Biomedical Engineering Letters (2024) 14:279–289
https://doi.org/10.1007/s13534-023-00343-2

ORIGINAL ARTICLE

Optimizing sepsis treatment strategies via a reinforcement learning model

Tianyi Zhang[1,2] · Yimeng Qu[4] · Deyong wang[1,2] · Ming Zhong[3] · Yunzhang Cheng[1,2] · Mingwei Zhang[1,2]

Received: 12 July 2023 / Revised: 28 October 2023 / Accepted: 13 November 2023 / Published online: 4 January 2024

Abstract

Purpose The existing sepsis treatment lacks effective reference and relies too much on the experience of clinicians. Therefore, we used the reinforcement learning model to build an assisted model for the sepsis medication treatment.

Methods Using the latest Sepsis 3.0 diagnostic criteria, 19,582 sepsis patients were screened from the Medical Intensive Care Information III database for treatment strategy research, and forty-six features were used in modeling. The study object of the medication strategy is the dosage of vasopressor drugs and intravenous infusion. Dueling DDQN is proposed to predict the patient's medication strategy (vasopressor and intravenous infusion dosage) through the relationship between the patient's state, reward function, and medication action. We also constructed protection against the possible high-risk behaviors of Dueling DDQN, especially sudden dose changes of vasopressors can lead to harmful clinical effects. In order to improve the guiding effect of clinically effective medication strategies on the model, we proposed a hybrid model (safe-dueling DDQN + expert strategies) to optimize medication strategies.

Results The Dueling DDQN medication model for sepsis patients is superior to clinical strategies and other models in terms of off-policy evaluation values and mortality, and reduced the mortality of clinical strategies from 16.8 to 13.8%. Safe-Dueling DDQN we proposed, compared with Dueling DDQN, has an overall reduction in actions involving vasopressors and reduces large dose fluctuations. The hybrid model we proposed can switch between expert strategies and safe dueling DDQN strategies based on the current state of patients.

Conclusions The reinforcement learning model we proposed for sepsis medication treatment, has practical clinical value and can improve the survival rate of patients to a certain extent while ensuring the balance and safety of medication.

Keywords Sepsis · Medication · Dueling DDQN · Off-policy evaluation · Mortality

1 Introduction

Sepsis, a serious infection with life-threatening acute organ dysfunction, is a leading cause of intensive care mortality [1]. Although international organizations have invested enormous efforts in the past 20 years to provide general guidance for the management of sepsis, clinicians still lack guidance on sepsis treatment strategies [2]. At present, the clinical treatment of sepsis mainly relies on comprehensive treatment methods such as fluid resuscitation and antibiotic application. Fluid resuscitation is one of the core measures of sepsis treatment [3]. In recent years, the update of sepsis fluid resuscitation treatment mainly focuses on the following aspects: the timing of treatment initiation, liquid selection, and the control of the amount of liquid [4]. However, some scholars have shown that this guideline has no obvious effect on actual treatment, and individualized treatment should be carried out under the guidance of monitoring indicators for different patients [5]. At the same time, the guidelines only provide general guidance for the early stage of treatment, do not provide effective and reliable references during other treatment windows, and basically rely on the experience of

✉ Mingwei Zhang
1294851516@qq.com

1 School of Health Sciences and Engineering, University of Shanghai for Science and Technology, Shanghai 200093, China

2 Shanghai Interventional Medical Device Engineering Technology Research Center, Shanghai 200093, China

3 Department of Critical Care Medicine, Zhongshan Hospital Affiliated to Fudan University, Shanghai 200032, China

4 Suzhou Medical College, Suzhou University, Suzhou 215031, China

doctors for medication in clinical practice, so research on personalized treatment strategies for sepsis heterogeneity is imminent. A study to quantify the heterogeneity of sepsis treatment found that the treatment heterogeneity of sepsis is obvious at the individual patient and group level, and the machine learning method can capture the significant heterogeneity of early sepsis hospitalized patients [16]. Vincent et al. [17] proposed a conceptual model for sepsis shock management, aiming at fluid management during the treatment of critically ill patients, divided sepsis shock treatment into 4 phases. On this basis, Malbrain et al. [18] discussed different fluid management strategies, including early adequate goal-directed fluid management, late conservative fluid management, and late goal-directed fluid removal, but there is no clear standard to explain how to precisely administer fluid therapy to the patient.

In recent years, the wide application of artificial intelligence technology has opened up a new way for the optimization of sepsis treatment [6, 7]. Especially with the in-depth study of reinforcement learning algorithms, reinforcement learning has been widely used in intelligent decision-making fields such as unmanned driving and medical decision-making [12–15]. Roggeveen et al. [15] developed a new reinforcement learning model for hemodynamic optimization of sepsis based on the Dueling-DQN network on MIMIC data and then introduced a new in-depth strategy examination method to analyze the interpretability of the strategies obtained by the model to evaluate the safety and reliability of the model. However, the author also mentioned that the model is only a clinical decision support system for hemodynamic optimization of sepsis, and the treatment strategy of the system is obviously different from the clinical, and its reliability needs to be verified. Jia et al. [20] proposed a "safety-driven design" approach, which can be used to guide the design to improve the safety of reinforcement learning models. Compared with the approach of designing first and then evaluating safety, this approach has a much lower failure cost, and it also provides an explanation of the learning model to help clinicians make informed decisions. The results show that this method can effectively identify the unsafe behavior of the machine learning model, especially the drastic changes in the dosage of vasopressors. Liang et al. [21] built a network named D3QN based on the Double DQN network with priority playback, and verified the model with MIMIC-III data set. The results showed that the evaluation value of the weighted double robust off-policy of the model was 26.3% higher than that of the clinician. However, due to the limited data and the model's imitation, the optimal treatment plan could not be obtained. Li et al. [22] optimized the disease treatment decision of reinforcement learning based on EHRs, and took the optimization of blood glucose control in DKA patients as an example to verify the effectiveness of the model. At the same time, the cooperative learning of linear value decomposition is used to simulate the cooperative therapy of multi-agents of different proportions, so as to improve the performance of the benchmark model. Jia et al. [23] used a deep reinforcement learning approach and evaluated whether a sudden major change was included in the recommended vasopressor dose, and then learned a safer strategy by setting the safety valve in combination with current clinical knowledge. Mehdi Fatemi et al. [24] built an ingenious algorithm to identify the "dead end" discrete states in the patient's treatment trajectory and suggested stopping the use of current strategies to avoid "fake" treatments with safety risks. Liu et al. [26] proposed a mixture policy to learn the transition model on key features of patients' physiological. Chan et al. [25] proposed a Bayesian DRL method that can infer the reward transition function. These model-based methods can effectively improve the sample efficiency in continuous state spaces.

We analyzed the current status of sepsis treatment strategies and established a reinforcement learning model for intravenous infusion and vasopressor drugs in sepsis treatment to compensate for the shortcomings of these two treatment strategies. At the same time, patient demographic information is integrated into the model to generate personalized treatment plans to improve patient survival. Among the decision models based on reinforcement learning, the DQN (Deep Q Network) model [9, 10] has a better performance in handling sudden anomalies than the model based on machine learning. However, the DQN model overestimates the target network. The problem makes the model often converge to the local optimal value and get suboptimal results when choosing actions. Therefore, we combined DDQN and Dueling DQN models to build a Dueling DDQN model, which solves the overestimation problem on the basis of DQN, improves the learning ability and speeds up the learning speed on the basis of Dueling DQN, and introduces experience playback in the training process. The learning speed and the convergence ability of the model are further improved.

We built a mortality assessment framework based on the SARSA algorithm to better evaluate the model's effectiveness for medication, and it combined with the dual robust off-strategy evaluation to compare the reinforcement learning model with the clinician's strategy. The results of this evaluation method are more intuitive.

In order to reduce the significant changes in the recommendation of vasopressor dosage in reinforcement learning models, which is clinically considered unsafe behavior, we proposed the safe Dueling DDQN model, which sets a safety mechanism in reinforcement learning to reduce the generation of this strategy.

In order to improve the guidance of clinically effective medication strategies to the model, we proposed a mixed

model of Dueling DDQN + expert strategies to realize the joint guidance of reinforcement learning and expert strategies to medication. The results show that the proposed model can be both effective and safe.

2 Materials and methods

2.1 Data source and processing

In this study, the data of patients with sepsis were screened in the Medical Information Mark for Intensive Care (MIMIC III) database [8]. The diagnostic criteria for sepsis adopted Sepsis 3.0, that is, infection is combined with organ dysfunction, and the sequential organ failure assessment (SOFA) score is ≥ 2 [11]. The MIMIC database has a total of 46,520 patients, of which 19,582 sepsis patients were extracted according to the Sepsis 3.0 standard for the research of medication strategies in this study. The data set was divided into the training set and test set in a ratio of 7:3, the data of 13,707 patients were used for model training, and the data of 5875 patients were used for model testing.

The data includes basic information, vital signs collected by bedside monitors, laboratory test data, microbiological test results, antibiotic usage, etc. In order to construct a complete patient treatment trajectory, we obtained treatment data and prognosis of patients at most 80 h, the patients who dropped out of treatment were excluded. The treatment strategies in this study were vasopressors and intravenous fluids. For the vital signs needed in this study, the K-nearest neighbor (KNN) interpolation method was used to complete missing values. We also calculated some derived features from existing data, such as oxygenation index (P/F), shock index (Shock Index), SOFA, SIRS, etc. The data recording time of patients may be quite different. In order to maintain the uniformity of the patient data sequence, the data needs to be encoded. We used a 2 h time step to encode the data. The flow chart of the whole data processing is shown in Fig. 1a.

The treatment strategies in this study include the dosage of vasopressors and intravenous infusion, which are shown in Fig. 1b. The current international guidelines on sepsis shock recommend norepinephrine as the first-line vasopressor and vasopressin as the second-line vasopressor. In clinical practice, due to drug availability, local practice variations, special settings, and ongoing research, several alternative vasoconstrictors and adjuncts are used in the absence of precise equivalent doses. Norepinephrine equivalence (NEE) is frequently used in clinical trials to overcome this heterogeneity and describe vasopressor support in a standardized manner. Intensive care studies use NEE as an eligibility criterion and also an outcome measure [19]. For vasopressors, we converted them into norepinephrine equivalents in the experiment, and the unit is ug/kg/min. Among them, 1 μg of epinephrine is converted into 1 μg of norepinephrine, 100 μg of dopamine is converted into 1 μg of norepinephrine, 2.2 μg of phenylephrine is converted into 1 μg of norepinephrine, and 1 unit of vasopressin is converted into 5 μg of norepinephrine. In this study, insulin administration, crystalloid infusion, colloid infusion, and blood products were selected for intravenous infusion. These different types of fluids were finally standardized according to the tension of the infusion rate. We used the total dosage of vasopressors and intravenous fluids for patients within a specified period of time to assist with medication.

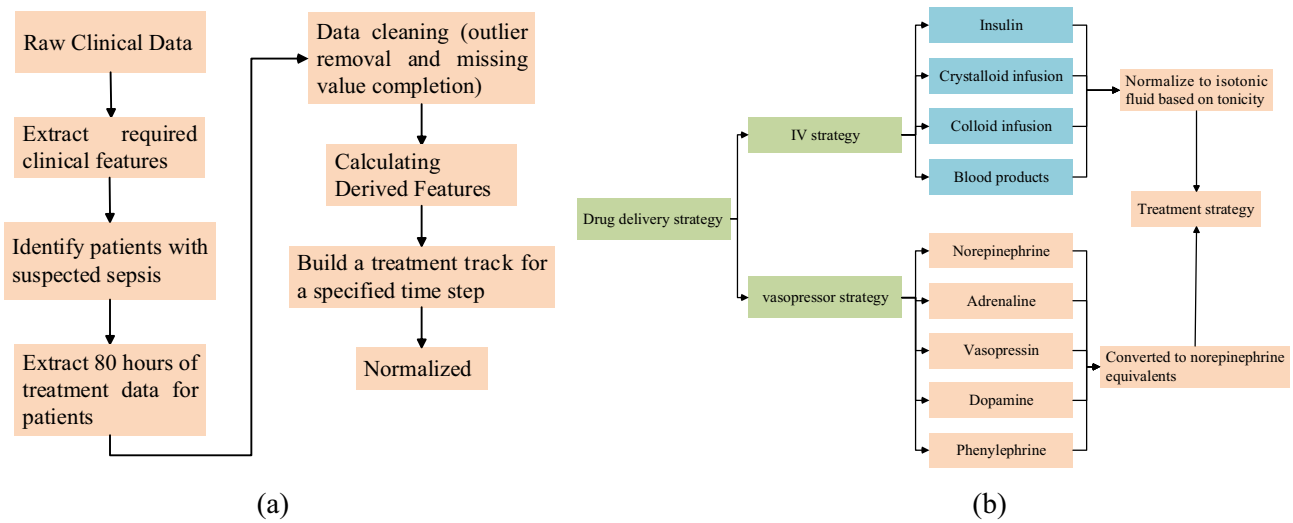

Fig. 1 Data processing flow and drug delivery strategy

2.2 State, action, and reward

The basic components of a reinforcement learning model are state, action, and reward. State: In this study, the 46 features are a set of statuses (Table 1). Includes the patient's basic information, vital signs, laboratory test data, and derived features.

2.2.1 Action

We discretized the medical intervention of the combination of intravenous fluids and vasopressors into 25 action spaces (Fig. 2). The dosage of each drug is represented by quartiles, so that there are five different dosages for each action of a single drug, and the combined action of two drugs constitutes 25 action spaces.

2.2.2 Reward

As an interactive model, the reinforcement learning model will be rewarded for taking corresponding actions in each state, and the reward value will be obtained according to the state of each patient at the last moment. The traditional reinforcement learning model generally adopts the discrete reward function standard, while the deep reinforcement learning assisted decision model established in this study adopts the continuous reward function standard. According to the diagnostic criteria of sepsis 3.0, three consecutive reward functions (Eqs. 1, 2, and 3) were set up in this study (Table 2).

$$r_1(s_t, a_t) = C_0(s_{t+1}^{SOFA} = s_{t+1}^{SOFA} \& s_t^{SOFA} > 0) + C_1(s_{t+1}^{SOFA} - s_t^{SOFA}) \tag{1}$$

$$\begin{aligned} r_2(s_t, a_t) =& C_0(s_{t+1}^{SOFA} = s_{t+1}^{SOFA} \& s_t^{SOFA} > 0) \\ &+ C_1(s_{t+1}^{SOFA} - s_t^{SOFA}) \\ &+ C_2 \tanh(s_{t+1}^{Lactate} - s_t^{Lactate}) \end{aligned} \tag{2}$$

$$\begin{aligned} r_3(s_t, a_t) =& C_0(s_{t+1}^{SOFA} = s_{t+1}^{SOFA} \& s_t^{SOFA} > 0) \\ &+ C_1(s_{t+1}^{SOFA} - s_t^{SOFA}) \\ &+ C_2 \tanh(s_{t+1}^{MAP} - s_t^{MAP}) \end{aligned} \tag{3}$$

$s_{t+1}^{SOFA} - s_t^{SOFA}$ is the change in SOFA, $s_{t+1}^{Lactate} - s_t^{Lactate}$ represents the change in arterial lactate, and $s_{t+1}^{MAP} - s_t^{MAP}$ represents the change in MAP. The change of each feature is updated with a single-step reward value based on the reverse difference. An increase in arterial lactate results in a larger negative reward to punish the treatment step. In contrast, an increase in MAP is generally beneficial to the patient, and a larger positive reward is used in this step. We used the tanh

Table 1 Extracted features

Nos.	Name
1	Gender
2	Mechvent
3	Re_admission
4	Age
5	Weight_kg
6	GCS (Glasgow Coma Scale)
7	HR (Heart Rate)
8	SysBP (Systolic Blood Pressure)
9	MeanBP (Mean Blood Pressure)
10	DiaBP (Diastolic Blood Pressure)
11	RR (Respiratory Rate)
12	Temp_C (Temperature Celsius))
13	FiO2
14	Potassium
15	Sodium
16	Chloride
17	Glucose
18	Magnesium
19	Calcium
20	Hb (Hemoglobin)
21	WBC_Count (White Blood Cell Count)
22	Platelets_count
23	PTT (Partial Thromboplastin Time)
24	PT (Prothrombin Time)
25	Arterial_pH
26	paO2
27	paCO2
28	Arterial_BE
29	HCO3
30	Arterial_lactate
31	SOFA
32	SIRS
33	Shock_Index (=HR/SysBP)
34	PaO2_FiO2 (=PaO2/FiO2)
35	SpO2
36	BUN (Blood Urea Nitrogen)
37	Creatinine
38	SGOT (Serum Glutamic-Oxaloacetic Transaminase)
39	SGPT (Serum Glutamic-Pyruvic Transaminase)
40	Total_bili (Total bilirubin)
41	INR (International Normalized Ratio)
42	Max_dose_vaso
43	Input_total
44	Input_hourly
45	Output_total
46	Output_hourly

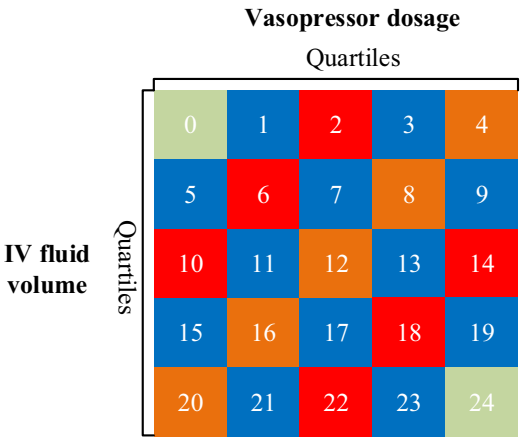

Fig. 2 Twenty-five action spaces of medication

Table 2 Combination of reward functions

Name	Indicator one	Indicator two
Reward 1	SOFA	None
Reward 2	SOFA	Arterial_lactate
Reward 3	SOFA	MeanBP

function to limit the absolute value of changes in arterial lactate to between 0 and 1, preventing differences in changes of different characteristics from adversely affecting the range of the reward function. After a round of medication, if SOFA increases or does not change in the next time step, a negative reward is also set.

2.3 Dueling DDQN model

The Dueling DDQN model based on the priority experience playback mechanism proposed in this study used different networks to implement the selection and evaluation of medication actions and divided the Q network into the value function that is only related to the state S and the advantage function related to state S and action A at the same time (Eq. 4).

$$Q(s,a;\theta,\alpha,\beta) = V(s;\theta,\beta) + \left(A(s,a;\theta,\alpha) - \frac{1}{|A|}\sum_{a\prime} A(s,a\prime;\theta,a)\right) \tag{4}$$

s is the state value, a is the action value, $V(s;\theta,\beta)$ is the value function, $A(s,a;\theta,\alpha)$ is the advantage function, $|A|$ is the number of actions, $a\prime$ is all actions that can be taken, θ is the network parameters of the public part, α and β is the network parameters of the value function and the advantage function respectively. Dueling DDQN consists of two neural networks (evaluation network and target network) with the same structure and different network parameters. The parameters of the evaluation network and the target network are represented by θ and θ^- respectively, and the evaluation network is used to estimate the optimal medication action for sepsis patients (Eq. 5).

$$Q(s,a;\theta) \approx Q^*(s,a) \tag{5}$$

$Q^*(s,a)$ is the value function of optimal medication action. It defines the maximum expected value when the patient is in the state s, takes a certain medication action a and follows the optimal strategy π^*. The patient's state value s_t at time t, the medication action value a_t, the reward value r_t returned from the medication result and the patient's state value s_{t+1} at the next moment $t+1$ are stored in the memory bank D_t as experience values e_t for training the evaluation network (Eqs. 6,7).

$$e_t = (s_t, a_t, r_t, s_{t+1}) \tag{6}$$

$$D_t = \{e_1, e_2, ..., e_t\} \tag{7}$$

In the i-th iteration, the evaluation network first extracted a sequence with a batch size of M from the memory bank, and used the stochastic gradient descent method to minimize the error of the Bellman equation by adjusting the network parameters, which is defined as the loss function $L_i(\theta_i)$ of the i-th iteration (Eq. 8):

$$L_i(\theta_i) = E_{s,a,r,s\prime}[(r + \gamma Q(s\prime, \arg\max Q(s\prime,a;\theta_i);\theta_i^-) - Q(s\prime,a;\theta_i))^2] \tag{8}$$

γ is the discount factor. In order to make the network update process more stable during model training, the parameters θ_i^- of the target network were updated with the Adam method, and then the parameters of the evaluation network were updated by error backpropagation. The overall network architecture is shown in Fig. 3.

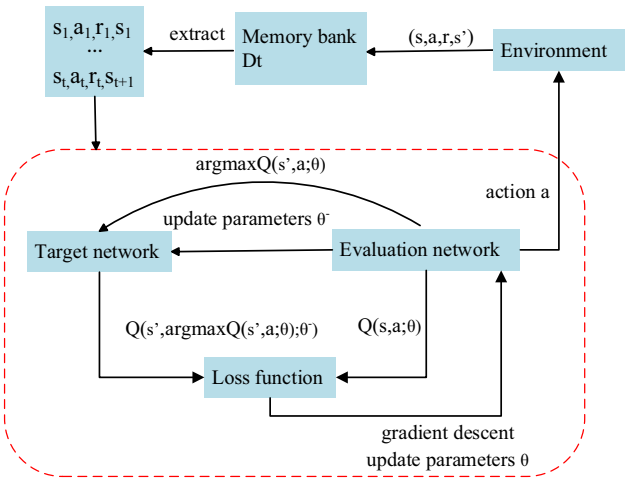

Fig. 3 The overall network architecture of the dueling DDQN

2.4 Security mechanism and hybrid model

Although deep reinforcement learning has made progress in clinically assisted medication, it still needs to be further integrated with clinical judgment, and careful clinical judgment should be exercised to guard against potentially high-risk actions introduced from pathologies in nonlinear function approximation. There is a need to improve the safety of medication strategies output by the model, especially since sudden dose changes in vasopressors can cause harmful clinical effects. Clinicians tend to give fewer vasopressors. To enable the reinforcement learning model to consider the difference in vasopressor dosage between the current step and the previous step while learning the best strategy. We optimized the training cost function on the basis of Eq. 8. Not only the regularization item is added to punish the output Q value if it exceeds the allowed threshold (|Q thresh|=20). A second regularization term is also added to penalize the output Q value if the vasopressor dose is higher or lower than the previous dose of 0.5 ug/kg/min (Eq. 9).

$$\begin{aligned} L(\theta) =& L_i(\theta_i) + \lambda \max(|Q(s,a;\theta) - Q_{thresh}|, 0) \\ &+ \lambda_1 \max(|V_{change}| - 0.5, 0) \end{aligned} \tag{9}$$

At the same time, in order to improve the guidance of clinical effective medication strategies to the model, we proposed a mixed model of safe-Dueling DDQN + expert strategies to optimize medication strategies. The expert strategy is to construct the expert decision set (the set of states and decisions in which the patient has a good prognosis in the training set), use Euclidean distance to calculate the nearest neighbor of the current states, and select the operation corresponding to the medication performed by the nearest neighbor (Fig. 4).

For patients with strong partial state heterogeneity, i.e., a large Euclidean distance from any neighbor, the expert strategy ultimately relies on neighbors that are less similar to the patient. The safe-Dueling DDQN strategy can be used to recommend a medication strategy that more aggressively uses vasopressors and fluids (with a safety mechanism in place to ensure overall medication safety). Our hybrid model switches between expert strategies and reinforcement learning strategies based on the patient's current state.

We examined several medical sources to determine which features might be most useful for medication decisions between experts [27]. Our final set of features were: Age, SOFA, FiO2, BUN, GCS, MeanBP. We set the threshold for searching the nearest neighbor in the expert policy to 1%. First, all features are normalized to between 0 and 1, and the gate is calculated. If the gate is less than 1%, the expert strategy is selected, otherwise the safe-Dueling DDQN strategy is selected (Eq. 10).

$$\text{gate} = \frac{Euclidean\ distance(nearest\ neighbor \text{ and } current\ states)}{len(current\ states)\ in\ Euclidean\ Space} \tag{10}$$

2.5 Model evaluation

The evaluation of the reinforcement learning model is significantly different from the machine learning model. The reinforcement learning models use the patient's treatment trajectory in the hospital, adjust the execution of the action through the interaction with the environment, and finally learn the best behavior strategy in the interaction with the system, in order to obtain the maximum reward value. Therefore, the test and evaluation of the model cannot be simply judged by the accuracy rate, recall etc. We used the off-policy evaluation and the mortality evaluation framework to evaluate the effect of reinforcement learning model. The off-policy evaluation method is when given a set of T-step trajectories ($M = \zeta(i)_{i=1}^{n}$) independently generated by the action policy π_b, the ultimate goal is to make a better estimate by evaluating the policy π_e. In the sepsis treatment, π_b represents the behavioral policy of the reinforcement learning algorithm, and π_e represents the target policy of the clinician. We used the Weighted Doubly Robust (WDR) off-policy evaluation to calculate the average cumulative return of patients on the strategy output by each model and introduced the patient mortality evaluation framework constructed by the SARSA algorithm (Fig. 5).

Figure 5 shows the overall flow of the mortality assessment framework for treatment strategies. We used

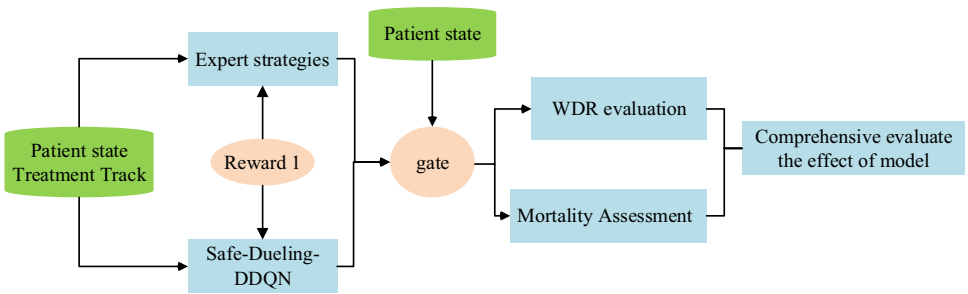

Fig. 4 The general design of the hybrid model

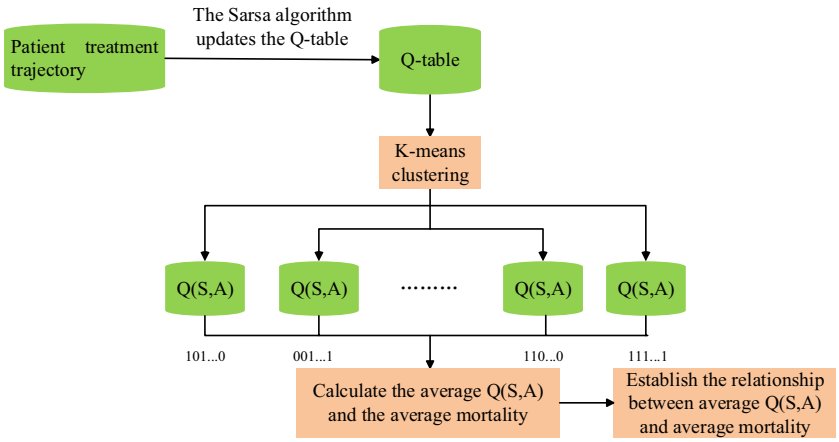

Fig. 5 Mortality assessment framework

Table 3 Off-policy evaluation results for different reward functions

Reward function	WDR value
Reward 1	9.97
Reward 2	10.57
Reward 3	10.18

the K-means algorithm to cluster the status of patients at each time step, used on-policy SARSA (State-Action-Reward-State-Action) algorithm to learn the relationship between $Q(s_t, a_t)$ and mortality. SARSA is an algorithm that learns strategies for Markov decision processes. The initial condition is $Q(s_0, a_0)$, by constantly updating the Q value, and then according to the new Q value to determine what action should be taken in a certain state (Eq. 11).

$$Q(s_t, a_t) = Q(s_t, a_t) + \alpha * [r + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)] \tag{11}$$

We randomly extracted the patient's treatment trajectory from the training set, so as to break the correlation between quintuple $< s_t, a_t, r, s_{t+1}, a_{t+1} >$, which can make the model more robust. After patient status clustering, $Q(s_t, a_t)$ in Table Q are grouped into corresponding buckets, and then the average mortality and average $Q(s_t, a_t)$ in each bucket are calculated. Then fit the linear relationship between average mortality and average $Q(s_t, a_t)$. Mortality assessment framework can be used as a tool in the model evaluation stage to estimate the possible mortality of the medication strategy obtained by the model, and finally evaluate the effectiveness of the model.

3 Results

3.1 Results of different reward function

We added the three reward functions into the Dueling-DDQN network and a total of 5000 epochs were trained.

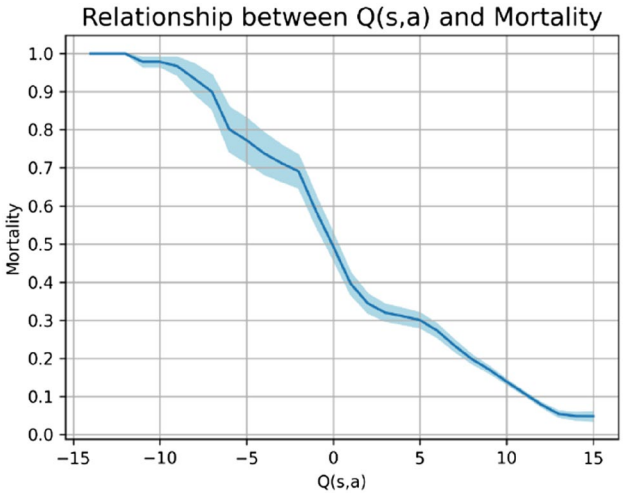

Fig. 6 Linear relationship of mortality assessment framework

We used WDR off-policy evaluation to properly evaluate each reward function (Table 3).

It can be seen that the off-policy evaluation value of the reward function composed of a single SOFA index is lower than that of the combined features. In the experiment, we found the convergence speed of Reward3 function is higher than that of Reward2, but the convergence error of Reward2 is smaller than reward3. In terms of off-policy evaluation values, Reward2 is also slightly higher than reward3. Finally, after comprehensively considering the convergence and off-policy evaluation values, we decided to select Reward2 as the standard of reward function during the training process of the reinforcement learning model.

3.2 Mortality assessment framework

We used SARSA algorithm to learn the relationship between mortality and $Q(s_t, a_t)$ in medication strategy. Figure 6 shows the relationship between $Q(s_t, a_t)$ and mortality, and it is

obvious that $Q(s_t, a_t)$ and mortality are negatively correlated. The higher the $Q(s_t, a_t)$ value, the lower the mortality rate, indicating that the design of the mortality assessment framework in this study is reasonable.

3.3 Model evaluation

We compared the proposed Dueling DDQN model with other reinforcement learning models (DQN, DDQN, Dueling DQN) in the expected return and mortality. All the models have introduced a priority experience playback mechanism, and compared the off-policy evaluation value and mortality in the test set (Fig. 7).

In the figure, the traditional reinforcement learning model and the deep reinforcement learning model are better than the clinical strategy in terms of off- policy evaluation values and estimated mortality. The WDR value obtained by the Dueling DDQN model we proposed is as high as 12.35, which is significantly higher than the 8.78 of the clinical treatment strategy, and the estimated mortality is also reduced from 16.8 to 13.8%. This experiment also proved that the combined feature of SOFA and arterial lactate have certain guiding significance for the medication strategy of sepsis, and can be used as an important research direction for the optimization of sepsis treatment strategies.

Then we quantitatively and visually analyzed the physician policy, Dueling DDQN policy, safe-Dueling DDQN policy and the hybrid model (Expert + Safe-Dueling DDQN) policy for medication decisions on the test set (Fig. 8).

As a complex function approximator, the Dueling DDQN model would recommend a more aggressive treatment strategy with vasopressors and fluids. In contrast to Expert policy, it can place a high emphasis on actions that the clinician rarely or never performs, recommending more moderate to high fluid volumes and vasopressor doses. Figure 8b shows a nearly three-fold increase in the frequency of Dueling DDQN action (action 24) corresponding to the highest levels of fluid and pressors compared to physician policy. These results suggest that although Dueling DDQN achieves a higher return value and a lower mortality rate, further introduction of clinical judgment is needed to prevent potentially high-risk behaviors. Figure 8c shows that safe-Dueling DDQN, compared with Dueling DDQN, has an overall reduction in actions involving vasoppressors. This result is clinically explained, and although vasoppressors are commonly used in ICU to increase mean arterial pressure, many patients with sepsis do not have hypotension. Therefore, vasopressors are not required, and vasopressors need to be increased slowly to reduce large dose fluctuations, so that the vasopressor treatment can be completed before reaching a large dose. Figure 8d shows that the Hybrid model medication strategy is adjusted between safe-Dueling DDQN and Expert policy according to patient status. In the neighbor of Expert policy, survivors are relatively healthier, so treatment is less aggressive. In the treatment of patients with heterogeneous states, Hybrid model will use vasopressors and fluids more actively.

4 Discussion

In recent years, many norms and guidelines have been developed on the use of intravenous fluids and vasopressor strategies in sepsis treatment, such as sepsis guidelines, early goal-directed therapy, etc. However, due to the high clinical heterogeneity of patients with different degrees of sepsis,

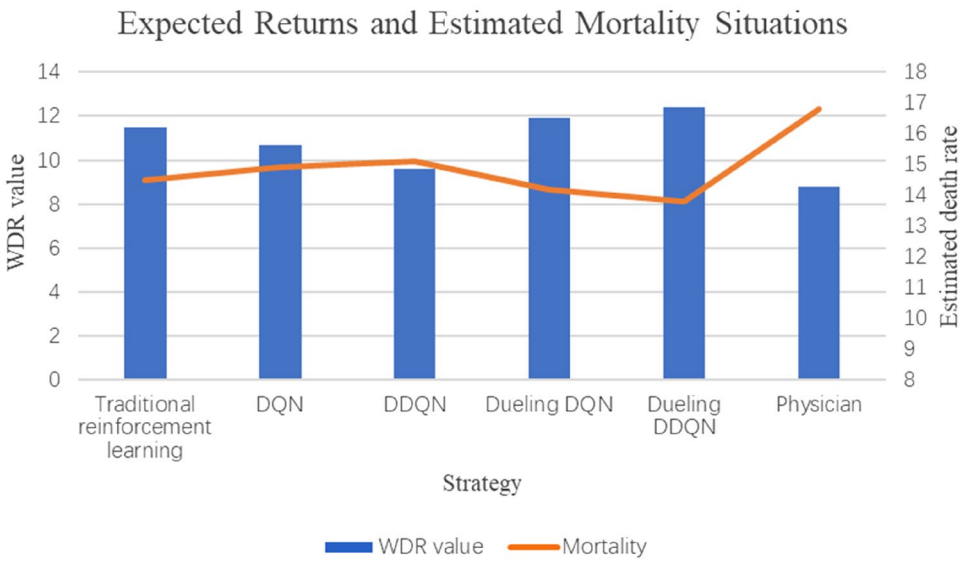

Fig. 7 Results of models on off-policy and mortality evaluation

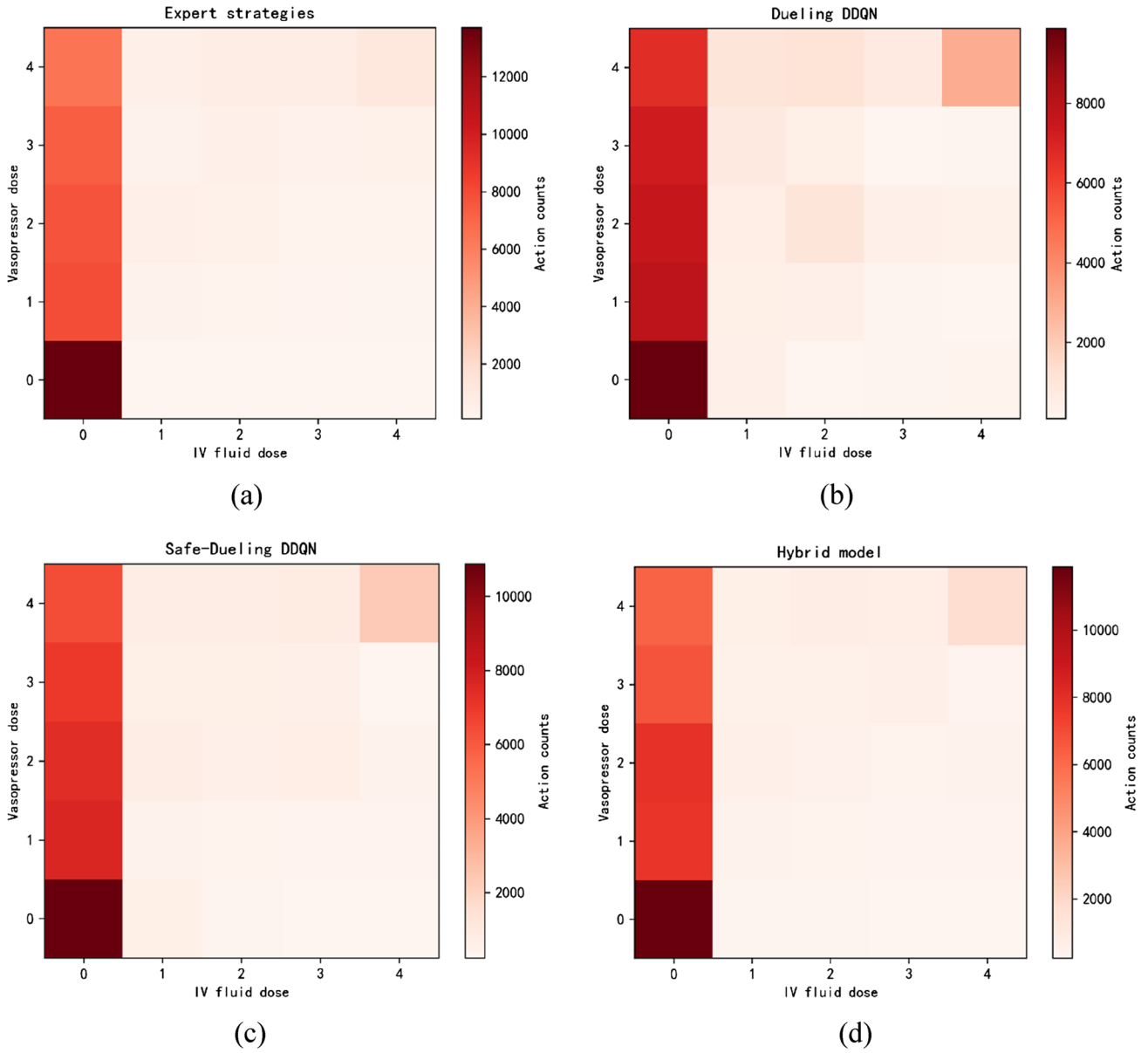

Fig. 8 Policies learned by the different models on different SOFA levels, as a 2D histogram, where we aggregate all actions selected by the physician and model on the test set over. The axes labels index the discretized action space, where 0 represents no drug given, and 4 represents the maximum of drug. **a** Expert policy; **b** Dueling DDQN policy; **c** Safe-Dueling DDQN policy; **d** Hybrid model (Physician + Safe-Dueling DDQN) policy

there has been no unified consensus on how to formulate guidelines for the amount of intravenous infusion and the dosage of vasopressors. At present, clinical treatment strategies mainly rely on the experience of clinicians, so research on personalized treatment strategies for sepsis heterogeneity is imminent.

The reinforcement learning model proposed in this study can well provide the direction of fluid therapy for the sepsis treatment, and guide the clinicians to adjust the fluid therapy strategy in the first place. We used the Dueling DDQN model as the framework, and the combined feature of SOFA and arterial lactate was used as the reward function, and 46 modeling features were included, including the patient's vital signs, laboratory tests, blood gas analysis indicators, demographic information and derivative indicators (such as Oxygenation index (P/F), shock index, SOFA, SIRS, etc.), and finally built a sepsis treatment strategy (vasopressor and intravenous infusion dosage) assisted decision-making model, the model can output relatively reliable and stable treatment strategy, which has a certain significance in

reducing the mortality of sepsis patients and reducing the burden on clinicians. We constructed a Dueling DDQN model with priority experience playback mechanism. Compared with the traditional reinforcement learning method, this model solved the problem that the limited patient state leads to unsatisfactory model results. Compared with the DQN and Dueling DQN network, it solved the problem of model overestimation. Experiments showed that the Dueling DDQN medication-aided model is superior to clinical strategies and other models in terms of off- policy evaluation values and mortality, there was a 3% reduction in mortality compared with the clinical strategy.

At the same time, we provided protection against the possible high-risk behaviors of Dueling DDQN, especially sudden dose changes of vasopressors can lead to harmful clinical effects. In order to improve the guiding effect of clinically effective medication strategies on the model, we proposed a hybrid model (safe dueling DDQN + expert strategies) to optimize medication strategies, switching between expert strategies and reinforcement learning strategies based on the current state of patients.

Therefore, the assisted decision-making model for the medication (vasopressor and intravenous infusion dosage) of sepsis patients solved some shortcomings in this research field, and the model has certain clinical value.

Limitations of this study: (1) The modeling and verification process was completed on the basis of the MIMIC III database, and whether it is applicable to other databases needs to be further confirmed; (2) This algorithm only provides direction for the fluid therapy of sepsis patients, only provides assisted guidance for the total dosage of vasopressors and intravenous infusion within a specified period of time, but cannot accurately predict the infusion rate. Therefore, the model can be used to assist medication of sepsis patients, but whether to follow the strategy of model should ultimately depend on the patient's response to clinical treatment and the actual situation.

5 Conclusions

In this study, we used reinforcement learning to build an assisted model for guiding the medication of sepsis patients (vasopressors and intravenous infusion dosage), theoretically solved the lack of effective reference for existing sepsis medication strategies and the problem of relying too much on clinician experience. There are still many areas worthy of exploration in the field of sepsis treatment. In the future, it is necessary to continue to improve and expand patient medical records to obtain more reliable and complete data, so as to help critical care medicine realize truly intelligent medical care.

Acknowledgements We thank all authors for their contributions to this work.

Author contributions All authors contributed to the study conception and design. Material preparation, data collection and analysis were performed by TZ, DW and MZ. The first draft of the manuscript was written by TZ and all authors commented on previous versions of the manuscript. TZ, YQ and MZ completed the revisions of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by the Academic Leader Program of Shanghai Public Health System Construction 3-Year Action Plan (2020–2022) (Grant Number: GWV-10.2-XD32); Shanghai "Science and Technology Innovation Action Plan" Biomedical Science and Technology Support Special Project (Grant Number: 20S31905100); Shanghai Engineering Technology Research Center Support Project (Grant Number: 18DZ2250900).

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

Ethics approval This article does not contain any studies with human participants or animals performed by any of the authors.

Consent to participate This article does not require the informed consent.

Consent for publications All of the authors confirmed the publication of this paper.

References

1. Seymour CW, Liu VX, Iwashyna TJ, Brunkhorst FM, Rea TD, Scherag A, Rubenfeld G, Kahn JM, Shankar-Hari M, Singer M, Deutschman CS. Assessment of clinical criteria for sepsis: for the third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA. 2016;315(8):762–74.
2. Rhodes A, Evans LE, Alhazzani W, Levy MM, Antonelli M, Ferrer R, Kumar A, Sevransky JE, Sprung CL, Nunnally ME, Rochwerg B. Surviving sepsis campaign: international guidelines for management of sepsis and septic shock: 2016. Intensiv Care Med. 2017;43:304–77. https://doi.org/10.1007/s00134-017-4683-6.
3. Gaieski DF, Edwards JM, Kallan MJ, et al. Benchmarking the incidence and mortality of severe sepsis in the United States. Crit Care Med. 2013. https://doi.org/10.1097/CCM.0b013e31827c09f8.
4. Levy MM, Evans LE, Rhodes A. The surviving sepsis campaign bundle: 2018 update. Intensiv Care Med. 2018. https://doi.org/10.1007/s00134-018-5085-0.
5. Jinxin Z, Kuo S, Dahai H, et al. (2022) Advances in early diagnosis and treatment of sepsis. Chinese journal of injury and repair (Electronic Edition)
6. Littman M. Reinforcement learning improves behaviour from evaluative feedback. Nature. 2015. https://doi.org/10.1038/nature14540.
7. Jeter R, Josef C, Shashikumar S, Nemati S. (2019) Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? arXiv preprint arXiv: 1902.03271. https://arxiv.org/abs/1902.03271

8. Johnson A, Pollard T, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data. 2016. https://doi.org/10.1038/sdata.2016.35.
9. Van Hasselt H, Guez A, Silver D. (2016) Deep reinforcement learning with double Q-Learning. National Conference on Artificial Intelligence, Beijing, China: IEEE. https://doi.org/10.1609/aaai.v30i1.10295.
10. Wang G, Schaul T, Hessel M, et al. (2016) Dueling network architectures for deep reinforcement learning. International Conference on Machine Learning, USA: IEEE. http://proceedings.mlr.press/v48/wangf16.pdf.
11. Singer M, Deutschman CS, Seymour CW, et al. The third international consensus definitions for sepsis and septic shock (Sepsis-3). JAMA. 2016. https://doi.org/10.1001/jama.2016.0287.
12. Raghu A, Komorowski M, Celi L A, Szolovits P, Ghassemi M. (2017) Continuous state-space models for optimal sepsis treatment: a deep reinforcement learning approach. Machine Learning for Healthcare Conference. https://proceedings.mlr.press/v68/raghu17a.html.
13. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Li-wei HL, Ross A, Faisal A, Doshi-Velez F. (2018) Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. American Medical Informatics Association (AMIA) Annual Symposium Proceedings. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6371300/.
14. Futoma, J, Lin, A, Sendak, M, Bedoya, A, Clement, M, O'Brien, C, Heller, K. (2018) Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks. https://openreview.net/forum?id=SyxCqGbRZ.
15. Roggeveen L, El Hassouni A, Ahrendt J, Guo T, Fleuren L, Thoral P, Girbes AR, Hoogendoorn M, Elbers PW. Transatlantic transferability of a new reinforcement learning model for optimizing haemodynamic treatment for critically ill patients with sepsis. Artif Intell Med. 2021. https://doi.org/10.1016/j.artmed.2020.102003.
16. Fohner AE, Greene JD, Lawson BL, Chen JH, Kipnis P, Escobar GJ, Liu VX. Assessing clinical heterogeneity in sepsis through treatment patterns and machine learning. J Am Med Inform Assoc. 2019;26(12):1466–77. https://doi.org/10.1093/jamia/ocz161.
17. Vincent JL, de Backer D. Circulatory shock. N Engl J Med. 2013. https://doi.org/10.1056/NEJMra1208943.
18. Malbrain ML, Van Regenmortel N, Saugel B, De Tavernier B, Van Gaal PJ, Joannes-Boyau O, Teboul JL, Rice TW, Mythen M, Monnet X. Principles of fluid management and stewardship in septic shock: it is time to consider the four D's and the four phases of fluid therapy. Ann Intensiv Care. 2018. https://doi.org/10.1186/s13613-018-0402-x.
19. Kotani Y, Di Gioia A, Landoni G, Belletti A, Khanna AK. An updated "norepinephrine equivalent" score in intensive care as a marker of shock severity. Crit Care. 2023. https://doi.org/10.1186/s13054-023-04322-y.
20. Jia Y, Lawton T, Burden J, Burden J, McDermid J, Habli I. Safety-driven design of machine learning for sepsis treatment. J Biomed Inform. 2021. https://doi.org/10.1016/j.jbi.2021.103762.
21. Liang D, Deng H, Liu Y. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. Appl Intell. 2022. https://doi.org/10.1007/s10489-022-04099-7.
22. Tianhao L, Zhishun W, Wei L, Zhang Q. Electronic health records based reinforcement learning for treatment optimizing. Inf Syst. 2022. https://doi.org/10.1016/j.is.2021.101878.
23. Jia, Yan, et al. (2020) "Safe reinforcement learning for sepsis treatment." 2020 IEEE International conference on healthcare informatics (ICHI). IEEE. https://doi.org/10.1109/ICHI48887.2020.9374403.
24. Fatemi M, Killian TW, Subramanian J, Ghassemi M. (2021) Medical dead-ends and learning to identify high-risk states and treatments. Adv Neural Inf Proces Syst. https://proceedings.neurips.cc/paper_files/paper/2021/hash/26405399c51ad7b13b504e74eb7c696c-Abstract.html.
25. Chan A J, van der Schaar M. (2021) Scalable Bayesian inverse reinforcement learning. International Conference on Learning Representations.https://doi.org/10.48550/arXiv.2102.06483.
26. Liu X, Yu C, Huang Q, Wang L, Wu J, Guan X. (2021) Combining Model-Based and Model-Free Reinforcement Learning Policies for More Efficient Sepsis Treatment. In Bioinformatics Research and Applications: 17th International Symposium, ISBRA. https://doi.org/10.1007/978-3-030-91415-8_10.
27. Beier K, Eppanapally S, Bazick HS, Chang D, Mahadevappa K, Gibbons FK, Christopher KB. Elevation of bun is predictive of long-term mortality in critically ill patients independent of normal creatinine. Crit Care Med. 2011;39(2):305.