An Optimal Control Framework for Personalized Chemotherapy: EKF State Estimation and DDPG Optimization

M. Ancona, B. Grosoli, D. Iacoviello, and F. Luzi

***Abstract*—Chemotherapy remains a cornerstone in the treatment of cancer, but its effectiveness is strongly limited by toxicity and the inability to adapt to patient-specific variability. This paper addresses the problem of personalized chemotherapy scheduling through a control-oriented framework based on state estimation and reinforcement learning. A nonlinear four-state tumor–immune–drug model is adopted to describe the interaction among tumor, normal, and immune cells under chemotherapy. To deal with noisy and biweekly clinical measurements, an Extended Kalman Filter (EKF) is used to estimate the unobservable states from biweekly tumor observations. The treatment design problem is then formulated as a discrete-time Markov Decision Process and solved through the Deep Deterministic Policy Gradient (DDPG) algorithm, in order to generate weekly dosing actions under clinical constraints. The reward function penalizes tumor growth, excessive drug administration, and unsafe conditions related to toxicity. Simulation results in representative patient scenarios show that the proposed EKF–DDPG strategy can effectively reduce tumor burden and adapt the treatment intensity to different patient conditions. These results support the potential of the proposed framework for adaptive chemotherapy scheduling under uncertainty.**

***Index Terms*—Cancer chemotherapy, optimal control, Kalman filter, deep reinforcement learning**

I. Introduction

Cancer remains one of the leading causes of mortality worldwide. Despite significant advances in diagnosis and treatment, chemotherapy continues to play a central role in cancer management. However, its effectiveness is strongly limited by toxicity and by the large inter-patient variability in tumor dynamics and drug response.

Modern oncology often relies on multimodal strategies combining surgery, radiotherapy, immunotherapy, and chemotherapy. Within this context, the design of optimal drug administration policies that balance tumor suppression with the preservation of healthy and immune cells is of primary importance [1], [2]. Excessive dosing may cause severe toxicity, whereas insufficient treatment may fail to control tumor progression.

From a systems and control perspective, tumor–immune interactions and drug effects are governed by nonlinear and uncertain dynamics. Conventional open-loop treatment schedules are therefore unable to guarantee robust performance across heterogeneous patient conditions. Closed-loop control strategies, instead, enable adaptive dose modulation based on the observed clinical response, improving robustness to model inaccuracies and parametric uncertainty.

Optimal control approaches have been applied to chemotherapy scheduling, [3], showing that mathematically optimized dosing policies can improve treatment outcomes. However, such methods are often highly sensitive to parameter uncertainty, limiting their robustness and clinical applicability.

Model Predictive Control (MPC), combined with state estimation techniques, has also been proposed to address unmeasurable states and model discrepancies [4]. Although effective in theory, these approaches may entail significant computational complexity, reducing their suitability for real-time clinical implementation.

More recently, Reinforcement Learning (RL) methods have been explored for optimizing dynamic treatment regimens, [5]. In particular, deep RL frameworks have demonstrated promising performance in continuous chemotherapy dosing problems, [6]. Nevertheless, existing approaches typically assume full state observability and frequent measurements, and often rely on continuous drug administration schemes that deviate from realistic clinical protocols.

In contrast, the present work introduces a DRL-based closed-loop framework that explicitly accounts for sparse clinical measurements and clinically feasible weekly dosing schedules. Parameter and measurement biases are incorporated during training to enhance robustness and generalization across heterogeneous patient conditions.

The paper is organized as follows: in Section II, the adopted model is presented, along with the Extended Kalman Filter and the Deep Deterministic Policy Gradient approach. In Section III, numerical results in realistic scenario are presented and discussed, whereas in Section IV some conclusions and future developments are outlined.

II. Methods

A. Model

For this mathematical framework, the four-state model proposed by De Pillis et al. [7] was selected. In the chosen model, tumor evolution is described by four differential equations representing the dynamics of normal cells $N(t)$, tumor cells $T(t)$, immune cells $I(t)$ and the concentration of chemo-drug $M(t)$ within the patient.

The corresponding state-space model is defined by the state variables $x_1(t) = N(t)$, $x_2(t) = T(t)$, $x_3(t) = I(t)$, $x_4(t) = M(t)$ capturing the interactions between the different cell populations and the effects of chemotherapy treatment.

Denoting by $u(t)$ the control input, the cell dynamics is

described as follows:

$$\begin{aligned}
\dot{x}_1 &= r_2 x_1(1 - b_2 x_1) - c_4 x_1 x_2 - a_3 x_1(1 - e^{-x_4}) \\
\dot{x}_2 &= r_1 x_2(1 - b_1 x_2) - c_2 x_3 x_2 - c_3 x_1 x_2 - a_2 x_2(1 - e^{-x_4}) \\
\dot{x}_3 &= s + \frac{\rho x_3 x_2}{\beta + x_2} - c_1 x_3 x_2 - d_1 x_3 - a_1 x_3(1 - e^{-x_4}) \\
\dot{x}_4 &= -d_2 x_4 + u
\end{aligned} \tag{1}$$

Normal and tumor populations follow logistic growth laws with competitive interactions and chemotherapy-induced cytotoxic effects. Immune cells are sustained by a constant influx and tumor-stimulated proliferation with saturation, while undergoing natural decay and drug damage. Drug dynamics are modeled as a first-order decay process with infusion rate $u(t)$ as control input.
The following table 1 summarizes the roles of every single component in the model (1).

Parameter	Meaning
r_1	Tumor cell growth rate
r_2	Normal cell growth rate
b_1	Reciprocal of tumor carrying capacity
b_2	Reciprocal of normal carrying capacity
c_1	Tumor-immune competition rate
c_2	Immune-tumor killing rate
c_3	Normal-tumor competition rate
c_4	Normal-tumor competition rate (reverse)
d_1	Natural death rate of immune cells
d_2	Drug decay rate
a_1	Drug effect on immune cells
a_2	Drug effect on tumor cells
a_3	Drug effect on normal cells
s	Influx rate of immune cells
β	Half-saturation constant for immune stimulation
ρ	Immune cell proliferation coefficient

TABLE I
MODEL PARAMETERS AND THEIR BIOLOGICAL MEANING.

To guarantee a more realistic approach, a bias has been introduced on several key parameters, specifically:

- r_1 *Tumor cell growth rate.* This parameter governs the intrinsic proliferation speed of tumor cells. Its variability reflects the heterogeneity in tumor aggressiveness observed across patients.
- a_1 *Fractional immune cell kill rate by chemotherapy.* This parameter captures the collateral damage of the different type of chemotherapy on immune cells.
- a_2 *Fractional tumor cell kill rate by chemotherapy.* This represents the cytotoxic effectiveness of the drug against tumor cells. Variability reflects chemosensitivity of the tumors.
- a_3 *Fractional normal cell kill rate by chemotherapy.* This describes chemotherapy-induced toxicity on healthy tissues. Its variability incorporates differences in patients' physiological ability to repair or tolerate drug-induced damage.

The parameters a_1, a_2, and a_3 were modeled as Gaussian-distributed variables, while r_1 was randomly sampled to account for inter-patient variability in tumor growth dynamics. Their uncertainty was propagated into the state equations and incorporated into the state-space model through a bias covariance matrix, used to adapt the process noise and improve the robustness of the Extended Kalman Filter (EKF) against parametric variations. The EKF design, including this bias modeling strategy, is discussed in Subsection B.

To qualitatively characterize the uncontrolled dynamics, the equilibrium points of the drug-free system were analyzed. In this cancer-immune model, different equilibria may arise, corresponding to clinically distinct regimes such as tumor eradication, tumor persistence, or uncontrolled proliferation.

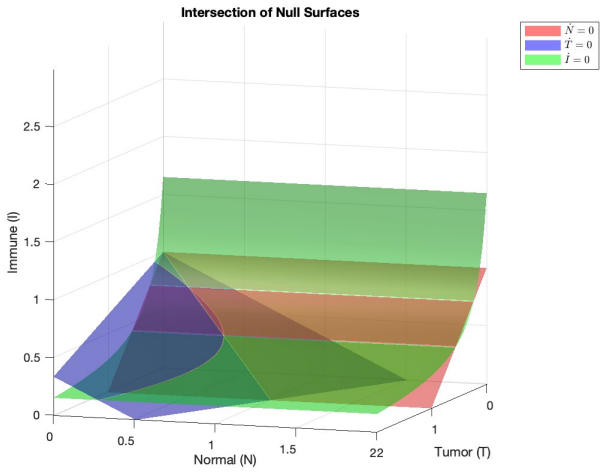

Fig. 1. Nullclines for the system (1) involving the states N, I, T.

These equilibrium points were determined from the nullclines of the state equations, whose intersections identify the possible steady-state configurations of the natural, tumor, and immune cell populations, see Fig. 1.

Under drug-free evolution, i.e., with zero control input, three main types of equilibria can be identified, corresponding to different patient outcomes:

- Tumor-free equilibrium: This condition corresponds to complete tumor eradication, in which the immune system maintains a dominant role and the tumor population is null. The associated equilibrium point is:

$$x_e^1 = \left(\frac{1}{b_2}\ 0\ \frac{s}{d_1}\right)^\top \tag{2}$$

- Death equilibrium: In this scenario, the normal cell population vanishes entirely, indicating systemic failure. Two subcases can be identified: one in which all populations, except immune cells, collapse and another in which tumor and immune cells persist, while normal cells are absent, where $f(\gamma)$ denotes the immune-cell equilibrium value associated with the tumor level γ, obtained by solving the steady-state conditions of the drug-free system with $x_1 = 0$. The equilibria are:

$$x_e^2 = \left(0\ 0\ \frac{s}{d_1}\right)^\top \quad \text{and} \quad x_e^3 = (0\ \gamma\ f(\gamma))^\top \tag{3}$$

- Co-existing equilibrium: These points represent clinically relevant scenarios where the tumor is present but kept under control, either in a dormant state or in balance with immune surveillance.

From a biological perspective, these equilibria correspond to situations in which none of the populations—tumor, immune, or normal cells—vanishes completely. Mathematically, they arise as interior equilibrium points resulting from the intersection of all three null-surfaces within the positive orthant of the state space. Depending on the values of key parameters—such as the immune stimulation rate ρ and the influx rate of immune cells s, the system may exhibit one or more coexisting equilibria, which can be stable or unstable. These equilibria are critical for designing control strategies, as they highlight the thresholds between complete remission and persistent, yet manageable, tumor dynamics.
The framework described in [7] delineates the existence and stability of these equilibria as functions of the immune response rate p and the immune source rate s. All parameter values are set to be equal to those used in the results section for the exception for the parameter affected by bias, for which the mean was used as a reference. For our parameter values with ρ = 0.01 and s = 0.33, the system exhibits two coexisting equilibria, as discussed in [7].
Since not all states are directly measurable in practice, state estimation is required, motivating the use of an Extended Kalman Filter.

B. EKF

In real clinical practice, monitoring the internal state of a tumor is typically limited to sporadic and partial measurements, obtained through imaging techniques such as Positron Emission Tomography (PET) or Computed Tomography (CT scans). These measurements are invasive, costly, and can only be performed at discrete time intervals (e.g., every few weeks or months).
In order to realistically model this constraint, an Extended Kalman Filter (EKF) [10] was implemented in the control framework. The EKF is not used to achieve full continuous state reconstruction, but to approximate the intermittent and noisy nature of clinical observations. Specifically, the EKF receives updates only at discrete sampling times, realistically every 14 days, simulating the clinical measurement schedule, and dynamically fuses these partial observations with the system's internal model predictions. This approach allows the estimation of hidden tumor dynamics between two consecutive exams, improving the control strategy's robustness against measurement sparsity and observation uncertainty.
The observer is designed to reduce the gap between theoretical models and the practical limits of clinical monitoring, ensuring that the proposed control strategies are both applicable and effective in realistic medical settings.
A central component in the design of the Extended Kalman Filter is the specification of the process noise covariance Q and the measurement noise covariance R. These matrices define the balance between trust in the internal tumor dynamics model and reliance on external clinical observations. In this work, both Q and R are kept constant throughout the simulations.
To improve the robustness of the EKF under sparse measurements and parameter uncertainty, an integral correction term was introduced, [11]. Let

$$e_2(t_k) = y(t_k) - \hat{x}_2^-(t_k) \tag{4}$$

be the tumor estimation error at the measurement instant t_k, where $y(t_k)$ denotes the available tumor measurement and $\hat{x}_2^-(t_k)$ the predicted tumor state before the EKF update. Between two consecutive measurements, the last available error is held constant, i.e.,

$$e_h(t) = e_2(t_k), \qquad t \in [t_k, t_{k+1}). \tag{5}$$

The correction signal is then generated as

$$\dot{z}_c(t) = e_h(t), \qquad c(t) = k_c\, z_c(t), \tag{6}$$

or equivalently,

$$c(t) = k_c \int_0^t e_h(\tau)\, d\tau. \tag{7}$$

This hybrid formulation preserves sensitivity to tumor growth dynamics while ensuring robustness under sparse measurements and parametric uncertainty, limiting bias accumulation and long-term drift at negligible computational cost. Only tumor burden and drug concentration are assumed measurable at discrete time instants, whereas healthy and immune cells remain unobservable. The controller therefore operates on EKF state estimates, reflecting realistic clinical monitoring constraints. The control action is bounded in $[0, 1]$ and mapped to a clinically feasible weekly piecewise-constant dosing schedule, consisting of an infusion phase followed by off-days. Between sessions, drug concentration evolves according to the pharmacokinetic model, avoiding unrealistic continuous administration. Dose optimization is performed via reinforcement learning, [12]. The treatment problem is formulated as a Markov Decision Process (MDP), where the agent iteratively updates its policy based on reward feedback from the tumor–immune dynamics.

C. Deep Deterministic Policy Gradient

To implement the control policy, it is adopted the Deep Deterministic Policy Gradient (DDPG), [14], an off-policy algorithm actor–critic method that combines the principles of Deep Q-Networks (DQN) and Deterministic Policy Gradients (DPG). To improve stability during training, the algorithm makes use of an experience replay buffer and target networks that are updated slowly to prevent divergence.
The training of the controller follows the typical iterative structure of reinforcement learning, where the agent continuously interacts with the environment in discrete time steps. At the beginning of each episode, the environment is initialized with the initial conditions of each patient, [7], [16]. At each weekly control step, the agent selects a dosing action based on the tumor state estimate provided by the EKF. This action

is then applied to the tumor–immune model. The updated system state generates two feedback signals: (i) a reward, which evaluates the quality of the action with respect to the therapeutic goals, and (ii) an "isDone" flag, which determines whether the episode should terminate.

The isDone function defines terminal conditions that prevent infinite episodes and exclude clinically unacceptable trajectories, such as uncontrolled tumor growth or excessive toxicity. In the proposed model, it directly encodes treatment feasibility and toxicity constraints, guiding the agent towards clinically acceptable dosing strategies.

The reward function [13] combines the treatment objectives into a single scalar value, balancing the tumor growth and minimizing the drug dose due to toxicity, to preserve the immune system of the patient. The reward function was defined as a quadratic penalty characterized by tumor progression, drug administration, and treatment feasibility.

This simple structure reflects the clinical objective of suppressing tumor growth while avoiding excessive chemotherapy exposure, and discourages treatment trajectories that end in clinically unacceptable states.

$$r_t = -\left(k_1\, x_2^2 + k_2\, u_{\text{pre}}^2 + 10 \cdot \text{isDone}\right) \tag{8}$$

The weighting coefficients were set to $k_1 = 1$ and $k_2 = 0.1$. The ratio $k_1/k_2 = 10$ reflects the clinical priority of tumor suppression over drug minimization, while still discouraging excessive chemotherapy exposure. The DDPG controller was trained for 1000 episodes, each consisting of 30 steps, corresponding to a treatment horizon of approximately seven months (210 days). The actor network was updated with a learning rate of 1×10^{-4}, while the critic network employed a learning rate of 1×10^{-3}. A batch size of 32 samples was used during training. Hyperparameter tuning was performed iteratively to improve performance and robustness. In particular, increasing the number of training episodes while progressively reducing the learning rate allowed the controller to achieve lower error rates and enhanced stability. This tuning strategy proved effective in refining policy performance and adapting the controller to inter-patient variability. A slower training process was found to improve robustness, as it reduced abrupt parameter updates and enabled the controller to generalize more effectively to new unseen scenarios.

III. Results

This section presents the numerical results obtained for the control problems introduced in Section II.

The model parameters and initial conditions were selected to reflect clinically plausible tumor dynamics. In particular, the baseline tumor size was chosen in accordance with the considerations reported by Del Monte [15], ensuring that the simulations are grounded in physiologically and clinically realistic scenarios.

Table 2 provides a summary of the parameters employed to characterize the tumor–immune model, including the adjustments to reflect inter-patient variability.

In addition to the model parameters, the equilibrium structure of the tumor–immune system plays a key role in interpreting the simulated dynamics. Using the parameter set reported in Table 2, the equilibrium points obtained from the nonlinear model are the following:

$$H = (1 \;\; 0 \;\; 1.65 \;\; 0)^T \tag{9}$$

$$C_1 = (0.42 \;\; 0.58 \;\; 0.42 \;\; 0)^T \qquad C_2 = (0.76 \;\; 0.24 \;\; 0.76 \;\; 0)^T \tag{10}$$

Here, H denotes the *healthy equilibrium*, corresponding to complete tumor eradication, whereas C_1 and C_2 represent *coexisting equilibria*, in which normal, tumor, and immune cell populations coexist at nonzero levels.

The equilibrium C_2 represents a more clinically favorable configuration, as it corresponds to a high level of normal cells and a relatively low tumor burden. However, its unstable nature prevents the system from remaining near this state without external intervention, making it unsuitable as a long-term outcome unless active control strategies are implemented.

In order to support state estimation under the sparsity and noise typical of clinical measurements, the Extended Kalman Filter described in Section II was implemented using fixed noise covariance matrices. The values were set equal to $Q = 10^{-3}$ and $R = 10^{-1}$, thus ensuring a robust yet computationally tractable observer capable of operating reliably across heterogeneous patient conditions. These settings garantee a balanced contribution between the internal model dynamics and the intermittent, noisy measurements available in realistic monitoring scenarios.

Parameter	Value	Unit
b_1	1	cell^{-1}
b_2	1	cell^{-1}
c_1	1	$\text{cell}^{-1}\text{day}^{-1}$
c_2	0.5	$\text{cell}^{-1}\text{day}^{-1}$
c_3	1	$\text{cell}^{-1}\text{day}^{-1}$
c_4	1	$\text{cell}^{-1}\text{day}^{-1}$
d_1	0.2	day^{-1}
d_2	1	day^{-1}
r_2	1	day^{-1}
s	0.33	$\text{cell}^{-1}\text{day}^{-1}$
β	0.3	cell
ρ	0.01	day^{-1}
r_1	[1.3, 1.7]	day^{-1}
a_1	mean 0.2, std 0.05	$\text{mg}^{-1}\text{day}^{-1}$
a_2	mean 0.3, std 0.05	$\text{mg}^{-1}\text{day}^{-1}$
a_3	mean 0.1, std 0.05	$\text{mg}^{-1}\text{day}^{-1}$

TABLE II
MODEL PARAMETERS USED IN THE TUMOR–IMMUNE SYSTEM SIMULATIONS [7].

In clinical practice, several factors influence the oncologist's decision when determining the appropriate chemotherapy dose, including patient age, sex, immune deficits or pre-existing conditions. Younger patients may tolerate higher doses, as their capacity to regenerate healthy and immune cells is generally greater. Conversely, elderly patients typically

require lower doses, since their regenerative potential is reduced. For this reason, an effective treatment strategy must ensure the preservation of healthy and immune cells throughout the therapy course.

The initial conditions of the state variables were sampled randomly within physiologically plausible ranges to account for inter-patient variability. All three populations, normal cells (x_1); tumor cells (x_2), and immune cells(x_3) were scaled to an order of magnitude of 10^9.

Variable	Min	Max
x_1	0.4	1
x_2	0.1	0.6
x_3	0.005	0.2

TABLE III
INITIAL RANGES FOR THE STATE VARIABLES USED IN THE SIMULATIONS.

To evaluate the proposed control strategy, two representative simulation scenarios were considered, corresponding to distinct tumor aggressiveness and treatment tolerance conditions. These cases illustrate the controller's ability to achieve tumor eradication in favorable conditions and stabilization in more constrained settings.

Case 1: The first scenario represents a young male patient characterized by favorable baseline conditions. The tumor cell population was initialized at a relatively low level ($x_2 = 0.30$), the immune response against the tumor was strong (high value of a_2), and no immunodeficiency was assumed, as reflected by the small value of $a_1 = 0.13$. The chemo-toxic effect on healthy cells was moderate ($a_3 = 0.05$), consistent with the generally higher drug tolerance observed in younger patients. The initial states were randomly sampled within the physiological ranges of Table IV, while the remaining parameters were taken from Table II.

x2	x4	a1	a2	a3	r1
0,3	0	0,13	0,3	0,05	1,48

TABLE IV
PATIENT 1 INITIAL CONDITION AND PARAMETERS' VALUE

In this scenario, the maximum admissible chemotherapy dose was set to $u_{\max} = 4.4\,\mathrm{mgl}^{-1}\,\mathrm{day}^{-1}$, reflecting the higher tolerance of younger patients to intensive treatment schedules.

Figure 2 illustrates the tumor dynamics for the young male patient in the absence of chemotherapy. Starting from an initial normalized value of $x_2(0) = 0.30$, the tumor population exhibits a monotonic growth pattern and gradually approaches a high equilibrium level. Despite the favorable immune responsiveness of this patient profile, the immune system alone is not sufficient to counteract tumor expansion, leading to progressive disease.

Figure 3 shows the evolution of the same patient under the proposed control chemotherapy strategy. The controller rapidly suppresses tumor growth, driving x_2 close to zero

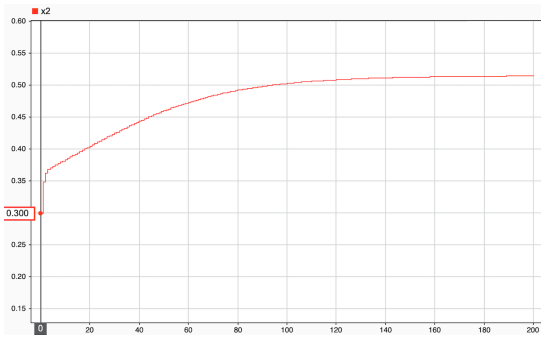

Fig. 2. Young man tumor evolution (x_2 cells) without control action

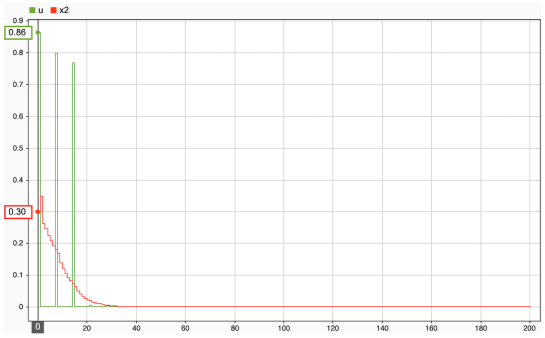

Fig. 3. Young man tumor evolution (x_2 in red) with control action (u in green)

within the first 40 days. The administration of chemotherapy, represented by the green pulses, is more intense in the early phase and progressively decreases as the tumor burden becomes negligible.

For Patient 1, the uncontrolled tumor burden reaches approximately $x_2(T) \approx 0.51$, while under the proposed control strategy it is reduced to a value close to zero, corresponding to an almost complete reduction of the tumor population.

Case 2: This scenario represents an elderly patient presenting both immunodeficiency and a more advanced tumor stage. The tumor burden was initialized at a relatively high value ($x_2 = 0.40$), while the immune system was significantly weakened, as indicated by the elevated value of $a_1 = 0.30$. The immune response against the tumor ($a_2 = 0.30$) was also reduced, and the chemo-toxic effect on healthy cells was high ($a_3 = 0.15$), reflecting the diminished treatment tolerance typically observed in older patients. The initial normal and immune cell populations were sampled within the physiological ranges reported in Table V, while all remaining parameters were taken from Table II.

x2	x4	a1	a2	a3	r1
0.4	0	0.30	0.30	0.15	1.51

TABLE V
PATIENT 2 INITIAL CONDITION AND PARAMETERS' VALUE

In this scenario, the maximum admissible chemotherapy

dose was limited to $u_{\max} = 1.9\,\mathrm{mg\,l^{-1}\,day^{-1}}$, reflecting the markedly reduced treatment tolerance of elderly patients. The therapeutic objective was therefore not tumor eradication but clinical stabilization, aiming to contain disease progression while avoiding additional damage to the already compromised immune system. Under these conditions, maximizing the patient's life expectancy becomes the primary goal.

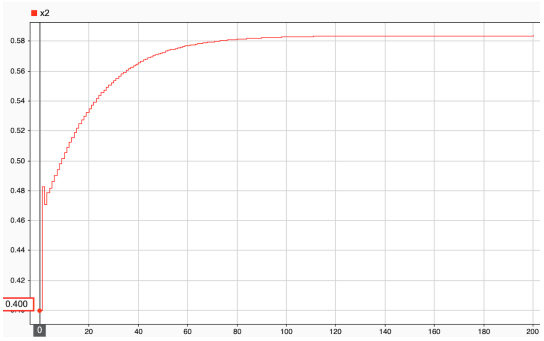

Fig. 4. Elder patient tumor evolution (x_2 cells) without control action

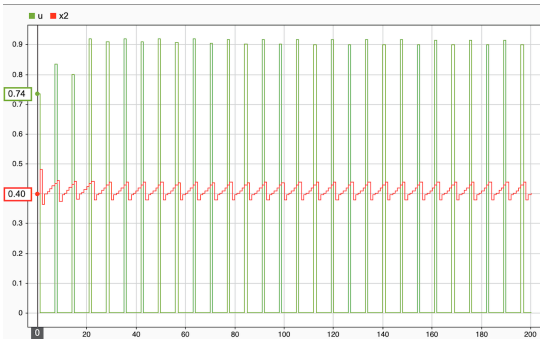

Fig. 5. Elder patient tumor evolution(x_2 in red) with control action (u in green)

Figure 4 illustrates the tumor dynamics for the elderly patient in the absence of chemotherapy. Starting from an initial value of $x_2(0) = 0.40$, the tumor population grows rapidly and then gradually approaches a high equilibrium level. Due to the combination of immunodeficiency and reduced immune effectiveness, the natural immune response is unable to contain tumor progression, leading to continuous disease worsening. Figure 5 presents the evolution of the same patient under the proposed chemotherapy-based control strategy. Because the maximum admissible dose is strongly limited, the controller administers frequent low-intensity treatment pulses. These doses partially slow tumor growth but are insufficient to eradicate the tumor.

Nevertheless, the therapeutic policy succeeds in stabilizing the tumor dynamics around a lower level compared to the uncontrolled case, achieving the clinical objective of disease containment while respecting the patient's limited tolerability. For Patient 2, the uncontrolled tumor burden reaches approximately $x_2(T) \approx 0.59$, whereas under the proposed control strategy it is reduced to $x_2(T) \approx 0.41$, corresponding to an approximate reduction of 30.5%.

IV. Summary and Conclusions

In this paper, a realistic scenario for personalized chemotherapy is proposed. The preliminary results of this study show that the proposed EKF–DDPG framework can support adaptive chemotherapy scheduling under noisy and biweekly measurements. By combining tumor–immune mathematical modeling, EKF-based state estimation, and reinforcement learning control, the approach generates clinically constrained dosing policies tailored to different patient conditions. In the simulated scenarios, the controller achieved either near-complete tumor suppression or bounded tumor stabilization, depending on the therapeutic constraints of the patient. Although further validation on broader patient cohorts is required, these results suggest that the proposed methodology is a promising control-oriented approach for personalized chemotherapy scheduling under uncertainty.

References

[1] E. M. Lee *et al.*, "Toxicities and quality of life during cancer treatment in advanced solid tumors," *Current Oncology*, vol. 30, no. 10, pp. 9205–9216, 2023.

[2] M. M. Hadjiandreou and G. D. Mitsis, "Mathematical modeling of tumor growth, drug-resistance, toxicity, and optimal therapy design," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 2, 2014, doi: 10.1109/TBME.2013.2280189.

[3] M. Engelhart, D. Lebiedz, and S. Sager, "Optimal control for selected cancer chemotherapy ODE models," *Mathematical Biosciences*, vol. 229, no. 1, pp. 123–134, 2011.

[4] T. Chen, N. F. Kirkby, and R. Jena, "Optimal dosing of cancer chemotherapy using model predictive control and moving horizon estimation," *Comput. Methods Programs Biomed.*, vol. 108, no. 3, pp. 973–983, 2012.

[5] C. Y. Yang *et al.*, "Reinforcement learning strategies in cancer chemotherapy treatments: A review," *Comput. Methods Programs Biomed.*, vol. 229, p. 107280, 2023.

[6] H. Mashayekhi *et al.*, "Deep reinforcement learning-based control of chemo-drug dose in cancer treatment," *Comput. Methods Programs Biomed.*, vol. 243, p. 107884, 2024.

[7] L. G. De Pillis and A. Radunskaya, "The dynamics of an optimally controlled tumor model," *Mathematical and Computer Modelling*, vol. 37, no. 11, pp. 1221–1244, 2003.

[8] J. Tellinghuisen, "Statistical error propagation," *J. Phys. Chem. A*, vol. 105, no. 15, pp. 3917–3921, 2001.

[9] University of California, Berkeley, *A quick guide to sketching phase planes*.

[10] K. Fujii, *Extended Kalman Filter*. ACFA-Sim-J Group.

[11] W. X. Zheng, "A bias correction method for identification of linear dynamic errors-in-variables models," *IEEE Trans. Autom. Control*, vol. 47, no. 7, pp. 1142–1147, 2002.

[12] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Comput. Surveys*, vol. 55, no. 1, Art. 5, 2023.

[13] A. Bonarini, C. Bonacina, and M. Matteucci, "An approach to the design of reinforcement functions in real world, agent-based applications," *IEEE Trans. Syst., Man, Cybern. B*, vol. 31, no. 3, pp. 288–301, Jun. 2001, doi: 10.1109/3477.931510.

[14] Y. Hou, L. Liu, Q. Wei, X. Xu and C. Chen, "A novel DDPG method with prioritized experience replay," 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Banff, AB, Canada, 2017, pp. 316-321.

[15] U. Del Monte, "Does the cell number 10^9 still really fit one gram of tumor tissue?," *Cell Cycle*, vol. 8, no. 3, pp. 505–506, 2009.

[16] J. Tordesillas and A. Arbelaiz, "Personalized cancer chemotherapy schedule: A comparative analysis of reinforcement learning and model-based methods," arXiv:1904.01200, 2019.