Reinforcement Learning For Sepsis Treatment: A Continuous Action Space Solution

Yong Huang YONGH7@UCI.EDU
Department of Computer Science
University of California, Irvine
Irvine, CA, USA

Rui Cao CAOR6@UCI.EDU
Department of EECS
University of California, Irvine
Irvine, CA, USA

Amir Rahmani A.RAHMANI@UCI.EDU
Department of Computer Science
University of California, Irvine
Irvine, CA, USA

Abstract

Sepsis is the leading cause of death in intensive care units. It is challenging to treat sepsis because the optimal treatment is still unclear, and individual patients respond differently to treatments. Recent attempts to use reinforcement learning to provide real-time personalized treatment recommendations have shown promising results. However, the discrete action design (i.e., discretizing the continuum of action space into coarse-grained decisions) poses problems in policy learning and evaluation, and limits the effectiveness of the treatment recommendations. In this work, we proposed a continuous state and action space solution inspired by the Deep Deterministic Policy Gradient (DDPG) algorithm. We performed qualitative evaluations and applied the direct method for off-policy evaluations. Our results match clinician performance and are more clinically reasonable and explainable than the state of the art.

1. Introduction

Sepsis is a life-threatening condition caused by the body's response to infection. It can cause a cascade of changes that damage multiple organ systems, leading them to fail, and potentially death (Singer et al., 2016). Sepsis is expensive to treat; In 2011 alone, the US spent 20.3 billion dollars on hospital care for septic patients (Pfuntner et al., 2014), and it remains one of the leading causes of death in intensive care units. Vasopressors and intravenous fluids (IV fluids) are two commonly used treatment strategies for septic patients besides antibiotics (Waechter et al., 2014). Specifically, IV fluids are used to treat hypovolemia and vasopressors are also used to deal with low blood pressure induced by vasodilation. Deciding the optimal dosage of vasopressor and IV fluids is crucially important because they directly impact patient outcomes (Waechter et al., 2014). However, it is

Table 1: Actual dosage range of each action after applying discretization strategy proposed in (Raghu et al., 2017). Different types of vasopressors are converted to noradrenaline-equivalent, and the unit is mcg/kg/min. IV fluids are corrected for tonicity and converted to a standard unit.

Action Number	**0**	**1**	**2**	**3**	**4**
Drug					
Vasopressors	0.00	(0.00, 0.08]	(0.08 , 0.20]	(0.20, 0.45]	(0.45 , 189.08]
IV fluids	0.00	(0.00, 50.00]	(50.00, 152.14]	(152.14,500.00)	(500.00, 10000.00]

also challenging because individual patients may respond differently to the same treatment strategy.

The recent advancements in machine learning and its various successful applications in the healthcare domain have drawn researchers' attention to reinforcement learning (RL), a sub-domain of machine learning specializing in optimizing sequential decision-making. Various efforts have been devoted to searching for personalized and real-time treatment strategies for sepsis to improve patient outcomes. Komorowski et al. (2016) are the first to formulate sepsis treatment optimization as a reinforcement learning problem and propose a discrete action and state solution to address this problem. Various follow-up works extend this idea with improvements in different aspects and demonstrated promising results (Yu et al., 2019; Raghu et al., 2017; Killian et al., 2020; Jia et al., 2020). However, several critical problems remain unaddressed hindering the real-world deployment of RL solutions for sepsis treatment.

Because most existing RL algorithms and quantitative evaluations are designed for discrete actions, previous works opted to recommend a coarse range of dosages instead of a precise recommendation of dosages (actions) (Raghu et al., 2017; Killian et al., 2020) . They discretized the action space into per-drug quartiles resulting in 5 options for each drug, with each option representing dosages in a particular range. However, it is more complicated in practice for clinicians to make decisions on what type of vasopressor to use and what dosage should be given (Rhodes et al., 2017). The most significant caveat of this design is that it may provide minimal clinical usability. For example, as you can see from Table 1, Action 4 for both IV fluids and vasopressors falls into a vast range due to the fact that dosages are exponentially distributed. Therefore, when a RL policy recommends Action 4 for vasopressors, the clinicians still need to decide the value ranging from as little as 0.45 to as large as 189.0. The same applies to the case of IV fluids. Moreover, the discrete action design poses problems in qualitative and quantitative evaluations of RL policy performance as it tends to introduce extra artifacts Gottesman et al. (2018). Another issue with previous work is that a RL policy often suggests minimal dosage for patients with very high acuity (Sequential Organ Failure Assessment score Lambden et al. (2019)). This behavior is reasonable for RL algorithms but is not logical and explainable from a clinical perspective.

In this paper, we propose an RL-based solution in continuous space for both states and actions. Conventional offline RL algorithms such as deep Q-learning do not support continuous actions. More recently, Lillicrap et al. (2015) proposed Deep Deterministic Policy

Gradient (DDPG), which is a model-free off-policy algorithm for learning continuous actions. For this reason, We first utilize a DDPG-based algorithm adapted for environments with continuous action and states to provide high-resolution dosage recommendations. Secondly, to address the issue with minimal dosage at high acuity, we modify the actor-network loss in DDPG by imposing an extra penalty term that forces the policy network to output actions similar to what clinicians would provide. Our qualitative results show that our RL policy overall resembles clinician policy with similar action distributions under different regimes. We perform off-policy evaluations based on a more direct and realistic metric instead of the commonly-used importance sampling-based methods in previous works (Killian et al., 2020; Raghu et al., 2017; Komorowski et al., 2016). This is because the importance ratio in these importance sampling-based methods is infeasible to be computed in a continuous action space. Our quantitative results demonstrate that our proposed solution matches the performance of clinician policy and outperforms random policy baselines.

Generalizable Insights about Machine Learning in the Context of Healthcare

- **Improving the resolution of decision support provides more meaningful results and effective recommendations**. Higher-resolution prediction generally translates into more meaningful clinical decision support and would be considered as a step up toward potential future deployment. It is imperative to consider improving the granularity of prediction since it provides more insights and helps improve the method.

- **Using RL algorithms to solve real-world problem is challenging and often requires ad-hoc adaptation**. It is often not feasible to directly apply an existing algorithm to the healthcare domain problems without considering the context. In many cases, the approach will fail. Carefully designing ad-hoc machine learning approaches by incorporating clinical insights is crucial to successfully applying machine learning in healthcare.

- **Extrapolation errors affects learning and evaluation in Offline RL algorithms**. Existing offline RL algorithms are known to perform poorly at extrapolating, which directly affects the policy learning and evaluation in previous studies as well as in our work. New advancements in machine learning research capable of addressing this issue would benefit efforts in RL in healthcare applications, especially in robust evaluations of the performance of RL policies.

2. Related Work

A spectrum of previous studies has explored applying RL algorithms to optimizing sepsis treatments. Komorowski et al. (2016) is one of the very first to formulate this as a reinforcement learning problem, and proposed a discrete action and state solution based on the value-iteration algorithm. Raghu et al. (2017) further extends this idea to the continuous state space and discrete action space by using Q-learning, besides several novel qualitative evaluation methods introduced in this work, it also formally uses off-policy evaluation to quantify the performance of their proposed RL policy. Following this work, Raghu et al.

(2018) and Peng et al. (2018) further investigated this area using more advanced RL methods to approach this problem ranging from using model-based RL algorithms to combing deep RL with kernel-based RL approaches. Another line of work focuses more on providing more clinically meaningful and interpretable results, such as in (Jia et al., 2020), they proposed a safer RL-based solution by incorporating current clinical knowledge and practice. More recently, Gottesman et al. (2018) surveyed the existing literature using RL for sepsis treatment, and by analyzing some common issues in previous works, they provided guidelines for clinical and computational researchers to help with designing and evaluating algorithms for new ways of treating patients. To our knowledge, we are the first to explore designing RL-based sepsis treatments in continuous action space. Inspired by several previous studies and the takeaway from Gottesman et al. (2018), we incorporate the safety constraints into our algorithm design and provide both qualitative and quantitative evaluations of our proposed solution.

3. Background

Reinforcement learning models a sequence of decisions as agents interact with an environment over time to maximize long-term rewards. The interactions between agents and the environment are commonly modeled as Markov Decision Process (MDP), where S is the state space, A is the action space, and P denotes the state transition dynamic. R is the reward function representing the intermediate reward an agent received from the environment, and is used to measure the goodness of actions, and γ is a discount factor. At each time step t, the agent observes a state s_t and chooses an action a_t according to the policy $\pi(a, s)$. In our case, the policy is deterministic. After the action is taken, the agent receives a reward r_t from the environment, and the state s_t transits to the next s_{t+1}. The objective of RL algorithms is to maximize the long-term accumulated rewards (formally, $E[\sum_t \gamma^t * r_t]$). RL algorithms can be divided into online and offline based on the nature of data sampling process. Unlike online RL, where the next batch of data must come from the newly updated policy whenever we improve the policy, off-policy algorithms evaluate and improve a target policy that is different from the observational policy used to generate the data. In our case the observational policy is the clinician policy. In the context of healthcare applications, it is not feasible to employ online RL algorithms due to safety and ethical issues. The most representative offline RL algorithm is Q-learning, which is built based upon the Bellman Equation $Q^*(s, a) = r(s, a) + \gamma * \max_a E[Q^*(s', a')]$, where $Q^*()$ is the optimal function approximator that estimates the future accumulate reward given the current state and action. Q-learning methods learn an optimal policy by minimizing the temporal difference (TD) error, defined as $r(s, a) + \gamma * Q^*(s', a') - Q^*(s, a)$.

In safety-critical situations, it is costly and risky to try a new policy in the real world without evaluating it. Off-policy evaluation (OPE) is a statistical framework that estimates the value function of a specific target policy, usually using information from another policy used to generate the data. Weighted Importance Sampling (WIS) Estimator and Doubly-Robust (DR) Estimator are the most commonly used OPE methods, and are widely applied in evaluating RL-based policy.

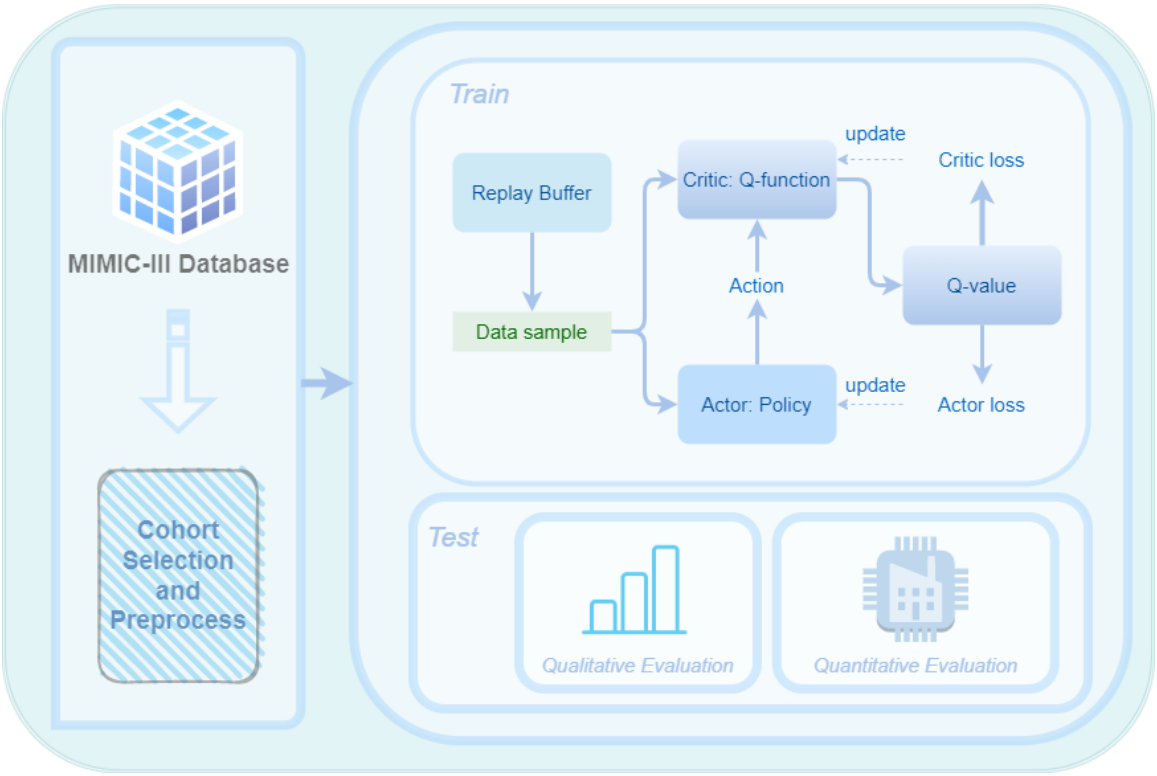

Figure 1: High level diagram of the proposed method

4. Methods

This section covers our proposed reinforcement learning algorithms and the quantitative evaluation methods for the RL policies, Figure1 depicts the high-level diagram of our approach.

4.1. DDPG with Clinician Supervision

Q-learning-based algorithms have been extensively explored to problems in the healthcare domain. However, vanilla Q-learning has its own limitation as it only works for discrete actions (Mnih et al., 2013). To allow policy learning in continuous action space, we applied DDPG algorithm, which extends Q-learning to the continuous action space. DDPG uses Bellman equation to learn the Q-function with bacthed data (Off-policy data), and the policy is then learned using the Q-function. The Q-function learning part is similar to Q-learning, where the goal is to learn an approximator to Q-function $Q^*(s,a)$, and the approximator is typically a neural network parameterized by ϕ. The Q-function is learned by minimizing the mean squared Bellman error function given below, which describes how well Q_ϕ satisfies the Bellman equation:

$$L(\phi, \mathcal{D}) = \underset{(s,a,r,s',d)\sim\mathcal{D}}{\mathrm{E}} \left[\left(Q_\phi(s,a) - \left(r + \gamma(1-d) \max_{a'} Q_\phi\left(s', a'\right) \right) \right)^2 \right] \tag{1}$$

Here d is a variable indicating if s' is a terminal state.

The major distinction between DDPG and Q-learning comes from the policy learning part. In Q-learning, the optimal action $a^*(s)$ can be found by solving

$$a^*(s) = \arg\max_a Q^*(s, a) \tag{2}$$

The argmax operation is convenient in discrete action space with a few action options. However, when the action space is large or continuous, it poses a problem since it would be computationally infeasible to exhaustively evaluate all possible actions. DDPG instead proposes to use a deterministic target policy network μ_θ parameterized by θ to replace the argmax operation, and it directly outputs the action that approximately maximizes the Q-function. We may assume the Q-function is differentiable with regards to parameter Θ, and is learned by maximizing the Q-function.

$$\max_\theta \mathop{\mathrm{E}}_{s\sim\mathcal{D}} \left[Q_\phi\left(s, \mu_\theta(s)\right)\right] \tag{3}$$

DDPG works well in many other applications, but still faces the same problem many other previous works have, that is, recommending small dosages when patients are critically sick. To make recommendations more clinically useful and meaningful, Gottesman et al. (2018) suggest limiting policies to be similar to physicians. This is because we cannot evaluate actions that clinicians never try. It is easy for algorithms to become overly optimistic about actions rarely performed in practice on particular patients (Gottesman et al., 2018). Taking this motivation into consideration, we propose to redesign the policy network training loss (Equation 3) by incorporating the divergence between the clinician action and the RL policy.

$$\max_\theta \mathop{\mathrm{E}}_{s\sim\mathcal{D}} [Q_\phi(s, \mu_\theta(s)) - \lambda * (a, \mu_\theta(s))^2] \tag{4}$$

Note that there is a time delay in this penalty term where we compare RL policy's next time step action $\mu_\theta(s)$ to clinicians' current time step action a, instead of the actual clinicians' following time step action a'. The reasons are two folds. First, we want to avoid degenerating DDPG policy learning into simple behavior cloning, which aims at mimicking clinicians' decision-making but, in theory, will not outperform the demonstrator. Secondly, this time delay could naturally serve as a safety constraint in decision-making. As pointed out by Fadale et al. (2014) and Allen (2014), a drastic change in vasopressor dosage is dangerous to some patients. It could lead to various aftereffects such as acute hypotension (arising from rapidly decreasing doses), hypertension, or cardiac arrhythmias (arising from rapidly increasing doses); therefore, it is essential also to consider the previous action when making action recommendations.

For more implementations details, please refer to Appendix section.

4.2. Off Policy Evaluation with Direct Method

Proper quantitative evaluation of learned policy is crucial before deployment, especially in healthcare. However, Off policy evaluation (OPE) with continuous treatment in the reinforcement learning context is challenging and remains an open research question due to its statistical sophistication.

Table 2: Demographic information of the selected sepsis cohort

Characteristic	Sepsis cohort
Female percentage	55.81%
Mean Age	64
Readmission Rate	7.31%
Mean weight in Kg	88.11
Mechanical Ventilation Rate	36.82%
N	19633

Previous work for discrete treatment/action spaces focuses on WIS and DR methods that use a rejection sampling approach for evaluation. In the continuous setting, this reduction is not applicable as we would almost reject all observations (Kallus and Zhou, 2018).

We apply another simple yet effective approach named the direct method to tackle this challenge. The direct method lends itself to continuous treatment since it completely circumvents the computation of importance ratio. The key idea is to use the Q-function estimation directly as accumulated reward estimation.

$$\hat{\rho}_{\mathrm{DM}} = \mathrm{E}_N \left[\mathrm{E}_{\pi^e} \left[\hat{Q}\left(s_0, a_0\right) \mid s_0 \right] \right] \tag{5}$$

where

$$\mathrm{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0, a_0 \right] = Q\left(s_0, a_0\right)$$

Here π_e is the policy distribution to be evaluated and N is the number of data points. Note that one caveat of applying the direct method is that the estimate could be inaccurate under model misspecification. In other words, when the learned Q-function is not an accurate representation of the accumulated reward, the results could be unreliable.

5. Cohort

We acquired a cohort of sepsis patients and their physiological data in electronic health records (EHR) from the MIMIC-III database (Johnson et al., 2016).

5.1. Cohort Selection

The cohort is selected from MIMIC-3 based on sepsis-3 criteria by employing the preprocess steps from (Killian et al., 2020), and this yields a list of 19633 patients who develop sepsis at some point during their ICU stays. Table 2 details demographical statistics of our selected cohort.

5.2. Data Extraction

Before selecting a sepsis cohort, we first extracted a list of relevant observations useful for cohort selection or state representation according to Killian et al. (2020)'s implementation. In total, we include 38 features which consist of demographics, vital signs, and lab tests.

The complete list of features we selected can be found in Appendix. We then discretize each patient's trajectory into 4-hour windows. The missing value is linearly interpolated or imputed (using K-nearest neighbor), depending on whichever is more appropriate. If there are multiple observations for a variable in that 4-hour time window, the values are averaged. The aggregated data is later used to select a cohort that satisfies sepsis-3 criteria and build their corresponding state representation.

5.3. MDP Formulation

After the sepsis cohort is selected, we formulate the aggregated data into MDP format such that it is learnable by RL algorithms. This process includes state representation, action formulation, and reward design.

5.3.1. State Representation

We first normalize observations and then, following conventions in previous literature, we truncate all long trajectories with only data in the first 21 time-steps retained. Various previous works have demonstrated that compressing the raw observations into latent representation is necessary and desirable for policy learning (Raghu et al., 2017; Killian et al., 2020). In this work, we apply an autoencoder to build the latent representation as described in Killian et al. (2020).

5.3.2. Action

A wide range of treatments could be employed to treat septic patients, including vasopressors, IV fluids, and antibiotics. In this work, we focus on recommending the dosage of vasopressors and IV fluids. Unlike previous works, which binned the continuous value of dosages based on a quantile, our actions are continuous, and any number within the safety range could be recommended.

5.3.3. Reward Formula

Ideally, the reward formulation should be clinically-informed. The reward should be positive when the patient's state improves and negative when the patient's condition deteriorates, and it is supposed to comprise the best indicators of patient health. Thus, we opted for a reward design based on the SOFA score as it is a good indicator of patients' health and is widely used in clinical settings. A practical reward function should penalize high SOFA scores and reward low SOFA scores. Moreover, the reward should decrease when SOFA scores increase between states. Motivated by this, our reward function is as follows:

$$r\left(s_t, s_{t+1}\right) = \lambda_1 tanh(s_t^{\text{SOFA}} - 6) + \lambda_2 \left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right)) \tag{6}$$

where $\lambda_0 = -0.25$ and $\lambda_1 = -0.2$. The first term is a base score determined by the current SOFA score. It is positive when the SOFA score is relatively low, and becomes negative when the SOFA score is high. The cut-off number 6 is determined based on previous literature on the association between SOFA score and mortality (Ferreira et al., 2001). The second term reflects the trend of patients' physiological state change, with

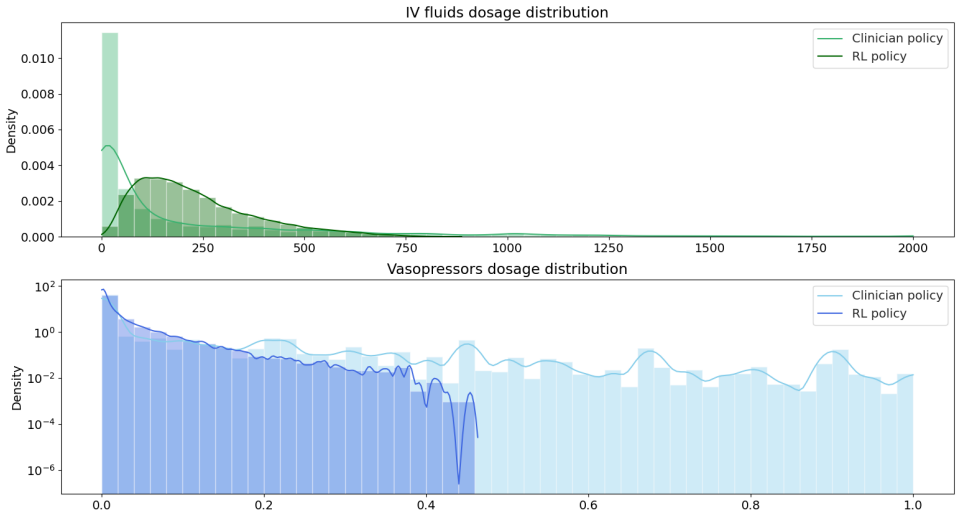

Figure 2: Distribution of IV fluids and vasopressors given by clinicians and RL policy, y axis for vasopressors is converted to log scale for better visualization.

positive values assigned when patients are recovering and negative values when conditions worsen.

6. Results

We conduct qualitative and qualitative evaluations to estimate the performance of the proposed RL policy and analyze the difference between actions recommended by the RL policy and clinicians. All evaluations are performed on a separate test set.

6.1. Qualitative Evaluations

This section provides qualitative results of the evaluation policy and its comparison with the clinician policy. Figure 2 shows the overall dosage distributions of IV fluids and vasopressors. For vasopressors, the distribution of RL policy resembles that of the clinician with minor differences; specifically, RL policy recommends slightly more zero dosage, and overall, large dosage is less often suggested than the clinician dosage. For IV fluids, our RL policy rarely recommends zero dosage and suggests higher dosage than clinicians.

We further investigate how RL policy behaves under different regimes compared to clinician policy. Specifically, we present the average dosage given by RL policy and clinicians as SOFA score changes (Figure 3). The RL and clinician policies follow a similar trend for vasopressors and IV fluids. That is the recommended dosage increases as the SOFA score increases. This is one of the major differences between this work and previous studies, where previous studies tend to give minimal dosage when SOFA scores are high.

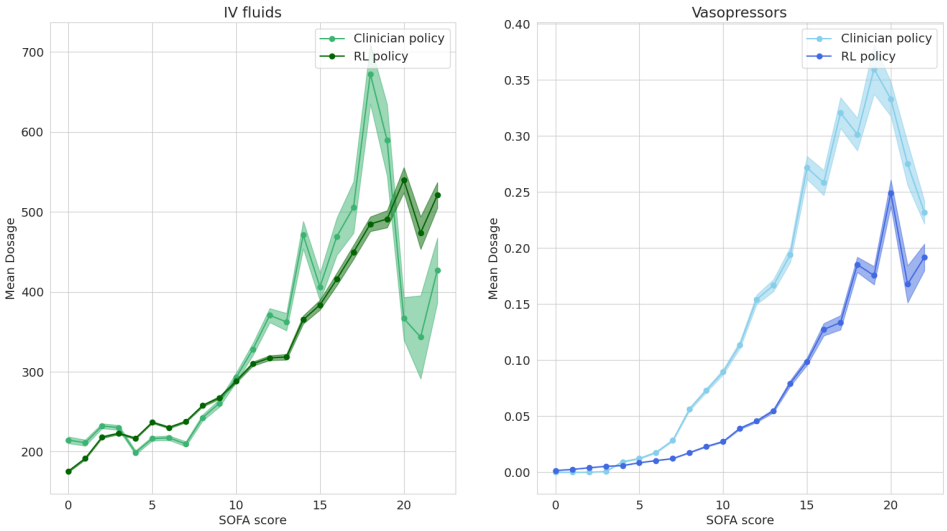

Figure 3: Dosage given by RL policy and clinicians under different sofa score, mean value along with stand errors are presented in this figure

Another commonly used heuristic-based measurement of performance is called the U-curve plot. The idea is to associate the difference between the clinician's policy and the evaluation policy with outcomes such as mortality. In this plot, observed mortality at different dosage deviations is aggregated. Its signature U-shape shows that the observed mortality reaches its lowest when RL policy dosage equals clinical dosage and the mortality increases when the dosage differences elevates. As pointed out by Gottesman et al. (2018), the U-shape could potentially be an artifact of how actions are binned. Note that the RL policies can only recommend actions in a rough range. Therefore, to calculate the exact difference between RL actions and clinicians' actions, an arbitrary number has to be chosen from the suggested range to represent the RL action. Thus, the results could be completely different when choosing a different binning strategy. In this work, we are providing continuous dosages recommendations, and eliminate this artifact. As we can see from Figure 4, the U-shape pattern still exists in continuous action space, with the observed mortality reaching its lowest when the dosage deviations are around zero. Similar to previous work (Gottesman et al., 2018), when we use the no-action policy, which is deemed a sub-optimal policy to compare with the clinician policy, the results are similar. For no-action policy, this trend reveals the association between observed mortality and the dosage given by clinicians. When patients are sicker and more likely to die in hospital, they tend to be treated more aggressively with larger dosages, and the U-shape pattern for the RL policy cases could also be for the same reason. Therefore, the validity of the U-curve plot method is questionable and further discussed in our discussion section.

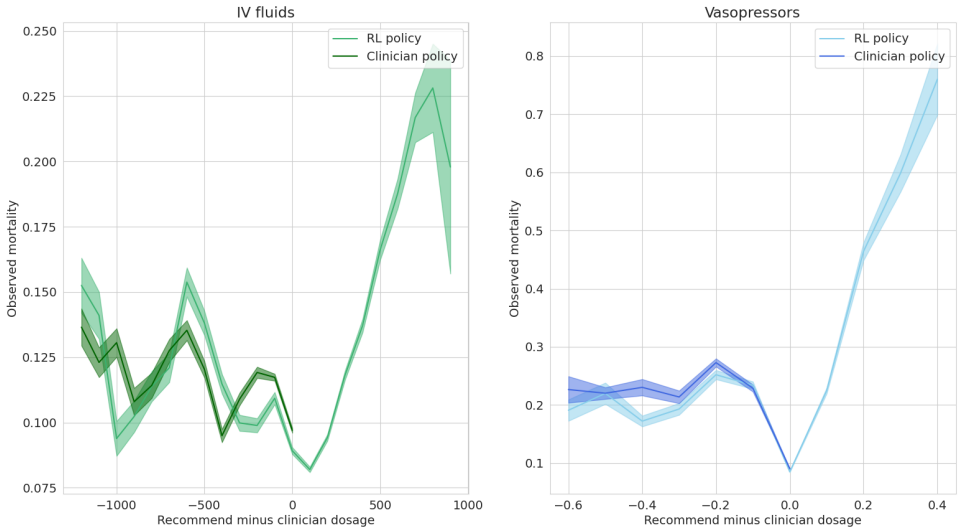

Figure 4: U-curve plot, y axis is the observed mortality under clinician policy and x axis is the dosage difference between clinician policy and evaluated policy, no-action policy refers to an action that only recommends zero dosage of vasopressors and IV fluids

6.2. Quantitative Evaluations

We perform off-policy evaluations with the direct method as described in Section Off Policy Evaluation with Direct Method to quantify the performance of our RL policy. We use clinician policy, random policy, and no-action policy as baselines. The clinician policy comprises actions from historical data which clinicians take. For random policy, actions is uniformly sampled from the 0 to safety upper bound range. Furthermore, for no-action policy, the results are the expected accumulated reward if no actions are taken at any given state. To account for randomness, we performed 30 experiments with a unique train/test set split in each experiment, and the results are presented in a box plot, as shown in Figure 5.

The quantitative results demonstrate that our proposed solution matched the performance of clinician policy while significantly outperforming the random action policy baseline. One unexpected observation in Gottesman et al. (2018) is that the no-action policy has the best quantitative performance. In this work, we also observed that the no-action policy was slightly superior to both the RL and clinician policies. To the best of our knowledge, this issue is likely to be caused by extrapolation errors in offline reinforcement learning, and we further discuss this matter in the following section.

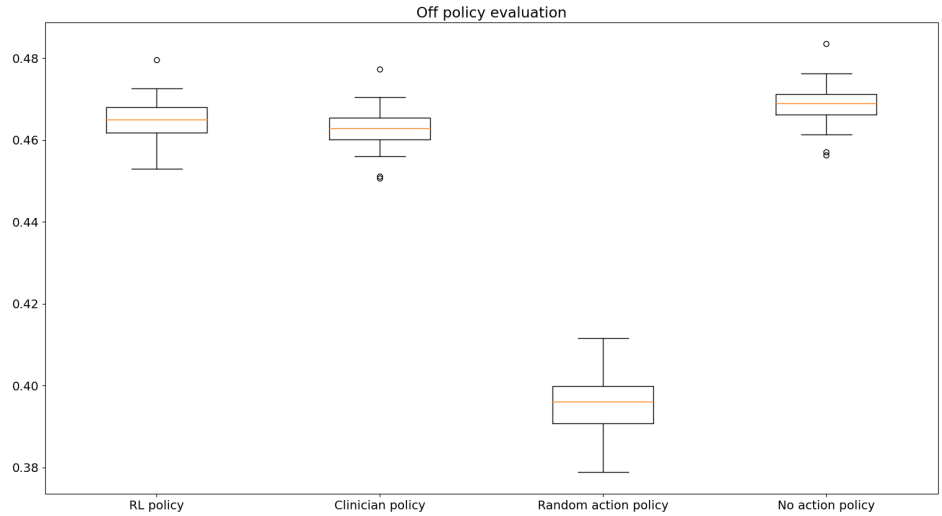

Figure 5: Off policy evaluation results on all evaluated policies

7. Discussion

Various pioneering studies have explored applying reinforcement learning algorithms to the search for optimal sepsis treatment (Raghu et al., 2017; Gottesman et al., 2018; Killian et al., 2020). These studies have demonstrated the potential of using RL to improve ICU patient outcomes. However, these studies are limited to a coarse-grained discrete action space that poses several policy learning and evaluation problems. In order to make personalized treatment design more clinically meaningful, we proposed a continuous action space RL solution that provides much more fine-grained clinical decision support. First of all, by extending actions to continuous space using DDPG family algorithms, we provide more meaningful and higher-resolution decision support to patients. This improvement also makes reinforcement learning-based sepsis treatment search closer to real-world deployment. We first addressed the problem from previous work that the actions recommended by RL algorithms are ambiguous in the large dose case. The vast range of possible dosage options makes it less applicable in the real world for clinician decision support. Besides that, the decision support at time steps when patients are recommended with a higher dosage is more critical compared to low dosage cases since patients who are given larger dosages are associated with worse outcomes. Secondly, the discrete action design oversimplifies the complex research problem and brings multiple challenges to the policy learning and evaluation of reinforcement learning algorithms. With the proposed solution, the actions are now learned in a clinical-guided fashion and under safety constraints, resulting in a dosage distribution close to clinician policy and more reasonable from a clinical perspective. Namely, when patients have higher SOFA scores and worse conditions, our algorithm no longer provides minimal drug dosages, which are common in previous works but do not make clinical sense.

As for qualitative evaluations, we carefully investigate the similarities and differences between the RL policy and the clinician policy. By eliminating the artifacts caused by discretizing actions, we provided a clearer picture of the associations between outcomes and RL-clinician policy agreements, which may help us determine the validity of the U-curve plot. Although the U-shape pattern extends to continuous action space and may indicate that when the clinician's decisions agree with the algorithm, the outcome is the best, we may not be able to make any conclusions on the efficacy of RL policy. First of all, when RL policy agrees with clinician policy, the decision-making is likely to be easier because, in those scenarios, the observed mortality is low, and the patients are usually in better conditions. As a result, the actual dosage given by clinicians is either zero or minimal. Secondly, at its core, the actual mortality can never be assessed without deploying the evaluation policy. When the clinician policy disagrees with the RL policy, one may argue that the outcome might be better if the clinician employs the algorithms' recommended dosage. However, the results could also be the opposite. Therefore, we believe that the U-curve plot provides little clinical significance when evaluating the performance of a RL policy. As for quantitative policy evaluation, we removed the biggest hurdle in off-policy evaluation with the direct method approach, which is essential in quantifying the performance of continuous action solutions. The direct method is intuitive and straightforward as it does not require importance sampling. The results from off-policy evaluations showes that the proposed the RL policy matches clinicians' performance and significantly outperforms the random policy baseline. However, the overestimation of zero policy remains a problem, and as mentioned, this problem resides deeply in the roots of offline reinforcement learning mechanisms, and its solution is out of the scope of this work.

Limitations Sepsis treatment guidelines are evolving rapidly in practice and our data is slightly outdated.

Fairly comparing the proposed method to discrete action methods is difficult because each method has its unique Q-value estimator that suits its own design, and qualitative evaluation based on outcome prediction is not well validated. Therefore, we did not perform a direct comparison between the proposed solution against discrete action methods. However, it is still important, and we will keep exploring alternatives in the future. Overestimation of the no-actions or minimal actions remains a significant problem due to the weakness of extrapolating in offline reinforcement learning algorithms. This is also why we observe that the no-action policy having better performance in OPE. Addressing this issue is outside the scope of this work, but it is important for future exploration from both policy learning and policy evaluation aspects. For example, we may consider calibrating the extrapolating errors in Q-learning-based algorithms using techniques proposed in Kumar et al. (2020).

Moreover, although off-policy evaluations using direct methods are simple and suit well for continuous action space, the soundness of the direct method relies on the accurate estimation of the Q-value. On the other hand, several other OPE methods with continuous treatments have been proposed. For example, Kallus and Zhou (2018) use kernel density estimation to solve the importance sampling problem while Cai et al. (2021) propose a discretization-based approach using deep jump learning to address continuous actions. Most of these efforts are designed for contextual bandit problems, but they could potentially be

extended to reinforcement learning problems, and we expect to explore these approaches for more reliable off-policy evaluations in future works.

References

John M Allen. Understanding vasoactive medications: focus on pharmacology and effective titration. *Journal of Infusion Nursing*, 37(2):82–86, 2014.

Hengrui Cai, Chengchun Shi, Rui Song, and Wenbin Lu. Deep jump learning for off-policy evaluation in continuous treatment settings. *Advances in Neural Information Processing Systems*, 34, 2021.

Kristin Lavigne Fadale, Denise Tucker, Jennifer Dungan, and Valerie Sabol. Improving nurses' vasopressor titration skills and self-efficacy via simulation-based learning. *Clinical Simulation in Nursing*, 10(6):e291–e299, 2014.

Flavio Lopes Ferreira, Daliana Peres Bota, Annette Bross, Christian Mélot, and Jean-Louis Vincent. Serial evaluation of the sofa score to predict outcome in critically ill patients. *Jama*, 286(14):1754–1758, 2001.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing function approximation error in actor-critic methods. In *International conference on machine learning*, pages 1587–1596. PMLR, 2018.

Omer Gottesman, Fredrik Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, et al. Evaluating reinforcement learning algorithms in observational health settings. *arXiv preprint arXiv:1805.12298*, 2018.

Yan Jia, John Burden, Tom Lawton, and Ibrahim Habli. Safe reinforcement learning for sepsis treatment. In *2020 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–7. IEEE, 2020.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

Nathan Kallus and Angela Zhou. Policy evaluation and optimization with continuous treatments. In *International conference on artificial intelligence and statistics*, pages 1243–1251. PMLR, 2018.

Taylor W Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare. In *Machine Learning for Health*, pages 139–160. PMLR, 2020.

Matthieu Komorowski, A Gordon, LA Celi, and A Faisal. A markov decision process to suggest optimal treatment of severe infections in intensive care. In *Neural Information Processing Systems Workshop on Machine Learning for Health*, 2016.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.

Simon Lambden, Pierre Francois Laterre, Mitchell M Levy, and Bruno Francois. The sofa score—development, utility and challenges of accurate assessment in clinical trials. *Critical Care*, 23(1):1–9, 2019.

Timothy P Lillicrap, Jonathan J Hunt, Alexander Pritzel, Nicolas Heess, Tom Erez, Yuval Tassa, David Silver, and Daan Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Xuefeng Peng, Yi Ding, David Wihl, Omer Gottesman, Matthieu Komorowski, Li-wei H Lehman, Andrew Ross, Aldo Faisal, and Finale Doshi-Velez. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In *AMIA Annual Symposium Proceedings*, volume 2018, page 887. American Medical Informatics Association, 2018.

Anne Pfuntner, Lauren M Wier, and Claudia Steiner. Costs for hospital stays in the united states, 2011: statistical brief# 168. 2014.

Aniruddh Raghu, Matthieu Komorowski, Imran Ahmed, Leo Celi, Peter Szolovits, and Marzyeh Ghassemi. Deep reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1711.09602*, 2017.

Aniruddh Raghu, Matthieu Komorowski, and Sumeetpal Singh. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602*, 2018.

Andrew Rhodes, Laura E Evans, Waleed Alhazzani, Mitchell M Levy, Massimo Antonelli, Ricard Ferrer, Anand Kumar, Jonathan E Sevransky, Charles L Sprung, Mark E Nunnally, et al. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. *Intensive care medicine*, 43(3):304–377, 2017.

Mervyn Singer, Clifford S Deutschman, Christopher Warren Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *Jama*, 315(8):801–810, 2016.

Jason Waechter, Anand Kumar, Stephen E Lapinsky, John Marshall, Peter Dodek, Yaseen Arabi, Joseph E Parrillo, R Phillip Dellinger, Allan Garland, Cooperative Antimicrobial Therapy of Septic Shock Database Research Group, et al. Interaction between fluids and vasoactive agents on mortality in septic shock: a multicenter, observational study. *Critical care medicine*, 42(10):2158–2168, 2014.

Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *2019 IEEE international conference on healthcare informatics (ICHI)*, pages 1–3. IEEE, 2019.

Appendix A. Appendix A.

A.1. Code availability

Our implementation can be found here: https://github.com/acneyouth1996/RL-for-sepsis-continuous

A.2. DDPG modifications

In our actual implementation, We utilized several improvements in twin delayed DDPG (TD3) (Fujimoto et al., 2018), which is an enhancement of DDPG to our final model to address some common problems of DDPG, these improvements are listed below.

- **Clipped Double-Q Learning**. TD3 learns two Q-functions instead of one and uses the smaller of the two Q-values as the targets in the Bellman error loss functions. This design addresses the problem of the overestimation of Q-function.
- **"Delayed" Policy Updates**. TD3 updates the policy network less frequently than the Q-function.
- **Target Policy Smoothing**. TD3 adds noise to the target action in training phase. This design allows the policy network to explore a broader range of actions.

A.3. Selected Features

A.3.1. Demographics

Gender, Mechanical ventilation, Readmission, Age, Weight.

A.3.2. Observations

Glasgow Coma Scale/Score; Heart Rate; Systolic Blood Pressure; Mean Blood Pressure; Diastolic Blood Pressure; Respiratory rate; Body Temperature, FiO2; Potassium; Sodium; Chloride; Glucose; Magnesium; Calcium; Hemoglobin; White Blood Cell Count; Platelets Count; PT - Prothrombin Time; PTT - Partial Thromboplastin Time; Arterial pH; PaO2; PaCO2; Arterial Blood Gas; HCO3; Arterial Lactate; PaO2/FiO2 ratio; SpO2; SGOT - Serum Glutamic-Oxaloacetic Transaminase; Creatinine; BUN - Blood Urea Nitrogen; SGPT - Serum Glutamic-Pyruvic Transaminase; INR - International Normalized Ratio; Total bilirubin.

A.4. Experiment configurations

A.4.1. AutoEnoder Hyperparameters

- Autoencoder number of layers: 3

- Autoencoder training epochs: 100
- Hidden size: 16
- Autoencoder learning rate: 0.001

A.4.2. **DDPG Hyperparameters**

- mini-batch size: 32
- Actor network number of layers: 3
- Critic nerwork number of layers: 3
- Actor network hidden size: 32
- Critic network hidden size: 32
- Replay buffer size: 350000
- Reward discount factor: 0.99
- Weight update parameter: 0.01
- Training iterations: 20000
- Evaluation iterations: 5000
- Actor network learning rate: 3e-3
- Critic network learning rate: 3e-5