Reinforcement Learning for Healthcare

Claire Coffey
Supervised by Prof. Christopher Yau

HDRUK PhD Spring Project - June 2021

---

1 INTRODUCTION

Reinforcement Learning (RL) is a machine learning technique that has exploded in popularity in recent years, especially in areas with dynamic worlds such as real-time gaming and robotics. It is yet to gain widespread popularity for use in healthcare, but there is plenty of research emerging in this area. Reinforcement learning has a multitude of variants, most of which fall into two distinct categories: those that generate deterministic policies, known as value-based techniques (such as Q-learning); and those that yield stochastic policies, known as policy-gradient techniques, which have been adapted to form actor-critic approaches. We will begin by introducing the various RL paradigms and surveying papers across these variants. We will discuss relevant literature in the context of RL for healthcare for each approach, giving examples of successes and challenges. Moreover, we will elaborate on offline RL [32], a promising new approach for use in healthcare, that has not yet been explored in much depth in this context. We will present a potential use case in the form of multimorbidity trajectory modelling, using the offline RL variant of conservative Q-learning. We will finish by summarising the trends, relevant challenges, and elaborating on opportunities. We hope that this literature review will be useful for future research projects, whether that be applying RL to a specific healthcare problem, or even extending existing RL methodologies to advance the field.

2 MARKOV DECISION PROCESSES

Reinforcement learning is built on the underlying statistical model of Markov decision processes (MDPs) [47], which model an environment as a set of states and actions. MDPs are the standard formalism used in sequential decision making problems from robotics to game playing. These MDPs can then be solved by an agent moving through the environment in an RL algorithm. RL algorithms generate policies, which are the sequence of states and actions taken when an agent moves through an environment. The agent receives a reward at each state. The RL algorithm uses this to learn the optimal policy, that is, the policy with the largest cumulative reward, solving the MDP.

3 REINFORCEMENT LEARNING PARADIGMS

Reinforcement learning approaches can generally be split into two categories: value-based and policy-gradient, both of which are suited to different problems. We will discuss these two paradigms, including examples of their most common variants (Q-learning and actor-critic). We will also summarise deep RL (and its popular variant, deep Q-networks), as this is regularly utilised in RL for healthcare. We will present Q-learning, actor-critic, and DQN in more depth, giving specific examples of their usages in healthcare. We chose to focus on these as according to a recent survey [10], these are the most popular approaches for RL in healthcare, as shown in Figure 1. We will finish this section by discussing offline reinforcement learning, specifically introducing its variant - conservative Q-learning. This is a new variant of RL that has yet to be utilised in healthcare, but theoretically provides a lot of promise for use in this area.

3.1 VALUE-BASED

Value-based RL algorithms yield deterministic policies. They use a value function to estimate a long-term expected value of each possible action given a state, and choose actions based on the highest expected cumulative reward over all successive steps to form a "greedy" policy. Example algorithms using this

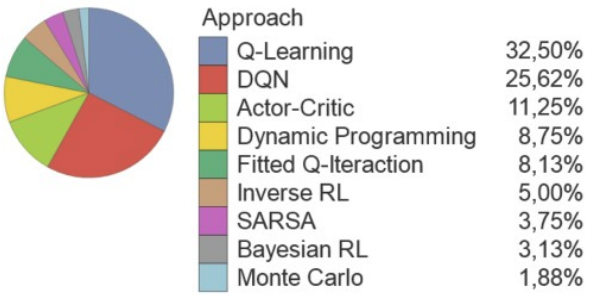

Figure 1: Most common RL approaches for healthcare applications. Figure from [10].

approach are SARSA (state-action-reward-state-action)[6] , dynamic programming, and perhaps most popularly, Q-learning [56], which will be discussed in the upcoming section. The value function can be defined in various ways - which are variant and problem specific - including mean squared error, residual gradient, temporal difference, and dynamic programming errors. Value functions for estimation do have some drawbacks; notably, that they find deterministic policies, ignoring probabilistic policies even when the optimal policy is stochastic. Therefore, it is not well-suited to many real-world stochastic problems [49], with the "optimal" policy potentially being quite suboptimal. Moreover, the policies chosen are quite brittle: a small change in an estimated value can cause a different action to be chosen, so small changes in the estimated value function can cause disproportionately large effects on the policy, which can cause difficulties with convergence [8, 4].

Q-LEARNING

As mentioned, Q-learning is a common value-based RL algorithm resulting in deterministic policies. It learns the optimal policy by learning the optimal Q-values for each state-action pair. This is calculated using the Q-function, which takes a state and an action and returns the expected discounted reward when following a given policy. Q-learning works using the concept of value iteration.

Value Iteration Each state-action pair has a Q-value. Q-learning iteratively updates this using the Bellman equation by moving through the possible states and actions, considering the explore/exploit tradeoff to decide which action to choose based on their corresponding rewards. It does this until the Q-function converges to the optimal Q-function, q*.

Explore/exploit trade-off If the agent chooses to exploit too soon, it may choose the same action over and over again, not exploring enough alternative options, getting stuck in local maxima. Moreover, if the agent chooses to explore forever, ignoring previous Q-values, it will not converge. Therefore, the tradeoff can be configured by defining the exploration rate, which is the probability of choosing to explore rather than exploit, and the learning rate, which is how quickly an agent decides to update a Q-value for a given state-action pair. More detailed information can be found in our jupyter notebook [9].

Applications Q-learning has been applied to many healthcare problems, examples of which are given in Table 1. It also forms the basis of an interesting offline RL approach for which there are potential healthcare applications: conservative Q-learning (CQL) - both offline RL and CQL will too be discussed later in this report.

Discussion Q-learning is undoubtedly the most popular RL technique used in healthcare, and it has been applied to many different healthcare problems across many diseases, as illustrated in Table 1 (meanwhile, as we will discuss in the upcoming section, actor-critic and DQN have a narrower set of applications). It is especially popular for use in personalised medicine and treatment optimisation and scheduling, including that of modelling and optimising treatment pathways [41, 5]. However, these are generally applied to the modelling and treatment of a single disease at a time (for example ADHD or obesity treatment). This trend of using Q-learning for personalised medicine has been suggested to be generalisable to any disease treatment, as in [17, 50] - however, they still treat only one disease at a time. Later, we will propose how we believe a recent offline extension of Q-learning, conservative Q-learning, could further extend this use case of personalised medicine to modelling multimorbidity trajectories and treatments - as this is clearly a gap in current research. In [53], it has been suggested that electronic health

Table 1: Healthcare applications of Q-learning

Paper	Year	Application
[15]	2005	Individualised treatment optimisation for anaemia
[48]	2006	Medical image segmentation
[58]	2009	Blood glucose control in diabetes
[61, 62]	2009, 2011	Designing individualised treatments in cancer clinical trials
[37]	2009	Optimising dosages for hemodialysis patients
[20]	2010	Chemotherapy scheduling optimisation
[39]	2011	Anaesthesia hypnosis control
[41]	2012	Personalised adaptation of treatment intervention pathways for ADHD
[17]	2012	Dynamic treatment optimisation for clinical trials with censored data
[40]	2014	Personalised anaesthesia optimisation
[30]	2014	Optimising treatment allocation
[13]	2014	Optimising anaemia treatment in hemodialysis patients
[50]	2014	Optimising personalised depression treatment in simulated clinical trial
[44]	2015	Anaesthesia and mean arterial pressure control
[51]	2015	Dynamic treatment regimen optimisation
[12]	2016	Optimising dynamic treatment of schizophrenia
[22]	2017	Personalising dosing strategies in simulated cancer trial
[28]	2017	Personalised treatments in control trials for blood and marrow transplants
[45]	2017	Optimising cancer drug dosage in chemotherapy
[43]	2018	Type 1 diabetes control
[5]	2018	Personalised treatment optimisation in control trials for obesity management

records capturing many measurements could be input into a Q-learning algorithm to aid doctor's decision making across multiple diseases. However, we struggled to understand the outcome of this paper and it left much room for improvement, so the gap of modelling multimorbidity trajectories remains unfilled. Furthermore, these usages of Q-learning are in simulated studies and control trials, including recent efforts such as [22], highlighting the difficulties of Q-learning usage in practice.

3.2 POLICY-GRADIENT

The other major RL paradigm is that of policy-based and policy-gradient approaches. These combat the issues caused by the deterministic policies of value-based approaches by generating stochastic policies. Instead of learning an approximation of the value function and basing the policy on the long term expected reward, these search the policy space directly and attempt to follow the *gradient* of the average reward - directly optimising the policy. These also generally converge more easily [7, 52]. However, these are not suited for all problem domains, with convergence time too slow due to large variances across gradient estimators, and high-levels of domain knowledge necessary [7], as well as each new gradient being estimated independently, so older information is not learnt. A popular variant to combat these issues is the actor-critic approach [27], which will be discussed below. Overall, it has been highlighted that although policy-gradient techniques are more theoretically sound than value-based, they do not usually perform well in real-world tasks, whereas value-function approaches have been shown to be successful in many domains.

ACTOR-CRITIC

Actor-critic methods are an attempt to combine the strong points of the "actor only" methodology of policy-gradient methods with the "critic only" value-based methods. They can be viewed as an extension of policy-gradient techniques, as they are a stochastic gradient algorithm on the parameter space of the actor. Simply, they work by learning a value function to update the actor's policy parameters in the direction of performance improvement. This potentially combats the difficulties faced with convergence in critic-only methods, and they converge more quickly than actor-only methods due to them having less variance.

Applications Due to the positive characteristics mentioned above, actor-critic approaches have been explored in various domains, including somewhat in healthcare; examples are given in Table 2.

Table 2: Healthcare applications of actor-critic

Paper	Year	Application
[23]	2004	Biological arm movement
[46]	2011	Optimising control of prosthetic limbs
[35]	2013	Personalised dosing of anaesthesia
[11]	2016	Optimising insulin infusion for diabetes glucose regulation
[24]	2017	Arm movement controller
[42]	2018	Controlling blood glucose for diabetes
[55]	2018	Dynamic treatment recommendation
[1]	2019	3-D anatomical landmark localisation
[33]	2020	Optimising sepsis treatment

Discussion As illustrated by Table 2, the majority of uses of actor-critic are related to robotics - in particular, the online control of prosthetic limbs as in [46, 24, 23]. This is especially common in older uses of actor-critic, such as [23], which is unsurprising as reinforcement learning has been a popular choice for learning in robotics for a long time - it is well suited to the exploring of live environments such as this. Another online use case popular for actor-critic is that of glucose management for diabetic patients. More recent trends show the expansion of use cases of actor-critic, such as in optimisation of treatments, as in [33, 55]. This aligns with the trend seen in uses of Q-learning (see Table 1) - a movement towards personalised medical treatments. This also agrees with our future trend forecast for the field of RL in healthcare, in that offline data such as electronic health records can hope to be utilised to generate personalised predictions - moving towards offline RL techniques as these are better suited to this problem. We will discuss this later.

3.3 DEEP REINFORCEMENT LEARNING

Simply, deep reinforcement learning works by combining neural networks with any reinforcement learning architecture. Neural networks are function approximators, so they are especially useful in RL when the state space or action space are not known or too large. Therefore, neural networks are used to approximate value or policy functions, to combine them with value-based or policy-gradient techniques; the network learns to map states to values or state-action pairs to Q-values. Building on our previous explanations, there exist deep variants of Q-learning (Deep Q-Networks) [38], and actor-critic models [19], however these are not specific to healthcare and it has been noted that the behaviour of these are often opaque and uncertain [21]. Interpretability is famously important in any deep learning for healthcare - the lack of this is a barrier to real-world use across many potential applications of deep RL.

DEEP Q-NETWORKS

Deep Q-networks (DQN) [38] are a popular Deep RL variant - they exploded in popularity after Deepmind's AlphaGo system beat a grandmaster in the game of Go - using DQN. The combination of neural networks with Q-learning, as mentioned above, allows for problems with larger state spaces to be solved. This is because Q tables no longer need to store all states and their values in DQN, and instead a neural network can be trained on samples from the state or action space to learn the value function. They have become more and more popular as computing power has increased, especially for use in situations for which no rules are defined, and the system learns in a live setting.

Applications The popularity of DQN have also expanded to use in healthcare, and it is claimed in [14] to be especially useful where learning requires physician demonstration such as in surgical robotics [26], where agents can learn from surgeons' physical motions through computer vision systems. Figure 1 highlights the popularity of DQN in healthcare, and Table 3 contains examples of their usage.

Discussion Deep Q-Networks are popular for use in image recognition in general, and this extends to popularity for their use in medical imaging in the healthcare space, as evidenced in Table 3. This is mostly made up of uses for image registration/standardisation [34, 36, 2] and landmark detection [16, 3]. It is also popular for natural language processing - handling medical freetext for various applications, for example, the interesting use cases of automatic disease diagnosis [25, 57]. DQN constitutes a large portion of usages of RL in healthcare (see Figure 1). We found that applications of DQN are all more recent

Table 3: Healthcare applications of DQN

Paper	Year	Application
[54]	2016	Symptom checking
[34]	2017	3-D image registration
[36]	2017	Image registration
[16]	2017	3D landmark detection in CT scans
[60]	2018	Vessel centerline tracing
[2]	2018	Standardising view planes in 3D images
[25]	2018	Disease diagnosis from symptoms
[57]	2018	Automated disease diagnosis dialogue system
[3]	2019	Anatomical landmark detection

in comparison to Q-learning or actor-critic approaches, as deep learning has increased in popularity and large datasets become more available. Moving aside from DQN in the area of deep RL, there are many deep variants of value-based, or policy-gradient approaches, as well as hybrid approaches that use deep learning for part of the training for example. This trend is also gaining traction, and so with increased data availability and computing power, it is likely that uses of both DQN and other deep RL frameworks will become more commonplace.

3.4 OFFLINE REINFORCEMENT LEARNING

The live data collection requirement of traditional RL methods mean that they are not applicable to many real-world problems, where only historical data is available. This includes most healthcare problems, such as treatment optimisations and modelling disease trajectories, using retrospective data. Offline RL [32, 31] is exclusively data-driven, and there is no live interaction with the environment needed - this hugely increases the potential possibilities. However, offline RL is very challenging, and traditional RL methods are not equipped for it. A number of offline variants have therefore been proposed, but these tend to have difficulties with function approximation and distributional shift. Distributional shift occurs because the original data must be deviated from (for example, in the form of counterfactual predictions) in order to produce good policies. However, if there is a significant disparity between the offline training data and the the states visited by the learned policy, the predictions become unreliable. RL algorithms tend to overestimate the value of unseen outcomes, which results in distributional shift. A promising new variant of offline learning - that addresses these issues - is conservative Q-learning (CQL) [29]. This is a simple extension of Q-learning, but it has not yet been shown to work in a healthcare context.

CONSERVATIVE Q-LEARNING

CQL addresses the distributional shift by giving a conservative estimate of the value of unseen outcomes. This ensures that the estimated policy executing these actions will not be overestimated, and hence, it will be reliable. The resulting value function learned by CQL is actually therefore pessimistic - no value for an unseen outcome is overestimated. It does this using regularisation. A regularisation parameter is added to Q-learning and it modifies the Q-function training objective. This regularisation is the difference between the actual states visited, and those from the initial data - we want the difference to be small for reliable policies, and so large disparities are penalised using this. The regularisation is applied to the Q-values of the unseen actions, to stop these being overestimated in the Q-function if they differ from the observed data, resulting in minimised Q-values for these. The Q-function is trained with a sum of this regularised value and the standard temporal difference error, and simultaneously maximises the expected Q-value for the existing dataset. This guarantees that the expected return of the learned policy is less than the actual performance. The regularisation parameter is easy to alter to make it applicable to different problems, theoretically including healthcare problems. It is simple to use on top of any standard (deep) RL algorithms - since all that must be done is to plug in the conservative policy estimate. However, it is difficult to get the regularisation parameter right - there is a trade-off in this respect: too little regularisation and the counterfactual values are overestimated; not enough and they are too conservative, so are ignored and the observed data only is used.

Application: multimorbidity trajectory modelling CQL has been evaluated [29], and shown to perform well by using offline data, massively outperforming alternative offline RL algorithms. In theory,

offline RL (including CQL) works with large, complex datasets. Health data, for example, electronic health records, are an example of this - suggesting many potential opportunities. It has not yet been applied to healthcare, but generalisable solutions are produced. We therefore suggest that it could be very powerful when applied to the problem of multimorbidity trajectory modelling using retrospective data.

4 DISCUSSION

In this section, we will summarise our findings by highlighting trends observed in the literature, current challenges identified, and mentioning potential opportunities.

4.1 TRENDS

Although RL is not established in healthcare the same way as it is for robotics or game-playing, there are many, many examples of its use. Recent surveys have been done [10, 59] including over 100 papers to compare many of these. These are done in an application-focused manner - emphasising potential use-cases; we have instead presented an approach-focused discussion of uses. Figure 2 - in the Appendix - highlights the breadth of potential application areas of RL in healthcare. The majority of examples we have seen fall under "automated medical diagnosis: medical image" using structured data, or "dynamic treatment regimes", especially in simulated settings with 'live' data collection. Moreover, refer back to Figure 1 for the popularity of different RL approaches in healthcare, and Tables 1, 2, and 3 for our methodology-based presentation of relevant examples. Here, we saw that different RL approaches are better suited to different tasks, as there were key application themes for each approach.

4.2 CHALLENGES

The major current challenge is the movement from theory to practice; very few of these systems are actually used in the clinic. There are many barriers, especially regarding interpretability (especially in deep RL) and patient and clinician trust, that play into this. Recent works have addressed this by proposing frameworks for pre-clinical validation of RL methods for patient treatment decisions [33, 18]. In particular, the focus of the guidelines in [18] are to encourage RL to be used in a safe manner in practice. However, the authors consider the many challenges to this, mentioning the importance of interpretability and presenting optimistic guidelines, but with a large dose of caution. Therefore, at this stage, it is difficult to imagine a reality wherein these systems are really used in practice.

4.3 OPPORTUNITIES

We foresee a trend towards modelling disease and treatment pathways, especially that of multimorbidities. This is as electronic health records and similar data are becoming more available, and there are new exciting approaches in offline RL, suited to utilising this retrospective data. Referring to Figure 2 - in the Appendix - this falls into the categories of "structured data" and "health management". Modelling multimorbidity trajectories is a difficult problem; the existing examples for optimising treatment intervention pathways, for example in [41], are for a single disease. This is a key difference as the problem becomes much more complex when modelling multimorbidities. We hope that new advances in offline RL, such as CQL, can be used to address this gap. There is plenty of opportunity in this exciting area, for furthering the development of RL algorithms themselves so they are well-suited to these problems, as well as implementing new ways to learn from vast datasets offline. Further, there s opportunity for creating new frameworks to increase trust and interpretability, narrowing the gap between theory and practice.

References

[1] Walid Abdullah Al and Il Dong Yun. "Partial policy-based reinforcement learning for anatomical landmark localization in 3d medical images". In: *IEEE transactions on medical imaging* 39.4 (2019), pp. 1245–1255.

[2] Amir Alansary et al. "Automatic view planning with multi-scale deep reinforcement learning agents". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 277–285.

[3] Amir Alansary et al. "Evaluating reinforcement learning agents for anatomical landmark detection". In: *Medical image analysis* 53 (2019), pp. 156–164.

[4] Leemon Baird. "Residual algorithms: Reinforcement learning with function approximation". In: *Machine Learning Proceedings 1995*. Elsevier, 1995, pp. 30–37.

[5] Abiral Baniya. "Adaptive interventions treatment modelling and regimen optimization using sequential multiple assignment randomized trials (SMART) and Q-learning". In: (2018).

[6] Andrew G Barto and Sridhar Mahadevan. "Recent advances in hierarchical reinforcement learning". In: *Discrete event dynamic systems* 13.1 (2003), pp. 41–77.

[7] Jonathan Baxter and Peter L Bartlett. "Infinite-horizon policy-gradient estimation". In: *Journal of Artificial Intelligence Research* 15 (2001), pp. 319–350.

[8] Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[9] Claire Coffey. *RL for Healthcare 1 - MDPs and Q-learning*. `https://github.com/clairecoffey/RL-proj/blob/main/RL_healthcare_MDP_QLearning.ipynb`. 2021.

[10] Antonio Coronato et al. "Reinforcement learning for intelligent healthcare applications: A survey". In: *Artificial Intelligence in Medicine* 109 (2020), p. 101964.

[11] Elena Daskalaki, Peter Diem, and Stavroula G Mougiakakou. "Model-free machine learning in biomedicine: Feasibility study in type 1 diabetes". In: *PloS one* 11.7 (2016), e0158722.

[12] Ashkan Ertefaie, Susan Shortreed, and Bibhas Chakraborty. "Q-learning residual analysis: application to the effectiveness of sequences of antipsychotic medications for patients with schizophrenia". In: *Statistics in medicine* 35.13 (2016), pp. 2221–2234.

[13] Pablo Escandell-Montero et al. "Optimization of anemia treatment in hemodialysis patients via reinforcement learning". In: *Artificial intelligence in medicine* 62.1 (2014), pp. 47–60.

[14] Andre Esteva et al. "A guide to deep learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 24–29.

[15] Adam E Gaweda et al. "Individualization of pharmacological anemia management using reinforcement learning". In: *Neural Networks* 18.5-6 (2005), pp. 826–834.

[16] Florin-Cristian Ghesu et al. "Multi-scale deep reinforcement learning for real-time 3D-landmark detection in CT scans". In: *IEEE transactions on pattern analysis and machine intelligence* 41.1 (2017), pp. 176–189.

[17] Yair Goldberg and Michael R Kosorok. "Q-learning with censored data". In: *Annals of statistics* 40.1 (2012), p. 529.

[18] Omer Gottesman et al. "Guidelines for reinforcement learning in healthcare". In: *Nature medicine* 25.1 (2019), pp. 16–18.

[19] Tuomas Haarnoja et al. "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 1861–1870.

[20] Amin Hassani et al. "Reinforcement learning based control of tumor growth with chemotherapy". In: *2010 International Conference on System Science and Engineering*. IEEE. 2010, pp. 185–189.

[21] Todd Hester et al. "Deep q-learning from demonstrations". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[22] Kyle Humphrey. "Using reinforcement learning to personalize dosing strategies in a simulated cancer trial with high dimensional data". In: (2017).

[23] Jun Izawa, Toshiyuki Kondo, and Koji Ito. "Biological arm motion through reinforcement learning". In: *Biological cybernetics* 91.1 (2004), pp. 10–22.

[24] Kathleen M Jagodnik et al. "Training an actor-critic reinforcement learning controller for arm movement using human-generated rewards". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 25.10 (2017), pp. 1892–1905.

[25] Hao-Cheng Kao, Kai-Fu Tang, and Edward Chang. "Context-aware symptom checking for disease diagnosis using hierarchical reinforcement learning". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018.

[26] Yohannes Kassahun et al. "Surgical robotics beyond enhanced dexterity instrumentation: a survey of machine learning techniques and their role in intelligent and autonomous surgical actions". In: *International journal of computer assisted radiology and surgery* 11.4 (2016), pp. 553–568.

[27] Vijay R Konda and John N Tsitsiklis. "Actor-critic algorithms". In: *Advances in neural information processing systems*. Citeseer. 2000, pp. 1008–1014.

[28] Elizabeth F Krakow et al. "Tools for the precision medicine era: how to develop highly personalized treatment recommendations from cohort and registry data using q-learning". In: *American journal of epidemiology* 186.2 (2017), pp. 160–172.

[29] Aviral Kumar et al. "Conservative q-learning for offline reinforcement learning". In: *arXiv preprint arXiv:2006.04779* (2020).

[30] Eric B Laber, Kristin A Linn, and Leonard A Stefanski. "Interactive model building for Q-learning". In: *Biometrika* 101.4 (2014), pp. 831–847.

[31] Sergey Levine. *Decisions from Data: How Offline Reinforcement Learning Will Change How We Use Machine Learning*. `https://medium.com/@sergey.levine/decisions-from-data-how-offline-reinforcement-learning-will-change-how-we-use-ml-24d98cb069b0`. 2021.

[32] Sergey Levine et al. "Offline reinforcement learning: Tutorial, review, and perspectives on open problems". In: *arXiv preprint arXiv:2005.01643* (2020).

[33] Luchen Li, Ignacio Albert-Smet, and Aldo A Faisal. "Optimizing medical treatment for sepsis in intensive care: from reinforcement learning to pre-trial evaluation". In: *arXiv preprint arXiv:2003.06474* (2020).

[34] Rui Liao et al. "An artificial agent for robust image registration". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1. 2017.

[35] Cristobal Lowery and A Aldo Faisal. "Towards efficient, personalized anesthesia using continuous reinforcement learning for propofol infusion control". In: *2013 6th International IEEE/EMBS Conference on Neural Engineering (NER)*. IEEE. 2013, pp. 1414–1417.

[36] Kai Ma et al. "Multimodal image registration with deep context reinforcement learning". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2017, pp. 240–248.

[37] José D Martın-Guerrero et al. "A reinforcement learning approach for individualizing erythropoietin dosages in hemodialysis patients". In: *Expert Systems with Applications* 36.6 (2009), pp. 9737–9742.

[38] Volodymyr Mnih et al. "Human-level control through deep reinforcement learning". In: *nature* 518.7540 (2015), pp. 529–533.

[39] Brett L Moore, Anthony G Doufas, and Larry D Pyeatt. "Reinforcement learning: a novel method for optimal control of propofol-induced hypnosis". In: *Anesthesia & Analgesia* 112.2 (2011), pp. 360–367.

[40] Brett L Moore et al. "Reinforcement learning for closed-loop propofol anesthesia: a study in human volunteers". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 655–696.

[41] Inbal Nahum-Shani et al. "Q-learning: A data analysis method for constructing adaptive interventions." In: *Psychological methods* 17.4 (2012), p. 478.

[42] Phuong D Ngo et al. "Control of blood glucose for type-1 diabetes by using reinforcement learning with feedforward algorithm". In: *Computational and mathematical methods in medicine* 2018 (2018).

[43] Phuong D Ngo et al. "Reinforcement-learning optimal control for type-1 diabetes". In: *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*. IEEE. 2018, pp. 333–336.

[44] Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. "Closed-loop control of anesthesia and mean arterial pressure using reinforcement learning". In: *Biomedical Signal Processing and Control* 22 (2015), pp. 54–64.

[45] Regina Padmanabhan, Nader Meskin, and Wassim M Haddad. "Reinforcement learning-based control of drug dosing for cancer chemotherapy treatment". In: *Mathematical biosciences* 293 (2017), pp. 11–20.

[46] Patrick M Pilarski et al. "Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning". In: *2011 IEEE international conference on rehabilitation robotics*. IEEE. 2011, pp. 1–7.

[47] Martin L Puterman. *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

[48] Farhang Sahba, Hamid R Tizhoosh, and Magdy MA Salama. "A reinforcement learning framework for medical image segmentation". In: *The 2006 IEEE International Joint Conference on Neural Network Proceedings*. IEEE. 2006, pp. 511–517.

[49] Satinder P Singh, Tommi Jaakkola, and Michael I Jordan. "Learning without state-estimation in partially observable Markovian decision processes". In: *Machine Learning Proceedings 1994*. Elsevier, 1994, pp. 284–292.

[50] Yousuf M Soliman. "Personalized medical treatments using novel reinforcement learning algorithms". In: *arXiv preprint arXiv:1406.3922* (2014).

[51] Rui Song et al. "Penalized q-learning for dynamic treatment regimens". In: *Statistica Sinica* 25.3 (2015), p. 901.

[52] Richard S Sutton et al. "Policy gradient methods for reinforcement learning with function approximation." In: *NIPs*. Vol. 99. Citeseer. 1999, pp. 1057–1063.

[53] M Swathib and KC Sreedhara. "A Novel Approach To Doctor's Decision Making System Using Q Learning". In: *European Journal of Molecular & Clinical Medicine* 7.11 (2021), pp. 4203–4209.

[54] Kai-Fu Tang et al. "Inquire and diagnose: Neural symptom checking ensemble using deep reinforcement learning". In: *NIPS Workshop on Deep Reinforcement Learning*. 2016.

[55] Lu Wang et al. "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*. 2018, pp. 2447–2456.

[56] Christopher JCH Watkins and Peter Dayan. "Q-learning". In: *Machine learning* 8.3-4 (1992), pp. 279–292.

[57] Zhongyu Wei et al. "Task-oriented dialogue system for automatic diagnosis". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. 2018, pp. 201–207.

[58] Sh Yasini, MB Naghibi-Sistani, and A Karimpour. "Agent-based simulation for blood glucose control in diabetic patients". In: *International Journal of Applied Science, Engineering and Technology* 5.1 (2009), pp. 40–49.

[59] Chao Yu, Jiming Liu, and Shamim Nemati. "Reinforcement learning in healthcare: A survey". In: *arXiv preprint arXiv:1908.08796* (2019).

[60] Pengyue Zhang, Fusheng Wang, and Yefeng Zheng. "Deep reinforcement learning for vessel centerline tracing in multi-modality 3D volumes". In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2018, pp. 755–763.

[61] Yufan Zhao, Michael R Kosorok, and Donglin Zeng. "Reinforcement learning design for cancer clinical trials". In: *Statistics in medicine* 28.26 (2009), pp. 3294–3315.

[62] Yufan Zhao et al. "Reinforcement learning strategies for clinical trials in nonsmall cell lung cancer". In: *Biometrics* 67.4 (2011), pp. 1422–1433.

Appendices

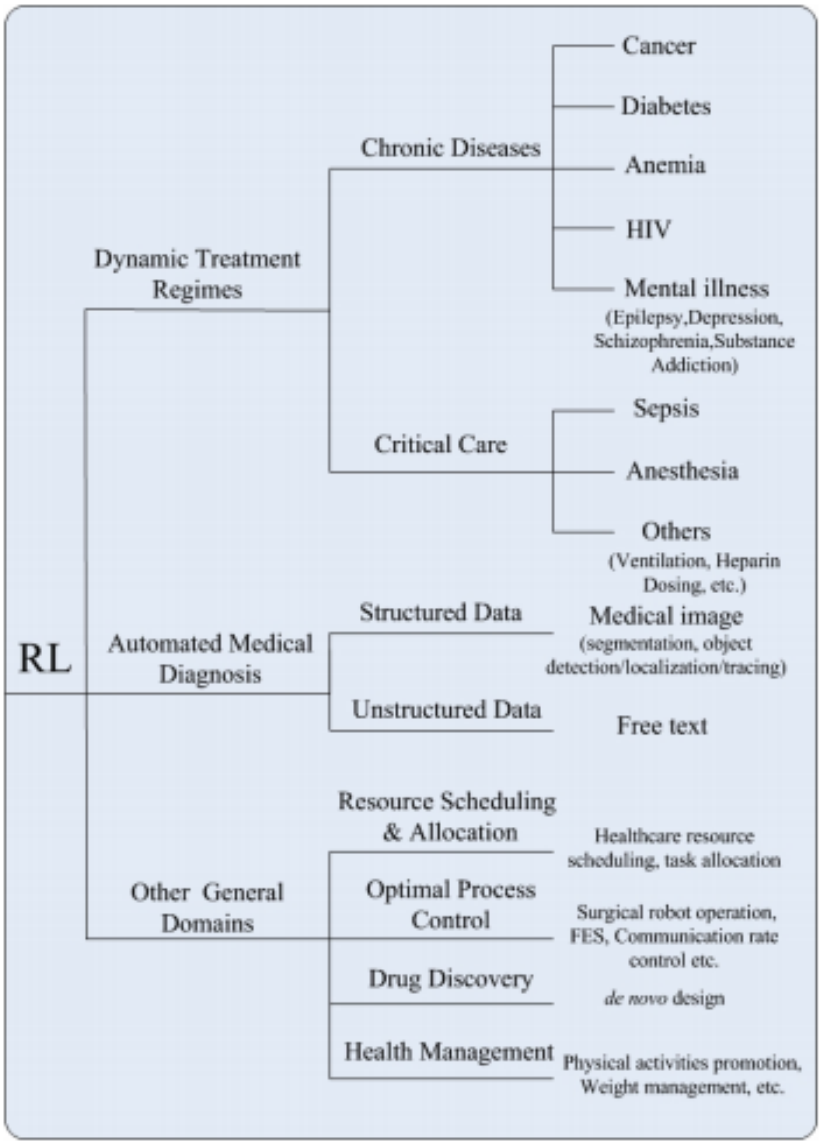

Figure 2: RL application areas within healthcare, figure from [59].