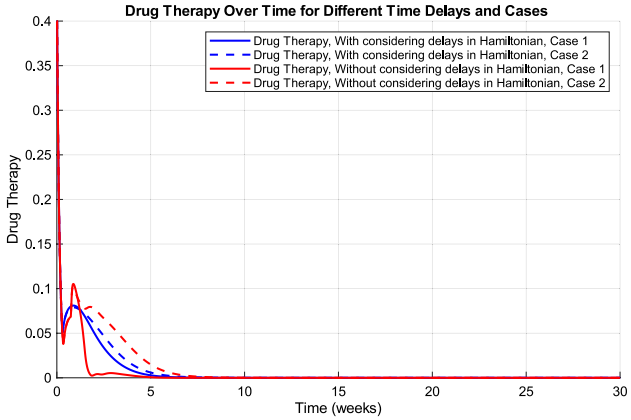

Fig. 11. The drug concentration.

Delay Duration	Drug Safety	Convergence
$d = 1$ week	Safe	✓ Stable
$d = 2$ weeks	Safe	✓ Stable
$d = 5$ weeks	Borderline	✓ Marginal
$d = 6$ weeks	Unreal	○ unstable

Table 3. Extended delay simulations.

the control policy from Eq. (32) with and without delay compensation–shown by solid (red and blue) and dotted (red and blue) curves, respectively. In Fig. 10, the dotted blue curve, which represents the case without delay compensation, exhibits slower convergence of the immune cell population compared to the corresponding solid blue curve. Additionally, the figure demonstrates that larger delays lead to slower convergence, as seen by comparing the red curve (two-weeks delay) with the blue curve (one-week delay). Figure 11 further illustrates that delay compensation contributes to a shorter drug administration period. The solid red and blue curves show the dose converging to zero more quickly, indicating improved dosing efficiency. These results underscore the advantage of accounting for delays in control design to enhance both treatment effectiveness and safety.

It should be noted that Case 3 represents the same control objective as Case 1 but with an added delay-compensation mechanism. As shown in Fig. 11, the delivered drug amount is considerably smaller than in Fig. 3 because the delay-compensated controller anticipates the system's response and avoids excessive actuation. This results in smoother drug administration and reduced overshoot, demonstrating the effectiveness of the proposed compensation scheme.

Table 3 summarizes the system behavior under different delay durations. As for the other part of your comment, we have considered it as follows:

From Table 3, it can be observed that as the delay increases, the system performance gradually degrades, and excessive delays may lead to marginal or unsafe behavior.

Remark 7 The days delay assumed in our simulations is a theoretical approximation to model the time between tumor measurement and chemotherapy administration, reflecting the technical challenges of real-time monitoring. Current clinical practices regarding chemotherapy scheduling and the necessity of online adjustment are not yet integrated into this study, as it focuses on establishing a proof-of-concept framework. We are currently establishing collaborations with hospitals to gather real-world data and assess the practical relevance of our method, including its applicability to aggressive and non-aggressive cancers, in future work.

Remark 8 The critic-only learning approach reduces computational complexity by focusing solely on updating the critic weights W_c to optimize the value function $V(\eta)$, eliminating the need for simultaneous actor updates as required in dual actor-critic methods, thus avoiding the overhead of concurrent policy adjustments. However, a significant challenge arises in selecting appropriate initial conditions for the critic weights. Unsuitable initialization can lead to slow convergence or instability, particularly when delays are present, making it critical to carefully determine suitable starting values to ensure effective learning.

Comparative results In this section, a comparison Table 4 is presented to summarize the control methods for delayed systems in reinforcement learning and related approaches. Furthermore, the proposed method in the simulation results is compared with the approach in[25] and a model predictive control (MPC) scheme.

To further demonstrate practical relevance, we added a comparison between RL and MPC , showing improved adaptability of the proposed RL method in short-delay scenarios.

Method	System type	Architecture	Delay handling	Stability analysis	Application
The work[40]	Continuous nonlinear	Model-based RL	Particle filtering	Not provided	Autonomous driving
The work[41]	Continuous/Discrete	Actor-critic	Hindsight resampling	Lyapunov (nominal)	MuJoCo control
The work[25]	Continuous nonlinear	Fuzzy RL	Discontinuous reward	Empirical	Cancer therapy
The work[24]	Continuous nonlinear	Model-free RL (Lyapunov-integrated)	Fixed delay compensation	Lyapunov	Cancer chemotherapy
Our work	Continuous nonlinear	Critic-only	Delay-explicit value function (integral term, no state augmentation)	BMI-based delay-dependent	Cancer therapy

Table 4. Comparison of control methods for delayed systems in reinforcement learning and related approaches.

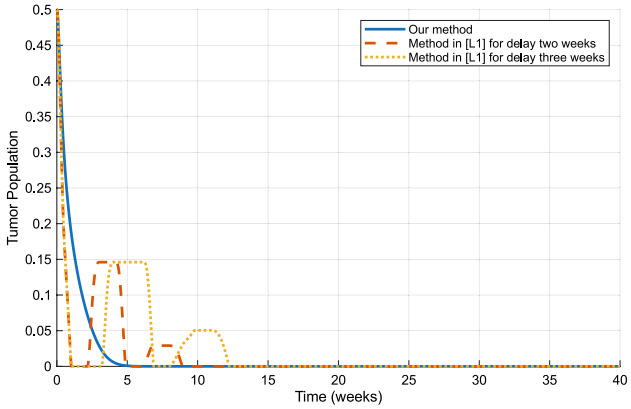

Fig. 12. Comparison of the performance of the proposed method and the method in [L1][25].

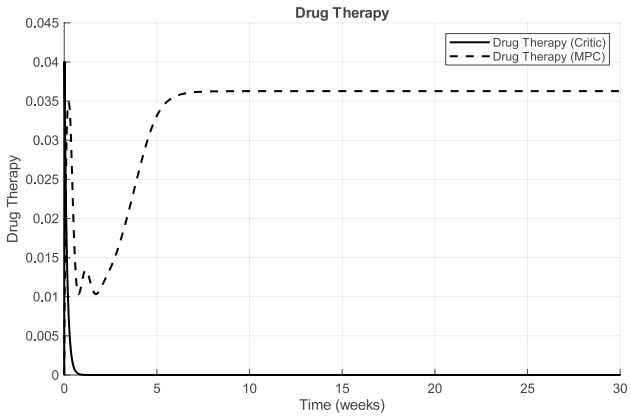

Fig. 13. Comparison of the performance of the proposed method and MPC approach.

From Fig. 12, it can be clearly observed that the proposed method exhibits a smooth and stable response without any noticeable oscillations or overshoots. This behavior indicates that the proposed control strategy effectively mitigates fluctuations and ensures a more consistent system performance. In contrast, the method presented in[25] shows significant oscillations and slower convergence, which demonstrates the superior transient and steady-state characteristics of our approach.

The MPC implemented here uses a prediction horizon of 5 steps (Hp = 5) to forecast future states and optimize drug dosing over a control horizon of 2 steps (Hu = 2), minimizing the quadratic cost function involving state deviations and control effort while respecting bounds on the input (u between 0 and 10). This setup provides a baseline for comparison with the critic-only learning method, demonstrating how MPC handles the nonlinear tumor dynamics without explicit delay compensation in its prediction, leading to potentially higher drug peaks.

From Fig. 13, it can be seen that the proposed method demonstrates superior performance in chemotherapy dosing by achieving complete convergence of drug concentration to zero after effectively eradicating the tumor,

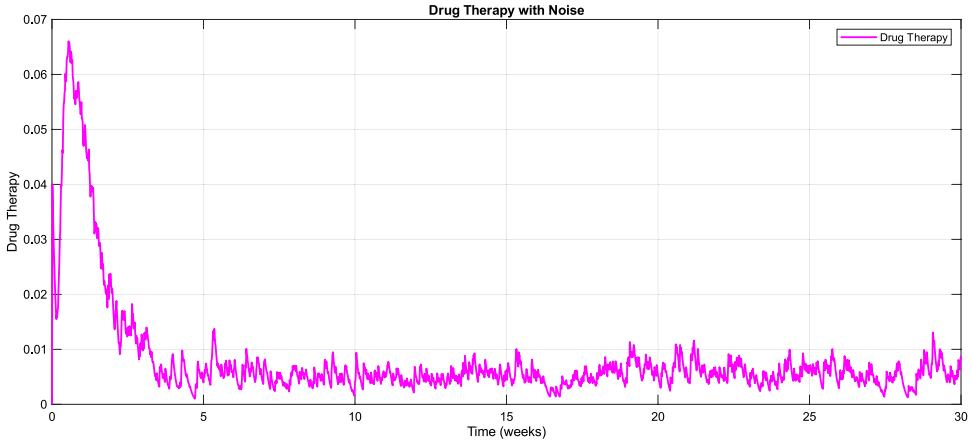

Fig. 14. Comparison of the performance of the proposed method and MPC approach.

ensuring minimal long-term toxicity for the patient. In contrast, the MPC approach, while maintaining stable cell populations, fails to fully taper off the drug dosage, resulting in persistent low-level administration that may increase cumulative toxicity risks over time.

For the noisy case, small fluctuations appear in the drug therapy input due to noise, but the signal remains bounded and at a low magnitude. The noise was modeled as

$$\tilde{\eta}(t) = \eta(t) + \nu(t), \quad \nu(t) \sim \mathcal{N}(0, \sigma^2), \tag{38}$$

with $\sigma = 0.01$, representing moderate sensor noise levels relative to the scale of the state variables. The noisy case was simulated for 30 weeks of therapy.

Figure 14 confirms that the algorithm is resilient to practical sensor noise conditions and validates its robustness in real-world applications. However, we need to prove this mathematically, which can be a great motivation for extending our proposed method. Software The numerical results discussed here were derived using MATLAB R2023a as the main computational tool, taking advantage of its built-in functions for optimization and solving differential equations. The simulations were implemented in MATLAB, with ODE solvers used to model the time-delay system defined by Eq. (36), and the YALMIP toolbox employed to express the stability criteria outlined in the Theorems. The code was written to carry out the online critic learning algorithm (Algorithm 1), iteratively solving HJB equation based on the initial conditions and parameters provided in Table 1. For the sake of transparency and reproducibility, the full source code and related documentation can be accessed via the Zenodo repository at [this link](https://doi.org/10.5281/zenodo.15088181).

Conclusion

This paper introduced an online critic learning method for controlling cancer chemotherapy drug dosing using an adaptive dynamic programming algorithm. We designed a novel value function that included state time delays, creating an effective control approach for handling delays between measuring tumor size and applying chemotherapy. We used a critic neural network structure to derive control laws and optimize drug dosing. We discussed the effects of time delay to ensure the stability of the proposed optimal controller, and simulation results showed these effects. We employed a straightforward mathematical approach to analyze the issues, demonstrating that the derived control laws stabilize the closed-loop system and compensate for time delays.

Several areas remain open regarding finding an optimal chemotherapy approach in cancer treatment. Exploring connections to other drug dosing optimization problems, such as anesthesia, could further extend the applicability of our approach. For instance, adaptive asymptotic tracking for uncertain switched positive compartmental models, as explored in[42], offers a promising direction.

Data availability

The repository includes MATLAB scripts for model definition, delay-aware control implementation, and visualization of simulation results. The simulation codes generated during the current study are available in the Zenodo repository, https://doi.org/10.5281/zenodo.15088181.

Received: 16 March 2025; Accepted: 17 October 2025
Published online: 10 December 2025

References

1. Turner, M. C. et al. Outdoor air pollution and cancer: An overview of the current evidence and public health recommendations. *CA: Cancer J. Clin.* **70**, 460–479 (2020).

2. Ghasemabad, E. S., Zamani, I., Tourajizadeh, H., Mirhadi, M. & Zarandi, Z. G. Design and implementation of an adaptive fuzzy sliding mode controller for drug delivery in treatment of vascular cancer tumours and its optimisation using genetic algorithm tool. *IET Syst. Biol.* **16**, 201–219 (2022).
3. Mokhtari, R. B. et al. Combination therapy in combating cancer. *Oncotarget* **8**, 38022 (2017).
4. DeVita, V. T., Lawrence, T. S. & Rosenberg, S. A. *DeVita, Hellman, and Rosenberg's cancer: principles & practice of oncology*, vol. 2 (Lippincott Williams & Wilkins, 2008).
5. Zhao, Y. et al. Systematic literature review on reinforcement learning in non-communicable disease interventions. *Artif. Intell. Med.* **154**, 102901 (2024).
6. Qods, P., Arkat, J. & Batmani, Y. Optimal administration strategy in chemotherapy regimens using multi-drug cell-cycle specific tumor growth models. *Biomed. Signal Process. Control* **86**, 105221 (2023).
7. Kuznetsov, M., Clairambault, J. & Volpert, V. Improving cancer treatments via dynamical biophysical models. *Phys. Life Rev.* **39**, 1–48 (2021).
8. Ahmadi, Z. & Razminia, A. Safe optimal control of cancer using a control barrier function technique. *Math. Biosci.* **369**, 109142 (2024).
9. Padmanabhan, R., Meskin, N. & Al Moustafa, A.-E. *Mathematical models of cancer and different therapies* (Springer, 2021).
10. De Pillis, L. G. & Radunskaya, A. The dynamics of an optimally controlled tumor model: A case study. *Math. Comput. Model.* **37**, 1221–1244 (2003).
11. Batmani, Y. & Khaloozadeh, H. Optimal chemotherapy in cancer treatment: state dependent Riccati equation control and extended Kalman filter. *Optim. Control. Appl. Methods* **34**, 562–577 (2013).
12. Khalili, P., Zolatash, S., Vatankhah, R. & Taghvaei, S. Optimal control methods for drug delivery in cancerous tumour by anti-angiogenic therapy and chemotherapy. *IET Syst. Biol.* **15**, 14–25 (2021).
13. Padmanabhan, R., Meskin, N. & Haddad, W. M. Optimal adaptive control of drug dosing using integral reinforcement learning. *Math. Biosci.* **309**, 131–142 (2019).
14. Hamdache, A., Saadi, S. & Elmouki, I. Nominal and neighboring-optimal control approaches to the adoptive immunotherapy for cancer. *Int. J. Dyn. Control.* **4**, 346–361 (2016).
15. Itik, M., Salamci, M. U. & Banks, S. P. Optimal control of drug therapy in cancer treatment. *Nonlinear Anal.: Theory, Methods & Appl.* **71**, e1473–e1486 (2009).
16. De Pillis, L. G. & Radunskaya, A. A mathematical tumor model with immune resistance and drug therapy: An optimal control approach. *Comput. Math. Methods Med.* **3**, 79–100 (2001).
17. Nasiri, H. & Kalat, A. A. Adaptive fuzzy back-stepping control of drug dosage regimen in cancer treatment. *Biomed. Signal Process. Control* **42**, 267–276 (2018).
18. Chen, T., Kirkby, N. F. & Jena, R. Optimal dosing of cancer chemotherapy using model predictive control and moving horizon state/parameter estimation. *Comput. Methods Programs Biomed.* **108**, 973–983 (2012).
19. Liu, C., Shi, S. & De Schutter, B. On the regret of model predictive control with imperfect inputs. *IEEE Control. Syst. Lett.* **9**, 601–606 (2025).
20. Mashayekhi, H., Nazari, M., Jafarinejad, F. & Meskin, N. Deep reinforcement learning-based control of chemo-drug dose in cancer treatment. *Comput. Methods Programs Biomed.* **243**, 107884 (2024).
21. Chen, L., Zhang, Y., Yang, P. & Jin, X. Event-triggered drug dosage control strategy of immune systems via safe integral reinforcement learning. *Eur. J. Control.* **82**, 101201. https://doi.org/10.1016/j.ejcon.2025.101201 (2025).
22. Maier, C., Hartung, N., Kloft, C., Huisinga, W. & de Wiljes, J. Reinforcement learning and Bayesian data assimilation for model-informed precision dosing in oncology. *CPT: Pharmacomet. & Syst. Pharmacol.* **10**, 241–254 (2021).
23. Jost, F. et al. Model-based optimal AML consolidation treatment. *IEEE Trans. Biomed. Eng.* **67**, 3296–3306 (2020).
24. Padmanabhan, R., Meskin, N. & Haddad, W. M. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Math. Biosci.* **293**, 11–20 (2017).
25. Treesatayapun, C. & Mu noz-Vázquez, A. J. Optimal drug-dosing of cancer dynamics with fuzzy reinforcement learning and discontinuous reward function. *Eng. Appl. Artif. Intell.* **120**, 105851 (2023).
26. Treesatayapun, C., Mu noz-Vázquez, A. J. & Suyaroj, N. Reinforcement learning optimal control with semi-continuous reward function and fuzzy-rules networks for drug administration of cancer treatment. *Soft Comput.* **27**, 17347–17356 (2023).
27. Niazmand, V. R., Raheb, M. A., Eqra, N., Vatankhah, R. & Farrokhi, A. Deep reinforcement learning control of combined chemotherapy and anti-angiogenic drug delivery for cancerous tumor treatment. *Comput. Biol. Med.* **181**, 109041 (2024).
28. Eisenhauer, E. A. et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *Eur. J. Cancer* **45**, 228–247 (2009).
29. Wu, C.-I., Wang, H.-Y., Ling, S. & Lu, X. The ecology and evolution of cancer: The ultra-microevolutionary process. *Annu. Rev. Genet.* **50**, 347–369. https://doi.org/10.1146/annurev-genet-112414-054842 (2016).
30. Bonate, P. L. *Pharmacokinetic-pharmacodynamic modeling and simulation* 2nd edn. (Springer, 2011).
31. Lewis, F. L., Vrabie, D. & Syrmos, V. L. *Optimal control* (John Wiley & Sons, 2012).
32. Werbos, P. Approximate dynamic programming for real-time control and neural modeling. *Handbook of intelligent control* (1992).
33. Prokhorov, D. V. & Wunsch, D. C. Adaptive critic designs. *IEEE Trans. Neural Netw.* **8**, 997–1007 (1997).
34. Ortega-Martínez, J., Santos-Sánchez, O. & Mondié, S. Lyapunov-Krasovskii prescribed derivative and the bellman functional for time-delay systems. *IFAC-PapersOnLine* **53**, 7160–7165 (2020).
35. He, Y., Wang, Q.-G., Xie, L. & Lin, C. Further improvement of free-weighting matrices technique for systems with time-varying delay. *IEEE Trans. Autom. Control* **52**, 293–299 (2007).
36. Fridman, E. *Introduction to time-delay systems: Analysis and control* (Springer, 2014).
37. Wu, Y., Wei, J. & Zhu, X. Online actor-critic algorithm to solve the approximate optimal adaptive control of continuous-time system with state delay. In *2022 34th Chinese Control and Decision Conference (CCDC)*, 2995–3000 (IEEE, 2022).
38. Lewis, F. L. & Liu, D. *Reinforcement learning and approximate dynamic programming for feedback control* (John Wiley & Sons, 2013).
39. Gagliato, D. M. et al. Clinical impact of delaying initiation of adjuvant chemotherapy in patients with breast cancer. *J. Clin. Oncol.* **32**, 735–744 (2014).
40. Chen, Z., Xu, J., Li, S. E. & Zhao, D. Delay-aware model-based reinforcement learning for continuous control. *Neurocomputing* **450**, 119–128 (2021).
41. Bouteiller, Y., Ramstedt, S., Beltrame, G., Pal, C. & Binas, J. Reinforcement learning with random delays. In *International Conference on Learning Representations (ICLR)* (2021).
42. Lv, M., De Schutter, B., Yu, W. & Baldi, S. Adaptive asymptotic tracking for a class of uncertain switched positive compartmental models with application to anesthesia. *IEEE Trans. Syst. Man Cybern.: Syst.* **51**, 4936–4942 (2019).

Author contributions

Farshad Rahimi developed the theoretical framework, conducted the data analysis, and drafted the initial manuscript. Mahdieh Samadi collected the data, performed the literature review, and contributed to the interpretation of results.

Funding
The authors claim that there are no funds, grants, or other support devoted to the preparation of this manuscript.

Declarations

Competing interests
The authors declare that they have no competing interests.

Additional information
Correspondence and requests for materials should be addressed to F.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.