* 1. Project Timelines

The RL4H project splits in three main steps. Each step investigates a question of major interest related to the application / improvement of the reinforcement learning techniques in the context of medical research. The scientific content of these steps is detailed in Section 2.4.

Step 1: Deep Q-learning to explore more complicated decision rules.

Investigation of Deep Q-learning (DQL) model

Research of the format, nature and volume of a relevant database for the implementation of DQL

Performance assessment based on data on treatment lines for Chronic Myeloid Lymphoma

Step 2: Hybrid data/expert driven approach for healthcare pathways analyses

Recommendations on the suitability of the model / objective / data

Adaptation of the algorithms to the context of healcare pathway analysis

Step 3: Data augmentation strategies for dealing with small databases

Use of Virtual patients’ generators (GPVs) (GANs or copulas) for the "data augmentation" approach from small databases

Identification of the thresholds of relevance of the algorithms according to the base and identification of the discriminating performance factors

The project timelines are summarized on the Gantt’s diagram Figure 1 page **4**.

**Project meetings** consist in step points to evaluate the evolution of the project.

* Meeting 1 is a kick off meeting devoted to the general organization of the project.
* Meetings 2 and 3 are devoted to discussions on the technical issues in the coding of the algorithms and discussions on the follow up of the theoretical research.
* Meeting 4 is devoted to agree on the editorial questions and on the organization of the closing workshop.

Finally, a **closing workshop** is planned in order to present our work, to present the perspective of these works in different therapeutic aera.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Project Month :** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** | **21** | **22** | **23** | **24** |
| Project Meetings | 1 |  |  |  |  | 2 |  |  |  |  |  | 3 |  |  |  |  |  | 4 |  |  |  |  |  |  |
| Step 1.a |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 1.b |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 1.c |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 2.a |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 2.b |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 3.a |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Step 3.b |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Closing  Workshop |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

**Figure 1:** Gantt’s chart of RL4H project

# Full scientific description of the project

* 1. Problems and scientific interests, brief bibliography

Reinforcement Learning

Reinforcement learning [1, 2] consists of learning, from data, the actions to be taken so that the reward acquired over time is maximum. We consider that the agent is immersed in an environment, and makes his decisions according to his current state. In return, the environment provides the agent with a reward (positive or negative). The agent seeks to learn a decisional behavior (called policy), a function associating with the current state the action to be executed in an optimal way, in the sense that the sum of the rewards over time is maximum.

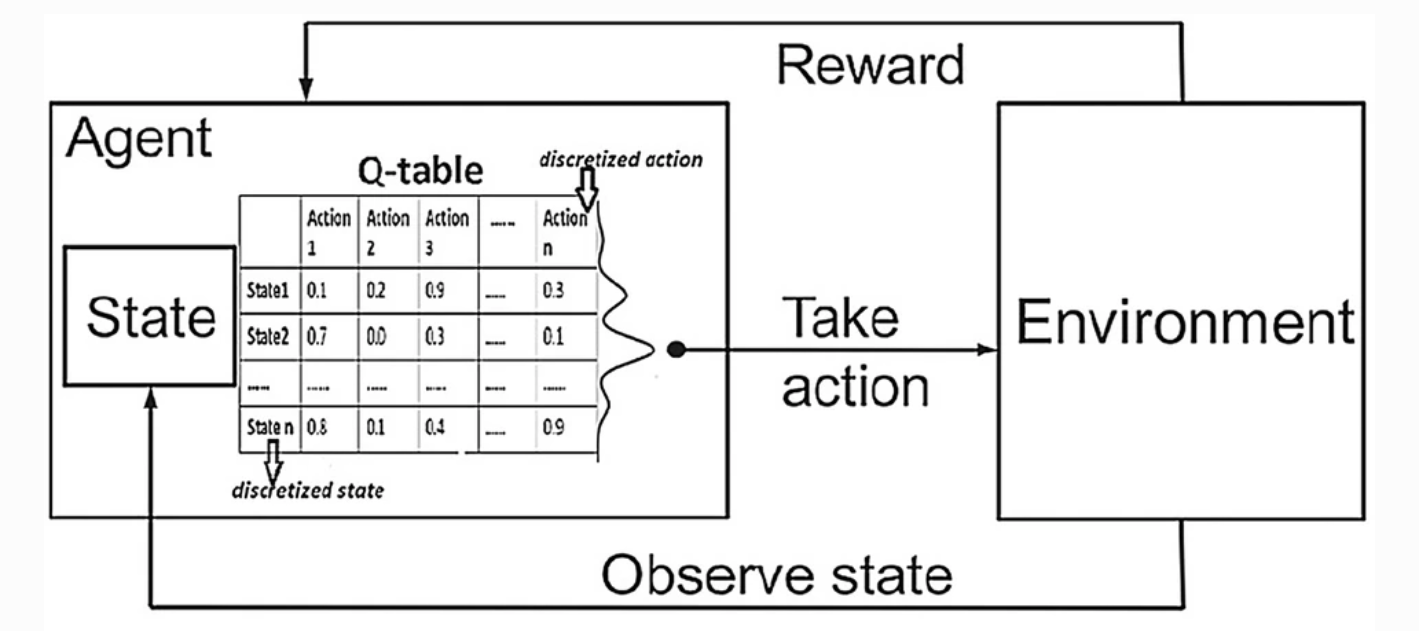
Basic reinforcement learning is modeled by a Markov decision process (MDP) involving

* **S** a set of environment and agent states,
* **A** a set of actions,
* the probability of transition (at time *t*) from state *s* to state *s’* under action *a*,
* the immediate reward after transition from state *s* to state *s’* under action *a*.

Agent interacts with its environment in discrete time steps. At a given time *t,* his current state and his reward are collected. The agent chooses an action and sends it to the environment which moves to state and reward according to the transition .

One aim on major interest is to learn a so-called optimal policy which is the map where and which maximizes the expected cumulated reward where .

Q-learning [3, 4] is the most popular model-free, off-policy algorithm to find optimal policy and is based on Bellman Equation. Usually applied to the context of discrete state / action spaces, the optimization procedure is based on Q-table as illustrated in Figure 1 below. In more elaborated setting, especially involving covariates, Q-table is replaced by an approximating Q-function fitted on the dataset.



**Figure 2:** Q-learning [4].

Application to Health: healthcare pathway analysis

Reinforcement Learning is a relevant way to deal with several aspects of healthcare pathways analysis for chronic disease. Indeed, a chronic disease is characterized by a sequence of clinical observations or a sequence of care. The natural study context is thus a trajectory in order to take into account the information contained in the observation dependency. The objectives of healthcare pathway analysis are numerous among which:

* To define “typical” trajectories and study the determinants of these,
* To define trajectories that lead to a given state of health and determine the risk factors of being in a given state at a given time,
* To describe and optimize the modalities of therapeutic management of certain patients,
* To understand and optimize the care of patients with chronic pathology,
* To measure the impact of intervention on the treatment pathway through simulation under given conditions,
* To model and standardize care,
* To evaluate the efficiency of support.

Reinforcement learning and especially Q-learning is used in this medical context of long-term patient care setting [5, 6] and can be described as follows. For each patient, the stages correspond to clinical decision points in the course of the patient’s treatment. At these decision points, actions (e.g., treatments) are chosen, and the state of the patient is recorded. As a consequence of a patient’s treatment, the patient receives a (random) numerical reward concretized by a numerical outcome (survival times for instance). The goal consists in finding an optimal decision rule in terms of long-term clinical outcome based on that individual's characteristics and history. Such a strategy is usually named Dynamic Treatment Regime (DTR) (AKA adaptive intervention, adaptive treatment strategy).

The experience of the consortium members: issues observed

DTR is a very active research aera essentially on the pulse of the research activity in Reinforcement Learning. Essentially supported by North America these methods are struggling to cross the Atlantic and remain little applied in France. Following several years of collaborations and several research stays, Erica Moodie and Michael Kosorok, among the foremost experts in DTR [5, 6], have agreed to join this consortium. It is a great opportunity to popularize DTR approach in France.

From a technical point of view, in a pilot study, addressing leukemia patients’ treatment pathways, we seek for an optimal therapeutic strategy, i.e. a combination of treatment lines, based on both the characteristics of the disease and of the patients, regarding the response to treatment and cancer relapse. This study has enabled us to highlight the strengths and the weaknesses of this techniques and the study of possible improvements to these techniques.

Issues to overpass

Three main weaknesses are currently observed in Reinforcement Learning approach [7] and will be investigated in this project:

* **How to approach more complex situations in terms of space of states and action?**

“Model-free” and “Model-based” is one of the most important branching points in an RL algorithm. “Model-based” approach consists in the setting where the agent has access to (or learns) a model of the environment, this means, a predictive model for state transitions and rewards. “Model-free” approach does not necessitate such a model, the strategy to get the optimal policy is thus purely data-driven. This is an essential distinction because the “model-free” approach has the undeniable advantage of being free from the estimation of the underlying Markov Decision Process and the inherent assumptions. In return the “model-free” approach is computationally intensive techniques and, to be relevant, requires a sufficient volume of data. Q-learning is a “model-free” approach, most the applications are constrained to simple topologies for the spaces of states and actions (limited number of states, discretized) to faced this computational issue.

The question of the extension to more complex situations in terms of state / actions spaces requires the use of learning techniques which are more complex. At the center of these learning techniques, deep Q-learning algorithms [8] are promising and will be explored in this project to evaluate their performances in the framework of healthcare pathways analyses.

* **How to quantify / overpass the so-called exploration-exploitation dilemma?**

“Model-free” approach is purely data-driven. The weakness of this approach comes from its propensity to explore the realm of possibilities, especially in a context of a continuous state / action space. This is known as the exploration-exploitation dilemma [9]. This dilemma is a fundamental problem in reinforcement learning. Agent is frequently faced when choosing between options, rather:

* pick something familiar in order to maximize the chance of getting what you wanted,
* pick something not tried and possibly learning more, which may (or may not) result in making better decisions in future.

This trade-off will affect either agent earn his reward sooner or he learns about the environment first then earn his rewards later. Second, in recent years, several more advanced exploration strategies have been proposed to deal with the question of the exploration-exploitation dilemma and will be explore as a second step of this project.

* **How to approach the issue of small databases?**

One of the main limitations in the application of these methods is, on the one hand, the volume of the data and, on the other hand, the computational constraints. Careful attention to small sample techniques is imperative. While still in its early stage, significant progress has been made in small sample learning research in recent years [10]. Several aspect of the so-called Data Augmentation strategy [11] will be explored in a third step of this project.

* 1. Description of possible interactions between mathematics and computer science

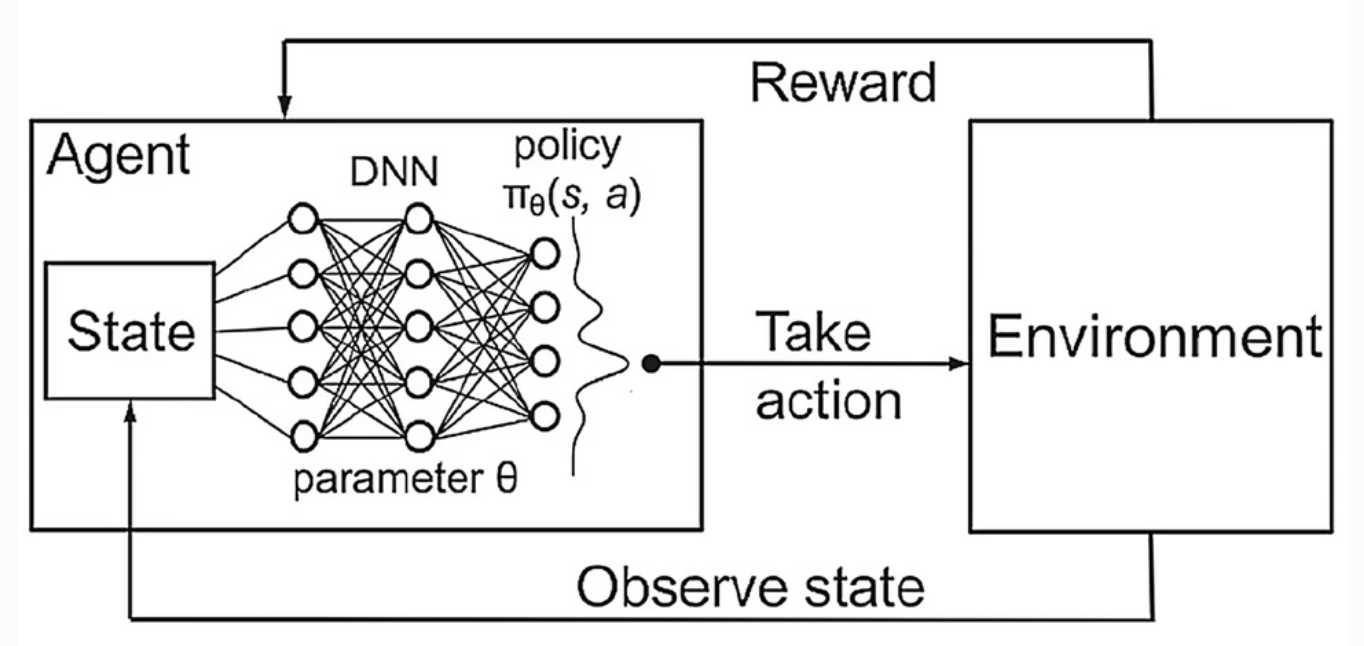
The interaction with computer sciences is clear. The tools this project aims to produce may be of interest for computer sciences. However, the upstream theoretical research has to be concluding and sufficiently relevant to propose such tools to computer science community. In the spirit of this project, this interaction will be the next step in the context of a wider project involving clinicians and computer scientists (see Section 3.4 or details).

* 1. Scientific programme and milestones.

RL4H project splits in three main steps according to the three questions asked in Section 3.1.

**Step 1: From Q-learning to Deep Q-learning**

The question of extension to more complex situations requires the use of learning techniques that are themselves more complex. As part of this project to explore the performance and feasibility of alternative methods at the heart of which deep Q-learning (DQL) [8]. In deep Q-learning, we use a neural network to approximate the Q-value function. The state is given as input and the Q-value of all possible actions is generated as output. Figure 3 illustrates the algorithm and can be compared to Figure 2 to identify the contribution of Deep Learning.

****

**Figure 3:** Deep Q-learning [4].

Thanks to this modeling, deep Q-learning allows to explore much more complicated setting as Q-learninq but like most deep learning algorithms, deep Q-learning presents many technical subtleties that we plan to explore within the framework of this project. The aim is to propose a clear methodology on the potential uses of this technique and its ramifications [12] in the context of the healthcare pathways analysis.

**Step 2: From a data-driven approach to a hybrid data / expert-driven approach**

The question of exploring the space of possible states and actions is of major importance. This problem arises at two levels: at the level of learning where only the domain defined by the data is used for learning and at the level of prediction, the optimal trajectory not necessarily being observed in the data and we use therefore a proxy of this trajectory called “off-policy evaluation”. This is all the more a problem in terms of health, the optimal trajectory may be unobservable (cost, risk of uncontrolled treatment, ….). There is now ample evidence that without explicit tailoring, learning can lead to biased, unsafe, and prejudiced solutions. The question of policy evaluation from observational data is therefore an important methodological issue to be discussed. To address this problem, we plan to use techniques allowing to introduce experts' a priori into learning techniques moving from a purely data-driven paradigm to a hybrid one mixing data-driven and expert-driven techniques. One method attracts our attention in this context and will be studied within the framework of this project is the PAC (Probably Approximately Correct) method [13]. In the context of PAC learning, the “learner” algorithm receives training data (“samples”) and must choose a function which generalizes these data. This function is chosen from a pre-established set possibly by experts’ a priori. The goal is that the function in question classifies new unknown data (distributed identically to the training data) with minimal error, and this with high probability. Finally, the question will benefit from advances in PAC techniques, in particular Probably Approximately Correct Constrained Learning [14].

It is in particular possible to introduce expert a priori or at least hypotheses in the learning methods in order to restrict the exploration space and to force the learning to be limited to the most probable area. These methods are part of the PAC (probably Approximately Correct) methods [13] guaranteeing a targeted exploration of the learning algorithm switching from purely data-driven paradigm to a hybrid data / expert driven one.

**Step 3: From a small to a large database.**

Like any data-based learning approach, the performance of these methods will be directly related to the volume of data. In the context of healthcare pathways, to ensure that we have data of quality, it is tempting to sacrifice massive databases and limit ourselves to small databases. To ensure the applicability of the techniques developed above, it is nevertheless necessary to think about strategies to overcome these constraints of dimension. To do this, a direct approach is the use of "data augmentation" strategies [15]. The idea is relatively naive and consists in artificially increasing the volume of the database by generating synthetic patient’s data, then conventional learning techniques remain to be applied. Methods of virtual patient generations have already been studied and discussed by the consortium [16, 17, 18, 19], in particular using R-vine-type copula methods. As part of this project we plan to enrich virtual patient models using GANs (Generative Adversarial Network [20]). A GAN is a machine learning technique based on the competition of two networks called "generator" and "discriminator". The generator is a type of convolutional neural network whose role is to create new instances of an object (a virtual patient in our setting). On the other hand, the discriminator is a "deconvolutive" neural network that determines the authenticity of the object or whether or not it is part of a data set. During the training process, these two entities compete and this is what allows them to improve their respective behaviors. This is called backpropagation. The objective of the generator is to produce outputs without being able to determine whether they are false, while the objective of the discriminator is to identify false ones. Thus, over the course of the process, the generator produces better quality outputs while the discriminator detects the false ones better and better. In fact, the illusion becomes more and more convincing over time. Initially used for images generation, GANs are more and more popular and is of capital interest for virtual patients’ generation. The main questions that we will investigate in this project is, first, to study the quality of the data generated by GANs especially in terms of structural dependence between covariates and, second, to study and compare these “data augmentation” approaches to understand the analysis of care pathways from small samples.

The counterpart of these investigations is the introduction of more and more sophisticated learning techniques and therefore more and more opaque black boxes. Particularly in the context of health data, it is necessary to attach particular importance to the interpretability of the results. Avenues are open [21, 22] and will be explored within the framework of this project.

# References

[1] R.S Sutton, A.G Barto. Reinforcement Learning: An Introduction (2nd edition). The MIT Press. 2018.

[2] S. Dash, B.R. Acharya, M. Mittal, *Deep Learning Techniques for Biomedical and Health Informatics.* Springer. 2019.

[3] C. Watkins, P. Dayan. *Q-Learning.* Machine Learning volume 8, pages 279–292, 1992.

[4] R. Jafari and M. Javidi. *Solving the protein folding problem in hydrophobic-polar model using deep reinforcement learning.* SN Appl. Sci. 2, 259, 2020.

[5] M.R. Kosorok and E.E.M. Moodie (Eds). *Adaptive Treatment Strategies in Practice*. ASA-SIAM Series on Statistics and Applied Mathematics, 2015.

[6] B. Chakraborty and E.E.M. Moodie. *Statistical Methods for Dynamic Treatment Regimes: Reinforcement Learning, Causal Inference, and Personalized Medicine*. Springer. 2013.

[7] Y. Chao, L. Jiming, and N. Shamim. *Reinforcement learning in healthcare: A survey*. arXiv preprint arXiv:1908.08796, 2019.

[8] V. François-Lavet, P. Henderson, R. Islam, M. G. Bellemare and J. Pineau, *An Introduction to Deep Reinforcement Learning*. Foundations and Trends in Machine Learning: Vol. 11, No. 3-4, 2018.

[9] R. Fruit. *Exploration-exploitation dilemma in Reinforcement Learning under various form of prior knowledge*. Phd Thesis, University of de Lille 1, 2019.

[10] S. W. Carden and J. Livsey, *Small-sample reinforcement learning: Improving policies using synthetic data*. Intelligent Decision Technologies, vol. 11, no. 2, pp. 167–175, 2017.

[11] S.Y. Li. Automating Data Augmentation: Practice, Theory and New Direction. The Stanford AI Lab Blog. <http://ai.stanford.edu/blog/data-augmentation/> 2020.

[12] K. Arulkumaran, M.P. Deisenroth, M. Brundage, A.A. Bharath. *A Brief Survey of Deep Reinforcement Learning.* IEEE signal processing magazine, 2017.

[13] J. Pazis and R. Parr, *PAC optimal exploration in continuous space Markov decision processes*. in AAAI, 2013.

[14] L.F.O. Chamon and A. Ribeiro. *Probably Approximately Correct Constrained Learning*. arXiv:2006.05487v1, 2020.

[15] J. Salamon and J. P. Bello, *Deep convolutional neural networks and data augmentation for environmental sound classification*. IEEE Signal Processing Letters, vol. 24, no. 3, pp. 279–283, 2017.

[16] N. Savy, P. Saint-Pierre, S. Savy, S. Julien, E. Pham, *”In Silico Clinical Trials”: a way to improve drug development?*. Proceedings of JSM 2019 – Biopharmaceutical Section, Denver, 2019.

[17] N. Savy, S. Savy, S. Andrieu, S. Marque. *Simulated Clinical Trials: Principle, Good Practices, and focus on Virtual Patients Generation.* Springer Proceedings in Mathematics & Statistics, chapter 21, 2018.

[18] P. Saint-Pierre, R. Demeulemeester, N. Costa, N. Savy. *Agent-Based Modeling in medical research – Application to Health Economics.* Precision Medicine, book chapter in preparation (2021).

[19] P. Saint-Pierre, N. Savy. *Agent-Based Modeling in medical research – Focus to virtual Patients’ generators.* In preparation for The International Journal of Biostatistics (2021).

[20] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, *Generative adversarial nets*. Advances in neural information processing systems, pp. 2672–2680, 2014.

[21] Z. Che, S. Purushotham, R. Khemani, and Y. Liu. *Distilling knowledge from deep networks with applications to healthcare domain.* arXiv preprint arXiv:1512.03542, 2015.

[22] M. Wu, M.C. Hughes, S. Parbhoo, M. Zazzi, V. Roth, and F. Doshi-Velez, *Beyond sparsity: Tree regularization of deep models for interpretability*. Thirty-Second AAAI Conference on Artificial Intelligence, 2018.