Reinforcement Learning and Fuzzy Logic Modelling for Personalized Dynamic Treatment

Marco Locatelli[1], Roberto Clemens Cerioli[1], Daniela Besozzi[1,2], Arjen Hommersom[3] and Fabio Stella[1,2]

[1]Department of Informatics, Systems and Communication, University of Milano-Bicocca, Milan, Italy
[2]Bicocca Bioinformatics Biostatistics and Bioimaging Centre - B4, University of Milano-Bicocca, Vedano al Lambro (MB), Italy
[3]Department of Computer Science, Open University, P.O. Box 2960, 6401DL Heerlen, The Netherlands

Abstract

The treatment of patients suffering from chronic diseases is a difficult problem to be tackled. Its complexity mainly originates from the following sources: the patient-specific response to the prescribed therapy, the impact of the interplay between disease and therapy on the quality of life of the patient and relatives, and the economic costs incurred by the healthcare system. Recently, there has been considerable interest in developing, studying, and applying artificial intelligence methods to diagnosis, prognosis and treatment personalization. This paper combines two techniques from artificial intelligence, namely fuzzy logic and reinforcement learning, to develop optimal dynamic treatment for patients suffering from a chronic disease. In this paper, we focus on cancer as a chronic disease and leverage a biologically validated fuzzy logic model from the literature. Different problem settings, of increasing complexity, are taken into account, presented and analyzed. Results of an extensive numerical experimental plan confirm the potential of non-myopic decision-making when treating chronic disease patients.

Keywords

Personalized dynamic treatment, Fuzzy Logic, Reinforcement Learning.

1. Introduction

Biomedical researchers, clinicians and healthcare practitioners are increasingly recognizing that the *one size fits all approach* is no longer acceptable [1, 2]. Indeed, biological interactions are known to happen at an individual level while this is not true for statistical interactions, which instead occur at a population level [3]. This heterogeneity stems from stochastic processes taking place at the molecular scale and inducing biological noise [4], which is not only well recognized as an indispensable trait in evolution [5] but also represents a critical feature that might be exploited in the identification of personalized treatments [6]. As highlighted in [7], it is commonly agreed upon that drawing conclusions, diagnosing, and making treatment decisions from population models may not be optimal for a specific patient. Thus, the scientific community is developing new approaches based on the rationale that each patient is unique.

Personalized medicine offers a tailored approach to healthcare by customizing medical treatments according to each patient's unique characteristics. A key component of personalized medicine are dynamic treatment regimes (DTRs), which involve sequences of decision rules adjusted over time depending on how the health condition of the patient evolves and the response to past treatments. The development of DTRs [8, 9, 10] relies on statistical methods, such as reinforcement learning and causal inference, to identify the optimal action, at each decision point, based on historical data. In recent years, statistical methods for optimizing DTRs have made significant progress. Techniques like Q-learning [11], A-learning [9], and structural nested models [10] allow researchers to estimate optimal treatment strategies from both observational and experimental data. These methods are designed to handle complex, multi-stage decision processes and account for time-varying confounders that can

HC@AIxIA 2024: 3rd AIxIA Workshop on Artificial Intelligence For Healthcare
m.locatelli60@campus.unimib.it (M. Locatelli)
0009-0000-2220-2023 (M. Locatelli); 0009-0003-8063-1311 (R. C. Cerioli); 0000-0001-5532-3059 (D. Besozzi); 0000-0003-0125-1680 (A. Hommersom); 0000-0002-1394-0507 (F. Stella)

affect both treatment decisions and outcomes. DTR algorithms have been successfully applied to develop personalized treatment plans for chronic diseases, including cancer, diabetes, anaemia, HIV, and various mental health disorders, as discussed in [12].

From a different yet complementary perspective, personalized medicine also benefits from the mathematical modeling and analysis of cellular processes, whose malfunctioning often lead to the onset of diseases. Though, the definition of models able to provide an accurate description of biological mechanisms still poses several challenges, which are mainly related to the effects of stochasticity, the paucity of precise quantitative data, and the technical limits in the observation or measurement of an organism as a whole. In this context, fuzzy logic [13] can be fruitfully exploited in the biomedical field to describe and analyze complex biological systems as well as multifactorial diseases. The most important features of fuzzy logic are the capacity to deal with uncertain data and loosely defined variables, not to mention the ease of interpretation that represent a fundamental characteristic in clinical scenarios.

To date, there are many applications of fuzzy logic as a means for clinical decision making within a personalized medicine approach. By way of example, in [14] an automatically defined fuzzy rule base was used to classify lung cancer patients based on transcriptomic data. In [15], a fuzzy classification model was exploited to aid in vasopressor administration in intensive care unit (ICU) patients: similar patients were grouped by a fuzzy clustering algorithm, and an ensemble fuzzy classifier was trained on ICU data to estimate the necessity of vasopressor administration. Nevertheless, these approaches fail to take into account the intricate dynamics of biological systems, which is due to a finely orchestrated mechanism of positive and negative feedbacks, and their descriptive power is thus limited. This kind of dynamics is usually analyzed using systems of ordinary differential equations (ODEs), whose downside is the need of well-defined crisp values for the model parameters. To bypass this issue, in [16], ODEs were complemented with fuzzy logic to predict the disease course of HIV patients taking into account the strength of their immune system. Fuzzy logic was also hybridized with other computational and statistical methods to make the most of them. For instance, in [17], fuzzy logic was integrated into causal inference to introduce fuzzy causal effect metrics able to take into account vague and imprecise data and, as a practical example, the approach was used to measure the impact of age and sodium intake on blood pressure.

In this work, we address the DTR problem by leveraging a combined framework based on fuzzy logic modeling and reinforcement learning. The paper is organized as follows. Section 2 gives the concept of personalized dynamic treatment, while reviewing the main approaches from the specialized literature. Dynamic fuzzy logic modelling, specifically tailored on biomedical processes, is introduced in Section 3, together with definitions and main concepts on reinforcement learning. Section 4 presents a proof of concept where these two methods are integrated to develop DTRs for optimizing the death of cancer cells by apoptosis. The results of three sets of numerical experiments are reported and commented on in Section 5. Conclusions and direction for future works are given in Section 6.

2. Personalized Dynamic Treatment

Managing chronic diseases, i.e., diseases which evolve over time with prolonged duration, requires designing and developing methods for assisting physicians to make personalized treatment decisions. In this work we informally define a personalized treatment decision as a treatment decision depending on the unique characteristics of the patient, with the goal of optimizing their long term outcome while reducing the risk of side-effects. Moreover, when treating chronic patients, the treatment decisions unfold temporally depending on the progression of the patient's illness state. In such a setting, dynamic treatment regimes (DTRs) offer a robust framework for making personalized treatment decisions over time by leveraging patient's characteristics and preferences, disease progression, as well as on the patient's response to past treatments.

Given the inherently sequential nature of chronic disease progression and treatment, reinforcement learning (RL) [18] offers an effective option for developing DTRs. Indeed, the ability of RL to solve sequential decision making problems naturally deals with the challenges of managing chronic diseases

with DTRs, as witnessed by several estimation methods developed for DTRs. As an example, Q-Learning [11] is a model-free approach that estimates the optimal DTR by learning the action-value function. A second approach is offered by A-Learning [9, 10], which directly estimates the advantage function, while offering efficiency gain when the baseline treatment is well-understood. Further advancements were made in [19], where regret functions are incorporated into a regression model for the final observed result, while in [20] a modification of the Q-learning approach was proposed, where both the observed rewards and the estimated loss due to sub-optimal actions are used. Following this, several machine learning methods were proposed in the specialized literature to approximate the Q-function.

It is also worth mentioning the great interest received by artificial neural networks (ANNs) [21], which resulted in the development of Deep Q-learning (DQL) [22], and its subsequent application to DTRs by means of feed-forward neural networks (FNNs) [23] or via convolutional neural networks (CNNs) to approximate the regret function [24]. More recently, adversarial networks were used to learn a policy mimicking successful treatments while avoiding unsuccessful ones [25]. Later on, the concept of pessimism and Bayesian machine learning (also Bayesian neural networks) methods were integrated to find the best treatment strategy [26].

An alternative approach aims to maximize the reward by learning the optimal policy or value directly. Inverse probability of treatment weighting (IPTW) was used in [27] and [28] to estimate the causal effect from observational data. This technique computes weights for each policy based on the value of covariates, thus allowing the estimation of the optimal decision rule as a weighted classification problem. Subsequently, modifications of this approach were proposed, mainly focused on the combination of Q-function or A-function approximations with the IPTW estimator [29, 30]. Deep learning techniques were also used to this extent: for instance, in [31] recurrent neural networks and multilayer perceptron networks were used to estimate the value function of a DTR. Over the years, other machine learning techniques have also been used, such as decision trees [32, 33] or methods focused on causality: [34] based its model on causal trees, while [35] and [36] overcame the problem of non-identifiability through causal-bounds and the use of instrumental variables, respectively.

Up to this point, the focus has been on the methods used in the standard DTR setting, while the problem of DTR has several facets. A main challenge is that of *competing outcomes*, i.e., the case when multiple and competing outcomes exist, which requires balancing treatment decisions over time [37, 38]. *Censored data*, where the full information about the outcome is not observed, further complicates the analysis, as does *missing data*, which lead to incomplete observations and potential biases [39]. An *infinite or indefinite time horizon* adds another layer of difficulty, as it requires considering the long-term effects of treatments beyond a finite period [40, 41]. Finally, challenges such as *time-to-event* analysis, which measures the duration until an event occurs [42], and *time-to-visit* optimization, which estimates the best timing for treatment interventions [43], further add complexity to developing effective DTRs.

To the best of the authors' knowledge, the only work using DTRs to optimize daily dosage adjustment of radiotherapy within a clinical decision support system is the one proposed in [44]. Nevertheless, many works validate their methods on data from real-world, though perhaps simplified, problems. The field of DTRs has seen recent growth, while at the current state of the art there is a significant research-practice gap, probably due to many of the presented methodologies being evaluated unevenly and because they often consider simplified problems when compared to those that would be found in real clinical settings. An additional limitation to the widespread adoption of these DTR optimization methodologies is the potential lack of acceptance by domain experts. To gain their trust, these methods need to provide interpretable results and clearly explain the decision-making process behind their treatment recommendation.

3. Methods

3.1. Fuzzy Logic Modelling of Biomedical Processes

Dynamic fuzzy models (DFMs) represent a mathematical formalism—based on the concepts of fuzzy logic [45, 46]—introduced in [47] to describe and analyze complex systems that consist of heterogeneous

components and are characterized by uncertainty, such as those related to biomedical contexts. DFMs can be straightforwardly used to simulate the evolution over time of the system state, therefore predicting its emergent behaviour in different conditions, such as physiological or pathological states of cellular systems and individuals. In this section we provide a brief description of DFMs, and refer the reader to [13, 48, 47, 49, 50] for further details.

The definition of a DFM requires the identification of a directed graph representing the system components $x_1, \ldots, x_n$ and their mutual interactions, which can correspond to either positive or negative regulations. Each component is formalized by means of a linguistic variable, that is, a list of linguistic terms together with the corresponding fuzzy sets and membership functions, defined over a proper universe of discourse. Formally, for each $i = 1, \ldots, n$, x_i is described by a linguistic variable $L_i = \{x_i, \mathcal{U}_i, \Lambda_i, M_i\}$, where:

- x_i is the name of the linguistic variable;
- $\mathcal{U}_i$ is the universe of discourse, that is, the range of values in which x_i has meaning;
- $\Lambda_i = \{\lambda_{i,1}, \ldots, \lambda_{i,P}\}$ is the set including the P linguistic terms that L_i can assume in $\mathcal{U}_i$;
- $M_i = \{\mu_{\lambda_{i,1}}, \ldots, \mu_{\lambda_{i,P}}\}$ is the set of P fuzzy sets in which the universe of discourse is partitioned, each one corresponding to a linguistic term appearing in Λ_i. Note that the symbol μ can be used to denote both the fuzzy set and its membership function: namely, a fuzzy set $\mu_{\lambda_{i,p}}$, $p = 1, \ldots, P$, is uniquely characterized by a membership function $\mu_{\lambda_{i,p}} : \mathcal{U}_i \to [0, 1]$, which maps every value $u \in \mathcal{U}_i$ to the so-called degree of membership of u to the fuzzy set $\mu_{\lambda_{i,p}}$.

The interactions among the system components are formalized as fuzzy rules, which are conditional statements consisting in an antecedent and a consequent written in the form "IF x_i IS $\lambda_{i,p}$ THEN x_j IS $\lambda_{j,q}$", where $\lambda_{i,p}$ and $\lambda_{j,q}$ are linguistic terms associated with variables x_i and x_j, respectively. In general, the antecedent of a fuzzy rule can contain any logic expression involving two or more linguistic variables connected by the fuzzy operators AND, OR, NOT. Figure 1 shows an example of fuzzy rules (panel B) and fuzzy sets (panel C) associated with two linguistic variables, x_i and x_k, which control a third variable, x_j (panel A).

The set of linguistic variables and fuzzy rules appearing in a DFM constitute a Fuzzy Inference System (FIS). The simulation of DFMs can be performed by means of Simpful [50], a Python library designed to easily define and analyse FISs based on Mamdani or Takagi–Sugeno fuzzy inference engines [46]. In this work, we rely on the 0-order Sugeno fuzzy reasoning, as implemented in Simpful. To determine the next state of any variable, this method calculates a weighted aggregation of the output values produced by all fuzzy rules having that variable in their consequent and being satisfied at the current time step [51]. To exemplify this step, in Figure 1 (panel D) we show how the next value σ of variable x_j will be calculated according to the fuzzy rules R_1, R_2 and R_4, R_5 that—given the current values of variables x_i and x_k, respectively—are the only rules whose degree of satisfaction (denoted by $w_1, \ldots, w_5$) is not equal to zero. The degree of satisfaction of a fuzzy rule is a number in $[0, 1]$ to which the current values of the (input) linguistic variables match the antecedent. In particular, for simple rules where no fuzzy operators appear in the antecedent (as those shown in Figure 1, Panel B), the degree of satisfaction is equal to the membership function value so that $w_1 = \mu_{\textit{Less-Functional}}(x_i) = 0.25$, $w_2 = \mu_{\textit{Medium-Functional}}(x_i) = 0.75$ and $w_3 = \mu_{\textit{More-Functional}}(x_i) = 0$ for $x_i = 4$, while $w_4 = \mu_{\textit{Slow}}(x_k) = 0.1$ and $w_5 = \mu_{\textit{Fast}}(x_k) = 0.9$ for $x_k = 0.8$ (see Figure 1, Panel D). The state of all variables is then updated, and the process is iterated until a maximum time step is reached. Panel E in Figure 1 shows the overall process of calculating the output of fuzzy rules, which includes fuzzification, fuzzy inference and defuzzification [46].

3.2. Reinforcement Learning

Reinforcement learning (RL) is a machine learning methodology where an agent learns to make decisions by trial-and-error interactions with its environment. In general, solving an RL task means learning an optimal way to choose the set of actions, or learning an optimal policy, so as to maximise the cumulative discounted outcome, denoted as R_t. This cumulative reward is calculated as the sum of discounted

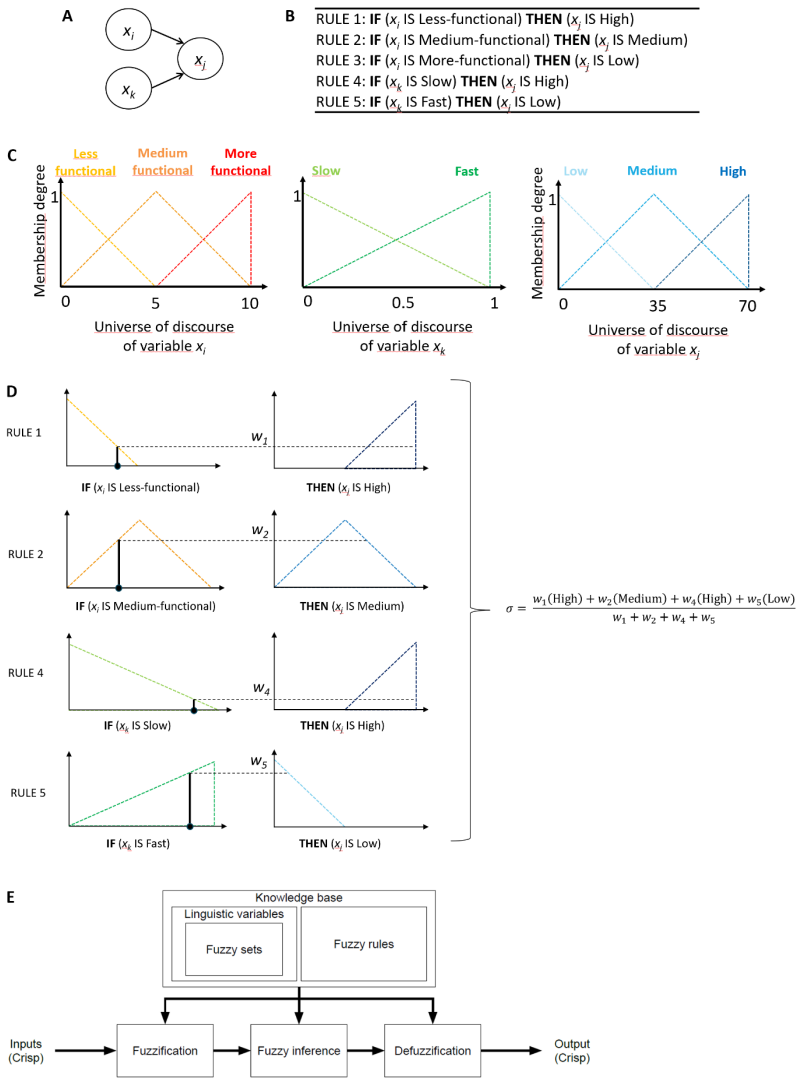


Figure 1: Panel A: graph representation of the interactions among linguistic variables x_i, x_k, x_j. Panel B: list of fuzzy rules associated with variable x_j, which is negatively regulated by x_i and x_k. Panel C: fuzzy sets, linguistic terms and universe of discourse of variables x_i, x_k, x_j. Panel D: weighted aggregation of the output value of variable x_j as produced by the four fuzzy rules whose degree of satisfaction $(w_1, \ldots, w_5)$ is not zero, namely: variable x_i, whose current value is 4, has a membership degree greater than zero only to the fuzzy sets Less-functional and Medium-functional (rules R_1, R_2), while variable x_j, whose current value is 0.8, has a membership degree greater than zero to both fuzzy sets Slow and Fast (rules R_4, R_5). Panel E: comprehensive definition and analysis framework of FISs.

future rewards, which can be represented as

$$R_t = \sum_{k=0}^{T-t-1} \gamma^k \, Y_{t+k+1}, \tag{1}$$

where $\gamma \in [0, 1]$ is the discounted factor, Y_t the outcome at t, and T is the final time step.

In the context of DTR, a policy is defined as a sequence of decision rules $\boldsymbol{d} = \{d_1, d_2, ..., d_T\}$ where each decision rule d_t corresponds to a treatment decision at time t. A policy maps from the history space $\mathcal{H}_t$ to the set of possible treatment actions $\mathcal{A}_t$. An optimal policy, denoted as d^*, can be defined as a

policy that returns the largest (or equal) cumulative outcome in comparison to other policies. Formally:

$$d^* = \arg\max_d \mathbb{E}\left[\sum_{k=0}^{T-t-1} \gamma^k \, Y_{t+k+1}\right]. \tag{2}$$

Another fundamental concept is the Q-function (or the action-value function), which provides the expected cumulative reward starting from a given patient history $\boldsymbol{h}_t$ at time t, after taking an action a_t and thereafter following the policy d. It essentially quantifies the value of taking action a_t given the patient history $\boldsymbol{h}_t$ and following the policy d at time t. Formally, the Q-function is defined as:

$$Q_t^d(\boldsymbol{h}_t, a_t) = \mathbb{E}\left[\sum_{k=0}^{T-t-1} \gamma^k \, Y_{t+k+1} | \boldsymbol{H}_t = \boldsymbol{h}_t, A_t = a_t\right]. \tag{3}$$

Subsequently, the optimal Q-function can be defined as a function that maximizes the expected return for each history-action pair under any policy d, formally:

$$Q_t^*(\boldsymbol{h}_t, a_t) = \max_{d_t} Q_t^d(\boldsymbol{h}_t, a_t). \tag{4}$$

It is possible to define Equation 4 recursively without making any reference to a particular policy. This key property of RL is given by the Bellman optimality equation [52] and can be expressed formally as follows:

$$Q_t^*(\boldsymbol{h}_t, a_t) = \mathbb{E}[Y_{t+1} + \gamma \max_{a_{t+1}} Q_{t+1}^*(\boldsymbol{h}_{t+1}, a_{t+1}) | \boldsymbol{h}_t, a_t]. \tag{5}$$

Q-learning [11, 53] is a popular RL algorithm used to learn the optimal action-value function. However, standard Q-learning can struggle in high-dimensional or continuous spaces, so approximation techniques, such as linear regression, are often employed to estimate Q-values. Suppose we use a Q-function approximation $Q_t(\boldsymbol{h}_t, a_t; \theta_t)$, where θ_t are the model parameters. By iteratively approximating Q_t for each time step, it is possible to determine the optimal treatment policy d^*.

4. Optimizing the apoptotic death of cancer cells

As a proof of concept of the effectiveness of combining fuzzy logic modelling with reinforcement learning, we consider a DFM describing cell death processes in oncogenic K-ras cancer cells as a result of progressive glucose depletion. Ras mutations are among the most frequent alterations in human cancers, especially in solid tumors, and are linked to marked metabolic rewiring, which has been recognized as one of the numerous hallmarks of cancer [54]. Oncogenic K-ras cancer cells have been considered undruggable for long, and available therapeutic strategies—especially the single-agent targeted ones, most often leading to acquired resistance—are still far from being fully curative [55]. Huge research efforts are thus in progress to discover effective treatments and improve the outcome of patients affected by Ras-mutant malignancies [56].

The DFM of oncogenic K-ras cancer cells, defined in [47], takes into account the main cellular components involved in energy production (e.g. glucose and glycolysis), different mitochondrial processes (e.g. the mitochondrial potential variation DeltaPsi and the activity of the mitochondrial Complex I (CI)), several processes and proteins involved in cellular adhesion, in the regulation of cell death (e.g. unfolded protein response (UPR)) and in survival mechanisms (e.g. protein kinase A (PKA)), as well as the phenotypes related to apoptosis and necrosis. Three variables in this DFM represent the input of the system: (1) glucose, whose consumption is modeled by a custom update function that drives the dynamical update of the system state; (2) Ras-GTP, which is always set to the High value to mimic the hyperactivation of K-ras in cancer cells; (3) PKA, which mainly regulates cancer cells fate: it is set to either the High or the Low state to mimic the ability of cancer cells to survive or die under glucose starvation, respectively. We refer to [47] for further details on the relevance and role of all model components.

The model was validated through an extensive comparison with experimental data measured on the MDA-MB-231 cancer cell line (human K-ras-mutated breast cancer) and on NIH3T3 K-ras cells (mouse fibroblasts transformed by oncogenic K-ras expression) [47]. In previous works [47, 49], global optimization algorithms were also combined with the simulation of this DFM to identify a set of treatments able to maximize cancer cell death by apoptosis—and possibly minimize necrotic processes too—while minimizing the total number of administered drugs. This approach allowed to identify treatments that were already validated in the literature, as well as novel potential therapeutic strategies.

The DFM of oncogenic K-ras cancer cell represents a valid case study for the novel methodology of personalized dynamic treatment that we are introducing with this work. Indeed, despite considering only processes occurring at the cellular level, it allows us to take into account the possible effects of biological noise that, by inducing a high heterogeneity among cells, can act as a proxy to mimic the status of patients at the individual level. In addition, it holds the added value of previous experimental validation for various drug administrations, some of which have also been considered in this work. Namely, here we focus on three different types of treatment:

- **Complex I inhibition:** disrupting Complex I in these cells generates harmful reactive oxygen species, leading to oxidative stress and triggering apoptotic pathways;
- **UPR activation:** UPR is a cellular mechanism triggered by significant stress, such as nutrient deprivation or oxidative damage. Prolonged activation of this process (i.e. "UPR is High") leads to apoptotic cell death;
- **Combination of inhibition and activation:** this will simulate a condition characterized by a high level of UPR activation (75%) and a lower level of concurrent Complex I inhibition (25%). This combination was not experimentally tested but might have the effect of increasing apoptosis.

In this study we applied a DTR approach to optimize the treatment strategy for inducing cancer cell death. In particular, a treatment strategy is represented as a policy, that is, a set of decision rules that guide treatment choices over multiple stages based on the evolution of the disease. At each decision point t, a treatment a_t is selected from a set of possible treatments. These decisions are determined by the policy, which maps the complete history of the disease progression to a treatment with the objective of maximising the desired outcome. In this case, the desired outcome is the induction of apoptosis in cancer cells. To achieve this, we first generated a dataset by simulating cancer cell behaviour using the fuzzy cellular model over a 72-hour period, with treatment decisions made each 12 hours. The dataset includes:

- the treatment assignment for each decision step. The choice of each treatment was based on three covariates: PKA, CI and UPR. The static rules used to select the treatment are based on previous results [47];
- the values of the three covariates at the beginning of the simulation and each decision step;
- the resulting state of the system at each decision step and at the end of the interval.

Using the generated dataset, we applied a DTR method to determine the optimal treatment strategy. The DTR algorithm considers the three covariates at each decision point and recommends the best treatment option based on the current state of the system. The recurring nature of the treatment assignment allows for adaptive decision making, accounting for the evolving state of the system (Figure 2).

To evaluate the effectiveness of the proposed approach we designed three numerical experiments with different goals:

1. **Maximizing apoptosis in cancer cells**: In the first experiment the goal is to maximize cancer cell death via apoptosis. We apply the DTR algorithm to our fuzzy cellular model, while using the current apoptosis value as the primary outcome measure.
2. **Balancing apoptosis and necrosis**: The second experiment aims to find a balanced treatment strategy that not only maximizes apoptosis but also minimizes uncontrolled cell death (necrosis), which can lead to undesirable inflammation effects. To achieve this, we modify our DTR algorithm to account for both apoptosis and necrosis in the outcome.

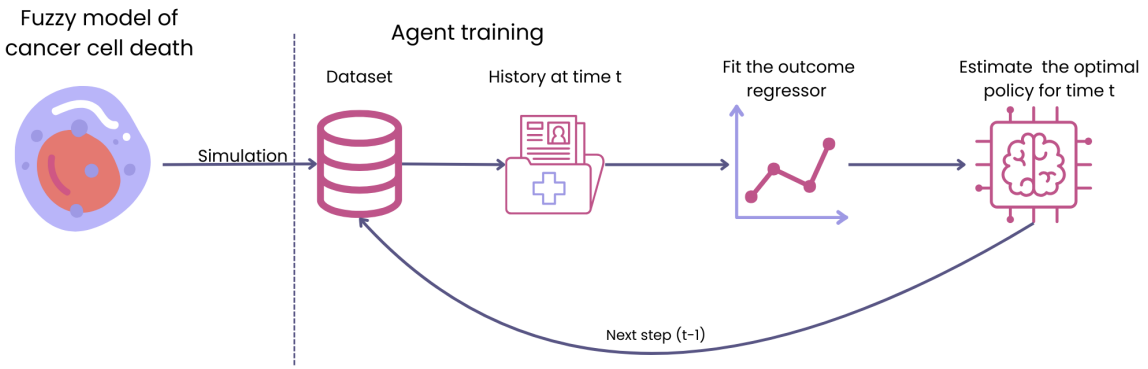


Figure 2: A fuzzy cellular model was employed to generate the dataset. At each time step t, the historical data of samples up to time t were used to fit a regression model. This model was then applied to identify the optimal action at time t. The process is repeated to estimate the best policy for each step, guiding the agent to the next step $(t-1)$.

3. **Impact of treatment intervals on DTR effectiveness**: The goal of this experiment is to investigate how different treatment intervals affect the performance of DTR strategies. We vary the decision-making intervals from the original 12 hours, used in the first two types of experiments, to include 6-hour and 24-hour intervals, generating a new dataset for each setting. For each interval setting, we run the DTR algorithm to determine the optimal strategy for maximizing apoptosis (as in the first experiment).

5. Experiments

5.1. Experimental setup

This section presents and discusses the results of the numerical experiments performed under three different experimental settings to evaluate the effectiveness of the proposed DTR approach[1]. For each experimental setting, we generated 30,000 samples and evaluated the results on a cohort of 50 samples, enabling direct comparison across different treatment strategies. The efficacy of the DTR was specifically assessed against static strategies, including the consistent inhibition of Complex I, the continuous activation of the UPR and the continuous usage of the combined treatment. The values of apoptosis and necrosis are derived from the fuzzy model, where the outcome is expressed as a fuzzy measure: for each individual sample, the value of apoptosis and necrosis will always range between 0 and 1. Thus, when considering a dataset of 50 samples, the maximum possible cumulative value for both apoptosis and necrosis is 50.

In particular, for each experimental setting, the available data consists of a set of finite horizon trajectories:

$$\{\text{PKA}, W_0, A_0, Y_0, \ldots, W_5, A_5, Y_5, Y_6\},$$

where each W_t consists of $\{\text{CI}_t, \text{UPR}_t\}$, Y_t is the value of the apoptosis at each decision step and Y_6 is the value of apoptosis after 72 hours. In order to maximize a sum of numerical rewards, we employed a recursive form of Q-learning, with $Q_t(\boldsymbol{H}_t, A_t)$ predicting:

$$\hat{R}_t = Y_{t+1} + \gamma \max_{a_{t+1}} \hat{Q}_{t+1}(\boldsymbol{h}_{t+1}, a_{t+1}), \tag{6}$$

where $\boldsymbol{h}_{t+1} = \{\text{pka}, w_0, a_0, y_0, \ldots, a_t, w_{t+1}, y_{t+1}\}$ and $\hat{Q}_{t+1}$ is the estimator of the Q-values. To obtain the estimator $\hat{Q}_t$ we utilize the Random Forest Regressor [57] for fitting Q_t backward and get $\{\hat{Q}_5, \hat{Q}_4, \ldots, \hat{Q}_0\}$. It is important to note that the performance of our model is affected by the choice of hyperparameters:

[1]The code is provided at: https://github.com/mLoca/FuzzyReinforcement.

- γ, the discount factor, quantifies how much importance we give to future rewards;
- *maximum tree depth*, a parameter that controls the complexity of individual decision trees and avoids overfitting.

The optimal values for the discount factor and the maximum tree depth were determined through a process of hyperparameter optimization.

5.2. Experimental results

5.2.1. Experiment 1: Maximizing Apoptosis in Cancer Cells

This experimental setting was designed to identify the most effective treatment strategy when the goal is to maximize apoptosis in cancer cells, an indicator that the treatment is successful to target cancer cells, a primary goal in the field of cancer treatment. The discount factor was set to $\gamma = 0.95$, while *max depth* was set to 5.

The results show that the proposed DTR approach is more effective than static treatment strategies. Indeed, DTR achieves significantly higher levels of apoptosis (Table 1) than all static treatment strategies. Furthermore, DTR shows a controlled and progressive increase in apoptosis rates (Figure 3), while also maintaining a high level of apoptosis throughout the 72-hour horizon. This indicates that DTR not only achieves optimal results at the end of the treatment horizon but also carefully manages the intermediate stages of the treatment, ensuring an effective therapeutic response.

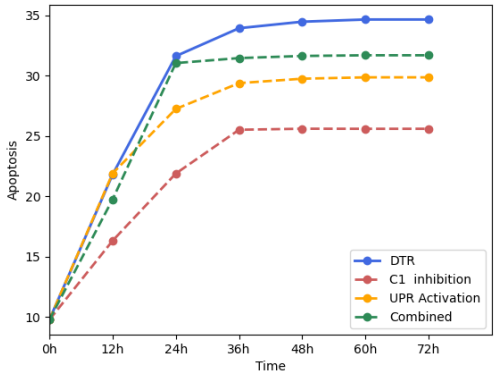


Figure 3: Graphical representation of the sum of apoptosis values across 50 samples over time for each treatment strategy

	Final Apoptosis	Mean (SD)
DTR	34.63	0.693 (0.066)
CI inhibition	25.66	0.513 (0.081)
UPR activation	30.09	0.611 (0.161)
Combined	31.54	0.631 (0.025)

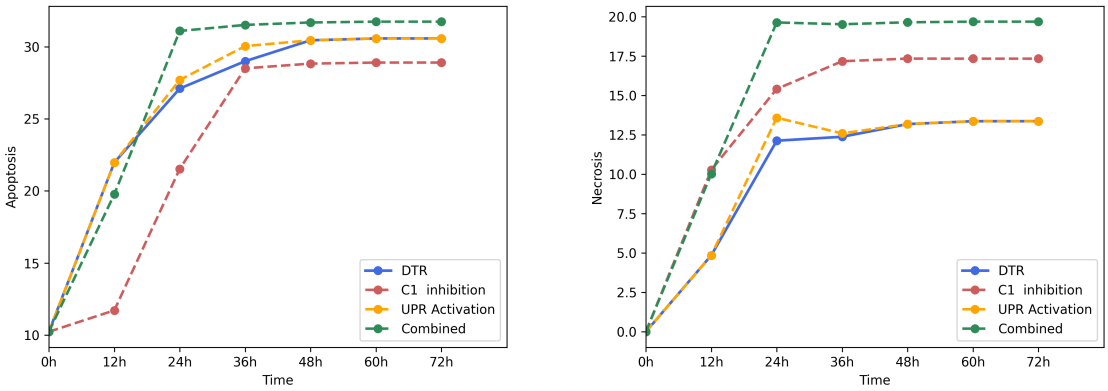
Table 1: Summary of final apoptosis values and corresponding mean (SD) for a sample of 50 data

By progressively increasing the apoptotic response over the 72-hour period, DTR demonstrates its capacity to balance short-term objectives with long-term goals, providing optimal results at every stage and maintaining high levels of apoptosis throughout the treatment process. This continuous progression highlights DTR's capacity to optimize the overall outcome without any compromise in efficacy at any point.

5.2.2. Experiment 2: Balancing Apoptosis and Necrosis

The experimental setting aims to identify the optimal strategy when the goal is to simultaneously maximize apoptosis and minimize necrosis. In the context of cancer therapy, the objective of maximizing apoptosis while minimizing necrosis is to improve outcomes by reducing the incidence of unwanted side effects.

In this context, the outcomes are Y_t^{apo} and Y_t^{nec}, denoting apoptosis and necrosis at time t, respectively. Two sets of Q-value estimators $\{\hat{Q}_t^{apo}, \hat{Q}_t^{nec}\}$ were used to estimate these outcomes. Specifically, the



(a) (b)

Figure 4: Graphical representation of the sum of apoptosis (a) and necrosis (b) values across 50 samples over time for each treatment strategy

	Apoptosis	Necrosis	Mean(SD) Apoptosis	Mean(SD) Necrosis
DTR	30.58	13.37	0.612 (0.164)	0.267 (0.172)
C1 inhibition	28.91	17.34	0.578 (0.025)	0.347 (0.006)
UPR activation	30.58	13.37	0.612 (0.164)	0.267 (0.172)
Combined	31.75	19.69	0.635 (0.026)	0.394 (0.022)

Table 2: Summary of final apoptosis and necrosis values and corresponding mean (SD) for a sample of 50 data

predicted values for apoptosis and necrosis are:

$$\hat{R}_t^{apo} = Y_{t+1}^{apo} + \gamma \max_{a_{t+1}} \hat{Q}_{t+1}^{apo}(\boldsymbol{h}_{t+1}, a_{t+1}) \text{ for apoptosis,}$$

$$\hat{R}_t^{nec} = Y_{t+1}^{nec} + \gamma \max_{a_{t+1}} \hat{Q}_{t+1}^{nec}(\boldsymbol{h}_{t+1}, a_{t+1}) \text{ for necrosis,}$$

where $\boldsymbol{h}_{t+1} = \{\text{pka}, w_0, a_0, y_0^{apo}, y_0^{nec}, \ldots, a_t, w_{t+1}, y_{t+1}^{apo}, y_{t+1}^{nec}\}$. The optimal policies are identified through the maximisation of the difference between apoptosis and necrosis:

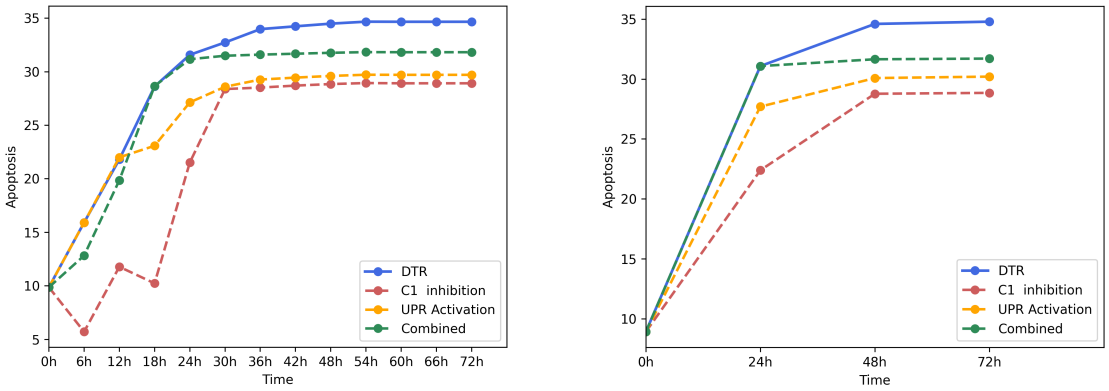
$$\hat{\pi}_t(\boldsymbol{h}_t) = \arg\max_{a_t} \hat{Q}_t^{apo}(\boldsymbol{h}_{t+1}, a_{t+1}) - \hat{Q}_t^{nec}(\boldsymbol{h}_{t+1}, a_{t+1}).$$

The results of this experiment, summarised in Figure 4 and Table 2 (with $\gamma = 0.9$ and *max depth* set to 5), show that DTR fails to identify a treatment combination which is better than the static UPR activation strategy.

However, CI inhibition leads to lower apoptosis and higher necrosis, while the combined treatment produces comparable apoptosis but increased necrosis, indicating a less favorable balance. In conclusion, these results demonstrate that DTR is not simply attempting combinations in an effort to improve the outcomes. Instead, it identifies that the static strategy of UPR activation remains the optimal approach. This underscores the algorithm's capacity to make well-informed decisions based on the data, thereby further confirming its potential for optimizing treatment strategies in cancer therapy.

5.2.3. Experiment 3: Impact of Treatment Intervals

The objective of this experimental setting is to evaluate the sensitivity of treatment strategies with respect to treatment assignment frequency. This experimental setting provides valuable information to



(a) (b)

Figure 5: Cumulative apoptosis in 50 samples over time for each treatment strategy. (a) Sum of apoptosis level with treatment administered every 6 hours. (b) Sum of apoptosis level with treatment administered every 24 hours.

optimize treatment regimens by investigating the effect of treatment frequency on their overall efficacy and behavior throughout the treatment horizon.

Specifically, we used $\gamma = 0.9$ and *max depth* set to 4 for the 6-hourly strategy, while $\gamma = 0.95$ and *max depth* set to 5 were used for the daily (24-hourly) allocation strategy. As shown in Figure 5, the outcomes of the DTR strategies at different frequencies are quite similar, with the 6-hourly treatment assignment generating an average result of 0.693 ± 0.070 of apoptosis value, and the daily assignment inducing 0.695 ± 0.065. The DTR strategy outperforms static treatment strategies in terms of efficacy and consistency across both dosing schedules (6-hourly and 24-hourly). The highest levels of cumulative apoptosis are achieved and maintained throughout the 72-hour. In general, all the strategies show comparable overall efficacy to 6-hourly dosing by 72 hours.

6. Conclusions and Future work

In this study, we explored the integration of fuzzy logic modeling and reinforcement learning to enhance personalized dynamic treatment regimes (DTRs), thereby offering a tailored approach to therapeutic treatments based on individual disease progression. This integration provides a robust framework for addressing the complexities of chronic disease management: fuzzy logic captures uncertainty and models complex biological systems, while reinforcement learning dynamically optimizes treatment plans over time. Our findings highlight the potential of combining these advanced computational techniques to improve treatment outcomes through non-myopic decision-making, leading to more effective and adaptive treatment strategies.

This work represents a significant advancement in the field of DTRs and a promising extension of the current state of the art. In contrast to conventional DTR approaches, our methodology employs dynamic fuzzy logic-based models (DFMs) to simulate how the biological processes involved in the onset and progression of diseases change over time, thereby enhancing the interpretability of disease-related mechanisms and their corresponding emergent behaviors at both local (cellular) and global (organism) levels. Subsequently, the reinforcement learning component maximizes outcomes through treatment decisions informed by the evolving system state. This novel combination of DFMs with reinforcement learning supports more adaptive and biologically informed decision-making, thereby offering a potential pathway for future validation in diverse clinical settings.

Despite these promising advancements, several limitations must be acknowledged. A key challenge lies in the limited range of treatments that were explored in the context of programmed death of cancer

cells. Expanding this approach to include additional drugs from the literature would provide a more comprehensive perspective. Even more important, it would be valuable to examine how the treatment strategy evolves in response to different glucose depletion dynamics. Moreover, while our models have shown efficacy in controlled experimental environments, their effectiveness in real-world clinical settings remains unverified and requires further validation. It is also worth stating that, although the framework demonstrated favorable outcomes with a synthetic dataset, the adequacy of the sample size for practical applications remains uncertain. Future research should determine the optimal sample size required to ensure the efficacy of the model in a healthcare environment, recognizing that small datasets may not entirely capture the complexity needed for reliable treatment personalization.

In conclusion, this research establishes a foundation for a more personalized approach to chronic disease management, emphasizing the critical role of individual characteristics in guiding treatment decisions. By addressing the current limitations and continuously refining these methodologies, we can advance toward more effective and tailored healthcare strategies.

Acknowledgments

This work was supported by Fondazione Regionale per la Ricerca Biomedica (Regione Lombardia), Project ERAPERMED2022-258, GA 779282. In particular the PhD Fellowship of M. Locatelli was funded by the MG-PerMed research project: *"Personalising myasthenia gravis medicine: from "one-fits-all" to patient-specific immunosuppression"* (ERAPERMED2022-258, GA 779282).

This work was partially supported by the MUR under the grant "Dipartimenti di Eccellenza 2023-2027" of the Department of Informatics, Systems and Communication of the University of Milano-Bicocca, Italy, and by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, converted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (ANTHEM). This work reflects only the authors' views and opinions; neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

References

[1] M. R. Kosorok, E. B. Laber, Precision medicine, Annu Rev Stat Appl 6 (2019) 263–286.
[2] D. Bzdok, G. Varoquaux, E. W. Steyerberg, Prediction, not association, paves the road to precision medicine, JAMA Psychiatry 78(2) (2021) 127–128.
[3] J. H. Moore, S. M. Williams, Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis, Bioessays 27(6) (2005) 637–46.
[4] A. Eldar, M. B. Elowitz, Functional roles for noise in genetic circuits, Nature 467 (2010) 167–173.
[5] H. Kitano, Biological robustness, Nature Reviews Genetics 5 (2004) 826–837.
[6] E. Sejdić, L. A. Lipsitz, Necessity of noise in physiology and medicine, Computer Methods and Programs in Biomedicine 111 (2013) 459–470.
[7] F. Melograna, Z. Li, G. Galazzo, N. van Best, M. Mommers, J. Penders, F. Stella, K. Van Steen, Edge and modular significance assessment in individual-specific networks., Scientific Reports 13 (2023).
[8] P. W. Lavori, R. Dawson, A design for testing clinical strategies: biased adaptive within-subject randomization, Journal of the Royal Statistical Society: Series A (Statistics in Society) 163 (2000) 29–38.
[9] S. A. Murphy, Optimal dynamic treatment regimes, Journal of the Royal Statistical Society Series B: Statistical Methodology 65 (2003) 331–355.
[10] J. M. Robins, Optimal structural nested models for optimal sequential decisions, in: Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data, Springer, 2004, pp. 189–326.
[11] C. J. Watkins, P. Dayan, Q-learning, Machine learning 8 (1992) 279–292.

[12] C. Yu, J. Liu, S. Nemati, G. Yin, Reinforcement learning in healthcare: A survey, ACM Computing Surveys (CSUR) 55 (2021) 1–36.
[13] P. Cazzaniga, S. Spolaor, C. Fuchs, M. S. Nobile, D. Besozzi, et al., Fuzzy logic for knowledge-driven and data-driven modeling in biomedical sciences, in: B. Carpentieri, P. Lecca (Eds.), Big Data Analysis and Artificial Intelligence for Medical Sciences, John Wiley & Sons, 2024, p. 17.
[14] N. Potie, S. Giannoukakos, M. Hackenberg, A. Fernandez, On the need of interpretability for biomedical applications: Using fuzzy models for lung cancer prediction with liquid biopsy, in: 2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE, 2019, pp. 1–6.
[15] C. M. Salgado, S. M. Vieira, L. F. Mendonça, S. Finkelstein, J. M. Sousa, Ensemble fuzzy models in personalized medicine: Application to vasopressors administration, Engineering Applications of Artificial Intelligence 49 (2016) 141–148.
[16] H. Zarei, A. V. Kamyad, A. A. Heydari, Fuzzy modeling and control of HIV infection, Computational and Mathematical Methods in Medicine 2012 (2012) 893474.
[17] A. Saki, U. Faghihi, Integrating fuzzy logic with causal inference: Enhancing the Pearl and Neyman-Rubin methodologies, 2024. arXiv:2406.13731.
[18] R. S. Sutton, Reinforcement learning: an introduction, A Bradford Book (2018).
[19] R. Henderson, P. Ansell, D. Alshibani, Regret-regression for optimal dynamic treatment regimes, Biometrics 66 (2010) 1192–1201.
[20] X. Huang, S. Choi, L. Wang, P. F. Thall, Optimization of multi-stage dynamic treatment regimes utilizing accumulated data, Statistics in medicine 34 (2015) 3424–3443.
[21] Y. Bengio, I. Goodfellow, A. Courville, Deep learning, volume 1, MIT press Cambridge, MA, USA, 2017.
[22] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al., Human-level control through deep reinforcement learning, nature 518 (2015) 529–533.
[23] Y. Liu, B. Logan, N. Liu, Z. Xu, J. Tang, Y. Wang, Deep reinforcement learning for dynamic treatment regimes on medical registry data, in: 2017 IEEE international conference on healthcare informatics (ICHI), IEEE, 2017, pp. 380–385.
[24] S. Liang, W. Lu, R. Song, Deep advantage learning for optimal dynamic treatment regime, Statistical theory and related fields 2 (2018) 80–88.
[25] L. Wang, W. Yu, X. He, W. Cheng, M. R. Ren, W. Wang, B. Zong, H. Chen, H. Zha, Adversarial cooperative imitation learning for dynamic treatment regimes, in: Proceedings of The Web Conference 2020, 2020, pp. 1785–1795.
[26] Y. Zhou, Z. Qi, C. Shi, L. Li, Optimizing pessimism in dynamic treatment regimes: A bayesian learning approach, in: International Conference on Artificial Intelligence and Statistics, PMLR, 2023, pp. 6704–6721.
[27] B. Zhang, A. A. Tsiatis, M. Davidian, M. Zhang, E. Laber, Estimating optimal treatment regimes from a classification perspective, Stat 1 (2012) 103–114.
[28] Y. Zhao, D. Zeng, A. J. Rush, M. R. Kosorok, Estimating individualized treatment rules using outcome weighted learning, Journal of the American Statistical Association 107 (2012) 1106–1118.
[29] B. Zhang, A. A. Tsiatis, E. B. Laber, M. Davidian, A robust method for estimating optimal treatment regimes, Biometrics 68 (2012) 1010–1018.
[30] B. Zhang, M. Zhang, C-learning: A new classification framework to estimate optimal dynamic treatment regimes, Biometrics 74 (2018) 891–899.
[31] Y. Zhang, M. van der Schaar, Gradient regularized v-learning for dynamic treatment regimes, Advances in Neural Information Processing Systems 33 (2020) 2245–2256.
[32] Y. Tao, L. Wang, D. Almirall, Tree-based reinforcement learning for estimating optimal dynamic treatment regimes, The Annals of Applied Statistics 12 (2018) 1914.
[33] Y. Sun, L. Wang, Stochastic tree search for estimating optimal dynamic treatment regimes, Journal of the American Statistical Association 116 (2021) 421–432.
[34] T. Blumlein, J. Persson, S. Feuerriegel, Learning optimal dynamic treatment regimes using causal tree methods in medicine, in: Machine Learning for Healthcare Conference, PMLR, 2022, pp.

146–171.
[35] J. Zhang, Designing optimal dynamic treatment regimes: A causal reinforcement learning approach, in: International Conference on Machine Learning, PMLR, 2020, pp. 11012–11022.
[36] S. Chen, B. Zhang, Estimating and improving dynamic treatment regimes with a time-varying instrumental variable, Journal of the Royal Statistical Society Series B: Statistical Methodology 85 (2023) 427–453.
[37] D. J. Lizotte, M. H. Bowling, S. A. Murphy, Efficient reinforcement learning with multiple reward functions for randomized controlled trial analysis., in: ICML, volume 10, 2010, pp. 695–702.
[38] E. B. Laber, D. J. Lizotte, B. Ferguson, Set-valued dynamic treatment regimes for competing outcomes, Biometrics 70 (2014) 53–61.
[39] Y.-Q. Zhao, R. Zhu, G. Chen, Y. Zheng, Constructing dynamic treatment regimes with shared parameters for censored data, Statistics in Medicine 39 (2020) 1250–1263.
[40] A. Ertefaie, R. L. Strawderman, Constructing dynamic treatment regimes over indefinite time horizons, Biometrika 105 (2018) 963–977.
[41] W. Zhou, R. Zhu, A. Qu, Estimating optimal infinite horizon dynamic treatment regimes via pt-learning, Journal of the American Statistical Association 119 (2024) 625–638.
[42] X. Wang, H. Lee, B. Haaland, K. Kerrigan, S. Puri, W. Akerley, J. Shen, A matching-based machine learning approach to estimating optimal dynamic treatment regimes with time-to-event outcomes, Statistical Methods in Medical Research 33 (2024) 794–806.
[43] W. Hua, H. Mei, S. Zohar, M. Giral, Y. Xu, Personalized dynamic treatment regimes in continuous time: a bayesian approach for optimizing clinical decisions with timing, Bayesian Analysis 17 (2022) 849–878.
[44] D. Niraula, W. Sun, J. Jin, I. D. Dinov, K. Cuneo, J. Jamaluddin, M. M. Matuszak, Y. Luo, T. S. Lawrence, S. Jolly, et al., A clinical decision support system for AI-assisted decision-making in response-adaptive radiotherapy (ARCliDS), Scientific Reports 13 (2023) 5279.
[45] G. J. Klir, B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice Hall, Upper Saddle River, New Jersey, USA, 1995.
[46] J. Yen, R. Langari, Fuzzy Logic: Intelligence, Control, and Information, Prentice Hall, Upper Saddle River, New Jersey, USA, 1998.
[47] M. S. Nobile, G. Votta, R. Palorini, S. Spolaor, H. De Vitto, P. Cazzaniga, F. Ricciardiello, G. Mauri, L. Alberghina, F. Chiaradonna, et al., Fuzzy modeling and global optimization to predict novel therapeutic targets in cancer cells, Bioinformatics 36 (2020) 2181–2188.
[48] D. Chicco, S. Spolaor, M. S. Nobile, Ten quick tips for fuzzy logic modeling of biomedical systems, PLOS Computational Biology 19 (2023) e1011700.
[49] S. Spolaor, M. Scheve, M. Firat, P. Cazzaniga, D. Besozzi, M. S. Nobile, Screening for combination cancer therapies with dynamic fuzzy modeling and multi-objective optimization, Frontiers in Genetics 12 (2021) 617935.
[50] S. Spolaor, C. Fuchs, P. Cazzaniga, U. Kaymak, D. Besozzi, M. S. Nobile, Simpful: a user-friendly Python library for fuzzy logic, International Journal of Computational Intelligence Systems 13 (2020) 1687–1698.
[51] M. Sugeno, Industrial Applications of Fuzzy Control, Elsevier Science Inc., New York, NY, 1985.
[52] R. Bellman, Dynamic programming, Science 153 (1966) 34–37.
[53] J. Clifton, E. Laber, Q-learning: Theory and applications, Annual Review of Statistics and Its Application 7 (2020) 279–301.
[54] D. Hanahan, Hallmarks of Cancer: New Dimensions, Cancer Discovery 12 (2022) 31–46.
[55] A. D. Cox, S. W. Fesik, A. C. Kimmelman, J. Luo, C. J. Der, Drugging the undruggable RAS: Mission possible?, Nature Reviews Drug Discovery 13 (2014) 828–851.
[56] S. R. Punekar, V. Velcheti, B. G. Neel, K.-K. Wong, The current state of the art and future trends in RAS-targeted cancer therapies, Nature Reviews Clinical Oncology 19 (2022) 637–655.
[57] L. Breiman, Random forests, Machine learning 45 (2001) 5–32.