npj | digital medicine

Review article

Published in partnership with Seoul National University Bundang Hospital

https://doi.org/10.1038/s41746-024-01316-0

# A Primer on Reinforcement Learning in Medicine for Clinicians

Check for updates

Pushkala Jayaraman[1], Jacob Desman[1], Moein Sabounchi[1], Girish N. Nadkarni[1,2,3,5] & Ankit Sakhuja[1,2,4,5]

Reinforcement Learning (RL) is a machine learning paradigm that enhances clinical decision-making for healthcare professionals by addressing uncertainties and optimizing sequential treatment strategies. RL leverages patient-data to create personalized treatment plans, improving outcomes and resource efficiency. This review introduces RL to a clinical audience, exploring core concepts, potential applications, and challenges in integrating RL into clinical practice, offering insights into efficient, personalized, and effective patient care.

In healthcare, making the right decisions is crucial, as professionals face complex choices daily, from diagnosis to treatment planning and resource allocation. Patient care needs strategies that consider both immediate actions and long-term consequences. Reinforcement Learning (RL), a branch of machine learning, offers a way to enhance decision-making by learning optimal strategies through trial and error and addressing sequential decision-making challenges. RL can model uncertainties, personalize treatments based on patient data, and adapt interventions in real-time, leading to improved patient outcomes, optimized resource utilization, and more efficient healthcare delivery.

RL is particularly well-suited for critical care due to availability of granular data allowing it to accurately model patient conditions, predict outcomes, and optimize treatment pathways using ICU data. However, RL can also be effectively applied in many other healthcare domains. RL's capacity to continuously learn from data can significantly improve clinical trial outcomes and healthcare practices.

This review introduces RL to clinicians and explores its applications for personalized treatment decisions across a range of clinical domains. It also emphasizes the unique challenges associated with integrating RL into medical research and practice. By doing so, it equips clinicians to critically evaluate the clinical relevance of RL research and highlights its transformative potential in shaping the future of healthcare delivery.

## RL – basic concepts

Reinforcement Learning (RL) is a machine learning approach which trains agents to learn decision-making strategies or functions ('policy') through continuous interaction with their environment. This interaction involves a process ('states') of trial ('action') and error[1] inspired by human learning, where the agent receives feedback—either in the form of rewards for successful actions or penalties for unsuccessful ones. Over time, this feedback allows the agent to improve its strategy and maximize expected cumulative rewards.

An intuitive analogy is learning to ride a bicycle. Initially, the learner may face difficulties in balancing, receiving negative reinforcement (e.g., falling) when balance is lost. As balance is achieved, positive reinforcement (e.g., staying upright) encourages repeating those successful actions. Similarly, RL agents refine their actions based on the feedback they receive, gradually developing a policy that enhances their decision-making.

Key components of RL include agent, environment, state, action reward and policy (Table 1).

Mathematically, RL can be described using the framework of Markov Decision Processes[2,3] (MDPs). An MDP is represented as a tuple (S, A, P, R, γ) where S is state, A is action, P is transition probability, R is reward and γ is discount factor (Table 2). MDPs offer a structured approach for modeling decision-making problems in which an agent interacts with an environment across discrete time steps. MDPs provide a flexible and widely applicable framework for modeling various decision-making problems, including robotics, game playing, finance, healthcare, and more. They serve as the foundation for many algorithms and techniques in reinforcement learning, allowing agents to learn effective decision-making strategies in complex and uncertain environments.

## Delving deeper into RL concepts

We have presented a detailed overview of RL concepts designed specifically for clinicians, equipping them with the tools to critically assess literature and understand how it can be integrated into clinical practice. For a deeper dive into the mathematical principles that underpin RL in clinical decision-making, the Supplementary Note 1 provides a more rigorous exploration (Supplementary Information). The main objective of RL is for an agent to learn a policy that maximizes cumulative rewards. This can be approached

[1]The Charles Bronfman Institute for Personalized Medicine (CBIPM), Icahn School of Medicine at Mount Sinai, New York, NY, USA. [2]Samuel Bronfman Department of Medicine Division of Data Driven and Digital Medicine (D3M), Icahn School of Medicine at Mount Sinai, New York, NY, USA. [3]Division of Nephrology, Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [4]Institute for Critical Care Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]These authors jointly supervised this work: Girish N. Nadkarni, Ankit Sakhuja. e-mail: girish.nadkarni@mountsinai.org; ankit.sakhuja@mssm.edu

**Table 1 | Key components of reinforcement learning**

| Terms | Definitions | Examples in the context of "learning to ride a bike" |
|---|---|---|
| **Agent** | The entity learning to make decisions and take actions within an environment. | The person learning to ride is the agent. |
| **Environment** | The external system with which the agent interacts. | The physical world within which the agent rides the bike or a simulated environment in the context of machine learning. |
| **State** | A representation of the current situation or condition of the environment. | The state could include factors such as the rider's speed, posture, and proximity to obstacles. |
| **Action** | The decision or choice made by the agent that affects the state of the environment. | Actions could include pedaling, steering, or braking. |
| **Reward** | Feedback from the environment that evaluates the goodness or badness of an action taken by the agent.<br>Rewards serve as signals to reinforce or discourage certain behaviors. | Falling off the bicycle could result in a negative reward, while successfully riding without falling could yield a positive reward |
| **Policy** | The strategy or rule that the agent follows to select sequence of actions based on its current state. | Learning to ride the bike could be constituted as the "learned policy". |

**Table 2 | Components of Markov Decision Processes**

| Component name | Description |
|---|---|
| **State (S)** | The set of States (s∈S) |
| **Action (A)** | The set of Actions (a∈A) |
| **Probability scores (P)** | Denotes transition probabilities, specifying the likelihood of moving from one state to another when the agent takes a particular action |
| **Rewards (R)** | Denotes the positive or negative rewards received upon transitioning from state s to s' (s, s'∈S) by taking an action a |
| **discount factor (γ)** | Determines the importance of future rewards compared to immediate ones in the calculation of cumulative rewards.<br>A discount factor of 0 would mean that only immediate rewards will be considered, while a discount factor of 1 would mean that future rewards would be valued equally to immediate rewards. |

through either model-based or model-free RL methods, as illustrated in Fig. 1.

## Model-based RL

In model-based learning, the agent learns the model of the environment's dynamics, including transition probabilities and expected rewards[4]. This model allows the agent to simulate possible future states and outcomes, facilitating efficient planning and decision-making. By using the model to simulate trajectories, the agent can anticipate the consequences of its actions without directly interacting with the environment. Let's take an example a robot learning to navigate a maze. In model-based RL, the robot would construct a model of the maze's layout and dynamics. This model might include information about the maze's structure, such as walls and corridors, as well as the outcomes of actions taken by the robot (e.g., moving forward, turning left or right). By simulating possible trajectories using this model, the robot can then anticipate the consequences of its actions and plan its path through the maze accordingly. AlphaZero[5] by DeepMind is an example of model-based RL algorithm implemented using the Monte Carlo Tree Search paradigm.

## Model-free RL

Another way for the robot to navigate the maze would be to learn directly from experience, updating its policy based on observed rewards and transitions without explicitly modeling the environment. This would be an example of model-free[6] RL that focuses solely on learning the value of states or state-action pairs through trial and error. This can be achieved in three ways - value-based, policy-based or hybrid manner.

**Value-based RL**. Value-based RL focuses on estimating the value of being in a particular state or taking a particular action, and then using these value estimates to make decisions that maximize cumulative rewards[7]. At the core of value-based RL is the concept of the value function, which is a measure of how "good" it is to be in a state i.e. it represents the expected cumulative reward achievable from a given state or state-action pair.

The agent can have two different learning strategies[8] in value-based RL: on-policy and off-policy; on-policy methods[9] update the agent's policy while it interacts with the environment. This means the data used for learning comes from the same policy being updated. In other words, the agent learns from its own experiences and updates its policy based on those experiences. For instance, think of a person learning to play a video game. They continuously adjust their strategy based on their current approach, refining their skills as they go.

An example of an on-policy method is SARSA[10] (State-Action-Reward-State-Action). In SARSA, the agent observes the current state, takes an action according to its current policy, receives a reward, observes the next state, updates the policy and then takes another action accordingly. The Q-values are updated based on these state-action-reward-state-action sequences, ensuring that the learning and acting policies are consistent. Off-policy methods[11], in contrast, separate the learning and behavior policies, allowing the agent to learn from experiences collected under a different policy than the one being improved. In other words, the agent learns from data generated by one policy while attempting to optimize a different policy. An analogy might be a chef learning to cook by studying recipes from various sources before developing their own unique style.

A classic example of off-policy RL is Q-learning[12]. In Q-learning, the agent maintains a table of Q-values, where each entry represents the estimated expected cumulative reward for taking a specific action in a specific state. The off-policy nature of Q-learning arises from the fact that the agent learns the value of state-action pairs under one policy (often an exploratory behavior policy) while executing a different policy (the target policy) to gather data. This behavior policy is typically exploratory, employing strategies like ε-greedy or Boltzmann exploration to balance exploration and exploitation. Importantly, the Q-values are updated independently of the behavior policy, enabling the agent to learn from experiences gathered under various policies. This decoupling of learning and behavior policies allows Q-learning to effectively explore the environment while learning optimal action-selection strategies.

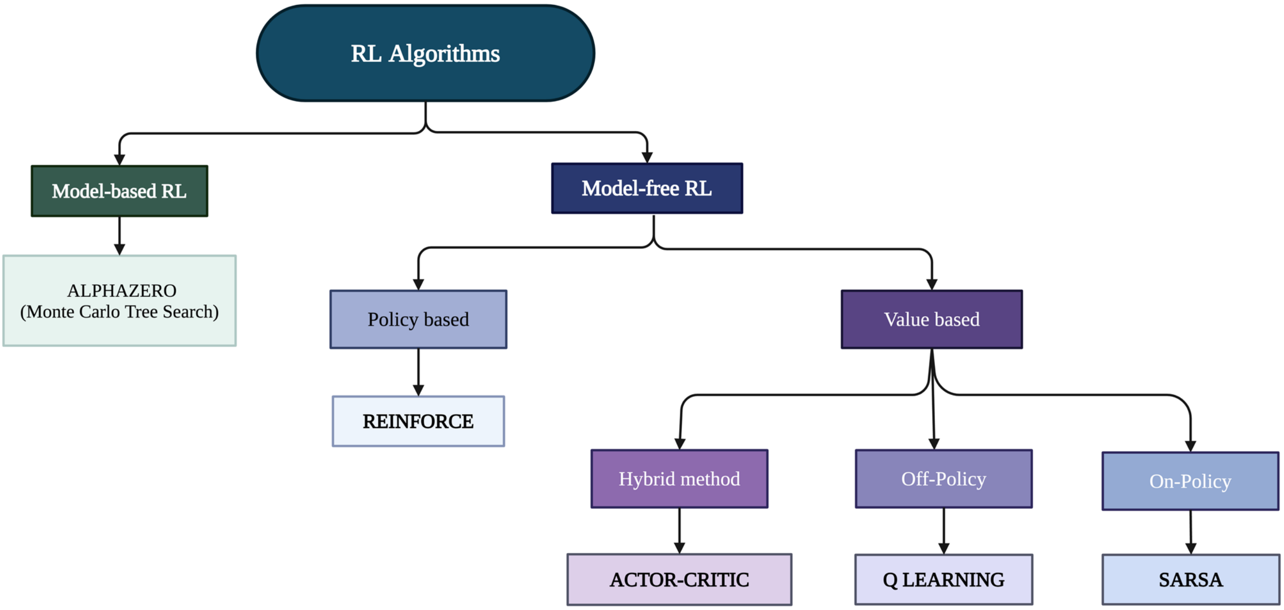

**Fig. 1 | Types of Reinforcement Learning.** An overview of the sub-categories of reinforcement learning is presented in Fig. 1. Each major sub-category has an example of a published RL model. **Source:** Original. **License:** Created with BioRender.com. Agreement number: **SF26VQJRKW** for NPJDM *(Issued on May 30th, 2024*). This document is to confirm that Pushkala Jayaraman has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications. All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com". BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan. For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide or contact BioRender Support at support@biorender.com.

**Policy-based RL**. Another approach is policy-based learning. Policy-based RL directly learns a policy, which is a mapping from states to actions, without explicitly estimating the value of states or state-action pairs[13]. The policy specifies the probability distribution over actions given to states. The objective is to find the policy that maximizes the expected cumulative rewards. For instance, in the REINFORCE algorithm[14], the policy parameters are updated based on the gradient of the expected return with respect to the policy parameters. Policy-based methods excel in handling stochastic policies and exploring the action space effectively. However, they may demand more data and computational resources to converge compared to value-based methods.

**Hybrid RL**. A third class of methods known as hybrid method incorporates both, value-based and policy-based approaches in the same framework. The Actor-Critic[15] class of methods are examples of this approach.

Thus, RL offers a diverse range of strategies for agents to learn optimal policies in complex environments.

## Applications of RL

RL has achieved notable real-world applications. DeepMind's AlphaZero[5] was an early success in mastering Atari games[16] like Breakout and Ms. Pac-Man, and widely-known for surpassing human-level performance using Deep Q-Networks (DQN) algorithms. DeepRL has also revolutionized board games like chess[17] and plays a crucial role in autonomous driving[18], where agents learn to navigate complex environments safely and efficiently. In robotics, RL has enabled robots to learn skills such as object grasping[19], assembly, and navigation in unstructured environments. RL's applications extend to large language models, including ChatGPT[20] where RL helps improve conversational abilities by learning from user feedback to generate more contextually relevant and coherent text over time. Additionally, RL has also been used for generating personalized recommendations[21] for subscription-based viewers, content moderation on social media[22] platforms, and smart grid systems[23]. In each of these examples, the RL agent interacts with an environment, receiving feedback based on its actions and using this feedback to learn and improve its decision-making over time. Such interaction is characteristic of 'online RL', where the agent learns directly from its experiences, making decisions, observing outcomes, and receiving immediate feedback. This feedback loop allows the agent to adjust its behavior in real-time, optimizing its actions to maximize cumulative rewards. Online RL algorithms are transforming digital health clinical trials[24] by personalizing treatment interventions using real-time participant data. These algorithms dynamically adapt to individual characteristics and responses, facilitating real-time decision-making. Their capacity for handling user heterogeneity, distribution shifts, and non-stationarity allows them to modify treatments based on continuous data inputs, facilitating real-time decision-making. However, in cases where real-time interactions are not feasible or ethical, such as in clinical settings, offline RL offers a safer alternative.

## Offline RL

While the previously discussed algorithms illustrate a class of learning based on iteratively collecting data by interacting with an environment and using those interactions to improve decision-making, many scenarios preclude this online interaction. In online learning, any adjustments to the learned policy requires collecting new data following this new policy, which may need to be done millions of times[25] during training. Safety, resource availability, cost, time, and lack of a simulation environment[26] may require data collection only once. Moreover, the increasing quantity of large, retrospective datasets that already exist can be repurposed to optimize agents on vast and diverse behaviors that would otherwise be inaccessible to obtain from live experimentation. These constraints provide significant motivation to effectively convert traditional online RL as an "offline" problem[27].

Offline Reinforcement Learning (RL) is designed for scenarios where direct interaction with an environment is limited or infeasible. This limitation removes the agent's ability to explore new actions, making it crucial for the dataset to contain high-reward behavior for effective learning. Additionally, offline RL algorithms must contend with counterfactual reasoning, where the aim is not to simply imitate past behaviors but to improve upon them. If the policy deviates significantly from the behavior policy learned from the generated dataset, it can cause a *distribution shift*[28], distribution shift, leading to discrepancies between expected and real-world outcomes. These challenges would necessitate careful algorithm design and benchmarking[29,30].

Offline RL has evolved through various algorithmic advancements and is an active field of research. Early approaches, like Behavioral Cloning (BC), focused on mimicking expert demonstrations via supervised[31] learning but struggled in generalizing to new or unseen scenarios. More sophisticated methods, such as Batch Constrained Q-learning (BCQ) and Conservative Q-Learning (CQL)[32], build on the Q-learning paradigm[33] while incorporating constraints to prevent the model from overestimating rewards for actions poorly represented in the dataset thereby allowing agents to cautiously propose new actions. Further innovations include Implicit Q-Learning[34] (IQL) which estimates optimal action-value functions indirectly to avoid challenges associated with explicit Q-value modeling. Additional models like Behavior Regularized Off-Policy Q-Learning[35], Behavior Regularized Off-Policy Q-Learning[36], Conservative Policy Iteration (CPI)[37], Dueling Network Architecture[38], Soft Behavior-regularized Actor Critic(SBAC)[39], and Adversarially Trained Actor-Critic (ATAC)[40] address specific challenges associated with learning from fixed datasets of offline experiences. While they offer promising avenues for applying RL in scenarios with limited or no interaction with the environment, they also come with their own limitations and trade-offs (Table 3).

After training a policy, assessing its performance (evaluation) without direct interaction in live environments or simulations poses challenges. However, Off-Policy Evaluation (OPE) methods such as Importance Sampling (IS) and Fitted-Q Evaluation[41] (FQE) offer a solution by predicting the trained policy's performance using historical data and probabilistically estimates a policy value. However, these methods are prone to bias or high variance, depending on the similarity between the evaluation policy and the behavior policy in the dataset. The Doubly Robust[42] (DR) method mitigates these issues by combining IS and model-based predictions, leveraging bias and variance mitigations of both methods. The newer DIstribution Correction Estimation (DICE)-based methods, such as DualDICE[43] and GenDICE[44], address evaluation by minimizing a divergence measure to target a core issue of distributional shift in OPE and offline RL. As offline RL continues to advance, refining these methods remains a key area of research.

## Applications of RL in medicine

The inherent trial-and-error methodology of training RL agents endows it with considerable efficacy, albeit presenting formidable hurdles for its integration into healthcare settings. As a result, most RL applications in healthcare rely on simulated environments or offline learning approaches[45]. For example, RL has been applied in simulated environments to optimize the personalized dosing of propofol, a sedative commonly used in intensive care units to ensure appropriate sedation[46]. RL has also been applied to identify individualized chemotherapy regimens[47–49] such as dynamic treatment regimens (DTRs)[50] for cancer patients, optimizing chemotherapy dosing strategies. Additionally, RL has been shown to optimize insulin dosing for type 1 diabetics[51] using the FDA approved University of Virginia/Padova type-1 diabetes simulator. RL has also shown potential in mitigating Parkinson's disease symptoms by optimizing combinations of medications[52] and in improving breast cancer screenings using envelope Q-learning[53]. Another prominent example includes applying Q-learning with expert-assisted rewards for diagnosing skin cancer[54]. These examples illustrate the use of RL's sequential decision-making capabilities across diverse medical conditions.

The expansion of RL applications in medicine has predominantly focused on offline RL, especially in the realm of critical care. With access to detailed and granular patient datasets[55–57], datasets, researchers have been able to leverage offline RL to optimize treatment decisions. A pioneering example is the work by Nemati et al., who utilized de-identified patient data to develop a deep RL algorithm for optimizing[58] heparin dosing policies in critical care settings. By using activated partial thromboplastin time as a measure of anticoagulant effect, they were able to dynamically adjust dosing based on patient-specific data, accounting for temporal changes observed in electronic medical records. Further work by Lin et al. expanded on this by employing a deep deterministic policy gradient framework[59] with continuous state-action spaces where they demonstrated that significant deviations between RL-recommended dosing and clinician-prescribed doses correlated with increased complications in patients.

The focus of offline RL in critical care primarily targets two crucial aspects: managing sepsis through fluid and vasopressor optimization, and ventilator management for critically ill patients. Initial studies like those by Komorowski et al. applied the SARSA algorithm to sepsis treatment, setting a foundation that was expanded upon by Raghu et al. using Dueling Double Deep Q-Networks[60] to refine state-action modeling. Subsequently, a sepsis dosing sample-efficient DRL treatment model[61] with episodic control utilized the MIMIC III dataset to increase the longevity of sepsis patients. Komorowski et al. further built on these advancements leading to the development[62] of "AI Clinician" to predict treatment strategies that outperformed real-world clinician decisions. Further innovations include Wu et al.'s introduction of weighted dueling double deep Q-networks with embedded human expertize (WD3QNE)[63] to align closer with clinician strategies (choosing a clinician-based Q function), particularly for patients with specific needs indicated by SOFA scores.

Offline RL has also been pivotal in developing weaning strategies for sedation in mechanically ventilated patients, improving safety and outcome metrics. Furthermore, efforts like Kondrup[64] et al.'s "DeepVent" - a Conservative Q-Learning (CQL) offline RL algorithm and Prasad et al.'s dueling network models integrate deep learning with traditional RL methods to wean sedation[65] for mechanically ventilated patients while ensuring hemodynamic stability and minimizing re-intubations. While Peine et al. focused on using Q learning to optimize mechanical ventilation settings[66] such as tidal volume, positive end expiratory pressure and FiO2 among critically ill patients, Kondrup et. al emphasized more conservative approaches to tackle the overestimation of Q values[67] and improving accuracy of clinical decision-making. Recently, Hengst et al. developed a guideline-informed RL framework for mechanical ventilation[68] that incorporated patient-specific masking actions and violation penalties, improving guideline adherence and outperformance in mortality prediction. However, challenges remain. Saghafian's work[69] highlights ambiguity in offline RL, presenting a model to manage "New Onset Diabetes After Transplantation" (NODAT) through augmented and safe V-learning, illustrating the need for more robust RL approaches when handling medical uncertainty.

Furthermore, RL in DTRs has faced issues with evaluation metrics and standard baselines. Luo et al. addressed these challenges, proposing DTR-Bench[70] a benchmarking framework that standardizes evaluation in areas such as diabetes, cancer chemotherapy, and sepsis treatment[71]. Additionally, RL's role in robotic-assisted surgery (RAS)[72] (RAS) continues to evolve, with RL-based systems enhancing adaptability and efficiency through computer vision and RL algorithms, particularly in automating surgical tasks such as knot-tying to save time and improve precision.

## Challenges of RL in medicine

The versatility of suitable tasks highlights RL's potential and the transformative impact it could have across various facets of healthcare. Through ongoing research and developments, RL will be instrumental in redefining how healthcare professionals' approach daily complex decision-making challenges and assist the discovery of new state-of-the-art care practices. Despite the promise of RL in healthcare, there exist several ongoing challenges and areas of research (Fig. 2) –

### State space formulation challenges

In offline RL, the state space is defined by the available data. Healthcare data, particularly electronic health records, can be high-dimensional presenting

**Table 3 | Examples of Offline Reinforcement Learning Algorithms**

| Offline RL algorithm | Conceptual framework | Key limitations |
|---|---|---|
| Behavioral Cloning (BC)[31] | BC learns a policy by imitating expert behavior from a fixed dataset of expert demonstrations | BC can suffer from compounding errors when the learned policy diverges from the expert behavior, leading to poor generalization and performance in new situations |
| Q-Learning[12] | Value-based off-policy method where the agent's goal is to find an optimal policy by maximizing the expected value of the total rewards through iterative interactions with the environment when the model is not known. | Q-learning selects the action that yields the highest expected value which results in selected actions having consistently overestimated values. |
| Deep Q-Network (DQN)[16] | DQN uses deep neural networks to represent the Q-function rather than a simple table of values. | DQN suffers from overestimation of action values and sensitivity to hyperparameters leading to computationally intensive training processes. |
| Double Deep Q-Network (DDQN)[85] | The DDQN is an improvement over the DQN algorithm as it reduces the overestimation of action values by decoupling the selection and evaluation of actions, using two value function estimates by employing two separate neural networks. | DDQN also suffers from potential overestimation bias due to shared target and online networks and increased computational complexity from maintaining two separate networks. |
| Batch Constrained Q-Learning (BCQ)[33] | It is as an off-policy algorithm that constrains exploration to improve policy learning and address overestimation bias in Q-learning. | BCQ's learned policy is akin to robust imitation learning rather than true reinforcement learning when exploratory data is limited. |
| Conservative Q-Learning[32] | CQL penalizes actions not well-supported by the dataset to mitigate overestimation bias, promoting safer policy learning in reinforcement learning scenarios. | CQL suffers from potential underestimation of action values due to conservative updates, leading to suboptimal policies, and increased computational complexity from the additional penalty term, impacting training efficiency. |
| Implicit Q-Learning (IQL)[34] | IQL addresses the challenges of traditional Q-learning methods by leveraging implicit estimation techniques for improved policy optimization. It estimates the optimal action-value function indirectly, without explicitly modeling Q-values. | IQL learned policy depends on the distributions of actions. The performance regresses when the data distribution is skewed toward sub-optimal actions in specific states. |
| Conservative Policy Iteration (CPI)[37] | CPI balances exploration and exploitation by penalizing deviations from observed behavior, aiming to converge to a risk-averse policy with improved performance in uncertain environments. | CPI suffers from conservative policies that may overly adhere to past behavior, potentially hindering exploration and innovation in dynamic environments. Additionally, CPI's computational complexity can escalate with larger datasets, impacting scalability in real-world applications. |
| Behavior Regularized Off-Policy Q-Learning (BRAC)[35] | BRAC introduces behavior regularization, which encourages the agent to prioritize actions that are consistent with the behavior observed in the dataset, leading to improved learning stability and performance. | The major limitations include difficulty in balancing exploration and exploitation, as well as challenges in effectively tuning the regularization parameter to achieve optimal performance across different environments and datasets. |
| Dueling Network Architecture (DNA)[38] | The DNA architecture consists of two streams of fully connected layers that represent the value and advantage functions separately. It enables more efficient learning by allowing the agent to focus on valuable state information while independently estimating the advantage of different actions. | DNA can suffer from increased complexity and increased computational requirements including a lack of interpretability and potential issues with generalizability. |
| Soft Behavior-regularized Actor-Critic (SBAC)[39] | SBAC incorporates behavior regularization by penalizing the policy for deviating from a behavior policy derived from past experience. This approach balances exploration and exploitation by leveraging previously collected data to improve learning efficiency. | Major limitations of SBAC include potential inefficiency in rapidly changing environments due to reliance on past behavior and the challenge of appropriately tuning the behavior regularization parameter, which can complicate the optimization process. |
| Adversarially Trained Actor-Critic (ATAC)[40] | In ATAC, 2 networks are trained – actor and critic. Actor, which is responsible for selecting actions, is trained against a worst-case behavior policy estimated by the critic network. This adversarial training enhances the actor's ability to perform well even under the most challenging conditions, promoting robustness and stability in learning. | ATAC suffers from need for increased computational complexity due to the adversarial training process, and potential challenges in effectively balancing the adversarial training objectives, which may affect convergence and performance stability. |

challenges in preprocessing and selecting relevant features. Data quality issues such as missing values, noise, and inconsistencies further complicate state space formulation. Healthcare data may also exhibit biases due to factors such as demographic disparities, clinical practice variations, and data collection methods[73]. These biases can lead to *distribution shifts* between the offline data and the target policy, affecting the generalizability and performance of learned policies. Finally, RL problems are formulated using MDPs which assume that the next state is only dependent on the current state and current action, which may not always be true in medicine.

### Reward formulation challenges

Designing reward functions in healthcare RL involves subjective judgments and complex trade-offs. Clinical outcomes, patient well-being, resource utilization, and adherence to medical guidelines are all factors that may need to be considered. Healthcare interventions often have long-term consequences, leading to sparse and delayed feedback on the efficacy of actions. Defining rewards that accurately capture the impact of actions over time while addressing the delay in feedback presents a significant challenge. This is seen in multiple studies that determine the dosing of vasopressors or management of mechanical ventilators based on all-cause in-hospital mortality[74]. Inverse reinforcement learning (IRL) offers a potential solution by deriving reward functions from observed behaviors or data, thus estimating rewards during the learning process[75]. However, IRL also necessitates extensive domain knowledge and often depends on heuristics, given the complexity of clinical data and uncertainties inherent in the decision-making process. Further research is necessary to fully realize the potential of this promising approach.

### Action formulation challenges

Healthcare interventions often span a spectrum from discrete choices (e.g., medication dosage, and treatment options) to continuous adjustments (e.g.,

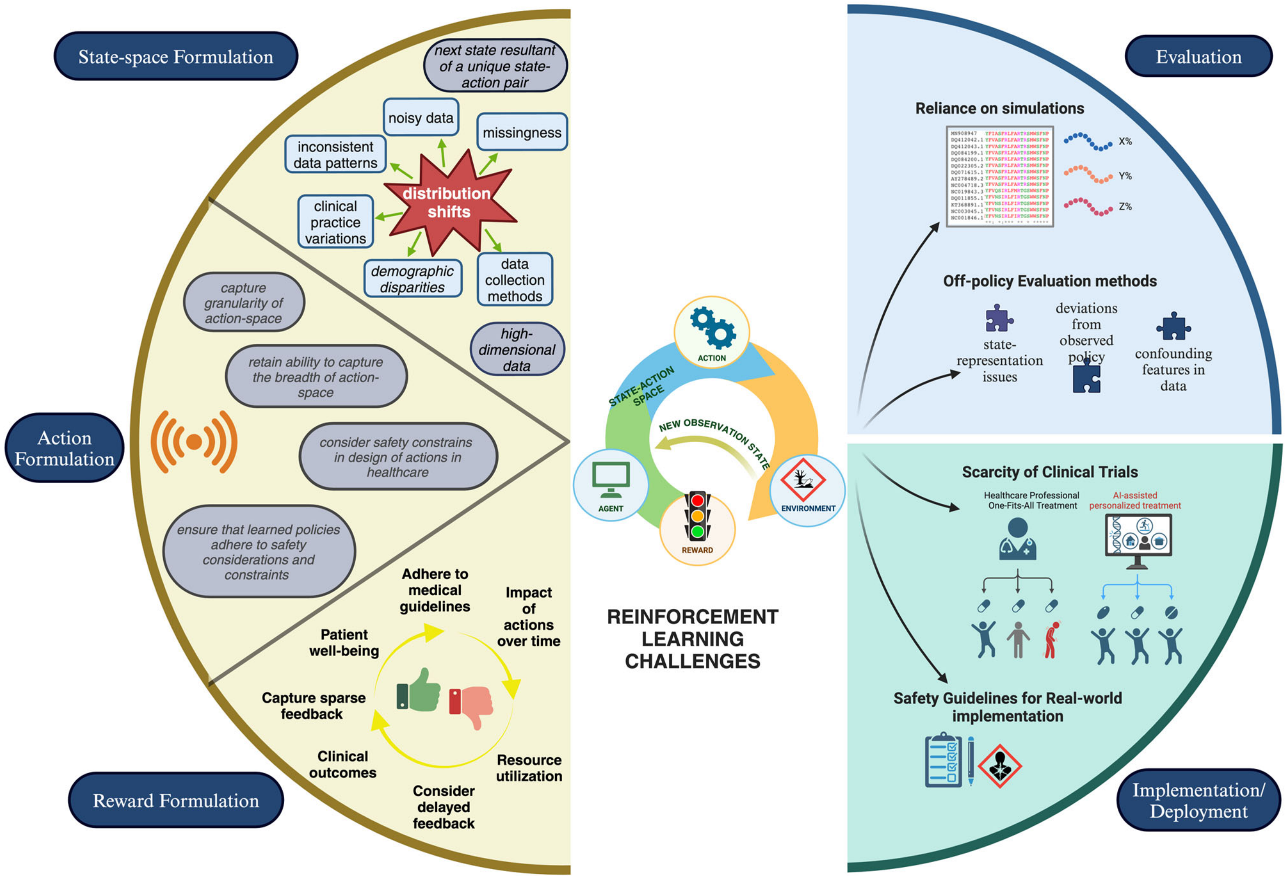
State-space Formulation
next state resultant of a unique state-action pair
noisy data
missingness
inconsistent data patterns
distribution shifts
clinical practice variations
demographic disparities
data collection methods
high-dimensional data
capture granularity of action-space
retain ability to capture the breadth of action-space
Action Formulation
consider safety constrains in design of actions in healthcare
ensure that learned policies adhere to safety considerations and constraints
Adhere to medical guidelines
Impact of actions over time
Patient well-being
Capture sparse feedback
Clinical outcomes
Resource utilization
Consider delayed feedback
Reward Formulation
ACTION
STATE-ACTION SPACE
NEW OBSERVATION STATE
AGENT
ENVIRONMENT
REWARD
REINFORCEMENT LEARNING CHALLENGES
Evaluation
Reliance on simulations
X%
Y%
Z%
Off-policy Evaluation methods
state-representation issues
deviations from observed policy
confounding features in data
Scarcity of Clinical Trials
Healthcare Professional One-Fits-All Treatment
AI-assisted personalized treatment
Safety Guidelines for Real-world implementation
Implementation/ Deployment

**Fig. 2 | Challenges of reinforcement learning.** A visual description of the ongoing challenges in RL is provided in Fig. 2. These could be categorized as challenges in formulation of state-space, action-space and reward, challenges in evaluation and challenges with deployment into a production environment. **Source:** Original. **License:** Created with BioRender.com. Agreement number: **SA26VQJVTH** for NPJDM *(Issued on May 30th, 2024)*. This document is to confirm that Pushkala Jayaraman has been granted a license to use the BioRender content, including icons, templates and other original artwork, appearing in the attached completed graphic pursuant to BioRender's Academic License Terms. This license permits BioRender content to be sublicensed for use in journal publications. All rights and ownership of BioRender content are reserved by BioRender. All completed graphics must be accompanied by the following citation: "Created with BioRender.com". BioRender content included in the completed graphic is not licensed for any commercial uses beyond publication in a journal. For any commercial use of this figure, users may, if allowed, recreate it in BioRender under an Industry BioRender Plan. For any questions regarding this document, or other questions about publishing with BioRender refer to our BioRender Publication Guide or contact BioRender Support at support@biorender.com.

infusion rates). Balancing the granularity of this high-dimensional action representation to capture the complexity of clinical decisions while allowing for optimization of learning algorithms is challenging. Healthcare actions are also subject to various constraints and safety considerations, such as drug interactions, physiological limits, and clinical guidelines. Ensuring that RL policies adhere to these constraints while still allowing for effective learning and adaptation presents another significant challenge.

## Evaluation and implementation challenges

Evaluation of RL algorithms remain a pivotal challenge in healthcare due to safety, ethics, and cost concerns. Applied RL practitioners often rely on simulated environments as testbenches before deployment. When simulators are unavailable, reliable Off-Policy Evaluation (OPE) methods become indispensable for selecting optimal policies across domains. Various innovative approaches, including fitted Q-estimation[41] (FQE) and importance sampling-based[42] methods, and marginalized sampling-based methods[43,76] have emerged. Despite their complexity, studies have demonstrated the suitability of relatively simple methods like FQE for policy[77,78] selection. However, challenges persist, and all methods can be affected by issues with state representations, deviations from observed behavior policies, and confounders. While OPE continues to evolve, real-world deployment is essential for comprehensive evaluation. Despite the abundance of literature applying RL to healthcare, there's a scarcity of clinical trials prospectively analyzing RL models. Notable examples include the REINFORCE trial of a RL-based text messaging program for type 2 diabetes treatment adherence[79] and a proof-of-concept trial of insulin control for type 2 diabetes[80]. However, conducting RL trials in high-risk domains remains challenging, as safety considerations take precedence.

## Future directions

The integration of large language models (LLMs)[81,82] offers significant potential for incorporating reinforcement learning (RL) into automated clinical decision-support systems in healthcare. A review of PubMed publications (Fig. 3) underscores the growing interest in RL within medical research. Although RL-related studies in healthcare remain relatively few, the steady rise in publications indicates a burgeoning recognition of RL's capacity to revolutionize clinical practice. RL's ability to optimize treatment decisions, manage resources, and improve patient outcomes through dynamic learning from clinical data is still underutilized but highly promising. RL also enables continuous adaptation, facilitating progressively sophisticated applications that enhance care quality. Inverse Reinforcement Learning (IRL)[75], when combined with Federated Learning (FL), can offer even greater advantages, particularly in maintaining patient privacy[83] while learning optimal action policies across hospitals. This approach is especially impactful in critical care settings like ICUs, where RL can optimize sedation and ventilation management while protecting data security. The goal of RL in healthcare extends

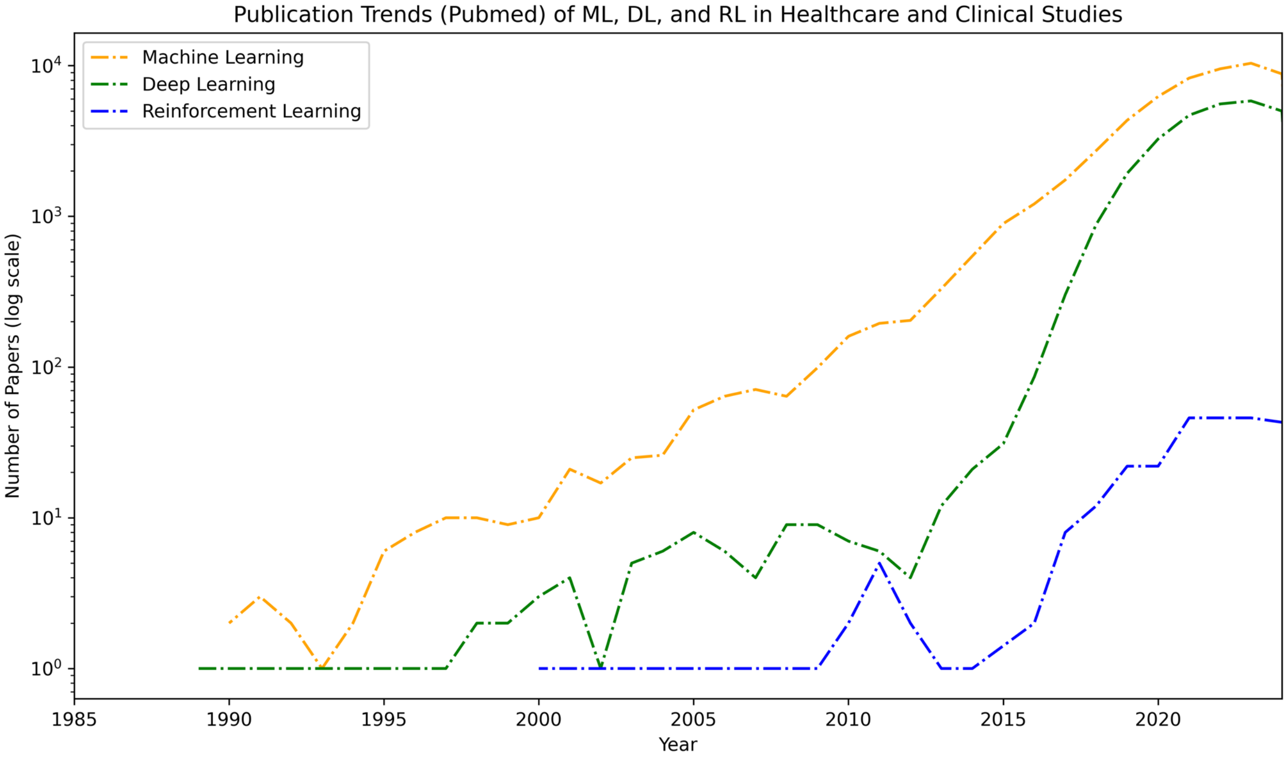

**Fig. 3 | Publication trends (PubMed) of ML, DL and RL in healthcare and clinical studies.** A review of the growth of PubMed-indexed publications underscores the growing interest in RL within medical research. Although RL-related studies remain relatively few, the steady rise in publications indicates a burgeoning recognition of RL's capacity to revolutionize clinical practice. **Source:** Original. Created from PubMed statistics. **License:** Publication Trends (PubMed) of ML, DL and RL in Healthcare and Clinical Studies © 2024 by Pushkala Jayaraman & Jacob Desman is licensed under Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International. To view a copy of this license, visit https://creativecommons.org/licenses/by-nc-nd/4.0/.

to optimizing interventions during clinical trials to enhance treatment efficacy and support comprehensive post-trial analyzes, contributing valuable insights for future interventions and policies. However, applying RL in areas like robotic-assisted surgery presents challenges, including the need for large, diverse datasets for imitation learning and the ability to adapt to unforeseen surgical anomalies. Developing effective model functions that can accurately interpret sensory data[72] and guide surgical actions remains a complex task. Additionally, integrating advanced computer vision models to interpret and learn from the surgical environment, automating repetitive surgical tasks reliably, and achieving precise instrument localization with pixel-wise segmentation are all ongoing challenges that are potential opportunities in the advancement of RL in surgical robotics. Resource variability and the need for domain-specific expertize also impact RL performance in healthcare, requiring models to evolve with varying practice patterns, patient demographics, and clinical standards. While more and more RL models are developed using real-time clinical data[84], iterative refinements are essential to capture additional complexities in RL models to improve treatment efficacy, yet balancing the simplicity of the model with comprehensive patient care remains a formidable challenge. Furthermore, the need for domain expertize to define reward functions and conduct clinical evaluations adds another layer of complexity, requiring specialized knowledge and considerable time investment. Additionally, enabling RL models to provide real-time decision support in critical care settings, while ensuring that decisions align with clinical best practices and meet patient-specific needs, continues to be a significant challenge[84]. Despite these challenges, RL continues to offer substantial opportunities, with future directions likely to include more personalized treatments, real-time adaptive interventions, and enhanced decision-making tools for clinicians, ultimately revolutionizing healthcare delivery.

## Conclusion

In conclusion, the integration of Reinforcement Learning (RL) into healthcare holds immense promise for transforming clinical decision-making and improving patient care through enhanced precision and personalization. Addressing the challenges in the development and deployment of RL algorithms is essential to fully realize its potential for more efficient and effective patient care. RL, with its sequential decision-making capabilities, is uniquely positioned to shift artificial intelligence in healthcare from predictive models to actionable, real-time tools, empowering clinicians to make data-driven, personalized decisions at the bedside.

Received: 8 June 2024; Accepted: 24 October 2024;
Published online: 26 November 2024

## References

1. Sutton, R. S. & Barto, A. G. *Reinforcement learning: An introduction, 2nd ed*, (The MIT Press, Cambridge, MA, US, 2018).
2. Bellman, R. A Markovian decision process. *J. Math. Mech.* **6**, 679–684 (1957).
3. Szepesvári, C. *Algorithms for Reinforcement Learning*, (Springer International Publishing, 2022).
4. Thomas, M. M., Joost, B., Aske, P. & Catholijn, M. J. *Model-based Reinforcement Learning: A Survey*, (now, 2023).
5. Silver, D. et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science* **362**, 1140–1144 (2018).
6. Huang, Q. Model-Based or Model-Free, a Review of Approaches in Reinforcement Learning. In *2020 International Conference on Computing and Data Science (CDS)* 219-221 (2020).
7. McKenzie, M. C. & McDonnell, M. D. Modern value based reinforcement learning: a chronological review. *IEEE Access* **10**, 134704–134725 (2022).
8. Poole, D. L. & Mackworth, A. K. *Artificial Intelligence: Foundations of Computational Agents*, (Cambridge University Press, Cambridge, 2017).

9. Rummery, G. A. & Niranjan, M. On-line Q-learning using connectionist systems. (1994).
10. Singh, S. P. & Sutton, R. S. Reinforcement learning with replacing eligibility traces. *Mach. Learn.* **22**, 123–158 (1996).
11. Prudencio, R. F., Maximo, M. & Colombini, E. L. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE Trans Neural Netw Learn Syst* **PP**(2023).
12. Watkins, C. J. C. H. & Dayan, P. Q-learning. *Mach. Learn.* **8**, 279–292 (1992).
13. Uehara, M., Shi, C. & Kallus, N. A Review of Off-Policy Evaluation in Reinforcement Learning. (2022).
14. Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach. Learn.* **8**, 229–256 (1992).
15. Konda, V. & Tsitsiklis, J. Actor-critic algorithms. *Advances in neural information processing systems* **12** (1999).
16. Mnih, V. et al. Human-level control through deep reinforcement learning. *Nature* **518**, 529–533 (2015).
17. Lai, M. Giraffe: Using Deep Reinforcement Learning to Play Chess. (2015).
18. Kiran, B. R. et al. Deep reinforcement learning for autonomous driving: a survey. *IEEE Trans. Intell. Transportation Syst.* **23**, 4909–4926 (2022).
19. Kober, J., Bagnell, J. A. & Peters, J. Reinforcement learning in robotics: a survey. *Int. J. Robot. Res.* **32**, 1238–1274 (2013).
20. Wu, T. et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* **10**, 1122–1136 (2023).
21. Li, L., Chu, W., Langford, J. & Schapire, R. E. A contextual-bandit approach to personalized news article recommendation. (ACM).
22. Gauci, J. et al. Horizon: Facebook's open source applied reinforcement learning platform. https://openreview.net/forum?id=SylQKinLi4 (2019).
23. Zhang, D., Han, X. & Deng, C. Review on the research and practice of deep learning and reinforcement learning in smart grids. *CSEE J. Power Energy Syst.* **4**, 362–370 (2018).
24. Trella, A. L., et al. Monitoring Fidelity of Online Reinforcement Learning Algorithms in Clinical Trials. *arXiv preprint arXiv:2402.17003* (2024).
25. Prudencio, R. F., Maximo, M. R. O. A. & Colombini, E. L. A Survey on Offline Reinforcement Learning: Taxonomy, Review, and Open Problems. *IEEE Transactions on Neural Networks and Learning Systems*, 1-0 (2024).
26. Jakobi, N, Husbands, P & Harvey, I. Noise and the reality gap: the use of simulation in evolutionary robotics. Springer, Berlin, p 704–720 (1995).
27. Levine, S., Kumar, A., Tucker, G. & Fu, J. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. abs/2005.01643. https://arxiv.org/abs/2005.01643 (2020).
28. Kuhn, D., Esfahani, P. M., Nguyen, V. A. & Shafieezadeh-Abadeh, S. Wasserstein distributionally robust optimization: Theory and applications in machine learning. In *Operations research & management science in the age of analytics* 130-166 (Informs, 2019).
29. Koh, P. W. et al. WILDS: A Benchmark of in-the-Wild Distribution Shifts. In *Proceedings of the 38th International Conference on Machine Learning*, 139 (eds. Marina, M. & Tong, Z.) 5637-5664 (PMLR, Proceedings of Machine Learning Research, 2021).
30. Fu, J., Kumar, A., Nachum, O., Tucker, G. & Levine, S. Datasets for deep data-driven reinforcement learning. https://openreview.net/forum?id=px0-N3_KjA (2021).
31. Bain, M. & Sammut, C. A Framework for Behavioural Cloning. in *Machine Intelligence 15* (1995).
32. Kumar, A., Zhou, A., Tucker, G. & Levine, S. Conservative q-learning for offline reinforcement learning. *Adv. Neural Inf. Process. Syst.* **33**, 1179–1191 (2020).
33. Fujimoto, S., Meger, D. & Precup, D. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning* 2052-2062 (PMLR, 2019).
34. Kostrikov, I., Nair, A. & Levine, S. Offline reinforcement learning with implicit q-learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=68n2s9ZJWF8 (2022).
35. Wu, Y., Tucker, G. & Nachum, O. Behavior regularized offline reinforcement learning. https://openreview.net/forum?id=BJg9hTNKPH, https://openreview.net/forum?id=BJg9hTNKPH (2020).
36. Kumar, A., Fu, J., Soh, M., Tucker, G. & Levine, S. Stabilizing off-policy q-learning via bootstrapping error reduction. *Advances in neural information processing systems* **32** (2019).
37. Vieillard, N., Pietquin, O. & Geist, M. Deep conservative policy iteration. in *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 6070-6077 (2020).
38. Wang, Z., et al. Dueling network architectures for deep reinforcement learning. in *International conference on machine learning* 1995-2003 (PMLR, 2016).
39. Xu, H., Zhan, X., Li, J. & Yin, H. Offline reinforcement learning with soft behavior regularization. abs/2110.07395. https://arxiv.org/abs/2110.07395 (2021).
40. Cheng, C.-A., Xie, T., Jiang, N. & Agarwal, A. Adversarially trained actor critic for offline reinforcement learning. In *International Conference on Machine Learning* 3852-3878 (PMLR, 2022).
41. Le, H., Voloshin, C. & Yue, Y. Batch policy learning under constraints. in *International Conference on Machine Learning* 3703-3712 (PMLR, 2019).
42. Jiang, N. & Li, L. Doubly robust off-policy value evaluation for reinforcement learning. in *International conference on machine learning* 652-661 (PMLR, 2016).
43. Nachum, O., Chow, Y., Dai, B. & Li, L. Dualdice: Behavior-agnostic estimation of discounted stationary distribution corrections. *Advances in neural information processing systems* **32** (2019).
44. Zhang, R., Dai, B., Li, L. & Schuurmans, D. Gendice: Generalized offline estimation of stationary values. In *international Conference on Learning Representations*. https://openreview.net/forum?id=HkxlcnVFwB (2020).
45. Yu, C., Liu, J., Nemati, S. & Yin, G. Reinforcement learning in healthcare: a survey. *ACM Comput. Surv. (CSUR)* **55**, 1–36 (2021).
46. Borera, E. C., Moore, B. L., Doufas, A. G. & Pyeatt, L. D. An Adaptive Neural Network Filter for Improved Patient State Estimation in Closed-Loop Anesthesia Control. in *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence* 41-46 (2011).
47. Zhao, Y., Kosorok, M. R. & Zeng, D. Reinforcement learning design for cancer clinical trials. *Stat. Med* **28**, 3294–3315 (2009).
48. Ahn, I. & Park, J. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Biosystems* **106**, 121–129 (2011).
49. Ebrahimi Zade, A., Shahabi Haghighi, S. & Soltani, M. Reinforcement learning for optimal scheduling of glioblastoma treatment with temozolomide. *Computer Methods Prog. Biomedicine* **193**, 105443 (2020).
50. Yang, C. Y., Shiranthika, C., Wang, C. Y., Chen, K. W. & Sumathipala, S. Reinforcement learning strategies in cancer chemotherapy treatments: A review. *Comput Methods Prog. Biomed.* **229**, 107280 (2023).
51. Visentin, R., Dalla Man, C., Kovatchev, B. & Cobelli, C. The university of Virginia/Padova type 1 diabetes simulator matches the glucose traces of a clinical trial. *Diabetes Technol. Ther.* **16**, 428–434 (2014).
52. Kim, Y., Suescun, J., Schiess, M. C. & Jiang, X. Computational medication regimen for Parkinson's disease using reinforcement learning. *Sci. Rep.* **11**, 9313 (2021).
53. Yala, A. et al. Optimizing risk-based breast cancer screening policies with reinforcement learning. *Nat. Med.* **28**, 136–143 (2022).

54. Barata, C. et al. A reinforcement learning model for AI-based decision support in skin cancer. *Nat. Med*. **29**, 1941–1946 (2023).
55. Johnson, A. E. et al. MIMIC-III, a freely accessible critical care database. *Sci. Data* **3**, 160035 (2016).
56. Johnson, A. E. W. et al. MIMIC-IV, a freely accessible electronic health record dataset. *Sci. Data* **10**, 1 (2023).
57. Pollard, T. J. et al. The eICU Collaborative Research Database, a freely available multi-center database for critical care research. *Sci. Data* **5**, 180178 (2018).
58. Nemati, S., Ghassemi, M. M. & Clifford, G. D. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Annu Int Conf. IEEE Eng. Med Biol. Soc.* **2016**, 2978–2981 (2016).
59. Lin, R., Stanley, M. D., Ghassemi, M. M. & Nemati, S. A deep deterministic policy gradient approach to medication dosing and surveillance in the ICU. *Annu Int Conf. IEEE Eng. Med Biol. Soc.* **2018**, 4927–4931 (2018).
60. Raghu, A. et al. Deep reinforcement learning for sepsis treatment. abs/1711.09602. http://arxiv.org/abs/1711.09602 (2017).
61. Liang, D., Deng, H. & Liu, Y. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. *Appl. Intell.* **53**, 11034–11044 (2023).
62. Komorowski, M., Celi, L. A., Badawi, O., Gordon, A. C. & Faisal, A. A. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat. Med*. **24**, 1716–1720 (2018).
63. Wu, X., Li, R., He, Z., Yu, T. & Cheng, C. A value-based deep reinforcement learning model with human expertise in optimal treatment of sepsis. *NPJ Digit Med*. **6**, 15 (2023).
64. Kondrup, F. et al. Towards safe mechanical ventilation treatment using deep offline reinforcement learning. *Proc. AAAI Conf. Artif. Intell.* **37**, 15696–15702 (2024).
65. Prasad, N., Cheng, L.F., Chivers, C., Draugelis, M. & Engelhardt, B. E. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. abs/1704.06300. http://arxiv.org/abs/1704.06300 (2017).
66. Peine, A. et al. Development and validation of a reinforcement learning algorithm to dynamically optimize mechanical ventilation in critical care. *NPJ Digit Med*. **4**, 32 (2021).
67. Hasselt, H. Double Q-learning. *Advances in neural information processing systems* **23** (2010).
68. den Hengst, F. et al. Guideline-informed reinforcement learning for mechanical ventilation in critical care. *Artif. Intell. Med.* **147**, 102742 (2024).
69. Saghafian, S. Ambiguous Dynamic Treatment Regimes: A Reinforcement Learning Approach. *Management Science* (2023).
70. Luo, Z., Pan, Y., Watkinson, P. & Zhu, T. Position: Reinforcement Learning in Dynamic Treatment Regimes Needs Critical Reexamination. In *Forty-first International Conference on Machine Learning*. https://openreview.net/forum?id=xtKWwB6lzT (2024).
71. Luo, Z. et al. DTR-Bench: An in silico Environment and Benchmark Platform for Reinforcement Learning Based Dynamic Treatment Regime. *arXiv preprint arXiv:2405.18610* (2024).
72. Esteva, A. et al. A guide to deep learning in healthcare. *Nat. Med* **25**, 24–29 (2019).
73. Challen, R. et al. Artificial intelligence, bias and clinical safety. *BMJ Qual. Saf.* **28**, 231–237 (2019).
74. Liu, S. et al. Reinforcement learning for clinical decision support in critical care: comprehensive review. *J. Med Internet Res* **22**, e18477 (2020).
75. Yu, C., Liu, J. & Zhao, H. Inverse reinforcement learning for intelligent mechanical ventilation and sedative dosing in intensive care units. *BMC Med Inf. Decis. Mak.* **19**, 57 (2019).
76. Nachum, O. et al. Algaedice: Policy gradient from arbitrary experience. abs/1912.02074. http://arxiv.org/abs/1912.0207 (2019).
77. Yang, M., Nachum, O., Dai, B., Li, L. & Schuurmans, D. Off-policy evaluation via the regularized lagrangian. *Adv. Neural Inf. Process. Syst.* **33**, 6551–6561 (2020).
78. Voloshin, C., Le, H. M., Jiang, N. & Yue, Y. Empirical study of off-policy policy evaluation for reinforcement learning. abs/1911.06854. http://arxiv.org/abs/1911.06854 (2019).
79. Lauffenburger, J. C. et al. The impact of using reinforcement learning to personalize communication on medication adherence: findings from the REINFORCE trial. *NPJ Digit Med*. **7**, 39 (2024).
80. Wang, G. et al. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nat. Med*. **29**, 2633–2642 (2023).
81. Karabacak, M. & Margetis, K. Embracing large language models for medical applications: opportunities and challenges. *Cureus* **15**, e39305 (2023).
82. Clusmann, J. et al. The future landscape of large language models in medicine. *Commun. Med (Lond.)* **3**, 141 (2023).
83. Gong, W. et al. Federated inverse reinforcement learning for smart icus with differential privacy. *IEEE Internet Things J.* **10**, 19117–19124 (2023).
84. Roggeveen, L. F. et al. Reinforcement learning for intensive care medicine: actionable clinical insights from novel approaches to reward shaping and off-policy model evaluation. *Intensive Care Med Exp.* **12**, 32 (2024).
85. Van Hasselt, H., Guez, A. & Silver, D. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, 30 (2016).

## Acknowledgements

This work was supported by NIH/NIDDK grant K08DK131286 (AS), Clinical and Translational Science Awards (CTSA) grant UL1TR004419 from the National Center for Advancing Translational Sciences (GNN) and the computational resources and staff expertize provided by Scientific Computing at the Icahn School of Medicine at Mount Sinai.

## Author contributions

Conceptualization: A.S, G.N.N, P.J. Methodology: P.J, J.D, M.S, A.S. Investigation: P.J, J.D, M.S, A.S. Visualization: P.J, J.D. Funding acquisition: A.S, G.N.N. Supervision: A.S, G.N.N. Writing: P.J, J.D, M.S, A.S. Revisions: P.J, J.D, M.S, A.S, G.N.N.

## Competing interests

GNN reports grants, personal fees, and non-financial support from Renalytix. GNN reports non-financial support from Pensieve Health, personal fees from AstraZeneca, personal fees from BioVie, personal fees from GLG Consulting, and personal fees from Siemens Healthineers from outside the submitted work. GNN is also serves as the Associate Editor for NPJDM. None of the other authors have any other competing interests to declare.

## Additional information

**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01316-0.

**Correspondence** and requests for materials should be addressed to Girish N. Nadkarni or Ankit Sakhuja.

**Reprints and permissions information** is available at http://www.nature.com/reprints