



**HAL**  
open science

# MEDICAL KNOWLEDGE INTEGRATION INTO REINFORCEMENT LEARNING ALGORITHMS FOR DYNAMIC TREATMENT REGIMES

Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre, Michael R Kosorok

► **To cite this version:**

Sophia Yazzourh, Nicolas Savy, Philippe Saint-Pierre, Michael R Kosorok. MEDICAL KNOWLEDGE INTEGRATION INTO REINFORCEMENT LEARNING ALGORITHMS FOR DYNAMIC TREATMENT REGIMES. 2025. hal-04919920

**HAL Id: hal-04919920**

**<https://hal.science/hal-04919920v1>**

Preprint submitted on 29 Jan 2025

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

---


# MEDICAL KNOWLEDGE INTEGRATION INTO REINFORCEMENT LEARNING ALGORITHMS FOR DYNAMIC TREATMENT REGIMES

---

A PREPRINT

 **Sophia Yazzourh**

Institut de Mathématiques de Toulouse  
UMR5219 - Université de Toulouse  
CNRS - UPS IMT  
F-31062 Toulouse Cedex 9, France  
sophia.yazzourh@math.univ-toulouse.fr

 **Nicolas Savy**

Institut de Mathématiques de Toulouse  
UMR5219 - Université de Toulouse  
CNRS - UPS IMT  
F-31062 Toulouse Cedex 9, France

 **Philippe Saint-Pierre**

Institut de Mathématiques de Toulouse  
UMR5219 - Université de Toulouse  
CNRS - UPS IMT  
F-31062 Toulouse Cedex 9, France

 **Michael R. Kosorok**

Department of Biostatistics  
University of North Carolina at Chapel Hill  
Chapel Hill, NC 27599, U.S.A.

## ABSTRACT

The goal of precision medicine is to provide individualized treatment at each stage of chronic diseases, a concept formalized by Dynamic Treatment Regimes (DTR). These regimes adapt treatment strategies based on decision rules learned from clinical data to enhance therapeutic effectiveness. Reinforcement Learning (RL) algorithms allow to determine these decision rules conditioned by individual patient data and their medical history. The integration of medical expertise into these models makes possible to increase confidence in treatment recommendations and facilitate the adoption of this approach by healthcare professionals and patients. In this work, we examine the mathematical foundations of RL, contextualize its application in the field of DTR, and present an overview of methods to improve its effectiveness by integrating medical expertise.

**Keywords** Expert Knowledge Integration · Precision Medicine · Adaptive Interventions · Medical Decision Making · Decision Process

## 1 Introduction

Modern medicine, with its remarkable advancements in care, drugs, and treatments, now seeks to enhance its ability to deliver personalized treatments for each individual patient. The paradigm of precision medicine [Kosorok and Laber, 2019] initiates a profound reflection on this question. Precision medicine aims to optimize the quality of healthcare by tailoring the medical approach to match the specific and continually changing health condition of every individual patient. The heterogeneity among patients' populations and sub-populations leads to distinct reactions and, consequently, necessitates different treatment approaches. Initially, this research domain introduced statistical models [Chakraborty and Murphy, 2014, Kosorok and Laber, 2019, Kosorok and Moodie, 2015] aimed at facilitating decision-making support. Naturally, with the advent of data storage and the computational power, machine learning methods [Coronato et al., 2020, Yu et al., 2021] have also begun to be applied to address this issue.

In this context, one of the growing interests of modern medicine is to adapt prescribed treatments to the individual data, unique characteristics and particular medical history of the patient. Precision medicine seeks to put the patient's own information at the center in order to improve their health. The motto behind is "The right treatment for the right patient (at the right time)". In a 2015 State of the Union address, President Obama announced a Precision Medicine

Initiative to revolutionize how we improve health, research, and treat disease. The initiative defines precision medicine as "an emerging approach for disease treatment and prevention that takes into account individual variability in genes, environment, and lifestyle for each person" [Terry, 2015]. In technical terms, Adaptive Treatment Strategies (ATS) or Dynamic Treatment Regimes (DTR) formalize the objective of enhancing the care pathway for patients by proposing an optimal and personalized treatment sequence. They aim to establish a decision rule at each stage of the care process. It conditions the treatment based on responses to previous prescriptions and medical history [Chakraborty and Murphy, 2014, Laber et al., 2014a]. The goal is to optimize the patient's long-term positive response to the sequence of treatment decisions while tailoring the treatment to their own medical information [Kosorok and Moodie, 2015].

In the past decades, machine learning has emerged as a solution to large-scale and high-complexity problems. When it comes to decision support, particularly in sequential scenarios, Reinforcement Learning (RL) [Sutton and Barto, 2018] offers the most effective solution. These methods excel in adapting to changing conditions and optimizing decisions over a series of steps, making them especially valuable in dynamic decision-making processes. The concept revolves around identifying a decision rule, referred to as policy, which is designed to optimize a long-term objective. This policy is crafted in order to make decisions over time that lead to the greatest cumulative benefit or outcome.

RL methods is thus an appealing candidate for precision medicine and has been intensely studied as a potential tool to guide medical decisions towards personalized medicine. First, the application of these methods to DTR is facilitated by modeling the underlying decision problem using a so-called Decision Process (DP), as detailed in Section 2. It is straightforward to express and establish connections between medical elements and its mathematical components. Second, the primary aim of RL is to identify this decision rule. In this context, there is a desire to establish this rule while maximizing long-term cumulative gains. In medicine, the effects of treatments and side effects are not immediate but can take several stages to manifest. The way the policy is constructed is a significant asset for precision medicine. Third, RL models have the capacity to simultaneously consider the extensive patient covariates data and address multi-stage decision problems. The scope of RL applications in precision medicine is in recent thematic reviews of major interest : a non-technical survey offering illustrations of RL applications in public health is proposed in Weltz et al. [2022]. More specifically, RL applications in the context of mobile health are presented in Deliu et al. [2022]. Two more technical reviews describe the methods for determining medical decision rules using off-policy RL approach [Uehara et al., 2022], or more specifically with the use of Q-learning [Clifton and Laber, 2020] and their empirical comparison with other estimation methods [Li et al., 2023].

While RL offers promising algorithms for sequential decision-making in healthcare, as detailed in the Supplementary Material, relying on a machine learning algorithm may create apprehension among all stakeholders in the process. This hesitation can originate from both the patient and the physician sides. In order to be operational in a clinical context, several points must be improved such as safer, more interpretative and efficient medical decision making [Eckardt et al., 2021]. One approach to enhance the application of RL in healthcare is the integration of expertise or human knowledge into the models. The concept is to create a partnership between both machine learning capabilities and domain experts [Holzinger, 2016, Maadi et al., 2021]. This "collaboration" would not only improve confidence in RL models and the recommendations they provide [Love et al., 2023] but also facilitate the utilization of this technology by healthcare professionals and patients within a clinical setting [Holzinger et al., 2019]. This fusion of machine learning and human expertise yields to improved results compared to RL in isolation or expert decisions alone [Arzate Cruz and Igarashi, 2020, Li et al., 2019a]. From a technical point, involving experts or medical knowledge also reduces the learning time, allowing for quicker adaptation and refinement of the methods, ultimately leading to more effective and patient-centered healthcare solutions.

The objective of this paper is to provide a comprehensive overview of RL applied to the optimization of treatment sequences. By facilitating an entry into this field for those interested in its practical application in precision medicine, we illustrate its mathematical framework and provide contextualization. This overview aims to help navigate the array of available algorithms. Additionally, we explore the integration of medical knowledge into RL models, highlighting considerations that could facilitate their clinical integration and application. We introduce these issues, offering initial questions and showcasing opportunities for further research in this area.

To achieve this objective, we structure the paper as follows. In Section 2, we delve into the mathematical foundations of RL approaches, specifically exploring decision processes and introducing key concepts and specific terms of the domain: policy, rewards, and value function. In Section 3, we contextualize this study within the realm of DTR, offering a more detailed explanation of how RL and the concept of precision medicine are intricately connected. We also explain the properties and classification of RL algorithms within our medical context. In Section 4, we provide an overview of methods to enhance reinforcement learning in the medical context by integrating expert knowledge. Various methods are presented and discussed. The paper ends by a concluding section 5.

## 2 Theoretical foundations of reinforcement learning

This section aims to outline the mathematical framework of RL applied in the DTR field. Typically, RL is explained in the context of a Markov Decision Process (MDP) and its evolution into a Partially Observable Markov Decision Process (POMPD). However, in this context, a return is made to a decision-making framework without the inclusion of Markov assumptions, which is referred to as a decision process. Subsequently, fundamental concepts are introduced : policy, value function, and the notion of optimality.

### 2.1 Decision process

#### 2.1.1 General statement

The modeling context revolves around the realm of decision-making. A foundation proposed is DP, which acts as the initial framework for DTR. It represents a dynamic system which evolves through time  $t \in \mathbb{T}$ . This system navigates within the space of states  $\mathbb{S}$  by executing actions within the realm of possibilities defined by the space of actions  $\mathbb{A}$ . The collection of non-empty measurable subsets of  $\mathbb{A}$ , denoted as  $\{\mathbb{A}(s)|s \in \mathbb{S}\}$ , represents the feasible actions that can be undertaken when the system finds itself in a specific state  $s \in \mathbb{S}$ .

**Definition 2.1** (Decision Process). A decision process  $(S, A, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$  on  $\mathbb{T}$  includes:

- a family  $S$  of  $\mathbb{S}$ -valued random variables  $\{S_t, t \in \mathbb{T}\}$ ,  $\mathbb{S}$  is called space of states.
- a family  $A$  of  $\mathbb{A}$ -valued random variables  $\{A_t, t \in \mathbb{T}\}$ ,  $\mathbb{A}$  is called space of actions.
- a family  $\{\mathbb{A}(s)|s \in \mathbb{S}\}$  of non empty measurable subsets of  $\mathbb{A}$ , the set of realizable actions when the system is in the state  $s \in \mathbb{S}$ . The requirement is for  $\mathbb{K} = \{(s, a)|s \in \mathbb{S}, a \in \mathbb{A}(s)\}$  to be a measurable subset of  $\mathbb{S} \times \mathbb{A}$ .
- a distribution  $\nu$  on  $\mathbb{S}$ .

**Remark 2.1.** DP is initially characterized for Borel spaces  $\mathbb{S}$  and  $\mathbb{A}$ . However, in most practical applications, these spaces typically have finite dimensions, context considered for the rest of the article.

**Remark 2.2.**  $S_t$  represents the state at time  $t$ , this variable may be a vector including the state and several covariates observed at the time  $t$ . In what follows no covariate are considered.

**Remark 2.3.** In full generalities,  $\mathbb{T}$  will be taken as continuous or discrete but for a sake of readability  $\mathbb{T}$  will be a discrete space denoted by  $\mathbb{T} = \{0 = t_0, t_1, \dots, t_n, \dots, \tau\}$ , with  $\tau$  representing either a finite ( $\tau = t_N < \infty$ ) or infinite ( $\tau = \infty$ ) value. For the sake of simplicity, the variables  $X_{t_n}$  will be indicated as  $X_n$  and  $X_\tau$  as  $X_\infty$  in infinite horizon setting.

**Definition 2.2.** For any  $n \in \mathbb{N}$ , an admissible history at time  $n$  is a vector which contains the states traveled by the system together with the actions taken up to time  $n$ . The set of admissible histories at time  $n$  is denoted:

$$\mathbb{H}_0 = \mathbb{S} \quad \mathbb{H}_n = \mathbb{K}^{n-1} \times \mathbb{S}$$

An element  $h_n \in \mathbb{H}_n$  writes  $(s_0, a_0, \dots, s_{n-1}, a_{n-1}, s_n)$  where for all  $0 \leq j \leq n-1$ ,  $(s_j, a_j) \in \mathbb{K}$ .

The point of main importance to deal with the decision process is to exhibit the probability to reach state  $s_{n+1}$  at time  $n+1$  given the history up to time  $n$  and the decision taken at time  $n$  this expresses as:

$$\mathbb{P}_\nu [S_{n+1} = s_{n+1} | H_n = h_n, A_n = a_n]. \quad (1)$$

In practice the computation of these probabilities requires significant computational resources because of the increasing length of the vector  $h_n$  as  $n$  increases. Rapidly working directly with such variable is intractable (usually when  $n \geq 4$ ).

#### 2.1.2 Markov decision process

To overpass this difficulty the Markov assumption is of particular interest. It consists in simplifying the dependence on the past by considering that all the necessary information for is contained in the current state.

**Definition 2.3** (Markov Decision Process). A Markov decision process on  $\mathbb{T}$  is a decision process  $(\mathbb{S}, \mathbb{A}, \{\mathbb{A}(s)|s \in \mathbb{S}\}, \nu)$  satisfying:

$$\mathbb{P}_\nu [S_{n+1} = s_{n+1} | H_n = h_n, A_n = a_n] = \mathbb{P}_\nu [S_{n+1} = s_{n+1} | S_n = s_n, A_n = a_n]. \quad (2)$$

A MDP is thus governed by a family of probability transitions

$$P_{a_n}(s_n, s_{n+1}) = \mathbb{P} [S_{n+1} = s_{n+1} | S_n = s_n, A_n = a_n].$$

which is the probability that action  $a_n$  in state  $s_n$  at time  $t_n \in \mathbb{T}$  leads to state  $s_{n+1}$  at time  $t_{n+1}$ .

The most traditional RL framework is MDP [Bellman, 1957, Garcia and Rachelson, 2013]. The majority of optimizing application complete their decision models with the memory-less Markov assumption.

**Remark 2.4.** Behind MDP modeling, there is a strong assumption that all the information necessary for the decisions observed. In reality, states space can be noisy or incomplete. To overpass this assumption, Partially Observable Markov Decision Process (POMDP) model introduced in Monahan [1982] provides a relaxation to this assumption. POMDP can be seen as a generalization of MDP and is broadly based on the same framework. The major difference comes from the expression of the state space. POMDP consider a distinction between observed data and unobserved data, whereas DP and MDP are based exclusively on the data which have been directly observed. Mathematically, POMDP defines as an MDP except  $S$  which is a family of  $\mathbb{S}^{obs} \times \mathbb{S}^{unobs}$ -valued random variables  $\{(S_n^{obs}, S_n^{unobs}), n \in \mathbb{N}\}$  where  $S^{obs}$  is observed and  $S^{unobs}$  is not.

## 2.2 Policy

The crucial concept in addressing dynamic programming is the notion of a policy, which is formalized as follows :

A policy is a sequence  $\pi = (\pi_n)_{n \in \mathbb{N}}$  of conditional distributions from  $\mathbb{A}$  given  $\mathbb{H}_n$  defined, for any  $\mathcal{A} \in \mathcal{B}(\mathbb{A})$  and all  $h_n \in \mathbb{H}_n$ , by:

$$\pi_n(\mathcal{A}, h_n) = \mathbb{P}[A_n \in \mathcal{A} \mid H_n = h_n],$$

satisfying for all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  :

$$\pi_n(\mathbb{A}(s_n), h_n) = 1,$$

and for all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and all  $a_n \in \mathbb{A}(s_n)$

$$\pi_n(a_n, h_n) > 0.$$

Decision-making is selecting an option based on environmental information. A policy represents a plan that establishes a sequence of actions. This strategy can be tailored to align with a specified objective. As a result, the focus will be on deriving the strategy that optimizes this objective. A policy  $\pi_n$  is a strategy that suggests, for every possible states  $s_n \in \mathbb{S}$ , an action  $a_n \in \mathbb{A}(s_n)$  taking to account the history  $h_n \in \mathbb{H}_n$  of the system.

**Theorem 2.1** (Hernández-Lerma and Lasserre [2012], Nivot [2016]). Given a policy  $\pi$  and the initial distribution  $\nu$ , there is a unique probability  $\mathbb{P}_\nu^\pi$  such that, for all  $\mathcal{B} \in \mathcal{B}(\mathbb{S})$ , the Borel algebra of  $\mathbb{S}$ , and  $\mathcal{A} \in \mathcal{B}(\mathbb{A})$ , the Borel algebra of  $\mathbb{A}$  :

$$\begin{aligned} \mathbb{P}_\nu^\pi [S_0 \in \mathcal{B}] &= \nu(\mathcal{B}), \\ \mathbb{P}_\nu^\pi [A_n \in \mathcal{A} \mid H_n = h_n] &= \pi_n(\mathcal{A}, h_n) \end{aligned}$$

In the following,  $\mathbb{E}_\nu^\pi$  denotes the expectation associated with the probability  $\mathbb{P}_\nu^\pi$  for an arbitrary policy  $\pi$  and an initial distribution  $\nu$ .

The following result is of major practical importance and expresses the likelihood to observe a trajectory  $h_n$  by means of the DP.

**Theorem 2.2.** Given  $(S, A, \{\mathbb{A}(s) \mid s \in \mathbb{S}\}, \nu)$  a decision process on  $\mathbb{T}$  and  $\pi$  a policy, we have for all  $n \in \mathbb{N}^*$  and all  $h_n \in \mathbb{H}_n$ ,

$$\mathbb{P}_\nu^\pi [H_n = h_n] = \prod_{j=1}^n \mathbb{P}[S_j = s_j \mid A_{j-1} = a_{j-1}, H_{j-1} = h_{j-1}] \pi(a_{j-1}, h_{j-1}) \nu(s_0)$$

In the framework of MDP, to follow the same lines as in the proof of Theorem 2.2, an additional assumption on the policy is needed yielding to the concept of Markov policy:

**Definition 2.4** (Markovian policy [Nivot, 2016]). A Markovian policy  $\pi = (\pi_n)_{n \in \mathbb{N}}$  is a policy satisfying for all  $n \in \mathbb{N}$ , all  $\mathcal{A} \in \mathcal{B}(\mathbb{A})$  and all  $h_n \in \mathbb{H}_n$ :

$$\mathbb{P}[A_n \in \mathcal{A} \mid H_n = h_n] = \mathbb{P}[A_n \in \mathcal{A} \mid S_n = s_n] = \pi_n(\mathcal{A}, s_n).$$

## 2.3 Rewards, valuation and optimization of policies

### 2.3.1 Rewards

As discussed in the Introduction, the aim of DP modeling is to find optimal policies associated to an objective. To do so, a criterion of optimality has to be introduced. This criterion is usually built by means of rewards functions which provides a temporal judgment of the desirability of a state-action pair and are formalized as follows:

**Definition 2.5.** Reward is defined as a family of bounded  $\mathbb{R}$ -valued random variables  $\{R_n, n \in \mathbb{N}\}$ . For a sake of simplicity, let us denote for a given  $n \in \mathbb{N}$ , for all  $h_n \in \mathbb{H}_n$ , all  $a_n \in \mathbb{A}$  and all  $s_{n+1} \in \mathbb{S}$ :

$$\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) = \mathbb{E}_\nu^\pi [R_{n+1} \mid H_n = h_n, A_n = a_n, S_{n+1} = s_{n+1}].$$

**Remark 2.5.** The concept of rewards functions are usually integrated in the definition of a decision process.

### 2.3.2 Valuation of policies and value-functions

State-value functions and state-action values functions are respectively known as V-function and Q-functions. These two concepts provide quantitative measures for evaluating policies, making meaningful policies comparisons and defining the optimal policy. These value-functions serve as qualitative evaluations for guiding strategic adaptations.

State-value functions allow to answer to : "How good is to be in state  $s$  after following the policy  $\pi$ ?" while action-value functions allow to answer to : "How good it is to have done the action  $a$  following policy  $\pi$  knowing that they were in state  $s$ ?". The key point is the evaluation is not assessing step-by-step evaluation but by means of the cumulative reward over time. In such a way, value functions focus on a long-term objective.

**Definition 2.6.** Given  $\gamma \in [0, 1]$  a discount parameter, the stage  $n$  long term discounted reward function is defined for all  $n \in \mathbb{N}$ , by:

$$G_n = \sum_{j=n+1}^{\infty} \gamma^{j-n-1} R_j$$

**Definition 2.7** (Value functions [Chakraborty and Murphy, 2014, Schulte et al., 2014]). Given  $(S, A, \{A(s) \mid s \in \mathbb{S}\}, \nu)$  a decision process on  $\mathbb{T}$ ,  $\{R_n, n \in \mathbb{N}\}$  a family of rewards,  $\pi$  a policy and  $\gamma \in [0, 1]$  a discount parameter.

- The stage  $n$  state-value function (V-function) for a history  $h_n$  is the total expected future rewards from stage  $n$  given by:

$$V_n^\pi(h_n) = \mathbb{E}_\nu^\pi [G_n \mid H_n = h_n].$$

- The stage  $n$  action-value function (Q-function) is the total expected future rewards starting from a history  $h_n$ , taking action  $a_n$  is given by

$$Q_n^\pi(h_n, a_n) = \mathbb{E}_\nu^\pi [G_n \mid H_n = h_n, A_n = a_n].$$

The crucial aspect to observe in these definitions is that, instead of a step-by-step evaluation, the approach aims to assess a long-term objective. The goal is to evaluate the cumulative reward over time. As a consequence of a decision, after each time step  $t_n$ , an immediate reward  $R_n$  is received which is the most distinctive feature of RL. The value functions represent the total expected future reward starting at a particular state  $s_0$  and thereafter choosing actions according to the policy  $\pi$ .

**Remark 2.6.** The discount factor  $\gamma$  introduced in the definition of the long-term reward at each step  $n$  aims to strike a thoughtful balance between immediate rewards and long-term rewards. It allows for a balancing between striving for the highest cumulative reward and the aim to reach substantial benefits within a reasonable time [Coronato et al., 2020]. This is also a mathematical trick to make the sum converge.

**Remark 2.7.** In the finite horizon case  $\tau = t_N$ , the values functions can be defined in a similar way by considering

$$G_n = \sum_{j=n+1}^N \gamma^{j-n-1} R_j$$

Notice that in this framework, the introduction of a discount parameter is not needed and is usually fixed to 1 from the definitions.

**Remark 2.8.** To consider valuation in infinite horizon, we have considered processes in infinite horizon and to do so, the Markov assumptions on the decision process and on the policy are necessary. The discount factor is now mandatory to insure the convergence of the long term discounted reward. The values functions can be defined in the same way by considering conditional to  $S$  expectations:

$$\begin{aligned} V_n^\pi(s_n) &= \mathbb{E}_\nu^\pi [G_n \mid S_n = s_n]. \\ Q_n^\pi(s_n, a_n) &= \mathbb{E}_\nu^\pi [G_n \mid S_n = s_n, A_n = a_n]. \end{aligned}$$

The following proposition highlights the link between V-functions and Q-functions.

**Proposition 2.1** ([Kosorok and Moodie, 2015, Schulte et al., 2014, Sutton and Barto, 2018]). For all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and  $a_n \in \mathbb{A}$ , we have:

$$V_n^\pi(h_n) = \sum_{a_n \in \mathbb{A}(s_n)} Q_n^\pi(h_n, a_n) \pi_n(h_n, a_n) \quad (3)$$

$$\begin{aligned} Q_n^\pi(h_n, a_n) &= \sum_{s_{n+1} \in \mathbb{S}} (\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma V_{n+1}^\pi((h_n, a_n, s_{n+1}))) \\ &\quad \times \mathbb{P}_\nu^\pi [S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n]. \end{aligned} \quad (4)$$

The remaining issue consists in the computation of the value functions. To do so, the result of major importance is the recursive form of the value functions which states that the value functions can be decomposed into immediate reward plus discounted value of successor state.

**Theorem 2.3** (Recursive form for value functions [Chakraborty and Murphy, 2014, Zhao et al., 2015]). For all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and  $a_n \in \mathbb{A}$ , we have:

$$\begin{aligned} V_n^\pi(h_n) &= \sum_{s_{n+1} \in \mathbb{S}} \sum_{a_n \in \mathbb{A}(s_n)} (\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma V_{n+1}^\pi(h_{n+1})) \\ &\quad \times \mathbb{P} [S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n] \pi(a_n, h_n) \end{aligned} \quad (5)$$

$$\begin{aligned} Q_n^\pi(h_n, a_n) &= \sum_{s_{n+1} \in \mathbb{S}} (\mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) \\ &\quad + \gamma \sum_{a_{n+1} \in \mathbb{A}(s_{n+1})} Q_{n+1}^\pi(h_{n+1}, a_{n+1}) \pi(h_{n+1}, a_{n+1})) \\ &\quad \times \mathbb{P} [S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n] \end{aligned} \quad (6)$$

Equations (5) and (6) are known as Bellman's equation. A policy being fixed, the Bellman equation can be solved, therefore making it possible to determine the values of the value functions and thus the values of Q-function. Indeed, in the case where the number of steps is finite, the Bellman equation actually hides a linear system of  $N$  equations to  $N$  unknowns. It can therefore be solved, once translated into a matrix equation, by a technique such as the Gaussian pivot.

### 2.3.3 Optimization of the policies

The key concern of the RL problem is to determine the optimal policy, denoted as  $\pi^*$ , which represents the optimal strategy for maximizing our long-term reward function. In other words, it is about finding the best way to make decisions in an environment to obtain the highest long-term rewards. The search for the optimal policy is based on the Bellman optimality principle developed below.

**Definition 2.8.** The optimal state-value functions ( $V_n^*$ ) are defined for all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  as the maximum value functions over all policies

$$V_n^*(h_n) = \max_{\pi} V_n^\pi(h_n)$$

The optimal action-value functions ( $Q_n^*$ ) are defined for all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and  $a_n \in \mathbb{A}$ , as the maximum action-value functions over all policies

$$Q_n^*(h_n, a_n) = \max_{\pi} Q_n^\pi(h_n, a_n)$$

**Definition 2.9.** Consider the partial ordering over policies defined by:

$$\pi' \geq \pi \quad \text{if and only if, for all } n \in \mathbb{N}, \text{ all } h_n \in \mathbb{H}_n, \quad V_n^{\pi'}(h_n) \geq V_n^\pi(h_n).$$

This partial ordering allows to define optimal policy in the following way:

**Proposition 2.2.** There exists an optimal policy  $\pi^*$  that is better than or equal to all other policies,  $\pi^* \geq \pi$  for all  $\pi$ .

**Theorem 2.4.** All optimal policies achieve the optimal value functions and the optimal action-value functions, for all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and  $a_n \in \mathbb{A}$ ,

$$V_n^{\pi^*}(h_n) = V_n^*(h_n) \quad \text{and} \quad Q_n^{\pi^*}(h_n, a_n) = Q_n^*(h_n, a_n).$$

**Theorem 2.5** (Bellman Optimality Equations for  $Q_n^*$ ). For all  $n \in \mathbb{N}$ , all  $h_n \in \mathbb{H}_n$  and  $a_n \in \mathbb{A}$ , we have

$$Q_n^*(h_n, a_n) = \sum_{s_{n+1} \in \mathbb{S}} \left( \mathcal{R}_{n+1}(h_n, a_n, s_{n+1}) + \gamma \max_{a \in \mathbb{A}(s_{n+1})} Q_{n+1}^*(h_{n+1}, a) \right) \times \mathbb{P}[S_{n+1} = s_{n+1} \mid H_n = h_n, A_n = a_n] \quad (7)$$

As a consequence of the Bellman Optimality Equation, we can claim that an optimal policy can be found by maximizing over  $Q_n^*(s, a)$  for all  $n \in \mathbb{N}$  and by considering the optimal policy defined as

$$\pi_n^*(s, a) = \begin{cases} 1 & \text{if } a \in \arg \max_{a \in \mathbb{A}(s)} Q_n^*(s, a) \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Note that this policy is deterministic.

## 2.4 Reinforcement Learning

The mathematical foundations established in the previous sections serve as the basis for building algorithms to determine decision rules. In the field of RL, numerous algorithms aim to learn optimal policies. We have chosen to present two of these algorithms to illustrate a first distinction between online and offline application contexts. Furthermore, the second algorithm presented has been widely adopted to meet our application context. A discussion on the different RL algorithms suitable for our context will be the subject of Section 3.5.

### 2.4.1 Q-learning Forward

Q-learning, proposed in 1989 by Chris Watkins [Sutton and Barto, 2018, Watkins and Dayan, 1992], is one of the most famous and widely used algorithms in RL. It was historically developed in the so-called online context where the algorithm can dynamically interact with its application context. This is associated with the notion of "agent" which is an entity capable of interacting with the environment while receiving rewards. The concept of interaction is related to the exploitation-exploration dilemma. The agent must, through trial and error, choose between exploiting acquired knowledge to maximize immediate rewards or exploring new actions to discover better long-term strategies [Sutton and Barto, 2018]. An excellent illustration of this problem is the  $\epsilon$ -greedy strategy presented in the following definition:

**Definition 2.10** ( $\epsilon$ -greedy Policy).

$$\pi_\epsilon(s) = \begin{cases} \text{random action from } \mathbb{A}(s) & \text{with probability } \epsilon \\ \arg \max_{a \in \mathbb{A}(s)} Q(s, a) & \text{with probability } 1 - \epsilon \end{cases}$$

where  $\epsilon \in [0, 1]$  is an hyperparameter called the exploration rate.

Q-learning relies on the recursive Bellman equations (2.3). The idea is to estimate value functions based on the differences between current and previous estimates, and then to derive an optimal strategy from Equation (8) of Bellman optimality.

---

#### Algorithm 1 Q-learning

---

**Initialisation** :  $Q(s, a)$  arbitrarily, set learning rate  $\alpha$ , discount factor  $\gamma$ , and exploration rate  $\epsilon$   
**for** each history to build **do**

    Initialize state  $s$

**while**  $s$  has not reached the terminal stage **do**

        Choose action  $a$  using policy derived from  $Q$  (e.g.,  $\epsilon$ -greedy)

        Take action  $a$ , observe reward  $r$  and new state  $s'$

        Update  $Q(s, a)$  using the Q-learning update rule:

$$Q(s, a) \leftarrow Q(s, a) + \alpha (r + \gamma \max_{a'} Q(s', a') - Q(s, a))$$

$s \leftarrow s'$

**end while**

**end for**

**Output**: The optimal decision rule is determined such as  $\pi^*(s, a) = \arg \max_a Q(s, a)$

---



### 2.4.2 Q-learning Backward

When exploration of the environment is challenging, learning can be conducted using existing data, allowing decision rules to be derived from a non-interactive environment. This is referred to as offline or batch-RL. In this context, the algorithm does not interact with its environment; learning relies on estimating value functions from pre-existing databases. This offline Q-learning [Ernst et al., 2005, Ormonet and Sen, 2002] follows a backward approach illustrated in Figure 1.

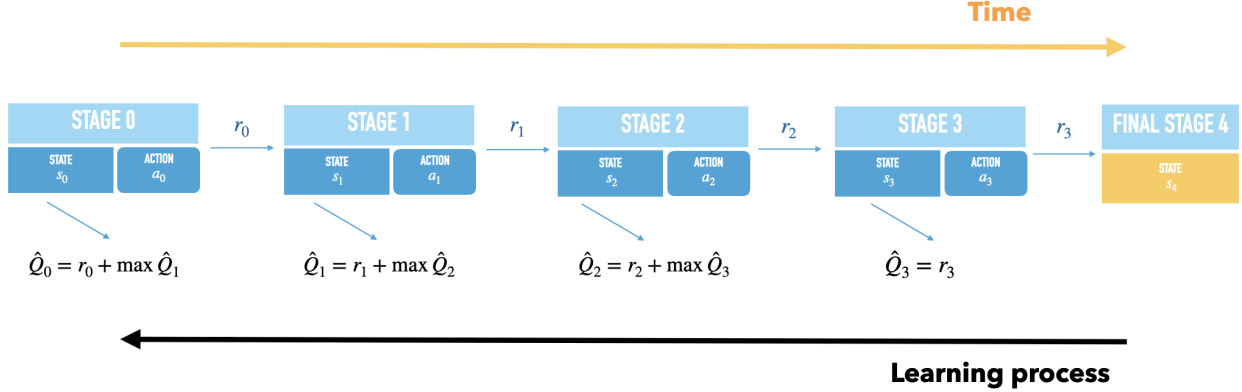


Figure 1: Illustration of the Backward Q-learning algorithm for estimating Q-values on a history with 4 steps.

The estimates of the Q-function are initialized at the terminal time and move backward in time step by step. This strategy allows for the consideration of a possible delay effect commonly observed in longitudinal data. To estimate the Q-functions, various regression algorithms can be used, such as linear regression, support vector machines, decision trees or by deep neural networks, among others.

---

#### Algorithm 2 Backward Q-learning

---

**Input:** A set of training offline data consists of patients admissible histories  $h_t$  and their associated indexed reward  $r_t, t = 0, \dots, \tau$  and a regression algorithm

**Initialisation :** Let  $t = \tau + 1$  and  $\hat{Q}_t$  be a function equal to zero everywhere on  $\mathbb{S} \times \mathbb{A}$

**while** until  $t = 0$  **do**

$t \leftarrow t - 1$  (Backward)

$Q_t$  is fitted with a regression algorithm though the following recursive equation :  $Q_t(s_t, a_t) = r_t + \max_{a_{t+1}} \hat{Q}_{t+1}(s_{t+1}, a_{t+1})$

**end while**

**Output:** Given the sequential estimates of  $\{\hat{Q}_0, \dots, \hat{Q}_\tau\}$ , the sequential optimal policies  $\{\hat{\pi}_0, \dots, \hat{\pi}_\tau\}$  can be determined

---

**Remark 2.9.** In an offline context, direct exploration is not present because decisions are made based on data collected in the database. Although there is no longer an exploration-exploitation dilemma as in the online context, it will be necessary to take into account a bias resulting from data where exploration-exploitation has already been performed.

## 3 Dynamic Treatment Regimes and Reinforcement Learning

### 3.1 Dynamic Treatment Regimes

Until the end of the 20th century, progress in medicine followed a "one-size-fits-all" approach. The search for the effect of a treatment or intervention was framed within evidence-based medicine on a target population. With the advent of massive data, particularly genomics, the paradigm has evolved. The volume of individual data collected has exploded, suggesting the possibility of integrating individual factors in the search for the effect of an intervention. The desired effect of treatment is no longer an average effect but a conditional effect on patient characteristics.

In this context, where the effect of an intervention is conditional to the variable characteristics of the patient which vary over time, the relevance of a treatment for a given individual may also vary over time. A central objective of precision medicine is to develop adaptive, and potentially optimal, intervention rules, where the definition of optimality must be clearly defined [Kahkoska et al., 2022a].

The search for adaptive (optimal) intervention rules is not a new question. A vast literature, primarily in the field of causal inference, exists and has real practical relevance. The foundational works in this context are attributed to Robins [1998], and the three extensions that allow for the effects of time-varying regimes in the presence of confounding variables: G-computation [Robins, 1986], the method of structural nested mean [Robins, 1994] models and G-estimation [Robins, 1992, 1989, 1998], as well as marginal structural models [Robins, 2000] and methods associated with inverse probability of treatment weighting [Chesnaye et al., 2022]. Subsequently, a number of methods have been proposed, both in frequentist and Bayesian frameworks. All estimate the optimal DTR based on distributional assumptions of the data generation process via parametric models. We can consider them as direct resolution methods. These methods will not be further developed in this article; an up-to-date review including direct methods can be found in Deliu and Chakraborty [2022].

In the following section, we will detail the parallel that can be drawn between DTR and RL, which helps overcome a major barrier of direct methods, namely the risk of misspecification of underlying assumptions [Zhao et al., 2015]. To address this limitation, in Murphy [2003], followed immediately by Robins [2004], semi-parametric methods were considered, marking the first examples of RL-based approaches in the literature on DTR. The innovations of RL have breathed new life into the search for optimal DTRs, gradually expanding its applicability domain.

### 3.2 Decision Process and Dynamic Treatment Regimes

In Section 2, we notably introduced decision processes, policy and rewards which forms the theoretical foundation for algorithms searching for optimal policies, namely reinforcement learning. To describe the contribution of RL algorithms in the medical context, we will begin by examining how the framework introduced and DTRs are linked.

As discussed in Section 3.1, an adaptive intervention involves making a treatment decision based on the patient’s characteristics and treatment history. An adaptive decision rule can thus be perceived as a policy in the theoretical sense presented in Section 2.2. To leverage the results of reinforcement learning, it is essential to define the applied framework of the underlying DP for DTRs.

Building upon the definition 2.1 of a decision process, it is natural to consider, in a medical context:

- The state space  $\mathcal{S}$  contains the selected covariates describing the patient’s state.
- The action space  $\mathcal{A}$  contains the selected treatments and their associated dosages.
- The subset  $\{\mathcal{A}(s) | s \in \mathcal{S}\}$  states that the treatments feasible or accessible for a patient depend on a given state.

**Remark 3.1.** It is worth noting that in our context, the variable  $S_t$  can be (and this is the usual situation) a vector containing both the patient’s health state and a set of covariates observed at time  $t$ , which may influence the transition probabilities from one state to another.

The observed histories  $h_t$  are then the care pathways of different patients. They contain health data and treatments administered up to decision  $t$ .

One of the key elements of RL is the reward. In the medical context, rewards are defined to address the clinical objective. This is a very important point as optimization relies on it. The notion of reward will be central in the discussion on the integration of medical expertise in Section 4. Indeed, for a given situation, different rewards can be associated depending on the expertise of the physicians, the specific objectives of the clinical trial, either proximally (directly after the decision) or distally (at the end of the follow-up).

### 3.3 Specificities of the Medical Context

DTRs find their primary application in medical contexts where multiple treatment lines are possible or in contexts with multiple possible decision points (see Figure 2). These adaptive strategies are particularly relevant in areas such as intensive care, chronic diseases, psychiatry, or oncology.

The medical context is known for the great heterogeneity of its data [Kahkoska et al., 2022b, Sperger et al., 2020], whether in terms of care pathways, treatment response, side effects, social factors, or lifestyle. In this regard, data-driven methods offer interesting perspectives by overcoming the issue of model misspecification. Precision medicine would thus offer a path to more equitable access to treatments. Moreover, the decision-making process can take into account



Figure 2: Illustration of medical history: treatment line for a patient.

variables such as resource availability, finances, and other socio-economic or discriminatory factors, leading to fairer decision rules.

The timing of decision-making moments is a central issue in the problem of adaptive interventions. Typically, these decision points are linked to patient visits to the practitioner. It is therefore natural to consider these moments as discrete and finite and to model them using a finite-horizon DP introduced in Section 2.1. Two issues arise: the time interval between two decision points and their frequency.

The issue of non-homogeneous time intervals between patients in the context of DTRs is typically addressed by considering the time between two visits as a covariate [Laber et al., 2014a]. In some scenarios, such as patient follow-up in oncology or diabetes care, the number of visits is indefinite and varies based on individual patient needs. These patients are regularly monitored through mobile Health (m-Health) initiatives, which operate in an online environment. Therefore, employing the Q-learning approach with backward induction, as explained in Section 2.4.2, becomes impractical. In Lockett et al. [2019], researchers identified optimal DTRs within an indefinite horizon framework using V-learning. This method aims to estimate the optimal policy from a predefined class of policies. Another approach, discussed in Ertefaie and Strawderman [2018], utilizes an inferential procedure for estimating Q-functions.

**Remark 3.2.** In the rapidly expanding field of m-Health research, online approaches are particularly suitable. Just-In-Time Adaptive Interventions (JITAI) have already been the subject of research efforts [Istepanian et al., 2007, Nahum-Shani et al., 2018, Rehg et al., 2017]. A synthesis of JITAI research is provided in Deliu et al. [2022], along with a comparative study with DTRs. This study addresses the technical aspect of making decisions about adaptive treatments in an interactive online environment. We will not cover these aspects further in the work, as the framework of DTRs on real data is discussed in Section 2.4.2, which is only feasible in the context of offline algorithms.

### 3.4 Real Data Application

The Supplementary Material provides an overview of the RL research conducted in the context of DTRs. It is important to note that decision points are typically few in real data application context; many studies consider two or three decision points. This choice is primarily driven by computational challenges: the more decision points there are, the more complex it becomes to integrate the patient's history into the models. An alternative approach is to impose a Markov assumption on the DP. However, in healthcare applications, this assumption is often unrealistic. The entire patient history can rarely be ignored or encapsulated in the current state.

As with any analysis on healthcare data, it is natural to question the biases inherent in the methods and the issue of causality [Hernan and Robins, 2023, Neuberger, 2003]. Since machine learning techniques are not causal inference methods, their use requires unbiased data. The issue typically arises in terms of "potential outcomes", and it is common to consider causal inference assumptions such as the "stable unit treatment value" assumption and the "no unmeasured confounders" assumption, as explained in [Chakraborty and Murphy, 2014, Chap. 2]. The question of causality in the field of reinforcement learning is also addressed more directly in the framework of "causal RL"<sup>1</sup> [Chakraborty and Murphy, 2014, Zhang, 2020]. The search for adaptive intervention rules relies on data with a specific longitudinal structure. Innovations in algorithms for finding optimal DTRs often begin with adjustments to existing observational databases.

The Medical Information Mart for Intensive Care (MIMIC) [Johnson et al., 2016] is a publicly accessible observational database containing information on 53,423 distinct admissions for patients in intensive care units between 2001 and 2012. It includes data on vital signs, medications, laboratory tests, measurements, caregiver notes, procedure and diagnostic codes, imaging reports, length of hospital stay, survival data, etc. Due to the wealth of available information and its longitudinal nature, MIMIC has been widely used by the RL community as a support for methods comparison (see [Roggeveen et al., 2021], Table 1 and Supplementary Material). It is also utilized as a training dataset for the development of data augmentation methods [Tseng et al., 2017] and the generation of interactive environment models [Peng et al., 2018, Raghu et al., 2017a].

<sup>1</sup>for details of "causal RL" initiative, see <https://crl.causalai.net/>

Reference	Model	State Space	Action Space	Rewards
Komorowski et al. [2016]	SARSA	Discretised state space	25 unique actions based on a 5 by 5 binning procedure of maximum vasopressor dose and sum of intravenous fluids per 4h time interval	Terminal reward at the end of each trajectory based on 90-day mortality
Raghu et al. [2017b]	Dueling DDQN	Ordinary and Sparse Auto-Encoders were used for latent state space representation	As paper Komorowski et al. [2016]	Terminal reward at the end of each trajectory based on in-hospital mortality
Raghu et al. [2017a]	Dueling DDQN	Continuous state space based on 4h aggregated features based on physiological parameters	As paper Komorowski et al. [2016]	Intermediate reward based on changes in critical care scores and lactate combined with a terminal reward for survival based on ICU mortality
Peng et al. [2018]	Dueling DDQN	Patient states are encoded recurrently using an LSTM autoencoder representing the cumulative history for each patient	As paper Komorowski et al. [2016]	The change in the negative mortality logodds of mortality between the current observations and the next observations.
Li et al. [2019b]	Actor-Critic	POMDP	As paper Komorowski et al. [2016]	As paper Komorowski et al. [2016]
Yu et al. [2019a]	Dueling DDQN	As paper [3]	As paper Komorowski et al. [2016]	Developed several reward functions based on 7 potential features most important during the treatment process

Table 1: Applications of RL algorithms on MIMIC database: highlighting various medical objectives with rewards design extract from Roggeveen et al. [2021].

Similarly to how randomized trials play a distinct role in clinical research and may be considered the gold standard for causal relationship investigation, the Sequential Multiple Assignment Randomized Trial (SMART) design [Cheung et al., 2015, Kosorok and Moodie, 2015] can be regarded as the gold standard for clinical trial design in the context of adaptive interventions. SMART designs involve an initial randomization of patients to various treatment options, followed by re-randomizations at each subsequent stage of some or all of the patients to another available treatment at that stage. With such a design, the stable unit treatment value assumption is "by design" fulfilled. However, SMART designs are challenging to implement, costly, and as a result, there is limited access to data from SMARTs. However, notable trials include :

- CATIE (Clinical Antipsychotic Trials of Intervention Effectiveness) is a SMART study involving 1,460 schizophrenia patients over 18 months aimed at evaluating the clinical effectiveness of specific sequences of antipsychotic medications [Shortreed et al., 2011].
- ADHD (Attention Deficit Hyperactivity Disorder) is a SMART study involving 150 simulated participants, aimed at evaluating an adaptive intervention for children with this disorder. This study integrates behavior modification treatment along with medication treatment [Chakraborty and Murphy, 2014, Laber et al., 2014a].
- STAR\*D (Sequenced Treatment Alternatives to Relieve Depression) is a SMART study involving 4,041 patients with major depressive disorders. This study evaluated the effectiveness of different treatment regimens [Chakraborty and Murphy, 2014, Laber et al., 2014b].

### 3.5 Properties of Reinforcement Learning Applied to Dynamic Treatment Regimes

There is a wide range of RL algorithms offering various methodological approaches tailored to specific contexts, as illustrated in the Supplementary Material Table. Figure 3 below provides a non-exhaustive overview of the most common RL algorithms. It presents many dichotomies, which will be explained in the following paragraph and contextualized in DTRs applications.

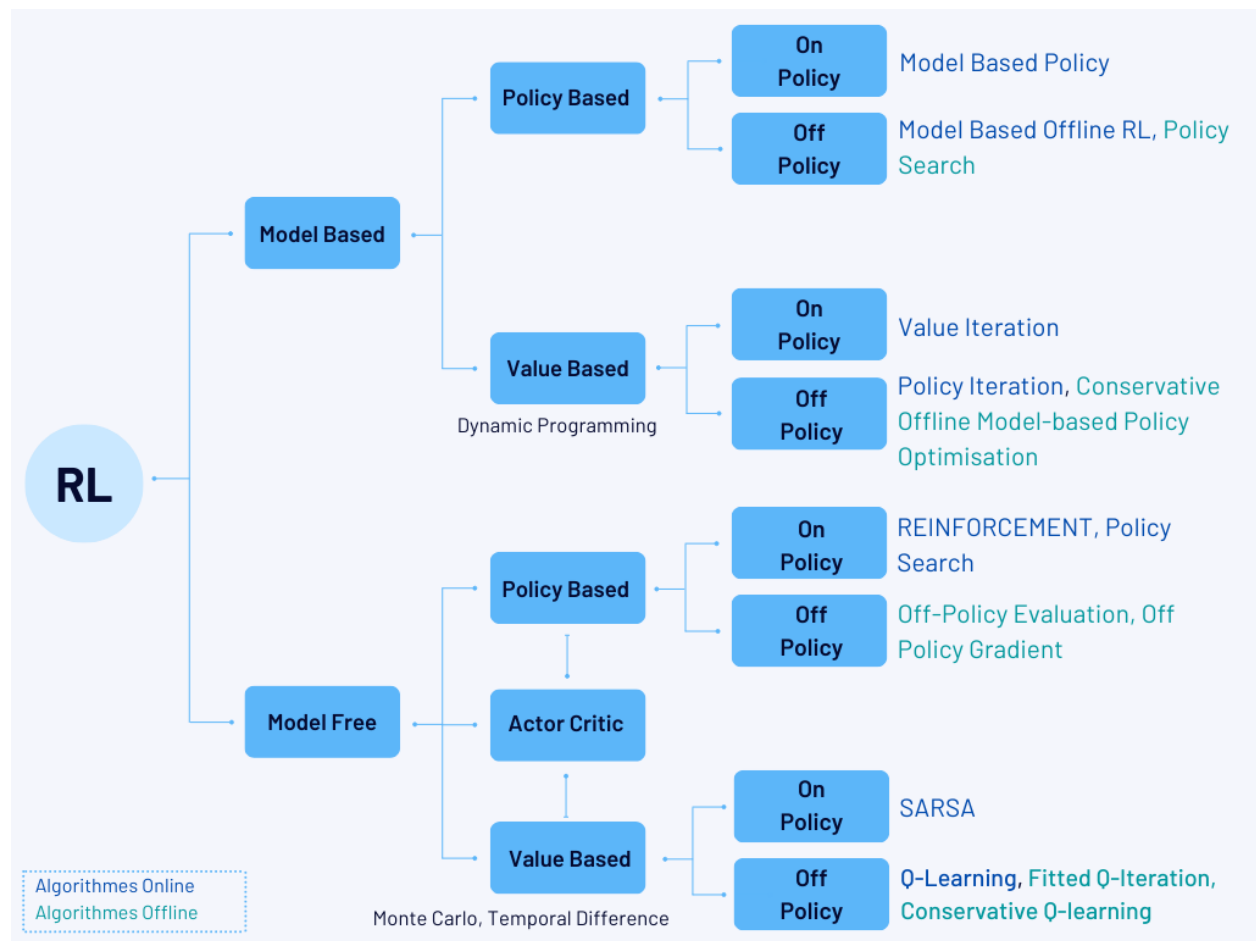


Figure 3: Classification of the most common RL algorithms.

#### 3.5.1 Model-based vs. Model-free

The first dichotomy in Figure 3 is based on the distinction between a model-based approach and a model-free approach. This distinction is related to the concept of transition probability defined by equation (1). A procedure is considered "model-based" when it relies on knowledge of all transition probabilities from a model, which means having access to

all dynamics of the system. A model-free method is able to bypass this model and is based on partial information of the associations between states and actions to determine the optimal strategy. In a model-based approach, all possible paths from an initial state  $s_0$  are explored, and an optimal policy is one that maximizes the objective.

However, in a medical context, exploring all possibilities from the same starting point is infeasible, mainly for clinical and ethical reasons. The environment is thus inherently partially observed. This reality inherently places us in a model-free framework. It is worth noting the existence of an application on simulated patient data based on MIMIC (see Section 1) in the model-based framework in Raghu et al. [2018].

### 3.5.2 Policy-based vs. Value-based vs. Actor Critic

The second distinction involves two different approaches to determine the best strategy: policy-based methods and value-based methods. The former aim to directly find the optimal policy by formalizing the RL problem through a family of policies, introduced in [Sutton and Barto, 2018, Chapter 13]. The latter seek the optimal policy through value functions, introduced in Section 2.3.2, and serve as the basis for algorithmic methods such as dynamic programming, Monte Carlo, and temporal-difference, also presented in the same book. These two approaches can be combined, thus forming actor-critic methods [Grondman et al., 2012, Sutton and Barto, 2018].

**Policy-based** Policy-based methods are direct approaches to finding the optimal policy that rely on a parametric form of the strategy  $\pi_\theta$  for  $\theta \in \Theta$ . Optimization can be typically achieved through gradient descent :

$$\theta_{n+1} = \theta_n + \nabla \mathbb{E}_{\pi_\theta} [G_n | \theta] \quad (9)$$

where  $G_n$  is the cumulative long-term reward introduced in Remark 2.7.

This method has been applied to simulated HIV data [Yu et al., 2019b] as well as in the intensive care domain [Raghu et al., 2018]. Note that the first application highlighted the challenges of converging to an optimal decision rule due to the simplification of simulation models. The main obstacle to using this method is the difficulty of convergence, which requires a large volume of data.

**Value-based** Value-based methods evaluate the optimal policy indirectly based on value functions  $V^\pi$  or  $Q^\pi$  introduced in Section 2.3.2. The general idea is to quantitatively evaluate states or action-state pairs using one of the value functions (Q-function or V-function). The optimal policy is then obtained by identifying actions that maximize these values. The success of these methods relies on the ability to model these value functions, as outlined in Section 2.4.2, through algorithms such as Backward Q-learning, making it a highly flexible approach.

The initial work was conducted by Murphy [2005a], who introduced an offline Q-learning, also known as batch learning, in a context of non-Markovian planning with a limited and restricted number of steps ( $n \leq 4$ ). This approach proves ideal for its application to DTRs and can serve as a starting point for many other applications. Research activity in this field quickly became significant, considering various parametric, semi-parametric, and non-parametric strategies to model the value function [Chakraborty and Murphy, 2014, Laber et al., 2014b, Murphy, 2005b, Tsiatis et al., 2019].

Value-based methods are better suited for application to DTRs. They enable the discovery of optimal decision rules in a non-Markovian framework with a small number of steps and data, unlike policy-based methods. This makes them easily applicable to real-world data. Moreover, they can offer a clearer interpretation, especially when Q-function estimation relies on a linear regression model [Laber et al., 2014b], thus providing interpretable decision rules. As shown in the Supplementary Material Table, this is the most widely used method in practice, particularly Q-learning approaches and its derivatives in the context of DTRs.

**Actor-Critic** A third approach to address the question of finding an optimal strategy is known as the 'Actor-Critic' method [Grondman et al., 2012]. It takes a hybrid approach by combining an Actor based on policy-based methods with a Critic based on model-based methods, thus integrating the advantages of both previous methods. The Actor refines the parameterized policy under the guidance of the Critic. The latter uses value functions, also parameterized  $V^{\pi_\theta}$  or  $Q^{\pi_\theta}$ , to guide learning. This third way of constructing decision rules was developed to correct biases in value-based methods and to counterbalance the high variability of the gradient part of policy-based methods in equation (9).

Actor-Critic methods have been applied to the MIMIC dataset. This compromise between policy-based and value-based methods converges towards a decision rule reducing patient mortality in [Wang et al., 2018] or providing a decision rule in line with physician's usual opinions in [Li et al., 2020, 2018]. This approach relies on gradient descent, similar to policy-based methods, thus necessitating databases containing a large number of individuals, often simulated data.

### 3.5.3 On-policy vs. Off-policy

This last dichotomy is closely related and sometimes confused with the concepts of offline and online algorithms presented in Section 2.4. DTRs on real data inherently operate in an offline context, seeking the optimal policy from previously collected data. Therefore, we are necessarily in an off-policy context, meaning that the strategy of generating the data ('behavior policy') is not necessarily optimal. The optimal strategy ('target policy') is deduced subsequently.

On-policy algorithms require an interactive online context where the strategy generating the data is optimized. The concepts of behavior and target policies are merged. The online framework can benefit from both on-policy algorithms, as is the case in the medical domain with Just-in-Time Adaptive Interventions (JITAI) discussed in Deliu et al. [2022], and off-policy algorithms (see Figure 3). Some online algorithms, both off-policy and on-policy, have been explored within the context of DTRs, but exclusively in simulated data settings, as indicated in the Supplementary Material table.

## 4 Integrating Medical Knowledge into Reinforcement Learning Models

The previous section has highlighted the variety of algorithms available for seeking optimal decision rules. Regardless of the method used, the construction of decision rules remains algorithmic and data-driven. Therefore, the legitimate question arises regarding the explainability of the obtained decision rule, both for the patient and the practitioner. A prerequisite for the clinical application of these decision rules will be to address these concerns. To do so, we will explore how medical expertise can intervene in the construction of these decision rules.

This section has two main goals: firstly, to outline how medical knowledge intersects with RL algorithms in the search for treatment decision rules, and secondly, to propose adjustments to these algorithms to better suit their application to DTR. These twin aims are aimed at enhancing the safety, interpretability, and relevance of tools for medical decision-making.

### 4.1 Medical Knowledge and Model Preparation

Like any machine learning method, the search for the optimal DTR depends on the data from which the method was trained. Data preparation is therefore an essential step. Medical knowledge is certainly involved in this process. Indeed, in this causal context, the choice of variables to collect and the selection of confounding factors are crucial. These decisions are primarily guided by medical expertise, drawn from the experience of practitioners and medical literature, as detailed in Section 3.3. The construction of the training dataset is thus the very first intervention of medical knowledge in RL models. It is primarily a methodological consideration that may bias the constructed optimal decision rule (Remark 2.9).

The second step in the preparation phase of applying RL in the context of searching for optimal DTRs involves selecting an algorithm from the various possibilities presented in Figure 3. This choice is primarily based on how the data were collected, the chronology of events, juxtaposed with the different characteristics of RL algorithms discussed in Section 3.5. The choice of method thus depends mainly on the application context and available data, and therefore, on underlying medical knowledge. Again, this is primarily a methodological issue, where the medical specialist collaborates with the machine learning specialist to make this choice or develop a new ad-hoc method. This discussion could follow the decision tree outlined in the figure titled "Overview of the guideline for the application of RL to healthcare" in Coronato et al. [2020].

### 4.2 Medical Knowledge and Rewards

One crucial aspect of learning optimal strategies is the formulation of rewards. This is a key component and one of the primary mechanisms for integrating medical knowledge into RL methods. In practical terms, commonly, the choice of reward is directly based on medical expertise. It is primarily a methodological issue closely linked to the study's objective. The selection of the reward is similar to choosing the primary outcome measure in the design of a clinical trial, with the same imperatives of precision and representativeness of the variable. Rewards mainly consist in scores or quantitative variables, such as changes in body mass index (BMI) in weight loss studies [Linn et al., 2015], or survival functions in critical care settings [Roggeveen et al., 2021]. Additionally, more complex rewards can be found, such as compromises or combinations of variables, as seen in oncology contexts [Zhao et al., 2009], where the reward is evaluated considering tumor size, treatment toxicity, patient well-being, and survival rates. In Table 1, an illustration of various reward functions is provided, each aiming to achieve a specific medical objective.

It is evident that selecting an ad-hoc reward for the problem under study can entail choices that are either too arbitrary or too context-specific, potentially leading to overly restrictive learning objectives. An alternative approach is to replace this choice of reward with reward shaping. Several approaches have been developed in this direction.

One way to generalize and automatically construct rewards is through inverse reinforcement learning. This method uses patient trajectories generated with expert medical decision-making to extract an estimate of the underlying reward function for these choices. Thus, it also seeks to highlight the characteristics that should be considered for its formulation. The latent medical knowledge will then be encapsulated in the estimation of the reward function. This approach has been used in the context of alcohol addiction management [Shah et al., 2022] for the search for a personalized decision-making rule. The application of inverse reinforcement Learning to the framework of DTRs is also explored in Lockett et al. [2017], where the objective of this study is to construct a reward function as a linear combination of covariates. Inverse reinforcement Learning allows for the determination of rewards from data, thereby accelerating the learning of a decision rule compared to manually constructed rewards. It is important to note that these methods assume that the physicians who generated the training data made decisions aimed at maximizing the interests of each patient. Thus, the constructed rewards are sensitive not only to the quality of the data but also to medical decisions.

Another approach involves the use of preference learning. This method relies on medical expertise, where the physician expresses preferences regarding patient trajectories. Patient histories are pairwise compared by an expert, thus creating a ranking. These comparisons are then used in a probabilistic model such as the Bradley-Terry to construct rewards by maximum likelihood estimation. These rewards are then integrated into RL algorithms. Preference Learning methods, described as on-policy by Frnkranz et al. [2012] and off-policy by Akrou et al. [2012], use preferences on trajectories on simulated data similar to the generic cancer scenario described by Zhao et al. [2009]. In Frnkranz et al. [2012], patient trajectories are compared using a partial order relation that considers survival, maximum toxicity over time, and final tumor size. Meanwhile, Akrou et al. [2012] formulate expert preferences by prioritizing trajectories with superior final outcomes, which include minimal tumor size and reduced toxicity levels. Preference Learning enables the construction of rewards based on expert preferences on trajectories, allowing learning to rely on explainable choices. However, the applications described in the articles are based on simulated cancer data and simulated preferences, and have been developed in an online framework, which is not suitable for direct clinical application.

Other methods for constructing rewards exist, such as human-centered reinforcement learning, which utilizes rewards directly provided by an expert. The agent interprets expert feedback as numerical rewards. These approaches are detailed in Li et al. [2019a], but they are generally applied in an online and on-policy context, which involves direct interaction of the agent with patients, thus raising ethical concerns and requiring a specific application framework beyond the scope of this article.

### 4.3 Medical Knowledge and Value Functions

The evaluation or estimation of value functions  $V_n^\pi$  and  $Q_n^\pi$  is also a key concept in RL. In the medical context, due to the complexity of environments and the volume of available data, these assessments often suffer from a lack of precision. Integrating medical expertise can be considered to improve results.

This is particularly true when medical expertise translates into knowledge of treatment response mechanisms. Indeed, these observations can then be integrated into RL methods to guide the learning of the optimal strategy. From a technical standpoint, it is conceivable to penalize the value function: decrease the value function when mechanisms identified by an expert indicate that the treatment is inappropriate and increase the value function when the treatment is considered relevant. Actions associated with a lower value function are less likely to be selected than those associated with a higher value function. This approach thus highlights actions considered more relevant by the expert and guides learning in the right direction. This approach was implemented for patients with renal failure in Gaweda et al. [2005]. Medical experts identified that patients who do not respond to standard treatment require higher doses. The authors constructed a DTR by incorporating this clinical fact into a Q-learning algorithm. When a patient does not respond to a treatment dose, the Q-values of lower doses are penalized, thus favoring higher doses. This approach offers the advantage of reducing the need for exploration and hence the learning time. However, it was developed in an online framework using simulated data, limiting its applicability to real-world data.

The integration of medical expertise can also occur through relay collaboration. The principle involves considering two concurrent value functions:  $Q$ , the usual value function, and  $Q^{clin}$ , the value function under the practitioner’s strategy in a given situation. The latter comes into play only when the patient is in a critical state, as evidenced by their vital signs. Subsequently, this decision and the patient’s response to treatment will be used to enrich the learning model through an enhanced value function, denoted as  $Q^+$ . Thus, the strategy for updating the value functions involves recommending treatments suggested by the RL model while seeking the expertise of physicians when the patient’s condition is deemed critical.  $Q^+$  can therefore be formalized as:

$$Q^+(s_t, a_t^+) = \begin{cases} Q^{clin}(s_t, a_t^{clin}) & \text{If the patient’s covariates indicate a critical state} \\ Q(s_t, a_t) & \text{Otherwise} \end{cases}$$



where  $a^{clin}$  is the treatment chosen by the clinician.

This approach has been deployed in the context of intensive care treatment in Wu et al. [2023] when the patient exhibits severe symptoms. In such situations, RL algorithms may propose aggressive treatment strategies to maximize reward, which can entail significant risk for the patient. In this study, a model based on value functions  $Q$  incorporates human expertise on the treatment of sepsis. Applied to the MIMIC database, this model is evaluated using a score reflecting the patient’s critical state. Expert intervention is triggered when the score is considered low. The application of this method demonstrates a higher survival rate compared to some similar methods without human expertise and also improves the estimation of the value function.

The principle of collaboration between the agent and the expert is also addressed in Sonabend et al. [2020] using the MIMIC database. It still impacts the  $Q$  functions, but now through a statistical test. The idea is to introduce exploration into an offline model by comparing risks between two strategies. One simulates standard medical decisions, while the other strategy suggests an alternative treatment. From a comparison test on state values associated with a policy, one of these strategies is adopted. The question is: when could a new treatment be better than conventional therapies? The solution seeks to balance choices of standard treatments with new options while assessing risks to discover promising alternatives that physicians have not considered.

This link between RL and medical expertise allows both supervision of treatments in complex cases and exploration of alternative treatments while assessing associated risks. Off-policy RL suffers from data biases that can be better controlled by these methods, providing critical evaluation of the strategy to be adopted. These are methods suited for real clinical applications.

#### 4.4 Medical Expertise and Objective Function

As we have just seen, value-based approaches can benefit from the integration of medical expertise in determining optimal strategies. Similarly, methods for incorporating medical expertise have been proposed for policy-based approaches, which directly modify on the objective function.

Supervised reinforcement learning merges two subfields of machine learning: Supervised Learning (SL) and RL. The fundamental principle of this method is to maximize a long-term objective, with the supervision of an expert, in order to maintain consistency with clinical treatment standards. Its ultimate goal is to predict an optimal treatment policy, minimizing deviations from medical expert recommendations. In this framework, the expert plays a crucial role as a reference for training the RL algorithm, using a database containing all medical decisions made within a cohort. This control affects the objective function in two ways. The latter is simplified into two parts: the first, derived from an actor-critic algorithm, aims to perfectly mimic the experts through its "critic" part (Section 3.5.2). The second part of supervised learning minimizes the difference between predicted treatments and those traditionally administered. This method, described notably in Yu et al. [2020], is applied in the intensive care domain using the MIMIC database and focusing on ventilation and sedation dosing. The primary objective is to provide optimal care that respects both short-term and long-term goals for patients, while adhering to best clinical practices. In this context, research shows that the supervised reinforcement learning approach outperforms the classical Actor-Critic approach in terms of convergence speed and alignment with usual medical decisions. In the study by Wang et al. [2018], the supervised reinforcement learning approach was applied to the MIMIC dataset. The treatment recommendations obtained would lead to a decrease in patient mortality rates. Supervised reinforcement learning, in its fundamental construction, aims to perfectly mimic the usual treatment practices, making it an excellent means of emulating practitioners. However, it does not allow for the proposal of alternative or less explored treatments compared to usual care methods.

#### 4.5 Medical Expertise and Policy

It is important to note that medical decision rules constructed within the framework of reinforcement learning recommend only a single action for a given state. The multiple policies approach involves proposing different equivalent or closely related strategies for a given patient state. Consequently, the specialist, relying on their expertise and the constraints of their environment, chooses the treatment from the selection of actions offered. This approach introduces the notion of quasi-equivalent actions that may take into account considerations such as side effects, less invasive treatments, and local availability. The general idea is essentially to train a set of policies evaluated by value functions, which learn a correspondence between each state and a collection of closely comparable actions. Subsequently, the approach involves restricting the choice of actions by evaluating the extent to which the deviation from optimal value is acceptable. This is the concept of worst-case value, referring to the expected gain in the worst possible scenario within the set of allowed actions. The level of deviation from optimality allowed will be controlled by a hyper-parameter.

The concept of multiple policies was introduced in Milani Fard and Pineau [2011] and applied in a simulated setting of sequential clinical trials for patients suffering from depression. It was developed within a model-based, on-policy, online framework with a finite horizon, not conducive to real data or real clinical applications. In the article Tang et al. [2020], the method evolved into a model-free and off-policy framework, still online using the Temporal Difference learning algorithm, and was applied in the simulated context of critical care based on MIMIC. Like the previous method, its development in an online environment does not align with our application context, but it establishes the foundation for a model-free approach, thus representing progress towards a model suitable for DTR.

Finally, the concept of multiple policies has also been employed in a multi-objective context, not based on expert opinion but on patient preferences, as detailed in Lizotte and Laber [2016]. By combining the notion of equivalent strategies with a multi-objective framework and Pareto dominance, and considering the preferences of patients, less restrictive solutions can be obtained. This approach, applied in the CATIE study specifically tailored to the DTR context, offers decision-makers increased choice by a larger class of optimal policies. These could provide the basis for an application that integrates experts' preferences and medical knowledge, thus addressing the issue outlined in this article.

## 5 Conclusions

This paper introduces and aims to facilitate the understanding of RL methods for precision medicine, especially its application to optimal DTR research. This topic is of major practical interest since it aims to determine an optimal decision rule for personalized treatments, with a large range of applications in areas such as intensive care, chronic diseases, psychiatry, and oncology. However, applying RL to medical research requires specific considerations and adaptations.

The main specificity arises from the data, typically derived from observational studies, which limits RL methods to offline applications. While an online setting is feasible, such as in m-health scenarios, for many cases, it is unethical to base treatment decisions solely on an algorithm. Therefore, since the data has already been collected beforehand, it is important to note that the well-known exploration-exploitation dilemma of online RL translates into an exploration-exploitation bias in offline RL settings. Section 3.5 details the properties of RL algorithms and helps identify the most desirable characteristics for an algorithm applied to DTR. First, due to clinical and ethical constraints, exploring all possibilities from the same starting point is impractical, necessitating the use of model-free algorithms. Secondly, value-based methods enable the discovery of optimal decision rules in a non-Markovian setting with limited steps and data, distinguishing them from policy-based approaches. Thirdly, off-policy algorithms are suited for offline contexts where data is already collected following a specific strategy, allowing for the determination of the optimal policy in a second phase. When these three characteristics converge, the result is an algorithm well suited for practical applications with real DTR data. Consequently, Backward Q-learning, also known as Fitted Q-Iteration, emerges as the most widely adopted and utilized algorithm in the realm of applying RL to DTR [Clifton and Laber, 2020].

Intimately linked to all work on observational data, the question of causality arises in the optimal DTR research context. A few research works directly focus on this challenge [Chakraborty and Murphy, 2014, Zhang, 2020], but most of the time causality is based on assumptions that are difficult to verify which make the results questionable. This limitation may be overcome by the experimental design relying on SMART designs but such designs are difficult and expensive to set up [Cheung et al., 2015, Kosorok and Moodie, 2015].

The classical formulation of RL relies on decision processes theory under the Markov assumption. However, this assumption is often too stringent in practical applications. Indeed, there is no guarantee that the current state under study contains all the necessary information to construct a precise decision. However most of the mathematical properties remain true without this Markov assumption by considering the entirety of the patient's history. In practice, that necessitates huge computational capacities and restricts to the applications the determination of adaptive strategies where the number of DTR steps is small (less than 4).

In addition to the previous issues, another problem emerges in the search for an optimal treatment strategy: the acceptability of the optimal DTR to both patient and practitioner. This raises concerns about how understandable the decision rules are for both patients and physicians, which is crucial for their clinical use. Integrating medical expertise into machine learning methods for personalized treatments is essential to improve safety, interpretability, and effectiveness in real-world scenarios.

One way to overpass this issue is to consider algorithms involving, one way or another, medical expertise or knowledge. We have seen that various approaches and studies demonstrate how medical expertise can be integrated into RL methods for sequential treatment decisions. This integration can be done at various stages of algorithm implementation.

First, the medical knowledge is often integrated before the study, at the design of the experiment. Indeed, physicians contribute to selecting the variables used for learning the decision rule. Similarly, algorithm selection involves collaboration between medical and machine learning expert, based on the application framework and available data.

Second, the medical knowledge can be integrated by acting on the rewards. Rewards is one of the main elements of a RL algorithm. Since they influence and guide the determination of the decision rule. Their design is thus crucial. Traditionally, a variable representative of the study’s objective is chosen. Methods such as inverse reinforcement learning and preference learning attempt to generalize their construction through expert input. Preference learning [Fürnkranz et al., 2012, Akrouf et al., 2012] and human-centered RL [Li et al., 2019a] directly incorporate expert knowledge into reward construction. However, this method suffers from being developed only in an online setup, which is not applicable to DTRs and real clinic application. Nonetheless, early research in this area can serve as a foundation for further exploration. On the other hand, inverse reinforcement learning is promising since it is developed within the offline context and it is well suited for real clinical application [Shah et al., 2022, Luckett et al., 2017].

Thirdly, the learning of decision rules can be achieved through value functions, allowing for the integration of medical expertise at this level. One approach is to incorporate observed medical mechanisms; specifically, the idea is to penalize the Q-values associated with non-decisive treatments [Gaweda et al., 2005]. However, this method was initially developed in an online context and requires reassessment for offline settings. A second idea is to establish a relay between human decisions and decisions proposed by the algorithm. In one scenario, the physician would take over when the patient is in critical conditions [Wu et al., 2023]. In another scenario, the algorithm would suggest alternative treatments to those traditionally proposed, along with associated risks [Sonabend et al., 2020]. These hybrid methods seem promising for real clinical applications, but concrete evidence of their implementation is currently lacking. In the policy-based methodological framework, the integration of expertise can occur through a method called supervised RL [Yu et al., 2020, Wang et al., 2018]. Its aim is to faithfully replicate common medical practices, offering precise emulation of physicians’ decisions. However, it does not allow for the discovery of alternative or underexplored treatments compared to conventional care methods.

Finally, the learning of decision rules can be approached methodologically through policy and it is worth noting that classical RL methods typically recommend only one policy, typically one treatment and one dose for each decision time. To enrich the context, multiple policies methods have been developed with the aim of offering an expert multiple equivalent treatment to choose from. The work of Lizotte and Laber [2016] is particularly suitable for application to real data-based DTRs, but it was developed within a framework of patient preferences and could be reassessed within an expert preference framework.

The integration of medical knowledge is a promising research field, exploring various innovative perspectives and methods. However, further research is needed to adapt them to the specific constraints and realities of precision medicine. These advancements have the potential to lead to practical clinical applications and significantly enhance daily hospital operations. This aligns with the broader challenge of applying mathematical solutions effectively in clinical practice. Particularly, the development of Health System Science enables the use of interdisciplinary skills to study the complexity of healthcare systems [Apostolopoulos et al., 2020, Kahkoska et al., 2022b]. Practically speaking, the aim is to ease the transition of laboratory discoveries into clinical practices [Gilliland et al., 2019], achieved by forming interdisciplinary teams within healthcare systems. Combining progress in both research areas could establish a tangible framework for applying RL alongside medical expertise, simplifying the treatment decision process for the benefit of all involved parties. We hope this study will encourage collaboration between machine learning researchers and healthcare professionals, by showing a framework that helps practically applying RL for DTR context.

## References

- Michael R Kosorok and Eric B Laber. Precision medicine. *Annual review of statistics and its application*, 6:263–286, 2019.
- Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.
- Michael R Kosorok and Erica EM Moodie. *Adaptive treatment strategies in practice: planning trials and analyzing data for personalized medicine*. SIAM, 2015.
- Antonio Coronato, Muddasar Naeem, Giuseppe De Pietro, and Giovanni Paragliola. Reinforcement learning for intelligent healthcare applications: A survey. *Artificial Intelligence in Medicine*, 109:101964, 2020.
- Chao Yu, Jiming Liu, Shamim Nemati, and Guosheng Yin. Reinforcement learning in healthcare: A survey. *ACM Computing Surveys (CSUR)*, 55(1):1–36, 2021.
- Sharon F Terry. Obama’s precision medicine initiative. *Genetic testing and molecular biomarkers*, 19(3):113, 2015.

- Eric B Laber, Daniel J Lizotte, Min Qian, William E Pelham, and Susan A Murphy. Dynamic treatment regimes: Technical challenges and applications. *Electronic journal of statistics*, 8(1):1225, 2014a.
- Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.
- Justin Wertz, Alex Volfovsky, and Eric B Laber. Reinforcement learning methods in public health. *Clinical therapeutics*, 44(1):139–154, 2022.
- Nina Deliu, Joseph Jay Williams, and Bibhas Chakraborty. Reinforcement learning in modern biostatistics: constructing optimal adaptive interventions. *arXiv preprint arXiv:2203.02605*, 2022.
- Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *arXiv preprint arXiv:2212.06355*, 2022.
- Jesse Clifton and Eric Laber. Q-learning: Theory and applications. *Annual Review of Statistics and Its Application*, 7: 279–301, 2020.
- Zhen Li, Jie Chen, Eric Laber, Fang Liu, and Richard Baumgartner. Optimal treatment regimes: a review and empirical comparison. *International Statistical Review*, 91(3):427–463, 2023.
- Jan-Niklas Eckardt, Karsten Wendt, Martin Bornhaeuser, and Jan Moritz Middeke. Reinforcement learning for precision oncology. *Cancers*, 13(18):4624, 2021.
- Andreas Holzinger. Interactive machine learning for health informatics: when do we need the human-in-the-loop? *Brain Informatics*, 3(2):119–131, 2016.
- Mansoureh Maadi, Hadi Akbarzadeh Khorshidi, and Uwe Aickelin. A review on human–ai interaction in machine learning and insights for medical applications. *International journal of environmental research and public health*, 18(4):2121, 2021.
- Tamlin Love, Ritesh Ajoodha, and Benjamin Rosman. Who should i trust? cautiously learning with unreliable experts. *Neural Computing and Applications*, 35(23):16865–16875, 2023.
- Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 9(4): e1312, 2019.
- Christian Arzate Cruz and Takeo Igarashi. A survey on interactive reinforcement learning: Design principles and open challenges. In *Proceedings of the 2020 ACM designing interactive systems conference*, pages 1195–1209, 2020.
- Guangliang Li, Randy Gomez, Keisuke Nakamura, and Bo He. Human-centered reinforcement learning: A survey. *IEEE Transactions on Human-Machine Systems*, 49(4):337–349, 2019a.
- Richard Bellman. A markovian decision process. *Journal of mathematics and mechanics*, pages 679–684, 1957.
- Frédéric Garcia and Emmanuel Rachelson. Markov decision processes. *Markov Decision Processes in Artificial Intelligence*, pages 1–38, 2013.
- George E Monahan. State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms. *Management science*, 28(1):1–16, 1982.
- Onésimo Hernández-Lerma and Jean B Lasserre. *Discrete-time Markov control processes: basic optimality criteria*, volume 30. Springer Science & Business Media, 2012.
- Christophe Nivot. *Analyse et étude des processus markoviens décisionnels*. PhD thesis, Bordeaux, 2016.
- Phillip J Schulte, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. Q-and a-learning methods for estimating optimal dynamic treatment regimes. *Statistical science: a review journal of the Institute of Mathematical Statistics*, 29(4):640, 2014.
- Ying-Qi Zhao, Donglin Zeng, Eric B Laber, and Michael R Kosorok. New statistical learning methods for estimating optimal dynamic treatment regimes. *Journal of the American Statistical Association*, 110(510):583–598, 2015.
- Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.
- Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6:503–556, 2005.
- Dirk Ormoneit and Šaunak Sen. Kernel-based reinforcement learning. *Machine learning*, 49:161–178, 2002.
- Anna R Kahkoska, Kristen Hassmiller Lich, and Michael R Kosorok. Focusing on optimality for the translation of precision medicine. *Journal of Clinical and Translational Science*, 6(1):e118, 2022a.
- James M Robins. Correction for non-compliance in equivalence trials. *Statistics in medicine*, 17(3):269–302, 1998.

- James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7(9-12):1393–1512, 1986.
- James M Robins. Correcting for non-compliance in randomized trials using structural nested mean models. *Communications in Statistics-Theory and methods*, 23(8):2379–2412, 1994.
- James Robins. Estimation of the time-dependent accelerated failure time model in the presence of confounding factors. *Biometrika*, 79(2):321–334, 1992.
- James M Robins. The analysis of randomized and non-randomized aids treatment trials using a new approach to causal inference in longitudinal studies. *Health service research methodology: a focus on AIDS*, pages 113–159, 1989.
- James M Robins. Marginal structural models versus structural nested models as tools for causal inference. In *Statistical models in epidemiology, the environment, and clinical trials*, pages 95–133. Springer, 2000.
- Nicholas C Chesnaye, Vianda S Stel, Giovanni Tripepi, Friedo W Dekker, Edouard L Fu, Carmine Zoccali, and Kitty J Jager. An introduction to inverse probability of treatment weighting in observational research. *Clinical Kidney Journal*, 15(1):14–20, 2022.
- Nina Deliu and Bibhas Chakraborty. Dynamic treatment regimes for optimizing healthcare. In *The Elements of Joint Learning and Optimization in Operations Management*, pages 391–444. Springer, 2022.
- Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(2):331–355, 2003.
- James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics: analysis of correlated data*, pages 189–326. Springer, 2004.
- Anna R Kahkoska, Nikki LB Freeman, and Kristen Hassmiller Lich. Systems-aligned precision medicine—building an evidence base for individuals within complex systems. In *JAMA health forum*, volume 3, pages e222334–e222334. American Medical Association, 2022b.
- John Sperger, Nikki LB Freeman, Xiaotong Jiang, David Bang, Daniel de Marchi, and Michael R Kosorok. The future of precision health is data-driven decision support. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 13(6):537–543, 2020.
- Daniel J Lockett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.
- Ashkan Ertefaie and Robert L Strawderman. Constructing dynamic treatment regimes over indefinite time horizons. *Biometrika*, 105(4):963–977, 09 2018. ISSN 0006-3444. doi:10.1093/biomet/asy043. URL <https://doi.org/10.1093/biomet/asy043>.
- Robert Istepanian, Swamy Laxminarayan, and Constantinos S Pattichis. *M-health: Emerging mobile health systems*. Springer Science & Business Media, 2007.
- Inbal Nahum-Shani, Shawna N Smith, Bonnie J Spring, Linda M Collins, Katie Witkiewitz, Ambuj Tewari, and Susan A Murphy. Just-in-time adaptive interventions (jitais) in mobile health: key components and design principles for ongoing health behavior support. *Annals of Behavioral Medicine*, pages 1–17, 2018.
- James M Rehg, Susan A Murphy, and Santosh Kumar. Mobile health. *Cham: Springer International Publishing*, 2017.
- M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023. ISBN 9781420076165. URL [https://books.google.fr/books?id=\\_KnHIAAACAAJ](https://books.google.fr/books?id=_KnHIAAACAAJ).
- Leland Gerson Neuberger. Causality: models, reasoning, and inference, by judea pearl, cambridge university press, 2000. *Econometric Theory*, 19(4):675–685, 2003.
- Junzhe Zhang. Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In *International conference on machine learning*, pages 11012–11022. PMLR, 2020.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.
- Luca Roggeveen, Ali El Hassouni, Jonas Ahrendt, Tingjie Guo, Lucas Fleuren, Patrick Thorald, Armand RJ Girbes, Mark Hoogendoorn, and Paul WG Elbers. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. *Artificial Intelligence in Medicine*, 112: 102003, 2021.
- Huan-Hsin Tseng, Yi Luo, Sunan Cui, Jen-Tzung Chien, Randall K Ten Haken, and Issam El Naqa. Deep reinforcement learning for automated radiation adaptation in lung cancer. *Medical physics*, 44(12):6690–6705, 2017.

- Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, H Lehman Li-wei, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 887. American Medical Informatics Association, 2018.
- Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017a.
- Matthieu Komorowski, A Gordon, LA Celi, and A Faisal. A markov decision process to suggest optimal treatment of severe infections in intensive care. In *Neural Information Processing Systems Workshop on Machine Learning for Health*, 2016.
- Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. In *Machine Learning for Healthcare Conference*, pages 147–163. PMLR, 2017b.
- Luchen Li, Matthieu Komorowski, and Aldo A Faisal. Optimizing sequential medical treatments with auto-encoding heuristic search in pomdps. *arXiv preprint arXiv:1905.07465*, 2019b.
- Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–3, 2019a. doi:10.1109/ICHI.2019.8904645.
- Ying Kuen Cheung, Bibhas Chakraborty, and Karina W Davidson. Sequential multiple assignment randomized trial (smart) with adaptive randomization for quality improvement in depression treatment program. *Biometrics*, 71(2): 450–459, 2015.
- Susan M Shortreed, Eric Laber, Daniel J Lizotte, T Scott Stroup, Joelle Pineau, and Susan A Murphy. Informing sequential clinical decision-making through reinforcement learning: an empirical study. *Machine learning*, 84(1-2): 109, 2011.
- Eric B Laber, Kristin A Linn, and Leonard A Stefanski. Interactive model building for q-learning. *Biometrika*, 101(4): 831–847, 2014b.
- Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.
- Ivo Grondman, Lucian Busoniu, Gabriel AD Lopes, and Robert Babuska. A survey of actor-critic reinforcement learning: Standard and natural policy gradients. *IEEE Transactions on Systems, Man, and Cybernetics, part C (applications and reviews)*, 42(6):1291–1307, 2012.
- Chao Yu, Yinzhaodong, Jiming Liu, and Guoqi Ren. Incorporating causal factors into reinforcement learning for dynamic treatment regimes in hiv. *BMC medical informatics and decision making*, 19:19–29, 2019b.
- Susan A Murphy. A generalization error for q-learning. *Journal of Machine Learning Research*, 6:1073–1097, 2005a.
- Susan A Murphy. An experimental design for the development of adaptive treatment strategies. *Statistics in medicine*, 24(10):1455–1481, 2005b.
- Anastasios A Tsiatis, Marie Davidian, Shannon T Holloway, and Eric B Laber. *Dynamic treatment regimes: Statistical methods for precision medicine*. Chapman and Hall/CRC, 2019.
- Lu Wang, Wei Zhang, Xiaofeng He, and Hongyuan Zha. Supervised reinforcement learning with recurrent neural network for dynamic treatment recommendation. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2447–2456, 2018.
- Luchen Li, Ignacio Albert-Smet, and Aldo A Faisal. Optimizing medical treatment for sepsis in intensive care: from reinforcement learning to pre-trial evaluation. *arXiv preprint arXiv:2003.06474*, 2020.
- Luchen Li, Matthieu Komorowski, and Aldo A Faisal. The actor search tree critic (astc) for off-policy pomdp learning in medical decision making. *arXiv preprint arXiv:1805.11548*, 2018.
- Kristin A Linn, Eric B Laber, and Leonard A Stefanski. iqlearn: Interactive q-learning in r. *Journal of statistical software*, 64(1), 2015.
- Yufan Zhao, Michael R Kosorok, and Donglin Zeng. Reinforcement learning design for cancer clinical trials. *Statistics in medicine*, 28(26):3294–3315, 2009.
- Syed Ihtesham Hussain Shah, Antonio Coronato, and Muddasar Naeem. Inverse reinforcement learning based approach for investigating optimal dynamic treatment regime. In *Workshops at 18th International Conference on Intelligent Environments (IE2022)*, pages 266–276. IOS Press, 2022.
- Daniel J Luckett, Eric B Laber, and Michael R Kosorok. Estimation and optimization of composite outcomes. *arXiv preprint arXiv:1711.10581*, 2017.

- Johannes Fürnkranz, Eyke Hüllermeier, Weiwei Cheng, and Sang-Hyeun Park. Preference-based reinforcement learning: a formal framework and a policy iteration algorithm. *Machine learning*, 89:123–156, 2012.
- Riad Akrouf, Marc Schoenauer, and Michèle Sebag. April: Active preference learning-based reinforcement learning. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II 23*, pages 116–131. Springer, 2012.
- Adam E Gaweda, Mehmet K Muezzinoglu, George R Aronoff, Alfred A Jacobs, Jacek M Zurada, and Michael E Brier. Incorporating prior knowledge into q-learning for drug delivery individualization. In *Fourth International Conference on Machine Learning and Applications (ICMLA'05)*, pages 6–pp. IEEE, 2005.
- XiaoDan Wu, RuiChang Li, Zhen He, TianZhi Yu, and ChangQing Cheng. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *npj Digital Medicine*, 6(1):15, 2023.
- Aaron Sonabend, Junwei Lu, Leo Anthony Celi, Tianxi Cai, and Peter Szolovits. Expert-supervised reinforcement learning for offline policy learning and evaluation. *Advances in Neural Information Processing Systems*, 33: 18967–18977, 2020.
- Chao Yu, Guoqi Ren, and Yinzhaodong. Supervised-actor-critic reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC medical informatics and decision making*, 20(3):1–8, 2020.
- M. Milani Fard and J. Pineau. Non-deterministic policies in markovian decision processes. *Journal of Artificial Intelligence Research*, 40:1–24, January 2011. ISSN 1076-9757. doi:10.1613/jair.3175. URL <http://dx.doi.org/10.1613/jair.3175>.
- Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In *International Conference on Machine Learning*, pages 9387–9396. PMLR, 2020.
- Daniel J Lizotte and Eric B Laber. Multi-objective markov decision processes for data-driven decision support. *The journal of machine learning research*, 17(1):7378–7405, 2016.
- Yorghos Apostolopoulos, Kristen Hassmiller Lich, and Michael K Lemke. *Complex systems and population health*. Oxford University Press, 2020.
- C Taylor Gilliland, Julia White, Barry Gee, Rosan Kreeftmeijer-Vegter, Florence Bietrix, Anton E Ussi, Marian Hajdich, Petr Kocis, Nobuyoshi Chiba, Ryutaro Hirasawa, et al. The fundamental characteristics of a translational scientist, 2019.
- Amin Hassani et al. Reinforcement learning based control of tumor growth with chemotherapy. In *2010 International Conference on System Science and Engineering*, pages 185–189. IEEE, 2010.
- Inkyung Ahn and Jooyoung Park. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *BioSystems*, 106(2-3):121–129, 2011.
- Yair Goldberg and Michael R Kosorok. Q-learning with censored data. *Annals of statistics*, 40(1):529, 2012.
- Inbal Nahum-Shani, Min Qian, Daniel Almirall, William E Pelham, Beth Gnagy, Gregory A Fabiano, James G Waxmonsky, Jihneeh Yu, and Susan A Murphy. Q-learning: a data analysis method for constructing adaptive interventions. *Psychological methods*, 17(4):478, 2012.
- Yufan Zhao, Donglin Zeng, Mark A Socinski, and Michael R Kosorok. Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer. *Biometrics*, 67(4):1422–1433, 2011.
- Robert Vincent. *Reinforcement learning in models of adaptive medical treatment strategies*. McGill University (Canada), 2014.
- Ying Liu, Yuanjia Wang, Michael R Kosorok, Yingqi Zhao, and Donglin Zeng. Robust hybrid learning for estimating personalized dynamic treatment regimens. *arXiv preprint arXiv:1611.02314*, 2016.
- Kyle Humphrey. *Using Reinforcement Learning to Personalize Dosing Strategies in a Simulated Cancer Trial with High Dimensional Data*. PhD thesis, The University of Arizona, 2017.
- Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. *Mathematical biosciences*, 293:11–20, 2017.
- Ammar Jalalimanesh, Hamidreza Shahabi Haghghi, Abbas Ahmadi, and Madjid Soltani. Simulation-based optimization of radiotherapy: Agent-based modeling and reinforcement learning. *Mathematics and Computers in Simulation*, 133: 235–248, 2017.

- Ying Liu, Brent Logan, Ning Liu, Zhiyuan Xu, Jian Tang, and Yangzhi Wang. Deep reinforcement learning for dynamic treatment regimes on medical registry data. In *2017 IEEE international conference on healthcare informatics (ICHI)*, pages 380–385. IEEE, 2017.
- Elizabeth F Krakow, Michael Hemmer, Tao Wang, Brent Logan, Mukta Arora, Stephen Spellman, Daniel Couriel, Amin Alousi, Joseph Pidala, Michael Last, et al. Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using q-learning. *American journal of epidemiology*, 186(2):160–172, 2017.
- Gregory Yauney and Pratik Shah. Reinforcement learning with action-derived rewards for chemotherapy and clinical trial dosing regimen selection. In *Machine Learning for Healthcare Conference*, pages 161–226. PMLR, 2018.
- Parisa Yazdjerdi, Nader Meskin, Mohammad Al-Naemi, Ala-Eddin Al Moustafa, and Levente Kovács. Reinforcement learning-based control of tumor growth under anti-angiogenic therapy. *Computer methods and programs in biomedicine*, 173:15–26, 2019.
- Amir Ebrahimi Zade, Seyedhamidreza Shahabi Haghighi, and Madjid Soltani. Reinforcement learning for optimal scheduling of glioblastoma treatment with temozolomide. *Computer methods and programs in biomedicine*, 193:105443, 2020.
- Salma Daoud, Afef Mdhaffar, Mohamed Jmaiel, and Bernd Freisleben. Q-rank: reinforcement learning for recommending algorithms to predict drug sensitivity to cancer therapy. *IEEE Journal of Biomedical and Health Informatics*, 24(11):3154–3161, 2020.
- Grégoire Moreau, Vincent François-Lavet, Paul Desbordes, and Benoît Macq. Reinforcement learning for radiotherapy dose fractioning automation. *Biomedicines*, 9(2):214, 2021.
- Chamani Shiranthika, Kuo-Wei Chen, Chung-Yih Wang, Chan-Yun Yang, BH Sudantha, and Wei-Fu Li. Supervised optimal chemotherapy regimen based on offline reinforcement learning. *IEEE Journal of Biomedical and Health Informatics*, 26(9):4763–4772, 2022.
- Pramod Kaushik, Sneha Kummetha, Perusha Moodley, and Raju S Bapi. A conservative q-learning approach for handling distribution shift in sepsis treatment strategies. *arXiv preprint arXiv:2203.13884*, 2022.



## Supplementary Material

Table 2: **Reinforcement Learning Applications for Sequential Decision in Healthcare.**

Ref, References; Environment, description of the medical application context; Data, Simulated or Real; Model, Decision Process (DP) or Markov Decision Process (MDP) or Partially Observable Markov Decision Process (POMDP); Stage, number of stages; Off/On, offline or online; Algorithm, standard reference algorithm of reinforcement learning without the paper specifications or innovation added.

Ref.	Environment	Data	Model	Stage	Off/On	Algorithm
Gaweda et al. [2005]	Simulated patient with anemia due to kidney failure	Real Data	MDP	24	Online	Q-learning
Zhao et al. [2009]	ODE Simulation of cancer trial for advanced generic cancer of treatment	Simulation	DP	6	Offline	Backward Q-learning
Hassani et al. [2010]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
Ahn and Park [2011]	ODE Simulation of cancer growth on a cell population level	Simulation	DP	N/A	Online	Actor-Critic
Shortreed et al. [2011]	CATIE	Real Data	DP	2	Offline	Backward Q-learning
Akrour et al. [2012]	Same as in Zhao et al. [2009]	Simulation	MDP	12	Online	Policy Search
Goldberg and Kosorok [2012]	Same as in Zhao et al. [2009]	Simulation	DP	3	Offline	Backward Q-learning
Nahum-Shani et al. [2012]	ADHD	Real Data	DP	2	Offline	Backward Q-learning
Fürnkranz et al. [2012]	Same as in Zhao et al. [2009]	Simulation	MDP	6	Online	Policy Iteration
Zhao et al. [2011]	Exponential distribution simulation of cancer for parties in phase III.	Simulation	DP	2	Offline	Backward Q-learning
Vincent [2014]	(Chapter 4) Electrical stimulation for epilepsy from in vitro experiments	Real Data	MDP	N/A	Offline	Backward Q-learning
Vincent [2014]	(Chapter 5) Electrical stimulation for Parkinson's disease	Simulation	MDP	2, 6, 9, 10	Online	SARSA, Temporal Difference
Vincent [2014]	(Chapter 5) Electrical stimulation for Parkinson's disease	Simulation	MDP	2, 6, 9, 10	Offline	Backward Q-learning
Vincent [2014]	(Chapter 6) Fractionation scheduling for radiation therapy	Simulation	MDP	4, 10, 30	Offline	Backward Q-learning
Laber et al. [2014a]	ADHD	Real Data	DP	2	Offline	Backward Q-learning
Laber et al. [2014b]	STAR*D	Real Data	DP	2	Offline	Backward Q-learning
Ertefaie and Strawderman [2018]	Simulated patients with diabetes	Simulation	MDP	Indefinite	Online	Q-learning
Cheung et al. [2015]	Comparison of depression interventions after acute coronary syndrom (SMART)	Real Data	DP	2	Offline	Backward Q-learning

Continued on next page

Table 2 – Continued from previous page

Ref.	Environment	Data	Model	Stage	Off/On	Algorithm
Liu et al. [2016]	STAR*D and ADHD	Real Data	DP	2	Offline	Outcome-Weighted Learning and Q-learning
Humphrey [2017]	Same as in Zhao et al. [2009]	Simulation	MDP	N/A	Offline	Backward Q-learning
Padmanabhan et al. [2017]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
Tseng et al. [2017]	114 patients used to construct synthetic data by GAN	Simulation	MDP	35	Online	Deep Q-learning
Jalalimanesh et al. [2017]	Model of tumor growth using Net-Logo package, Agent-based simulation	Simulation	DP	N/A	Online	Q-learning
Liu et al. [2017]	Registry data from 6021 AML patients who underwent allogeneic stem cell transplantation	Real Data	DP	5	Offline	Deep Q-learning
Krakov et al. [2017]	Nonrandomized registry data from 11,141 patients who underwent allogeneic stem cell transplantation	Real Data	DP	2	Offline	Backward Q-learning
Raghu et al. [2017b]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
Raghu et al. [2017a]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
Yaune and Shah [2018]	Linear and ODE Simulation of cancer trial	Simulation	MDP	N/A	Online	Deep Q-learning
Peng et al. [2018]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
Yazdjerdj et al. [2019]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
Luckett et al. [2019]	Simulated patients with diabetes	Simulation	MDP	Indefinite	Online	V-learning
Li et al. [2019b]	MIMIC	Real Data	POMDP	N/A	Offline	Deep Q-Learning
Zade et al. [2020]	ODE Simulation of cancer growth on a cell population level	Simulation	MDP	N/A	Online	Q-learning
Daoud et al. [2020]	Real dataset of breast cancer	Real Data	MDP	N/A	Online	Q-learning
Tang et al. [2020]	MIMIC	Real Data	MDP	750	Offline	Temporal Difference
Sonabend et al. [2020]	MIMIC	Real Data	MDP	N/A	Offline	Q-learning
Moreau et al. [2021]	Simulate tumor development inside healthy tissue and the effect of radiation therapy	Simulation	MDP	N/A	Online	Deep Policy Gradient
Wu et al. [2023]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning
Shiranthika et al. [2022]	40 patients of stage-four colon cancer	Real Data	MDP	6	Offline	Deep Q-Learning
Kaushik et al. [2022]	MIMIC	Real Data	MDP	N/A	Offline	Deep Q-Learning