A Markov Decision Process Framework for Optimal Cancer Chemotherapy Dose Selection

A O Adeyiola¹, S I Rabia¹, A Elsaid¹, S Fadel² and A Zakaria³

¹ Institute of Basic and Applied Sciences, Egypt-Japan University of Science and Technology, P O Box 179, New Borg El-Arab, 21934, Alexandria, Egypt
² Department of Clinical Oncology and Nuclear Medicine, Alexandria University, 22, El-Guish Road, El-Shatby, 21526, Alexandria, Egypt
³ Alexandria Pediatric Oncology Center, Borg El-Arab University Hospital, New Borg El-Arab, 21934, Alexandria, Egypt

Abstract. Chemotherapy is arguably the most effective treatment for treating cancer. However, it triggers toxic effects on patients due to its interaction with normal cells. In this paper, we propose a finite-horizon Markov decision process framework for optimal chemotherapy dose selection during cancer treatment. In contrast to many other research works, our model accommodates all possible patient’s clinical health states that could be considered during treatment. We study the use of our model by applying it on Wilms tumor. We explore three different cases of varying transition probabilities and rewards. Our numerical results agree with clinical intuition and demonstrate the potential applicability of our model.

1. Introduction
Chemotherapy is arguably the most effective and extensively used modality for most cancers [1]. Although it is highly useful for treating cancer, it triggers toxic effects due to the interaction of the drug with normal cells. This prevents sufficient doses from being administered to obtain a complete cure [2]. Thus, it is important to design treatment strategies for cancer that will ensure a desired rate of tumor cells are damaged without overdosing the host.

Several studies have used optimal theory to model chemotherapy treatment problems. Nanda et al. [3] proposed a mathematical model that seeks to find the treatment regimens that minimize cancer cell count and deleterious effects of drug therapy on patients. They represented the disease dynamics by a system of ordinary differential equations and solve them using the fourth-order Runge-Kutta method. Pillis et al. [4] considered the dynamics of the interaction of the triplet: tumor, immune system, and anti-cancer drug. They represented the dynamics with a coupled differential equation and evaluated two optimal control strategies: a quadratic control, and a linear control. Ledzewicz and Schaeftler [5] proposed a mathematical model that takes into account the heterogeneity of the tumor. They considered the interaction between two cancer cell sub-populations of different sensitivities with respect to chemotherapeutic agent and proved that administration of lower dose chemotherapy is a viable option for population with varying chemotherapeutic sensitivities.

As can be noted, all of these works mainly considered the problem of optimal chemotherapy treatment strategy that minimizes the size of tumor population without quantifying the benefit/disutility of the patient over the treatment period. In addition, the methods used are highly deterministic and do not capture effectively the stochastic dynamics of cancer chemotherapy treatment process. Therefore, we propose a Markov decision process (MDP) framework for the optimum strategy during chemotherapy...
treatment. MDPs are powerful analytical tools used to solve complex, stochastic and dynamic problems [6]. MDP models have found sizeable usability in modelling common chronic diseases that involve time to intervention and treatment strategy. Shechter et al. [7] considered the problem of optimal initiation of HIV treatment based on patient CD4 count that maximizes the patient’s quality-adjusted lifetime. They proposed a stationary, infinite-horizon MDP model and proved the existence of a control-limit policy in terms of the patient’s CD4 count. Alagoz et al. [8] presented an MDP model that seeks to find the optimal timing of living-donor liver transplantation based on the patient health state which is described by the Model for End-stage Liver Disease (MELD) score. Akhavan-Tabatabaei et al. [9] proposed a finite-horizon MDP model for cervical cancer screening decision. The optimal policy was formulated in terms of current diagnosis, age, human papillomavirus (HPV) contraction risk, and screening test result. They further evaluated their model by analyzing the trade-off between cost of screening and early treatment procedures, against the risk of letting the disease advance. Furthermore, several studies have designed MDP models for optimal treatment strategy for cancer. Kim et al. [10] developed a finite-horizon MDP model for optimal fraction of radiotherapy based on number of organs at risk, and targeted tumor volume. They built the model mainly on medical intuition by considering some key features, as well as state-action transition. Maass et al. [11] proposed a finite-horizon MDP model for multi-modality cancer treatment based on tumor progression and normal tissue side effect. They categorized the treatment options into three types in terms of tumor reduction, risk to normal tissue, and repeatability. More recently, Bazrafshan et al. [12] used MDP framework to propose a model for optimizing cancer chemotherapy in clinical trials which captured the treatment dynamics. However, the authors only considered the toxicity level of the patient. They did not consider the utility of chemotherapy with respect to its primary purpose. Moreover, the treatment regimens are for gastric and gastroesophageal cancers and cannot be used for other types of cancers.

In this study, we propose a general MDP-based framework for cancer chemotherapy treatment. Our contributions to the existing literature on cancer treatment are:

1) We propose a novel approach for optimum chemotherapy dose selection that can be used for treatment of any type of cancer.

2) Our proposed model does not only consider the tumor progression, but also considers the clinical health states of the patient.

3) We propose a multi-objective model that solves the trade-off problem between chemotherapy treatment utility and disutility.

The remainder of this paper is organized as follows: In section 2, we present our MDP model and give a detailed description of its components. In section 3, we present numerical results. Finally, we highlight our main conclusions and possible extensions in section 4.

2. MDP model
We construct a discrete-time finite-horizon MDP model with decision epochs $t \in T = \{1, 2, \ldots, E\}$. At each decision epoch $t$, the decision maker (i.e., oncologist) observes the patient clinical health states by conducting some clinical tests, such as complete blood count (CBC), temperature, weight, etc. Afterwards, the decision maker considers the tumor progression level. If an absorbing state is attained (e.g., cancer free, death, etc.) treatment is stopped and the decision process terminates. Otherwise, the decision maker decides (based on the patient clinical health state and tumor progression level) on the optimal dose of chemotherapy to be given (if any). This decision continue till an absorbing state is attained or the treatment period ends. For each action the decision maker takes, there is an immediate reward representing the treatment utility (if the tumor progression is expected to improve based on this action) or penalty to represent action disutility (if this action is expected to get the patient health state worse). Our goal is to solve the trade-off problem between maximizing the total chemotherapy treatment utility and minimizing its total disutility. A formal definition of the core components of our MDP model is given as follows:
2.1. States
The patient state is composed of two state variables: clinical health state (vector), and tumor progression state (scalar). The clinical health state is defined by the pre-assessment tests done by the decision maker at each decision epoch $t$, and is denoted by $c_t \in C = \{(c_{t1}, c_{t2}, ..., c_{tN}) : c_i \in \{1, 2, ..., l_i\}\}$, where $N$ is the number of pre-assessment tests, and $l_i$ is the number of possible outcomes of the $i^{th}$ pre-assessment test. For any test state $c_i$, we assume that $c_i = 1$ represents the best clinical health state, while $c_i = l_i$ represents the worst clinical health state.

Let $\tau_t \in T = \{t_0, t_1, ..., t_M\}$ denote the tumor progression observed at decision epoch $t$. $T$ is ordered such that $\tau_t = t_0$ is the best patient state (e.g., cancer-free) while $\tau_t = t_M$ is the worst patient state (death due to tumor progression). We note that $t_0$ and $t_M$ are absorbing states. The patient state at time $t$ is given by $s_t = (c_t, \tau_t) \in S$, where $S = C \times T$.

2.2. Actions
We denote the action space by $A = \{a_0, a_1, ..., a_K\}$, where $a_0$ represents wait, $a_K$ represents full dose, and $a_1, ..., a_{K-1}$ are different fractions of chemotherapy. We assume that if $a_0$ is chosen, the patient's clinical health state can either improve or remain the same, while the tumor progression state can either remain the same or get worse. For any other action $a_k, k = 1, 2, ..., K$, the patient's clinical health state can either get worse or remain the same, while the tumor progression state can either get better or remain the same. We assume that for any action $a_k$, the clinical health and tumor progression states can only improve or get worse by at most one step.

2.3. Transition probabilities
When the decision maker chooses action $a_t \in A$ at decision epoch $t$ for a patient in state $s_t$, the patient state moves to $s_{t+1}$ at $t + 1$ with probability $P_t(s_{t+1} | s_t, a_t)$. We assume that

\[ P_t(s_{t+1} | s_t, a_t) = P_t^C(c_{t+1} | c_t, a_t) \times P_t^T(\tau_{t+1} | \tau_t, a_t), \]

where $P_t^C(c_t, a_t)$ and $P_t^T(\tau_t | \tau_t, a_t)$ are the transition probabilities for the clinical health state and tumor progression state, respectively. This assumption is consistent with our intuition that clinical health and tumor progression do not depend on each other but depend on the decision maker’s action only.

Moreover, we order $A$ in terms of its effect on the patient's clinical health state and tumor progression state. More specifically, we assume that

\[ P_t^C(c_{t+1} | c_t, a_4) > P_t^C(c_{t+1} | c_t, a_2) > \cdots > P_t^C(c_{t+1} | c_t, a_K), \]

where $c_{t+1}$ is a better clinical health state than $c_t$. Also,

\[ P_t^T(\tau_{t+1} | \tau_t, a_4) < P_t^T(\tau_{t+1} | \tau_t, a_2) < \cdots < P_t^T(\tau_{t+1} | \tau_t, a_K), \]

where $\tau_{t+1}$ is a better tumor progression state than $\tau_t$.

We note that $a_0$ is excluded from the above ordering because its effect on the patient's clinical health state and tumor progression state is reversed compared to any other action.

Since $t_0$ and $t_M$ are absorbing states, we have

\[ P_t^T(t_0 | t_0, a_t) = P_t^T(t_M | t_M, a_t) = 1. \]

2.4. Reward functions
Our model includes two reward functions: immediate reward, and terminal reward. A formal description of the two rewards is given as follows:

- The immediate reward function $r_t(s_t, a_t)$ reflects the utility/disutility the decision maker incurs when the patient health state is $s_t$ and action $a_t$ is taken at decision epoch $t$. It is defined as

\[ r_t(s_t, a_t) = \sum_{i=1}^{N} r_{t,i}(c_t, a_t) + r_t(\tau_t, a_t), \]

where $r_{t,i}(c_t, a_t)$ refers to the immediate reward regarding clinical health state $c_t$, and $r_t(\tau_t, a_t)$ is the the immediate reward regarding tumor progression state $\tau_t$.

- The terminal reward function quantifies patient's utility/disutility at the end of the treatment period and is denoted by $r_{T+1}(s)$ where $s$ is the terminal state. This function is defined as
where \( r_{T+1}(c_i) \) is the terminal reward regarding clinical health state \( c_i \), and \( r_{T+1}(\tau) \) is the terminal reward regarding \( \tau \).

Recall that the two state variables, \( c_t \) and \( \tau_t \), are ordered according to their benefit to the patient. Let \( c_i \) and \( c_i' \) denote two different patient clinical health states where \( c_i > c_i' \), and \( \tau \) and \( \tau' \) denote two different tumor progression states where \( \tau > \tau' \). Hence, the corresponding reward functions are assumed to follow this inequality for all \( t \):

\[
 r_t(c_i, a_t) < r_t(c_i', a_t), r_t(\tau, a_t) < r_t(\tau', a_t). \tag{7}
\]

### 2.5. Value function

The goal of our MDP model is to find the optimal strategy for cancer chemotherapy treatment. That is, we seek a rule for taking action at each state that will maximize the expected total reward of a patient over the treatment period. This can be achieved by solving the following Bellman’s recursive equations for all \( s_t \in S \) and \( t = 1, 2, \ldots, T \):

\[
 V_t(s_t) = \max_{a_t \in A} \left\{ r_t(s_t, a_t) + \sum_{s_{t+1} \in S} P_t(s_{t+1}|s_t, a_t)V_{t+1}(s_{t+1}) \right\}, \tag{8}
\]

where \( V_t(s_t) \) is the patient’s maximum expected total reward at the decision epoch \( t \) when the patient is in state \( s_t \), with the boundary condition \( V_{T+1}(s) = r_{T+1}(s) \).

### 3. Results and discussion

In this section, we present numerical studies to illustrate the applicability of our proposed MDP model. Specifically, we base our numerical examples onWilms tumor (WT). We chose WT because it is the most common primary renal tumor of all childhood cancer [13]. The incidence of WT in the United States is approximately 7 new cases for every 1 million children with top occurrence between 2 and 3 years [14]. Meanwhile, there have been some recent advances in management of WT with more than 85% overall survival rate [13]. There is need to further improve its treatment over the current risk stratification to better direct therapy beyond its present limitation [13].

For simplicity, we consider complete blood count (CBC) as the only patient’s clinical health state. We denote this variable by \( c \in C = \{1, 2\} \), where 1 and 2 represent normal and abnormal CBC, respectively. Furthermore, we represent the tumor progression states by \( \mathcal{T} = \{0, 1/4, 1/2, 3/4, 1\} \), and recall that \( \mathcal{T} \) is an ordered set with 0 representing the best state (cancer free) and 1 representing the worst state (death). The action space is denoted by \( A = \{W, 33\%, 66\%, F\} \), where \( W \) denotes wait, and 33%, 66%, and \( F \) denote 33%, 66%, and full dose, respectively. We consider a treatment period of length 4 weeks, i.e., \( E = 4 \). Backward induction was used to solve (8) and compute the optimal treatment policy over the treatment period. This was implemented on MATLAB using the MDP toolbox.

The remainder of this section is organized as follows: In section 3.1, we present a base case example for numerical illustration, including the transition probabilities, rewards, as well as some general assumptions made to make them clinically applicable. In section 3.2, we investigate the sensitivity of our model by exploring how changes in transition probabilities and reward function affect our optimal policy. In all our numerical examples, we consider stationary probabilities and rewards, i.e., \( P_t(s_{t+1}|s_t, a_t) = P(s_{t+1}|s_t, a_t) \), and \( r_t(s_{t+1}|s_t, a_t) = r(s_{t+1}|s_t, a_t) \) for all \( t \).

### 3.1. Base case

Tables 1-3 contain the state-action transition probabilities for both tumor progression, and CBC. The CBC immediate rewards for all the possible actions are contained in Table 4. We penalize the decision maker for not giving full dose when the patient's CBC is normal. Moreover, if full dose is given when the patient's CBC is abnormal, the decision maker is penalized. The maximum reward/penalty is taken to 100/-100, respectively. Since in WT treatment, the tumor progression is usually measured once after the surgery (by the end of treatment period), we assign 0 immediate reward for
tumor progression and 0 terminal reward for CBC. We then define a linear terminal reward function for tumor progression which is given by

\[ r_{T+1}(\tau) = 100(1 - 2\tau), \tau = 0, 1/4, 1/2, 3/4, 1. \]  

(9)

The resulting optimal policy for our base case example is presented in Table 5. The structure of the policy is quite simple and follows the medical intuition. If at any decision epoch the patient's CBC is normal, \( f_{ull} \) dose should be selected. Meanwhile, at any decision epoch in the absorbing states, \( wait \) is the only available action. For the intermediate tumor progression states, 33\% should be selected if the patient’s CBC is \( abn_{ormal} \) for all decision epochs, except in the first decision epoch when the tumor progression is still considerably good (\( \tau = 1/4 \)).

| \( \tau/\tau' \) | \( 0 \) | \( 1/4 \) | \( 1/2 \) | \( 3/4 \) | \( 1 \) |
|---|---|---|---|---|---|
| \( 0 \) | 0 | 0 | 0 | 0 | 0 |
| \( 1/4 \) | 0 | 0.4 | 0.6 | 0 | 0 |
| \( 1/2 \) | 0 | 0 | 0.4 | 0.6 | 0 |
| \( 3/4 \) | 0 | 0 | 0 | 0.4 | 0.6 |
| \( 1 \) | 0 | 0 | 0 | 0 | 1 |

Table 2. Tumor progression state-action transition matrix for \( wait \) and 33\% action.

| \( \tau/\tau' \) | \( 0 \) | \( 1/4 \) | \( 1/2 \) | \( 3/4 \) | \( 1 \) |
|---|---|---|---|---|---|
| \( 0 \) | 0 | 0 | 0 | 0 | 0 |
| \( 1/4 \) | 0 | 0.7 | 0.3 | 0 | 0 |
| \( 1/2 \) | 0 | 0.6 | 0.4 | 0 | 0 |
| \( 3/4 \) | 0 | 0 | 0.5 | 0.5 | 0 |
| \( 1 \) | 0 | 0 | 0 | 0 | 1 |

Table 3. CBC state-action transition matrix for all actions.

| \( c' \) | \( c' \) | \( c' \) | \( c' \) |
|---|---|---|---|
| \( 1 \) | 0.9 | 1 | 0.8 | 0.2 |
| \( 2 \) | 0.8 | 0.2 | 2 | 0.5 |

Table 4. CBC immediate rewards for all actions.

| \( c \) | \( wait \) | \( 33\% \) | \( 66\% \) | \( full \) |
|---|---|---|---|---|
| \( 1 \) | -100 | -75 | -50 | 100 |
| \( 2 \) | 100 | 75 | 50 | -100 |

Table 5. Tumor progression state-action transition matrix for \( wait \) and 33\% action.

3.2. Sensitivity analysis

In this section, we investigate how changes in our model parameters affect the structure of the optimal policy. First, in subsection 3.2.1, we demonstrate how change in the immediate rewards for CBC affects the optimal policy (case 1). Then, we consider how slight increment in the effect of chemotherapy on tumor progression could affect the optimal policy (case 2).

3.2.1. Case 1. In contrast to the assumptions made on the immediate reward in section 3.1, we penalize the decision maker for selecting any dose of chemotherapy in an \( abn_{ormal} \) CBC state. Meanwhile, the decision maker receives a reward for administering chemotherapy when the patient’s CBC is \( n_{ormal} \), and penalized for taking action \( wait \). The immediate rewards used for this example are
contained in Table 6. We maintain all other parameters and assumptions used in our base case example.

As shown in Table 6, the structure of the optimal policy is particularly simple. Irrespective of the patient’s tumor progression state, the optimal action is full when the patient’s CBC is normal for all decision epochs. In addition, whenever the patient is in an abnormal CBC state, the optimal action is wait for all decision epochs and all tumor progression states. This agrees with the general chemotherapy guidelines of the Renal Tumor Study Group of the International Society of Paediatric Oncology (SIOP-RTSG) 2016 Umbrella Protocol [15].

3.2.2. Case 2. Here, we consider a scenario whereby the effect of chemotherapy on tumor progression states is slightly higher than initially assumed in the base case. We therefore increased the probability that the patient’s tumor progression state will increase by 0.1. Thus, the following transition probabilities are changed: \( P(0|1/4,33\%) = 0.7 \), \( P(1/4|1/2,33\%) = 0.6 \), \( P(1/2|3/4,33\%) = 0.5 \), \( P(0|1/4,66\%) = 0.8 \), \( P(1|4|1/2,66\%) = 0.7 \), \( P(1/2|3/4,66\%) = 0.6 \), \( P(0|1/4,F) = 0.9 \), \( P(1/4|1/2,F) = 0.8 \), \( P(1/2|3/4,F) = 0.7 \).

We maintain all other assumptions and parameters used in our base case example, including the state-action transition matrix for wait. We note that the structure of the optimal policy (Table 7) is similar to that of our base case example, except that the optimal action is wait in the first and second decision period when patient’s CBC state is abnormal and the tumor progression is in a considerably good state.

As can be noted in Tables 5-7, the optimal action is to give full dose whenever the clinical health state of the patient is normal. This is consistent with the clinical intuition that as much as the patient is in a good clinical health state, full dose chemotherapy should be initiated. Furthermore, we note that the proposed approach gives a policy that prioritizes treating the patient for Wilms tumor, but with a careful approach by initiating 33% dose when the patient’s clinical health state is abnormal.

4. Conclusion

With cancer being one of the major life threatening diseases worldwide, many patient's characteristics are considered during its treatment. In this study, we demonstrated the potential of using mathematical framework for solving complex sequential decision making problems in cancer chemotherapy treatment by proposing a finite horizon MDP model. Specifically, our model accommodates all patient's clinical factors that could be considered during chemotherapy treatment. This will enhance decision making during cancer chemotherapy treatment and can be personalized for each individual patient based on their profile. As a case study, we applied our model to Wilms tumor. Although we considered only one clinical factor, our results are consistent with clinical intuition and demonstrate the potential applicability of our model. In the future, we could apply our model to other types of cancers, such as leukemia, colorectal cancer, lymphoma, etc. We note that the problem of curse of dimensionality may arise as the patient clinical health state increases. This problem can be solved by approximate dynamic programming techniques [16].
Table 7. Optimal policy for case 2.

| Decision period | $\tau = 0$ | $\tau = 1/4$ | $\tau = 1/2$ | $\tau = 3/4$ | $\tau = 1$ |
|-----------------|-----------|-------------|-------------|-------------|-----------|
| $c$             | 1         | 2           | 3           | 4           | 1         |
|                 | 2         | 3           | 4           | 1           | 2         |
| $F$             | F         | F           | F           | F           | F         |
| $W$             | W         | W           | W           | W           | W         |

References

[1] Abbas Z and Rehman S 2018 An overview of cancer treatment modalities *Neoplasim* 1 139–57
[2] Iliadis A and Barbolosi D 2000 Optimizing drug regimens in cancer chemotherapy by an efficacy–toxicity mathematical model *Computers and Biomedical Research* 33 211–26
[3] Nanda S, Moore H and Lenhart S 2007 Optimal control of treatment in a mathematical model of chronic myelogenous leukemia *Mathematical Biosciences* 210 143–56
[4] de Pillis L G, Gu W, Fister K R, Head T, Maples K, Murugan A, Neal T and Yoshida K 2007 Chemotherapy for tumors: An analysis of the dynamics and a study of quadratic and linear optimal controls *Mathematical Biosciences* 209 292–315
[5] Ledzewicz U and Schättler H 2014 On optimal chemotherapy for heterogeneous tumors *Journal of Biological Systems* 22 177–97
[6] Schaefer A J, Bailey M D, Shechter S M and Roberts M S 2004 Modeling medical treatment using Markov decision processes *Operations Research and Health Care: A Handbook of Methods and Applications* ed M L Brandeau, F Sainfort and W P Pierskalla (Boston, MA: Springer US) pp 593–612
[7] Shechter S M, Bailey M D, Schaefer A J and Roberts M S 2008 The optimal time to initiate HIV therapy under ordered health states *Operations Research* 56 20–33
[8] Alagoz O, Maillart L M, Schaefer A J and Roberts M S 2007 Choosing among living-donor and cadaveric livers *Management Science* 53 1702–15
[9] Akhavan-Tabatabaei R, Sánchez D M and Yeung T G 2016 A Markov decision process model for cervical cancer screening policies in Colombia *Medical Decision Making* 37 196–211
[10] Kim M, Ghate A and Phillips M H 2009 A Markov decision process approach to temporal modulation of dose fractions in radiation therapy planning *Physics in Medicine & Biology* 54 4455
[11] Maass K and Kim M 2020 A Markov decision process approach to optimizing cancer therapy using multiple modalities *Mathematical Medicine and Biology: A Journal of the IMA* 37 22–39
[12] Bazrafshan N and Lotfi M M 2020 A finite-horizon Markov decision process model for cancer chemotherapy treatment planning: an application to sequential treatment decision making in clinical trials *Annals of Operations Research* 295 483–502
[13] Oostveen R M and Pritchard-Jones K 2019 Pharmacotherapeutic management of Wilms tumor: An Update *Pediatric Drugs* 21 1–13
[14] Lorenzo A and Lopes R I 2017 Recent advances in the management of Wilms’ tumor *F1000Research* 6 670
[15] Vujanić G M, Gessler M, Ooms A H A G, Collini P, Coulomb-l’Hermine A, D’Hooghe E, de Krijger R R, Perotti D, Pritchard-Jones K, Vokuhl C, van den Heuvel-Eibrink M M and Graf N 2019 Publisher Correction: The UMBRELLA SIOP–RTSG 2016 Wilms tumour pathology and molecular biology protocol (Nature Reviews Urology, (2018), 15, 11, (693-701), 10.1038/s41585-018-0100-3) *Nature Reviews Urology* 16 563
[16] Powell W B 2007 *Approximate Dynamic Programming: Solving the curses of dimensionality (Wiley series in probability and statistics)* (USA: Wiley-Interscience)