Proceedings of the 3rd International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/5/20230527

# Analysis of reinforce learning in medical treatment

**Ningyan Zhang**

University of California, Irvine, CA, 92697

ningyanz@uci.edu

**Abstract.** As human approaches the big data period, artificial intelligence becomes dominating in almost every domain. As part of machine learning, reinforcement learning (RL) is intended to utilize mutual communication experiences around the world and assess feedback to strengthen human ability in decision-making. Unlike traditional supervised learning, RL is able to sample, assess and order the delayed feedback decision-making at the same time. This characteristic of RL makes it powerful when it comes to exploring a solution in the medical field. This paper investigates the wide application of RL in the medical field. Including two major parts of the medical field: artificial diagnosis and precision medicine, this paper first introduces several algorithms of RL in each part, then states the inefficiency and unsolved difficulty in this area, together with the future investigation direction of RL. This paper provides researchers with multiple feasible algorithms, supported methods and theoretical analysis, which pave the way for future development of reinforcement learning in medical field.

**Keywords:** Artificial diagnosis, Precise Medicine, Reinforcement learning, Upper confidence bound, Thompson sampling.

## 1. Introduction

With tremendous data and calculating models and algorithms coming to the world, artificial intelligence has increased the functionality of medical treatment in the past decades. This trend makes people more and more interested in reinforcement learning and modern strategies of data analysis. In the problems of RL, the agent would choose its action in every current state and then receive and assess the feedback from the environment for the next state. The goal of the agent is to make the optimal decision to maximize its cumulated rewards. Hence, during the learning process, the agent will not receive any direct command of decision making but to have repeated dynamic assessments of the environment to get the optimal decision [1].

Normally, medical treatment has two parts. The first part is diagnosis. For one thing, a single human doctor's knowledge is limited, and humans could make fatal errors due to multiple reasons. For another thing, many people would not choose to go to the hospital for some light symptoms and people in poor areas cannot afford a real doctor. Hence, online artificial doctors would be a time-saving and money-saving alternative. An artificial human doctor would assess the patient's symptoms to give a medical certificate. The second part is precise medicine. After diagnosis, the patient needs to take medical treatment for some courses. Especially for some chronic diseases, like cancers and mental illnesses, when to take pills, how many pills to take at each course need a dynamic treatment to monitor the

DOI: 10.54254/2755-2721/5/20230527

patients' physical conditions. The application of RL for precise medicine can help pick up the optimal treatment at each period.

This paper focuses on the analysis and application of multiple algorithms fit in both artificial diagnosis and precise medicine. This paper also makes the comparison between different algorithms and their efficacy and deficiency, which makes proper suggestions for future innovations and improvement on reinforcement learning for medical field.

## 2. Some algorithms applied to the medical field

### *2.1. Algorithms for artificial diagnosis*

*2.1.1. Dialogue System for Medical Automatic Diagnosis (DSMAD).* For the artificial diagnosis part, people always face a dilemma: insufficient tests would result in a high risk of mistakes, too many tests would be a waste of medical resources. Sometimes it is even hard for people in poor areas to get a reliable medical diagnosis. Hence, the development of DSMAD, the dialogue system for medical automatic diagnosis, which is intended to mimic a human doctor, is necessary [2]. MAD is intended not only to increase the accuracy of diagnosis but also to reduce the cost of collection of symptom data.

MAD includes three steps: (1) patients report their symptoms to the artificial doctor, which is the self-report part; (2) system as a doctor would communicate with the patient for detailed problems related to the symptoms, for example, what activities the patient recently took that could potentially lead to the illness, then allocates the patients to pathological examination like blood test, X-ray, MRI, etc. (3) When a system has sufficient information, it would provide the patient with the final diagnosis [3].

As stated in the previous paragraphs, the importance of MAD is to determine when to stop the symptom check to get sufficient information about the patient but also be money saving. MAD focuses on Dialogue Management (DM), whose model is given below in Figure 1.

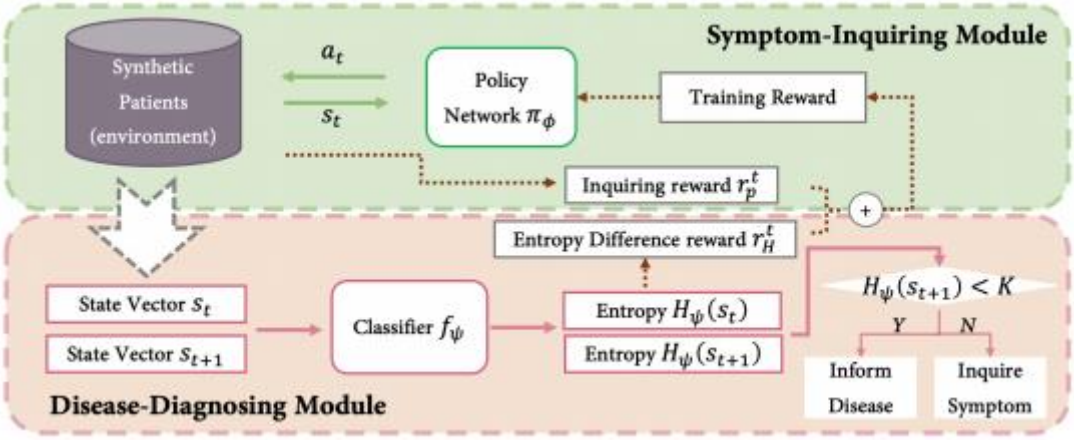


**Figure 1.** Model of DS-MAD [3].

*Source: https://blog.csdn.net/qq_43058281/article/details/122903372*

This model includes the query agent, the disease diagnosis module, and cease standard module. MAD set the threshold value K, by calculating the entropy $H_\psi(S)$, if $H_\psi(S)$ is larger than K, then MAD will stop the symptom check. K is a dynamic value that would keep updating if $H_\psi(S_{t+1}) < K$.

*2.1.2. Upper Confidence Bound (UCB).* UCB algorithm is assigning a UCB index to each arm and then chooses the arm with the largest value. The key to the UCB algorithm is to keep optimistic about the uncertainty of each arm, which means always selecting the arm with the largest UCB index and keep updating the UCB index as more exploration takes place on each arm. As shown in Figure 2, the

DOI: 10.54254/2755-2721/5/20230527

expected value falls between the confidence intervals, the more exploration of the arm, the closer the confidence intervals, and the closer the expected value to the UCB index.

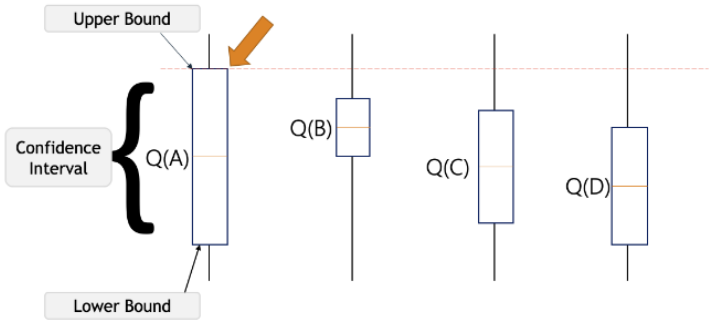


**Figure 2.** Example of UCB algorithm [4].

*Source: Upper Confidence Bound Algorithm in Reinforcement Learning*

In Onur Atan and Mihaela van der Schaar's paper, they invented CDS, a clinical decision system, which refers to RS-UCB, the Relevant Source Upper Confidence Bound algorithm [5]. RS-UCB includes two parts, first to learn what sources to access for specific decisions and second to learn what decisions to make. Since the algorithm needs to include as much data from the internet as possible, and some data could be irrelevant to patients' symptoms, it is important to sift useful information. The RS-UCB algorithm helps derive a lower bound to determine the probability that the policy would choose the wrong source, and the RS-UCB algorithm would have logarithmic regret through time, which regrets are scaled with a number of relevant sources. They proved RS-UCB algorithm has logarithmic regret which means the more sources that the system detected, the less the regret is. Of course, the information that the decision maker gets is only the relevant which is chosen from the first step.

### *2.2. Algorithms for precise medicine and clinical trials*

After looking into the diagnosis part, it is time to shift the focus to medicine precision. Precise medicine is a dynamic notion. It requires the algorithm to keep monitoring the patient's physical condition to make in-time choices of medicine dosage to get the optimal outcome. It is noticeable that even for the same disease, the condition is different for each individual in each period. Hence, it is crucial for each patient to receive their customized medical treatment.

*2.2.1. MAB (Multi-armed Bandit).* The multi-armed bandit algorithm that fits into the precise medicine is introduced. A bandit problem is a sequential game between a learner and an environment. The goal of MAB is to balance the exploration-exploitation trade to gain the maximized cumulative rewards. To be specific, for example, the exploration part could be finding out the treatment that fits the patient best, and the exploitation part is to find what usage of medicine could benefit the patient most. The goal is to build a reliable, dynamic, and adaptative system to achieve optimal outcomes.

*2.2.2. TS (Thompson Sampling).* In Aziz, Kaufmann, and Riviere's study, they use the Thompson Sampling for dose-finding trials [6]. They use the multi-armed bandit model that was first introduced in the 1930s for a 3 phases of clinical trial. Phase 1 is to find the most appropriate dose level that could be used in further phases, which is also called MTD, the Maximum Tolerated Dose. Phase 2 is estimating the smallest dose that could satisfy the efficacy, which is called MED, the Minimal Effective Dose. Phase 3 is an alternative treatment.

Compared with other bandit algorithms which require the knowledge of the value assigned to each arm and have to select the optimal arm, they prefer to use the straighter forward Thompson Sampling as alternative, since in their study, their focus is only to identify the arm with mean closest to the threshold

DOI: 10.54254/2755-2721/5/20230527

value, not to find the arm with the largest mean value. The characteristic of Thompson Sampling is an algorithm that only requires to define the notion of optimal arm.

They defined K arms as doses of different levels, $p_k$, the unknown toxicity probability of dose k and MTD is the dose with $p_k$ closet to the target. They use TS to choose a dose with posterior probability to be MTD randomly knowing the distribution priorly, the posterior distribution could be computed through t times' observations. Then they will choose the doze with the optimal value to be the value of MTD.

Also facing the inventory of constraining to find the optimal treatment selection rule, Zhijin Zhou, Yingfei Wang and Hamed Mamani determine the genetic difference of the patients to build a personalized treatment for each patient using the TS algorithm [7]. They apply the TS algorithm to the cytogenetic information from the clinical data related to each patient to get precise medicine. They apply a multilevel Bayesian linear model to sample the unknown model parameters from the posterior distributions. As more observations were done for the patients, the posterior distribution would be close enough to the true mean. Then they would choose the arm with the optimal value.

*2.2.3. Q-Learning.* Compared to other algorithms, Q-learning is outstanding for its simplicity. It requires no specific model, but only depends on the reward mechanism. Given the expected result, the agent would learn how to reach the expected result with maximum rewards.

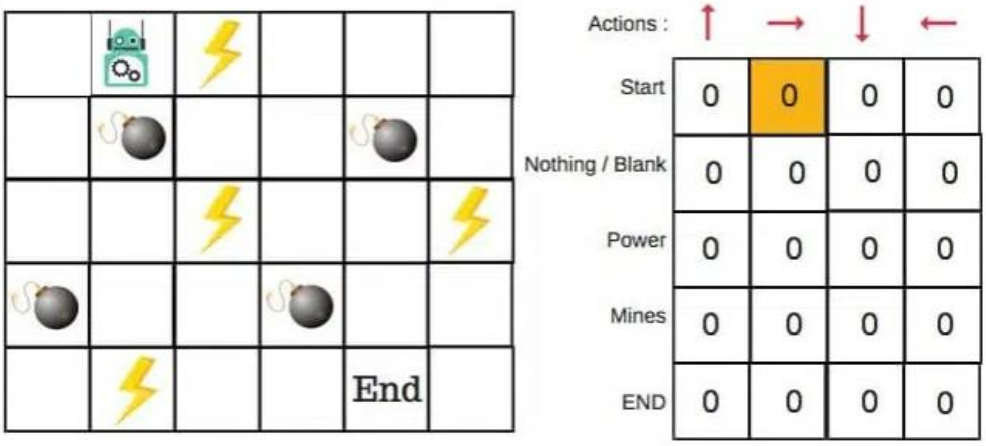


**Figure 3.** Example of Q-learning algorithm [8].

*Source: FreeCodeCamp*

For example, as shown in Figure 3, to reach the end of the maze with booms on the way. If the robot step on the lighting or booms on the route, the reward would be deducted. Hence, to find the shortest route with minimum regrets, the bouts would learn from repeated trials.

In Yufan Zhao, Michael R. Kosorok, and Donglin Zeng's early study, they use Q-learning for precision medicine. They use ODE differential equations for the clinical data of the growth pattern of tumors. With SVG and ERT applied, Q-learning could derive the optimal strategy from clinical data [9].

In addition, in Padmanabhan, Meskin and Haddad's work, they innovatively proposed some reward equations for Q-learning to provide different diplomas for patients with different symptoms [10].

## 3. Applications of reinforcement learning

It is noticeable that RL is adaptative and dynamic which makes it suitable for some chronic diseases. For example, cancer, diabetes, HIV, etc. All of these diseases need the patient to be monitored over different phases of medical treatment.

Here this paper focuses on the studies of cancer. The treatment of cancer includes chemotherapy, surgery and radiotherapy. For chemotherapy, Zhao, etc. proposed the Q-learning method with ODE to

DOI: 10.54254/2755-2721/5/20230527

quantify the growth pattern of tumor from the clinical data [9]. Also, in Ahn and Park's work, they use ODE to study NAC's feasibility for chemotherapy for cancer [11].

In Zhijin Zhou, Yingfei Wang and Mamani's paper, they use the Thompson Sampling for Multiple myeloma, an incurable cancer of bone marrow plasma cells. TS algorithm looks into patients' genetic profiling to observe genetic differences between patients and makes personalized precision medicine for them. The incorporation of cytogenetic information improved the model performance by 19.75% [7].

## 4. Comparison between RL algorithms

With all these innovative algorithms applied to medical treatment, it is hard to determine a single solution that outperforms the others. The goal of the artificial diagnosis part is to use minimum cost but to have access to as much relative information as possible, which makes DSMAD much more powerful than other algorithms. With all the useful information gathered, the UCB algorithm can help select the solution with the optimal expectations while keeping greedy about the uncertainties. For the precise medicine part, researchers would need the dosage to be suitable for the patient at each phase of medical treatment, not the dosage with the highest expected value. This prerequisite makes popular algorithms like UCB less useful. To be mentioned, it is simpler to use the Q-learning method to get a much more general solution quickly. However, which its simplicity, it is hard to take all the circumstances into account. Hence, the use of Thompson sampling would keep updating the value of each dose with access to the posterior distribution.

We cannot judge one algorithm's feasibility in one single scenario since they all have their strengths and weaknesses. It is proper to utilize the strength of each algorithm to the place that suit it most. The combination and innovation of algorithms would make them complementary to each other.

## 5. Conclusion

Although most of the algorithms have access to all the clinical data and medical information available on the internet, and many of them developed approaches to sift useful information from massive data, most of the data is still fragmented in each medical system. With the combination of structured and unstructured output, it slows down the speed to capture relative useful data. The consideration of ethics and economics also slows down the search speed. The algorithm only considers whether the treatment is best for the patients, but omits their feelings as human beings. Moreover, there are no specific standards or laws that regulated artificial intelligence in medical treatment. Cybercriminals could use the data that they get from the medical system and people's privacy is not promised.

Although with all these concerns existed, reinforcement learning still place an important role in medical treatment. Its application is inevitable and still has a bright future. People could say that the emergence of artificial intelligence for medical treatment is subversive and epic. Since the development is still in the early stage of applying AI to the medical field, its limit is almost infinite. People could see all the contributions that former researchers brought to our eyesight and we could make all the innovations and adaptations based on that.

## References

[1] Chao Yu, Jiming Liu, Shamim Nemti: Reinforcement Learning in Healthcare: A Survey. IEEE, arXiv, 2019, pp.1-23. doi: https://arxiv.org/pdf/1908.08796v4.pdf

[2] Hongyi Yuan, Sheng Yu: Efficient Symptom Inquiring and Diagnosis via Adaptive Alignment of Reinforcement Learning and Classification. arXiv. 2021, pp.1-7. doi: https://arxiv.org/pdf/2112.00733.pdf.

[3] Wwwilling, Efficient Symptom Inquiring and Diagnosis via Adaptive Alignment of Reinforcement Learning and Classification. CSDN, 2022. doi: https://blog.csdn.net/qq_43058281/article/details/122903372.

[4] Samishawl, Upper Confidence Bound Algorithm in Reinforcement Learning. Geeksforgeeks, 2020. doi:https://www.geeksforgeeks.org/upper-confidence-bound-algorithm-in-reinforcement-learning/

DOI: 10.54254/2755-2721/5/20230527

[5] Onur Atan, Mihaela van der Schaar, Discover Relevant Sources: A Multi-armed Bandit Approach. UCLA medianetlab, 2015, pp.1-4. doi: http://medianetlab.ee.ucla.edu/papers/Sourceselection.pdf

[6] Maryam Aziz, Emilie Kaufmann, Marie-Karelle Riviere: On Multi-armed Bandit Designs for Dose-Finding Clinical Trials. Journal of Machine Learning Research, 2021, pp.1-6. doi: https://www.jmlr.org/papers/volume22/19-228/19-228.pdf.

[7] Zhijin Zhou, Yingfei Wang, Hamed Mamani: How do Tumor Cytogenetics Inform Cancer Treatments? Dynamic Risk Stratification and Precise Medicine Using Multi-armed Bandits. SSRN, 2019, pp.1-8.

[8] ADL: An introduction to Q-learning: reinforcement learning. FreeCodeCamp, 2018. doi: https://www.freecodecamp.org/news/an-introduction-to-q-learning-reinforcement-learning-14ac0b4493cc/.

[9] Yufan Zhao, Michael R, Kosorok, Donglin Zeng: Reinforcement learning design for cancer clinical trials. Onlinelibrary, 2009. doi: https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3720.

[10] Regina Padmanabhan, Nader Meskin, Wassim M. Haddad: Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment. Sciencedirect, 2017. doi: https://www.sciencedirect.com/science/article/abs/pii/S0025556417304327.

[11] Inkyung Ahn. Jooyoung Park: Drug scheduling of cancer chemotherapy based on natural actor-critic approach. Sciencedirect, 2011, pp.121-129. doi: https://www.sciencedirect.com/science/article/abs/pii/S0303264711001365.