Human-Centric Intelligent Systems (2025) 5:63–76
https://doi.org/10.1007/s44230-025-00093-7

RESEARCH ARTICLE

Offline Safe Reinforcement Learning for Sepsis Treatment: Tackling Variable-Length Episodes with Sparse Rewards

Rui Tu[1] · Zhipeng Luo[1] · Chuanliang Pan[2] · Zhong Wang[2] · Jie Su[2] · Yu Zhang[3] · Yifan Wang[1]

Received: 20 December 2024 / Accepted: 16 February 2025 / Published online: 26 February 2025

Abstract

In critical medicine, data-driven methods that assist in physician decisions often require accurate responses and controllable safety risks. Most recent reinforcement learning models developed for clinical research typically use fixed-length and very short time series data. Unfortunately, such methods generalize poorly on variable-length data that can be overlong. In such as case, a single final reward signal appears very sparse. Meanwhile, safety is often overlooked by many models, leading them to make excessively extreme recommendations. In this paper, we study how to recommend effective and safe treatments for critically ill septic patients. We develop an offline reinforcement learning model based on CQL (Conservative Q-Learning), which underestimates the expected rewards of rarely seen treatments in data, thus enjoying a high safety standard. We further enhance the model with intermediate rewards by particularly using the Apache II scoring system. This can effectively deal with variable-length episodes with sparse rewards. By performing extensive experiments on the MIMIC-III database, we demonstrated the enhanced performance and robustness in safety. Our code of data extraction, preprocessing, and modeling can be found at https://github.com/OOPSDINOSAUR/RL_safety_model.

Keywords Offline reinforcement learning · Intermediate rewards · Variable-length time series · Sepsis treatment

1 Introduction

In critical care medicine, sepsis is a severe systemic reaction to infections, commonly observed in patients with acute or critical conditions who have sustained trauma, burns, shock, or infections [1]. The symptoms are characterized by a systemic inflammatory response, a sudden drop in blood pressure, and multi-organ dysfunction. In clinical practice, sepsis frequently results in septic shock, multi-organ failure, and even life-threatening conditions. It is estimated that nearly 50 million individuals worldwide have been diagnosed with sepsis, with an estimated 11 million deaths; the mortality rate of severe sepsis is approximately 30% to 50%, and it dramatically increases to ≥ 50% for septic shocks [2]. In 2017, the UK Sepsis Trust stated that the burden of sepsis on the UK healthcare system resulted in 44,000 deaths and an estimated 250,000 cases of sepsis each year [3]. In the past decades, although mortality rates have declined, the increase in the number of sepsis cases has doubled [4]. The diagnosis and treatment of sepsis have consistently represented a significant challenge within the medical field.

Besides clinical effort, a substantial amount of data-driven studies has been devoted to improving the diagnosis and treatment of sepsis. Electronic health records (EHRs) have laid the data foundation, and many machine learning methods demonstrated superior decisions than physicians [5–8]. Reinforcement learning (RL), due to its decision-orientation nature, has been widely used in the research of data-driven sepsis treatment recommendations. This includes administering various types of antibiotics and vasopressors, whose dosage requires physician expertise. RL regards decision time series (i.e. episodes) as a Markov decision process (MDP) where *actions* (e.g. clinical decisions) can affect subsequent *states* (e.g. patient states) and the associated *rewards* (e.g. treatment effects). Moreover,

✉ Zhipeng Luo
zpluo@swjtu.edu.cn

Rui Tu
turui@my.swjtu.edu.cn

[1] School of Computing and Artificial Intelligence, Southwest Jiaotong University, Chengdu 611756, Sichuan, China

[2] Department of Intensive Care Units, The Third People's Hospital, Chengdu 610031, Sichuan, China

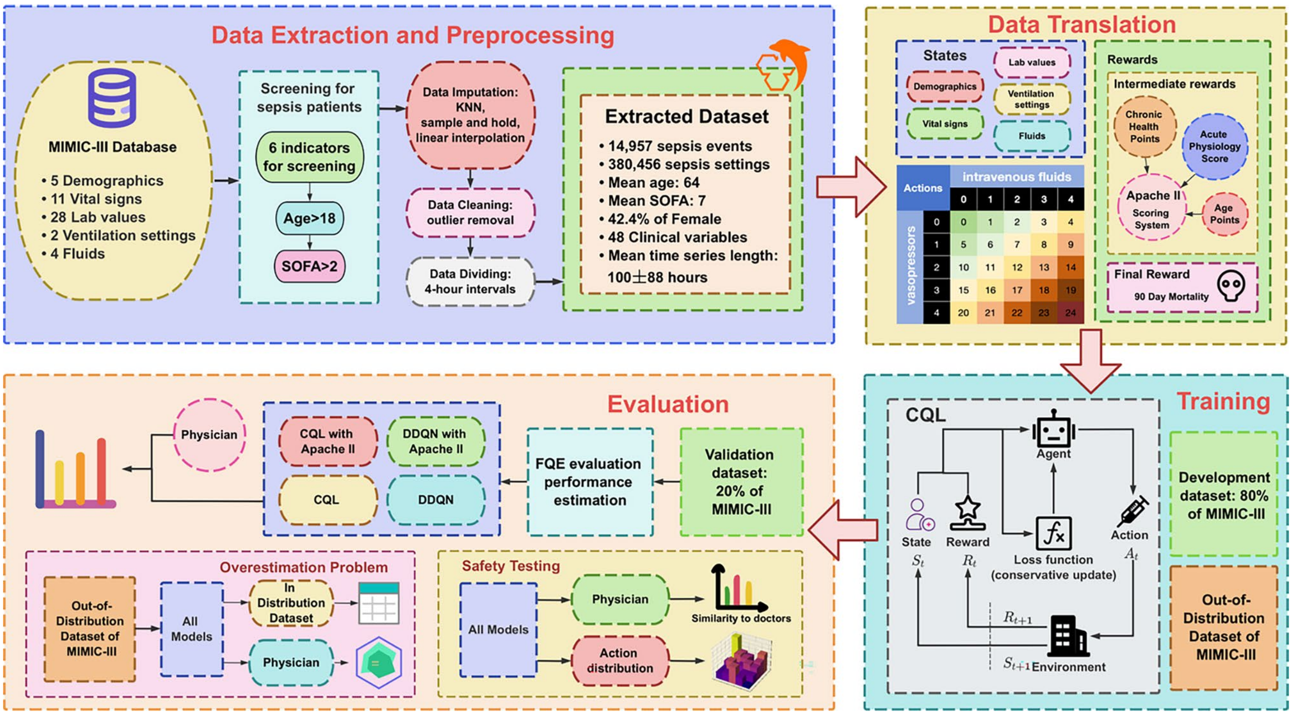
[3] Department of Intensive Care Units, Tangshan People's Hospital, Tangshan 063001, Hebei, China

most clinical RL methods are *offline*, which means learning an optimal policy from *retrospective* datasets that were the recordings of real physicians' behaviors. EHRs can provide rich data on patient care and outcomes, allowing us to learn entirely from static datasets without interaction to the online clinical environment.

However, existing offline RL methods designed for sepsis treatment typically have two limitations. First, their retrospective datasets often use fixed-length episodes among all patient cases (this was initialized by a 2018 Nature Medicine benchmark study [9] and then followed by many others [10, 11]). This stems from their data extraction procedure which we argue is problematic: they first flag out an initial infection point either by an antibiotic or culture use and expand from this point backward and forward a fixed time window to extract data. However, such truncated processing ignores the variations of different patients, resulting in either under-use or over-use of the true inflection data, which can dramatically harm the model training. Second, vanilla offline RL methods, guided solely by rewards (e.g. treatment effects), have been shown to suffer from distributional shifts between the learning policy and the behavior policy, tending to recommend infrequent, unseen actions, or extreme actions (e.g. overdoses) that can deviate from their reasonable ranges [12, 13]. As of 2021, a record 81,000 sepsis patients in the United States have died from drug overdose [13].

To address the problems mentioned above, we propose and develop an offline safe reinforcement learning method that can handle variable-length sepsis data. Figure 1 provides an overview of our study. Our dataset was extracted from MIMIC-III [14], a large-scale open-sourced ICU database. We prepared 14,957 sepsis cases for a total of 380,456 sepsis records. Each case's average length is 100 h with a standard deviation of 88 h. Our methods prioritize both treatment efficacy and safety, aiming to bridge the gap between research findings and practical applications. Our contributions can be summarized as follows:

- *Data extraction*. We argue that the current data extraction process benchmarked by the 2018 Nature Medicine study [9] is obsolete—it produced fixed-length episodes for all patient cases, either introducing data noises or losing precisions. We, under the advice of critical medicine physicians, systematically extracted and preprocessed raw data from MIMIC-III, which produced variable-length time series data closely aligned with patients' actual infection time. This produces a larger quantity of high-quality data, thereby improving the model training performance and enhancing its reliability. Our code of data extraction and preprocessing can be found at Github.
- *Safe reinforcement learning*. We propose and develop an offline safe reinforcement learning model based on con-



Actions		intravenous fluids				
		0	1	2	3	4
vasopressors	0	0	1	2	3	4
	1	5	6	7	8	9
	2	10	11	12	13	14
	3	15	16	17	18	19
	4	20	21	22	23	24

Fig. 1 The general workflow of our proposed reinforcement learning framework for sepsis treatment recommendation. First, a comprehensive sepsis time series dataset was extracted from MIMIC-III and preprocessed. Then the data are translated into the reinforcement learning syntax. Finally, the dataset was partitioned into 80% for training and 20% for model evaluation

servative Q-learning (CQL) [15]. CQL underestimates the Q-value of infrequent or unseen actions in data and can effectively regulate the agents from making actions rarely used by physicians. Our experiment results suggest that our method, compared to usual RL methods such as Double Deep Q-Learning (DDQN) [16], can recommend actions more consistent with the real actions seen in data. This can avoid making extreme actions such as overdosing and thus enhances the safety of data-driven methods.

- *Intermediate rewards*. A notorious problem of RL methods for sepsis treatment is the sparse reward issue. Usually, only 90-day mortality is used as reward signals which can hardly guide the action exploration process. We showed that indeed both CQL and DDQN underperform physicians if without using any other rewards. Therefore, we propose using Apache II scores (Acute Physiology and Chronic Health Evaluation II) [17] as intermediate patient assessments and rewards. The Apache II scoring system offers more comprehensive assessments that have incorporated multiple physiological and health indicators. Our experiments demonstrate significant improvements for both CQL and DDQN when using Apache II. Furthermore, CQL performs more stable on out-of-distribution datasets, thus with demonstrated robustness.

2 Background & Related Work

2.1 Variable-Length Time Series

Time series data occur almost everywhere in our lives, including healthcare [18, 19], fraud detection [20], and transportation [21, 22]. In the medical field, precision medicine is trending in modern healthcare [23]. It aims to provide more effective and precise medical services to individuals. This yet brings a challenge that patient-level situations vary significantly and additional efforts for precision care are needed. When it comes to developing data-driven methods, one often confronts dealing with variable-length clinical time series with different densities. However, the majority of current methods for sepsis [9, 24] still use truncated, fixed-length time series, which is obsolete for precision medicine.

Modern models to process variable-length time series include recurrent neural networks (RNNs) [25, 26] and their variants such as Long Short-Term Memory Networks (LSTMs), Gated Recurrent Units (GRUs), and temporal convolutional networks (TCNs) [27]. Recently, Transformer models have also gained attention for their ability to handle long-range dependencies in time series data through self-attention mechanisms [28, 29].

Furthermore, interpolation and filling methods (e.g., linear interpolation, Gaussian processes) and self-encoders (especially convolutional self-encoders) are frequently employed to process and compress time series data and to address issues such as zero-filling [30, 31].

These techniques have been used widely in the processing of variable-length time series data, particularly within the domain of healthcare and precision medicine [32].

2.2 Offline Reinforcement Learning

Offline Reinforcement Learning (RL) is a branch of RL methods that works for scenarios where real-time interaction is impractical or expensive. Online RL [33, 34] collects training experience from real-time interaction with environments, while offline RL utilizes pre-collected datasets to train agents. This is particularly suited to the medical field where retrospective data are relatively abundant while training agents in a real clinical environment is infeasible [9, 35, 36]. Other areas using offline RL methods include autonomous driving [37, 38], finance [39], etc.

The mathematical foundation of RL is Markov Decision Process (MDP). An MDP is defined as (S, A, P, R, γ), which are states S, actions A, transition probability $P(s' \mid s, a)$, and reward function $R(s, a)$ with a discount factor γ. The goal in RL is to find an optimal policy $\pi(a|s)$ that can maximize the expected cumulative reward $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right]$. In offline RL, the training data is a pre-collected dataset of episodes $\mathcal{D} = \{(s_i, a_i, r_i, s_{i+1})\}_{i=1}^{N}$, where s_i denotes a state, a_i an action, r_i a reward, and s_{i+1} the next state. We usually say that the retrospective dataset was generated by a behavioral policy (or a physician policy in the medical context). The goal of offline RL is to learn a different policy that may outperform the behavioral policy.

Because of its offline nature, offline RL is particularly suited for the medical field, often used to optimize personalized clinical plans. For example, Harry Emerson et al. [40] employed offline RL to train an agent capable of assisting in the treatment of type I diabetes patients. They demonstrated that, in addition to preventing hypoglycaemic episodes, offline RL could also effectively address common and challenging treatment scenarios such as wrong dosage of medication administered, irregular mealtimes, and execution errors.

2.3 Safe Reinforcement Learning

The safety concern of offline reinforcement learning often hinders its use in practice [41]. The primary reason lies in over-estimating of the Q-values of unseen actions in data. In certain applications, some actions are prohibited because of domain knowledge. However, such knowledge is implicitly reflected in data, so naively trained agents can choose risky and unseen actions that may break the safety rules.

In light of this, a body of research work has been devoted to developing safe reinforcement learning models. Though focusing on different aspects, their objective is to guarantee that specific security constraints are satisfied during the learning and execution of policies. For instance, constrained Markov decision process (CMDP) [42] aims to optimize agents' behaviors towards a maximum reward while also restricting their actions on certain states [43–45]. Some work also leverages external knowledge for MDP optimization [46, 47]. Additionally, some studies utilize policy and value-based techniques [48, 49] to develop safe RL methods. Safe RL methods are now also being incorporated into many domains such as autonomous driving [50, 51], finance [52], and healthcare [53–56].

2.4 Intermediate Rewards

Many reinforcement learning scenarios confront sparse rewards, that is, final reward signals are provided only after the agent has taken numerous actions. For example, in game playing, the true signal is the final win-loss; in clinical decisions, the treatment effects are usually delayed after a while. Overly sparse signals can result in very inefficient or even biased training of agents. To mitigate this issue, intermediate rewards are used, which are artificially designed rewards provided *during* the execution of a task prior to its completion [57]. Though not identical to final rewards, intermediate rewards are often effective in guiding agents to the optimal direction of exploration.

There are already many intermediate reward designs in medical RL methods. For instance, in the context of heparin dosage management, the combined use of APTT (activated partial thromboplastin time) and a scaling function serves as an intermediate incentive, facilitating optimal dose adjustment [58]. The utilization of intermediate rewards based on the maintenance of clinically acceptable and safe ranges for mean arterial pressure (MAP) and the Riker Sedation-Agitation Scale (SAS) facilitates the development of more effective sedation management strategies [59]. In the management of sepsis, the improvement in patient status is evaluated through the assessment of changes in Sequential Organ Failure Assessment (SOFA) scores and lactate levels (a measure of cellular hypoxia in sepsis patients) [11]. While these methods are demonstrated useful, they are typically based on a single or two metrics and can only be applied in specific domains. Therefore, we develop intermediate rewards that can provide both effective feedback and comprehensive medical knowledge.

3 Methods

3.1 Data Extraction and Pre-processing

3.1.1 Data Source

Our data are extracted from MIMIC-III (Medical Information Mart for Intensive Care III) [14], a publicly accessible database of electronic health records (EHRs) of patients admitted to intensive care units (ICUs). The MIMIC-III database was created by the Massachusetts Institute of Technology (MIT) Computer Science and Artificial Intelligence Laboratory (CSAIL) in collaboration with Beth Israel Deaconess Medical Center. It comprises clinical data from approximately 60,000 ICU patients between 2001 and 2012, including demographic information, vital signs, laboratory test results, medication utilization records, diagnostic codes, nursing records, imaging reports, and more.

3.1.2 Sepsis Determination

Sepsis is defined as a life-threatening organ dysfunction resulting from an uncontrolled host response to infection. In accordance with the Sepsis-3 definition, organ dysfunction can be identified by an infection-induced total SOFA (Sequential Organ Failure Assessment) score of 2 or greater [60]. Accordingly, the individual indicators utilized to calculate the SOFA score for each adult patient were initially selected and subsequently combined in parallel. We calculated the SOFA values by using eight clinical indicators, which were scored as shown in Table 1. The time points with SOFA scores of 2 or above were calculated and screened for each patient. The time points with a difference of less than 48 h between the preceding and following time intervals were merged into a single time period. The complete sepsis time series was constructed using the time period that included all time points from the first 24 h to the last 48 h (illustrated in Fig. 2). Consequently, some patients may have multiple episodes.

3.1.3 Discretization

Based on the time windows defined, each patient's other data such as vital signs, laboratory values, demographics, elixhauser premorbidity status[61], pressor-boosting medications received, and fluid balance were extracted. Then data were further structured as multidimensional discrete time series with a temporal resolution of 4 h. In each 4-h time step, values appearing in the same step would be averaged (vital signs, tests), summed (e.g., fluids, medications), or otherwise operated as needed. Compared to the

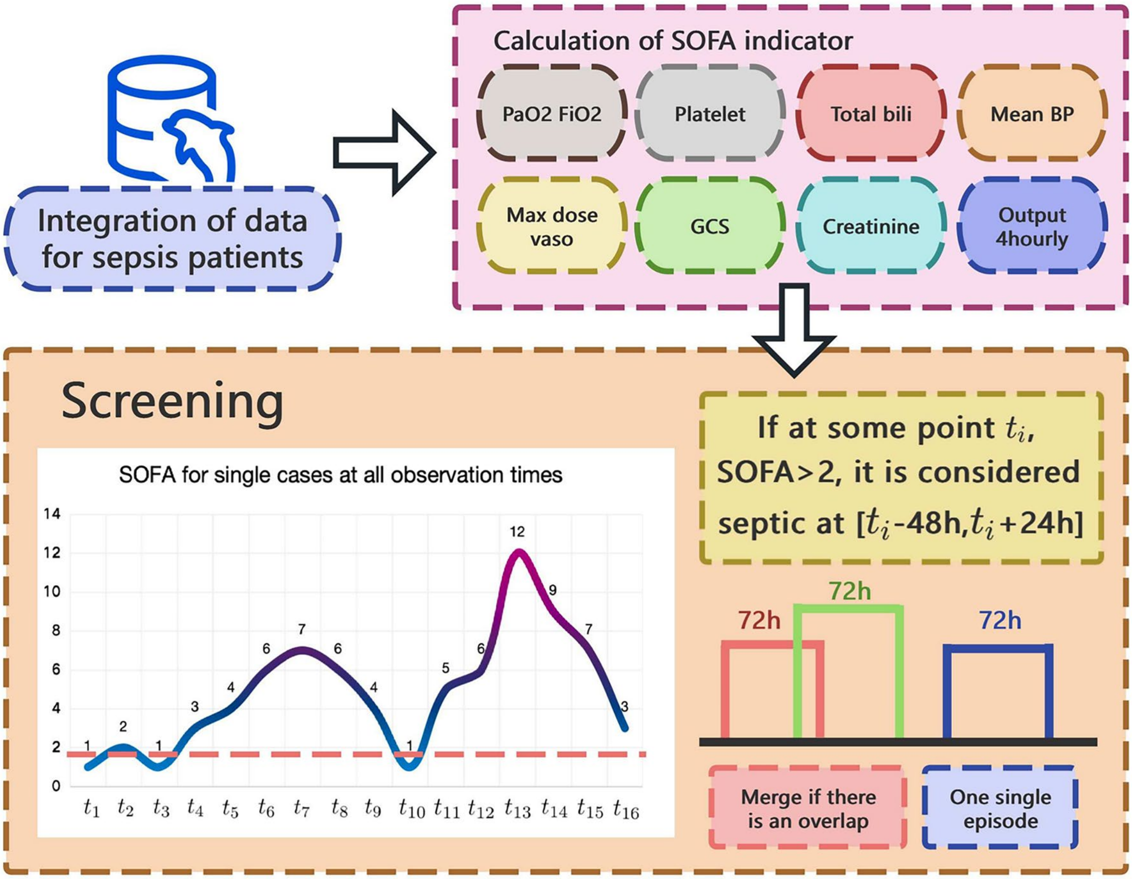

Integration of data for sepsis patients
Calculation of SOFA indicator
PaO2 FiO2
Platelet
Total bili
Mean BP
Max dose vaso
GCS
Creatinine
Output 4hourly
Screening
SOFA for single cases at all observation times
14
12
10
8
6
4
2
0
1
2
1
3
4
6
7
6
4
1
5
6
12
9
7
3
t_1 t_2 t_3 t_4 t_5 t_6 t_7 t_8 t_9 t_{10} t_{11} t_{12} t_{13} t_{14} t_{15} t_{16}
If at some point t_i, SOFA>2, it is considered septic at [t_i-48h,t_i+24h]
72h
72h
72h
Merge if there is an overlap
One single episode

Fig. 2 Procedure of assessing and selecting ICU septic patients using the SOFA scoring system. First, patients' data are extracted, preprocessed, and integrated from MIMIC-III, followed by SOFA score calculation. If the SOFA score at any time t_i is greater than two, the period around that time (from $t_i - 48$ h to $t_i + 24$ h) is considered a sepsis episode. Note that overlapping periods are combined into a single sepsis episode

Table 1 SOFA rating standard

Clinical indicators	4 Points	3 Points	2 Points	1 Points	0 Points
PaO2_FiO2	< 100	100–200	200–300	300–400	> 400
Platelet (K/uL)	< 20	20–50	50–100	100–150	> 150
Total_bili (mg/dL)	> 12.0	6.0–12.0	2.0–6.0	1.2–2.0	< 1.2
Mean_BP (mmHg)			0–65	65–70	≥ 70
Max_dose_vaso (mcg/kg/min)	> 0.1	0–0.1			
GCS (points)	≤ 5	5–9	9–12	12–14	> 14
Creatinine (mg/dL) or Output 4 hourly (mL)	> 5 or < 34	5–3.5 or < 84	2–3.5	1.2–2	< 1.2

current extraction of corresponding itemid values from MIMIC-III for various indicators, and based on the specific needs for sepsis-related indicators, I communicated with professional doctors to perform a new round of filtering and cleaning to ensure the extracted data was more relevant to sepsis and its clinical context.

3.1.4 Missing Value Imputation

We employed a parameter-specific sampling-and-holding strategy within a finite-time window to fill in incomplete data. This approach is widely used in the analysis of medical time series data [62]. For variables with a low percentage of missing values, linear interpolation was used [63]; for those with a mild missing rate between 30% and 70%, K-nearest-neighbour interpolation [64] was used; for those of which the missing rate exceeds 70%, we deleted the whole variable.

Part of the information on the resulting cohort is described in Table 2.The table provides a comparative analysis of survival and non-survival groups, highlighting key demographic and clinical differences. Non-survivors tend to be older (68.4 vs. 58.8 years) and include a slightly higher proportion of female patients (43.8% vs. 40.6%). The average time series length is longer in the non-survival group (27.02 vs. 23.43 h), potentially reflecting prolonged ICU monitoring. Additionally, non-survivors exhibit higher initial (8.46 vs. 7.29) and final SOFA scores (8.16 vs. 5.15), indicating more severe organ dysfunction and a poorer prognosis. They also have lower platelet counts, mean blood pressure, and Glasgow Coma Scale (GCS) scores, alongside elevated creatinine and total bilirubin levels, suggesting greater renal and hepatic impairment. Furthermore, the PaO2_FiO2 ratio and four-hourly urine output are lower in non-survivors, pointing to more severe hypoxia and potential kidney failure. These findings underscore the greater disease burden and critical condition of non-survivors compared to survivors.

3.1.5 Reward Balancing

An overly high death rate (i.e. negative reward) would detrimentally hurt the model training. To address this issue, we performed undersampling on the death episodes by only selecting 6,000 ones randomly, while all 6611 episodes of survivors have been preserved. The final episodes were randomly divided into a training set (80%) and a validation set (20%) in multiple runs.

3.2 Conservative Q-Learning

Conservative Q-Learning (CQL) [15] is an offline reinforcement learning method that features reducing the risk of overestimation of unseen or rarely seen actions by employing a conservative estimation of the Q-values. The essence is to use a conservative regularization to suppress the Q-values of infrequent state-action pairs. This property can effectively reduce the likelihood of selecting unacceptable or risky actions. Because of this, CQL has been widely used in the medical field, and we thus adopt it as the backbone reinforcement learning model. In detail, the following equation shows its iterative Q-learning procedure indexed by k:

$$\begin{aligned}\hat{Q}^{k+1} \leftarrow \arg\min_{Q}\ \alpha \cdot \Big(&\mathbb{E}_{s\sim\mathcal{D}, a\sim\mu(a|s)}[Q(s,a)] \\ &- \mathbb{E}_{s\sim\mathcal{D}, a\sim\hat{\pi}_\beta(a|s)}[Q(s,a)]\Big) \\ &+ \frac{1}{2}\mathbb{E}_{s,a,s'\sim\mathcal{D}}\Big[\big(Q(s,a) - \mathcal{B}^{\pi}\hat{Q}^{k}(s,a)\big)^{2}\Big]\end{aligned} \tag{1}$$

The above formula consists of two parts. The first minimization term is the conservative regularization that underestimates the Q-value of unseen actions. Here, $\hat{\pi}_\beta$ refers to the empirical behavioral policy that generated the dataset $\mathcal{D}$, and $\mu(a|s)$ is a dynamically chosen conditional distribution of actions that maximizes the current Q-function, in favor of seen actions that lead to positive rewards. Based on this setting, we can see that this minimization term conservatively chooses the Q-values closest to what the actual data suggests. The second term serves to update and improve the accuracy of the Q function by minimizing the error between the Q value and the target Q value. $\mathcal{B}^{\pi}(s,a)$ represents the standard Bellman operator with respect to the current policy π, discounting the Q-value of the subsequent state to provide an estimate for the current state-action pair:

Table 2 Cohort statistics for the final dataset after preprocessing

	% of female	Mean age	Total number	Time series length(4 h)	SOFA score at series start	SOFA score at series end	PaO2_FiO2 mean
Survival	40.6%	58.8	6611	Max: 245, Avg: 23.43	Min: 0, Max: 20, Avg: 7.29	Min: 0, Max: 22, Avg: 5.15	312.58
Non-survival	43.8%	68.4	8346	Max: 311, Avg: 27.02	Min: 0, Max: 22, Avg: 8.46	Min: 0, Max: 24, Avg: 8.16	271.59
	Platelet	Total_bili	Mean_BP	Max_dose_vaso	GCS	Creatinine	Output_4 hourly
	Mean (K/uL)	Mean (mg/dL)	Mean (mmHg)	Mean (mcg/kg/min)	Mean (points)	Mean (mg/dL)	Mean (ml)
Survival	225.04	1.43	81.00	0.08	11.74	1.30	360.61
Non-survival	208.92	2.37	77.36	0.15	10.41	1.71	267.66

Table 3 48 clinical indicators to represent patient states

Group	Features
Demographics	Age, Gender, Weight, Elixhauser Score, Readmission to the ICU
Vital Signs	SOFA, SIRS, GCS, Heart Rate, SysBP, DiaBP, MeanBP, Shock Index, SpO2, Temperature, Respiratory Rate Lab Values Potassium, Sodium, Chloride, Glucose, Creatinine, Magnesium, Calcium, SGOT, Ionized Calcium, Carbon Dioxide, SGPT, Total Bilirubin, Albumin, Hemoglobin, White Blood Cells Count, PTT, PT, BUN, Platelets Count, INR, pH, PaO2, PaCO2, Base Excess, Bicarbonate, Lactate, PaO2/FiO2 Ratio, HCO3
Ventilation parameters	Mechanical Ventilation, FiO2
Fluids	Urine Output, Cumulative Fluid balance

Table 4 Doses of drugs into discretized actions

Discretized action	IV fluids (mL/4 h)	Vasopressors (mcg/kg/min)
1	0	0
2	(0, 50]	(0, 0.08]
3	(50, 180]	(0.08, 0.22]
4	(180, 530]	(0.22, 0.45]
5	>530	>0.45

$$B^{\pi}Q(s,a) = r(s,a) + \gamma \mathbb{E}_{a' \sim \pi(a'|s')}[Q(s',a')] \tag{2}$$

where the expectation is taken over actions in the next state s', reward $r(s,a)$ is the immediate reward received from taking action a in state s, and γ is a discount factor.

3.3 Building the Reinforcement Learning Model

Now we present how the preprocessed septic patients' data are incorporated into the CQL model. First, to build the state space S, 48 clinical variables reflecting patients' states are used, shown in Table 3. Second, to build the action space A, we focused on two key treatments for septic patients in ICU: intravenous fluids and vasopressor drugs. The experimental design employs a 4-h time step to monitor and compare the effects of different treatment conditions. To ensure comparability, we standardized the various treatment types by dividing each therapeutic dose into five ranges, thereby creating a treatment strategy space of 25 possible actions. This is illustrated in Table 4. Third, for reward signals, we used 90-day survival records from the patients' records, where a final state reward of +1 was given if the patient survived, and −1 was given if they died. However, experiments revealed that a strategy that rewards only this has limited lift compared to the physician's strategy in variable-length time series. Accordingly, we devised an intermediate reward mechanism that permits a more comprehensive evaluation of the patient's condition.

The intermediate rewards in our reinforcement learning framework are derived from the Apache II score (Acute Physiology and Chronic Health Evaluation II) [17], a well-established system for assessing the severity and prognosis of ICU patients. This score is computed using multiple physiological and laboratory indicators to estimate the risk of mortality. Specifically, we selected nine key physiological parameters (e.g., temperature, blood pressure, heart rate, etc.), each assigned a score based on its deviation from normal ranges, where higher scores indicate greater severity of illness. The final Apache II score is obtained by summing the individual parameter scores, with an additional adjustment based on the difference between the full Glasgow Coma Scale (GCS) score and the actual score. The detailed scoring criteria are presented in Table 5.

Table 5 Apache II rating standard

Clinical indicators	4 Points	3 Points	2 Points	1 Points	0 Points
Temperature(°C)	$> 41, \leq 30$	39–41, 30–32	32–34	38.5–39, 34–36	36-−38.5
Mean blood pressure (mmHg)	$> 160, \leq 50$	130–160	110–130, 50–70		70–110
Heart rate (bpm)	$> 180, \leq 40$	140–180, 40–55	110–140, 55–70		70–110
Arterial pH (units)	$> 7.7, \leq 7.15$	7.6–7.7, 7.15–7.25	7.25–7.33	7.5–7.6	7.33–7.5
Sodium (mEq/L)	$> 180, \leq 111$	160–180, 111–120	155–160, 120–130	150–155	130–150
Potassium (mEq/L)	$> 7, \leq 2.5$	6–7	2.5–3	5.5–6, 3–3.5	3.5–5.5
Creatinine (mg/dL)	> 305	170–305	130–170, ≤ 53		53–130
White blood cell count (K/uL)	$> 40, \leq 1$		20–40, 1–3	15–20	3–15

To integrate patient condition dynamics into the reinforcement learning process and enable more personalized treatment strategies, we incorporated the change in Apache II score between consecutive time steps S_t and S_{t+1} as an intermediate reward. Since sepsis is a severe and rapidly progressing condition, fluctuations in the Apache II score often reflect changes in the patient's clinical state. To reinforce learning towards safer and more effective treatment actions, we assigned greater weight to positive Apache II score changes, encouraging actions that contribute to patient stabilization and recovery. Additionally, to ensure numerical stability and consistency in reward scaling, the score change was normalized by dividing it by the total possible score range. By combining this intermediate reward with terminal rewards-determined by patient survival or mortality-the final reward function $r(s_t, a_t, s_{t+1})$:

$$\begin{cases} +1 & if\ t+1 = l \text{ and } m_{t+1} = 0 \\ -1 & if\ t+1 = l \text{ and } m_{t+1} = 1 \\ \frac{\mathcal{A}_t - \mathcal{A}_{t+1}}{\max_{\mathcal{A}} - \min_{\mathcal{A}}}\left(1 + 0.75 \cdot H\left(\mathcal{A}_t - \mathcal{A}_{t+1}\right)\right) & \text{otherwise} \end{cases} \tag{3}$$

where $\mathcal{A}_t$ is the modified Apache II score for patient at timestep t, m_t equals 1 if patient is dead at timestep t and 0 otherwise, l represents the length of patient's stay in the ICU, $\max_{\mathcal{A}}$ and $\min_{\mathcal{A}}$ are the maximum and minimum values of the modified Apache II score, H denotes Heaviside step function

3.4 Out-of-Distribution Dataset

The presence of out-of-distribution (OOD) [65] data in the overall data set may result in a decline in model performance. This data is often a contributing factor in the failure of neural networks [66]. Accordingly, in order to comprehensively evaluate the model's performance under extreme data scenarios, we constructed an OOD patient set. The OOD dataset comprises data that the model has never encountered during the training process and that exhibit a markedly disparate feature distribution. In particular, we define an outlier patient as one who exhibits at least one state characteristic (including demographic characteristics, vital signs, laboratory test values, or fluid balance) in the top or bottom 1% of the distribution at the outset of the time series. Such outlier patients may exhibit state characteristics that markedly diverge from the typical clinical profile. This approach enables an assessment of the model's performance in contexts characterised by exceptional and unconventional circumstances. The objective of evaluating the performance of the model on the OOD dataset is to conduct a more comprehensive assessment of the model's generalisation ability and robustness. This is done to ensure that the model is capable of dealing with a diverse range of patient populations and unseen clinical scenarios in real-world applications.

3.5 Baselines and Hyperparameters

Two baseline approaches were employed: the physician strategy and the Double Deep Q-Network (DDQN) model [16]. Physician actions are based on all successive action combinations in the collected dataset. This approach relies on the experience and rules of the medical experts, and thus these action combinations are believed to effectively simulate the decision-making process of the attending physicians in the MIMIC-III dataset. DDQN is an enhanced deep reinforcement learning algorithm that builds upon the foundations of classical Q-learning [67] and Deep Q-Network (DQN) [68]. DDQN is designed to address the issue of overestimation of Q, which is a common challenge in traditional DQNs, by reducing estimation bias through the utilisation of two independent neural networks. One of the networks (the behavioural network) is responsible for selecting the action, while the other (the target network) evaluates the Q-value associated with that action.

Before training, a grid search was conducted to determine the optimal values for the learning rate (η) and scaling factor (α) in CQL and the learning rate (η) in DDQN through pre-training. The scaling factor (α) plays a crucial role in controlling how much the algorithm penalizes actions that deviate from the observed behavior in the offline dataset, thus balancing exploration and exploitation. The values of the learning rate, denoted by η, were set to the following values: $[1^{-7}, 1^{-6}, 1^{-5}, 1^{-4}]$, while the scaling factor, denoted by α, was set to the following values: [0.05, 0.1, 0.5, 1, 2]. A total of 500,000 steps were trained by combining the aforementioned parameters in a separate manner. The results demonstrate that the optimal learning rate η for DDQN is 1^{-7}, while the optimal learning rate η for CQL is 1^{-5}. Additionally, the best scaling factor for CQL was found to be 0.05.

Regarding the neural network architecture, a structure with two hidden layers, each containing 256 nodes, was selected. This configuration strikes a balance between model complexity and computational efficiency. The hidden layers used the ReLU (Rectified Linear Unit) activation function, which is commonly chosen for its simplicity and its ability to mitigate the vanishing gradient problem during training. This architecture was chosen for its effectiveness in handling non-linear relationships and enabling the model to learn complex representations of the data.

3.6 Off-Policy Evaluation

Off-policy evaluation (OPE) [69] is a technique for evaluating reinforcement learning policies offline. It employs empirical data generated by a behavioral policy to estimate the expected return of a target policy. In the medical domain, directly testing new treatment plans on patients can be very risky and unsafe. Instead, using OPE can assess novel strategies in a secure and cost-effective environment. At present, many RL methods proposed for the treatment of sepsis utilize OPE [9, 11, 36, 70].

According to a recent comparative study [71], Fitted Q Evaluation (FQE) [72] consistently exhibited superior performance than many other OPE methods, so we adopt it in our study.

4 Results

We conducted an extensive study to evaluate our proposed model compared to alternative methods. In detail, we evaluated the following methods: CQL with or without intermediate rewards, DDQN with or without intermediate rewards, and a physician policy. The physician policy is simply the empirical action distributions conditioned on each state, which can most closely mimic the actual physicians' behaviors. To assess safety risks, a statistical comparison was conducted to illustrate the action distributions of the five methods. Furthermore, to demonstrate the superiority of our model in extreme healthcare configurations, we also tested the methods on OOD data (out-of-distribution data). Below we present the results of the experiments.

4.1 Test Performance Comparison

We performed a fivefold cross-validation to assess different policies' performance in terms of FQE values using the final reward, i.e. 90-day mortality. Each of the five runs consisted of 20 training rounds (about 100,000 steps) and 25 evaluation rounds (about 100,000 steps). The test results are shown in Fig. 3, with higher FQE values being better.

As suggested by the results, directly applying previous methods CQL and DDQN to variable-length medical datasets does not perform well. CQL was shown to be even worse than the physician policy, while DDQN performed slightly better. This is due to the sparse signal issue, which can be even worse when encountering overly long time

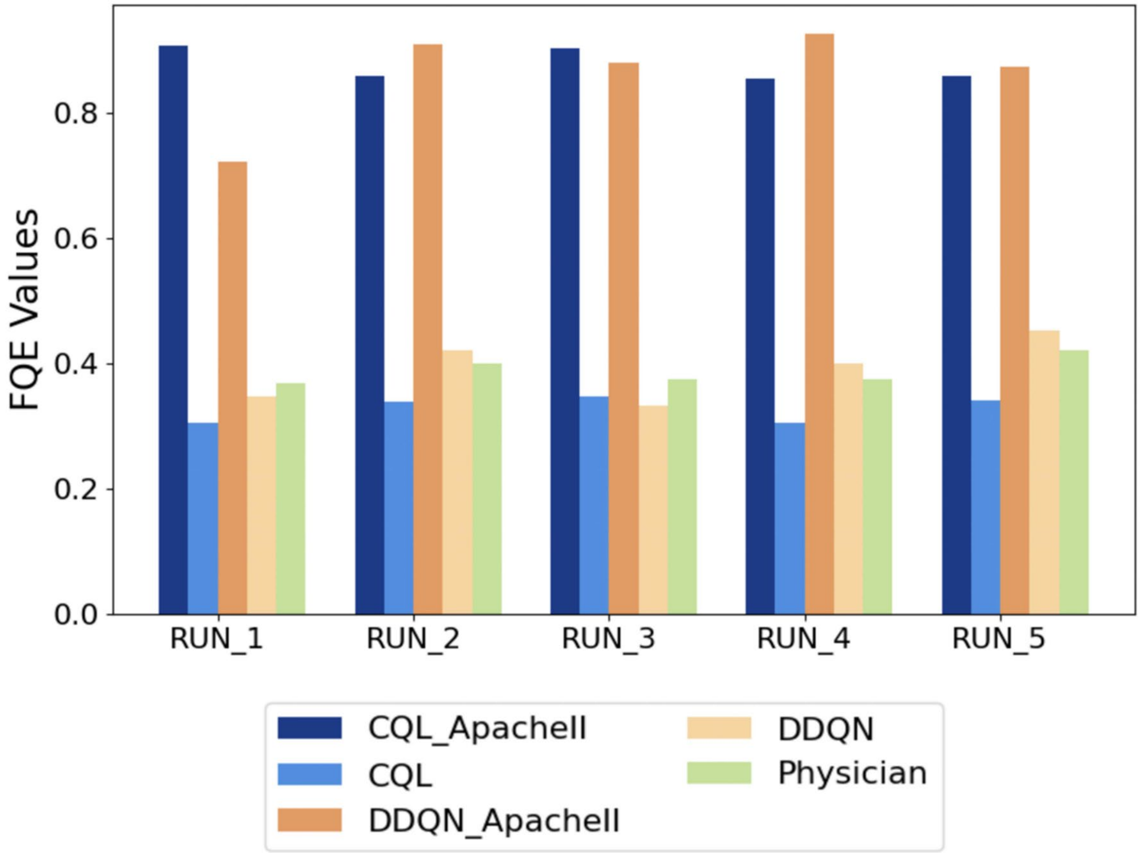


Fig. 3 Test performance of different RL models on variable-length episodes. Our method, CQL_ApacheII, overall achieves the highest FQE value of 0.88 ± .0005, and DDQN_ApacheII follows with 0.84 ± .0053. The second-tier results are without the intermediate rewards: 0.33 ± .0003 for CQL, 0.39 ± .0020 for DDQN, and 0.38 ± .0003 for the physician policy

series. Besides, CQL's underestimation strategy can also hurt its performance.

In contrast, our designed intermediate rewards can dramatically boost CQL and DDQN's performance. The Apache II scoring system offers a more comprehensive evaluation of patients' conditions and is consistent with the final rewards. Therefore, the intermediate rewards can efficiently guide the training of CQL and DDQN especially when dealing with long episodes. Consequently, the learning efficiency and outcomes both improve.

To compare CQL and DDQN with intermediate rewards, we see that DDQN overall performs better but shows weaker stability. It is as expected since DDQN is reward-driven and its only goal is to maximize the expected return. CQL, on the other hand, also balances the safety risk by underestimating the rarely seen actions. This can lead to suboptimality but with stronger safety and stability. Below we are to assess the methods' safety risk.

4.2 Safety Risk Comparison

Besides rewards, we also need to check the safety risks of different RL methods. A very rigorous definition of safety risk in our medical setting might be infeasible, as the exact risk is patient-dependent and can be hard to assess even for physicians. Instead, a more convenient way is to compare a policy's suggested actions with the actions recorded in data, as were given by physicians. So, we rely on the assumption that the safety risk is high if the suggested actions deviate too much from the real actions. To this end,

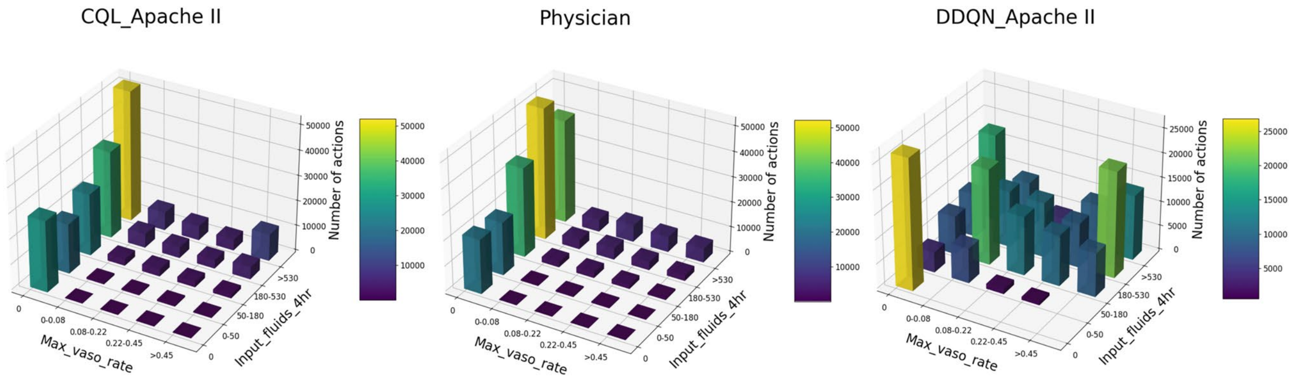


Fig. 4 Distribution of action combinations for different RL methods

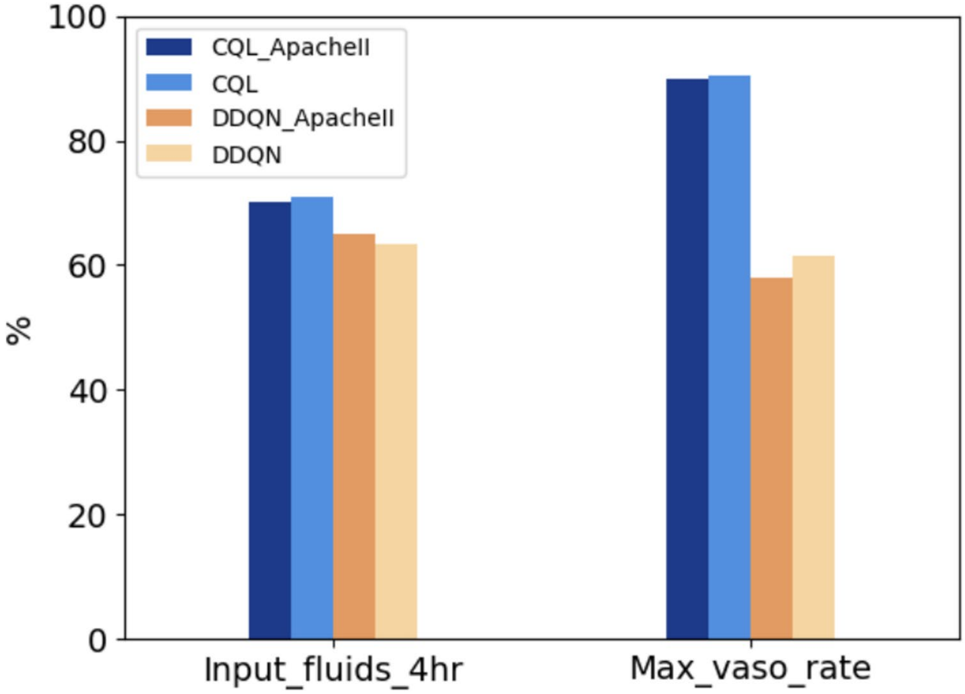


Fig. 5 Similarities between the actions suggested by RL methods and the actions adopted by physicians. CQL and CQL_ApacheII models achieved around 70% and 75%, respectively, on the IV input volumes per 4 h; and on the pressor drug dose, both CQL methods achieved about 95% of similarities

we plot in Fig. 4 the action distributions given by CQL, DDQN (both with intermediate rewards), and the physician policy. We also calculate their overall similarities to provide a quantitative comparison, illustrated in Fig. 5.

In Fig. 4 we present the distributions of twenty-five discretized action combinations of intravenous fluids (IVs) and vasopressors, where the intervals can be found in Table 4. From this chart, firstly, we see that the physician policy indicates a preference for intravenous fluids as a primary treatment, with volumes typically more than 180 ml, while the use of vasopressors tends to be conservative. Pressor drugs are good at maintaining patients' blood pressure and improving organ perfusion; however, excessive use may result in adverse effects [73]. Overusing pressors may result in an increased cardiac load, which in turn increases the risk of cardiac events, particularly in patients with sepsis, a group with an elevated risk of underlying cardiovascular disease [74]. Therefore, in the treatment of sepsis, healthcare professionals typically exercise greater caution in selecting the type and dosage of pressor drugs, aiming to achieve a balance between therapeutic efficacy and potential adverse effects.

When it comes to RL agents CQL and DDQN, we clearly see that CQL's suggested actions are much more consistent with the physicians', while DDQN's actions appear more explorative. DDQN demonstrated a preference for using medium- to high-dose (0.22–0.45) pressor drugs, which elevated the risk of adverse effects, particularly in septic patients. Furthermore, DDQN also demonstrated a broader distribution of IV use but failed to effectively concentrate on high-dose IVs, indicating its deficiencies in balancing therapeutic effectiveness and safety.

In comparison, CQL effectively mimicked physicians' behavior by focusing on the utilization of low-dose (≤0.08) and medium-dose (0.08–0.22) pressor medications, and employing high-dose IV fluids (>530) to maintain patients' hemodynamics, thus having reduced the reliance on high-dose pressor medications. This strategy not only effectively balances therapeutic effectiveness and safety, but also reduces the risk of potential adverse effects by decreasing the use of high-dose pressor drugs.

In Fig. 5, we further provide quantitative similarity scores of actions given by different policies. We see that the similarities of CQL-based models in terms of input volume per 4 h and maximum pressor drug dose were highly consistent with the physician's treatment strategy, almost mimicking the physician's strategies. In contrast, the DDQN and DDQN_ApacheII models performed poorly on the maximum pressor drug dose, with only about 60% of the doses approaching the physician's strategy, showing a higher risk of safety concerns.

4.3 Evaluation with OOD Data

In the medical field, the generalization and robustness of a model are important for evaluating its applicative use. One important aspect is that a well-trained model (or even a human physician) can occasionally encounter patients with extreme states. To this end, we held out an out-of-distribution (OOD) dataset that was unseen in the model training phase. We assessed different methods on this OOD dataset and reported

Fig. 6 FQE values of different models when tested on out-of-distribution data

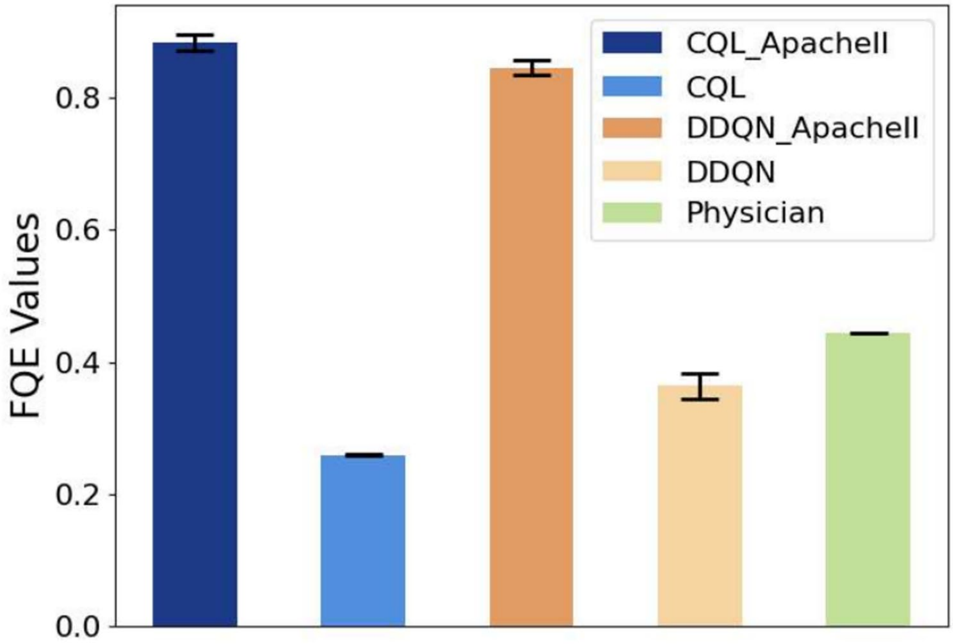


Table 6 FQE Values of methods when tested with ID and OOD data

	CQL+	CQL	DDQN	DDQN+
In distribution	0.88 ± .0005	0.33 ± .0003	0.39 ± .0020	0.84 ± .0053
Out of distribution	0.88 ± .0024	0.26 ± .0002	0.36 ± .0040	0.84 ± .0021

their FQE values in Fig. 6. The results suggest that in general, the use of intermediate rewards helped the RL models achieve much superior performance, indicating enhanced generalization and robustness. If without the guidance of intermediate rewards, we can see that both CQL and DDQN demonstrated certain extents of over-fitting and very unstable generalization, even worse than the physician. This indicates that RL-based approaches may have inherent limitations when confronted with complex and variable clinical data.

To further validate the contribution of Apache II scores, we compared the performance of each model on in-distribution versus out-of-distribution data and reported the FQE values in Table 6. The boost of performance indicates that the incorporation of Apache II scores as an intermediate bonus (with +) resulted in a notable improvement in the performance of these two models in the out-of-distribution data test, which approached that observed in the in-distribution data. This further substantiates the efficacy of Apache II scores in enhancing model robustness and adaptability.

5 Discussion & Conclusion

In this study, we aimed to develop a data-driven model that can provide effective and safe sepsis treatment recommendations for patients in intensive care units. We proposed to use Conservative Q-learning (CQL), a reinforcement learning model to train safe agents with offline clinical data. To deal with variable-length time series episodes, we combined CQL with the Apache II scoring system, a set of patient state assessment criteria as the intermediate rewards, to improve the efficacy of the model's recommendations. Extensive comparative experiments in both in-distribution and out-of-distribution data demonstrate that our model significantly outperforms physicians and other comparative algorithms, exhibiting superior performance and safety robustness.

Our future work will focus on several key areas to further enhance the model's performance and clinical applicability. First, we plan to integrate additional clinical indicators and data sources. Sepsis is a complex disease, and expanding the range of data, including more physiological parameters, lab results, and patient demographics, will improve the model's robustness and prediction accuracy, allowing for a more comprehensive understanding of sepsis. Second, we intend to collaborate with multiple healthcare organizations to validate the model across diverse datasets. This will help assess its generalizability and adaptability in different clinical environments. Working with hospitals and medical centers will provide valuable insights, refine the model, and ensure it performs well across varying patient populations and data conditions.

Moreover, we aim to incorporate continual learning into the system. This will allow the model to continually learn and update itself with new data, enabling it to adapt to evolving clinical practices and emerging data. Through continual learning, the system will dynamically adjust its strategies and model parameters, continuously improving its performance and utility in the clinical setting.

Additionally, we plan to expand the scope of our comparisons across different fields, such as exploring how similar approaches can be applied to other critical diseases or conditions. This will help assess the potential for the model's generalization to broader medical contexts. By integrating expertise from various disciplines, we aim to refine the system further and enhance its versatility and robustness in addressing a wider range of healthcare challenges.

Acknowledgements The work presented was supported by the National Natural Science Foundation of China (No. 62302405), the Natural Science Foundation of Sichuan Province (Grant No. 2024NSFSC0488), China Postdoctoral Science Foundation (Grant No. 2023M732914), and the Fundamental Research Funds for the Central Universities (No. 2682023ZT007). The content of this paper is solely the responsibility of the authors and does not necessarily represent the official views of the abovementioned funding agencies.

Author Contributions All authors are contributed equally.

Funding The work presented was supported by the National Natural Science Foundation of China (No. 62302405), the Natural Science Foundation of Sichuan Province (Grant No. 2024NSFSC0488), China Postdoctoral Science Foundation (Grant No. 2023M732914), and the Fundamental Research Funds for the Central Universities (No. 2682023ZT007).

Data Availability The MIMIC-III data used in this study were accessed from the freely available Medical Information Mart for Intensive Care database under the identifier https://doi.org/10.13026/C2XW26. The data can be obtained through a request to the PhysioNet repository after completing the required data use agreement and training in human subjects research.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

Ethics Approval and Consent to Participate Not applicable.

Consent for Publication Not applicable.

References

1. Matot I, Sprung CL. Definition of sepsis. Intensive Care Med. 2001;27(14):3–9.
2. Paoli CJ, Reynolds MA, Sinha M, Gitlin M, Crouser E. Epidemiology and costs of sepsis in the united states-an analysis based on timing of diagnosis and severity level. Crit Care Med. 2018;46(12):1889–97.
3. Teggert A, Datta H, Ali Z. Biomarkers for point-of-care diagnosis of sepsis. Micromachines. 2020;11(3):286.
4. O'Brien JM Jr, Ali NA, Aberegg SK, Abraham E. Sepsis. Am J Med. 2007;120(12):1012–22.
5. 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;112:102003.
6. Ju S, Kim YJ, Ausin MS, Mayorga ME, Chi M. To reduce healthcare workload: Identify critical sepsis progression moments through deep reinforcement learning. In: 2021 IEEE International Conference on Big Data (Big Data). IEEE; 2021. pp. 1640–1646.
7. Yang M, Li R, Hao T, Ma C, Li J, Liu C, Raising high-risk awareness in hemodynamic treatment with reinforcement learning for septic shock patients. In: 2022 Computing in Cardiology (CinC), vol. 498. IEEE; 2022. p. 1–4.
8. Su L, Li Y, Liu S, Zhang S, Zhou X, Weng L, Su M, Du B, Zhu W, Long Y. Establishment and implementation of potential fluid therapy balance strategies for ICU sepsis patients based on reinforcement learning. Front Med. 2022;9:766447.
9. Komorowski M, Celi LA, Badawi O, Gordon AC, Faisal AA. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. Nat Med. 2018;24(11):1716–20.
10. Liang D, Deng H, Liu Y. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. Appl Intell. 2023;53(9):11034–44.
11. Do TC, Yang HJ, Yoo SB, Oh I-J. Combining reinforcement learning with supervised learning for sepsis treatment. In: The 9th international conference on smart media and applications; 2020. pp. 219–223.
12. Adam N, Kandelman S, Mantz J, Chrétien F, Sharshar T. Sepsis-induced brain dysfunction. Expert Rev Anti-Infective Ther. 2013;11(2):211–21.
13. Kimball SL, Levy MM. Sepsis and the opioid crisis: integrating treatment for two public health emergencies. Crit Care Med. 2021;49(12):2151–3.
14. Johnson AE, Pollard TJ, Shen L, Lehman L-H, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci Data. 2016;3(1):1–9.
15. Kumar A, Zhou A, Tucker G, Levine S. Conservative q-learning for offline reinforcement learning. Adv Neural Inf Process Syst. 2020;33:1179–91.
16. Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 30; 2016.
17. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. Apache II: a severity of disease classification system. Crit Care med. 1985;13(10):818–29.
18. Che Z, Purushotham S, Cho K, Sontag D, Liu Y. Recurrent neural networks for multivariate time series with missing values. Sci Rep. 2018;8(1):6085.
19. Duchesne P, Pacurar M. Evaluating financial time series models for irregularly spaced data: a spectral density approach. Comput Oper Res. 2008;35(1):130–55.
20. Yang J, Rahardja S, Rahardja S, Click fraud detection: Hk-index for feature extraction from variable-length time series of user behavior. In: 2022 IEEE 32nd International Workshop on Machine Learning for Signal Processing (MLSP). IEEE; 2022. pp. 1–6.
21. Xu D, Cheng W, Zong B, Song D, Ni J, Yu W, Liu Y, Chen H, Zhang X. Tensorized lstm with adaptive shared memory for learning trends in multivariate time series. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34; 2020. pp. 1395–1402.
22. Rajeh T M, Luo Z, Javed M H, et al. A clustering-based multi-agent reinforcement learning framework for finer-grained taxi dispatching[J]. IEEE Trans Intell Transport Syst. 2024.
23. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat. 2011;39(2):1180.
24. Wang Z, Zhao H, Ren P, Zhou Y, Sheng M. Learning optimal treatment strategies for sepsis using offline reinforcement learning in continuous space. In: International Conference on Health Information Science. Berlin: Springer; 2022. pp. 113–124.
25. Wu X, Huang C, Robles-Granda P, Chawla NV. Representation learning on variable length and incomplete wearable-sensory time series. ACM Trans Intell Syst Technol. 2022;13(6):1–21.
26. Morid MA, Sheng ORL, Dunbar J. Time series prediction using deep learning methods in healthcare. ACM Trans Manag Inf Syst. 2023;14(1):1–29.
27. Lea C, Flynn MD, Vidal R, Reiter A, Hager GD. Temporal convolutional networks for action segmentation and detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. pp. 156–165.
28. Ashish V. Attention is all you need. In: Advances in neural information processing systems. 2017;30.
29. Lim B, Zohren S. Time-series forecasting with deep learning: a survey. Philos Trans R Soc A. 2021;379(2194):20200209.
30. Daberdaku S, Tavazzi E, Di Camillo B. A combined interpolation and weighted k-nearest neighbours approach for the imputation of longitudinal icu laboratory data. J Healthc Inf Res. 2020;4(2):174–88.
31. Jazayeri A, Liang OS, Yang CC. Imputation of missing data in electronic health records based on patients' similarities. J Healthc Inf Res. 2020;4(3):295–307.
32. Bashiri FS, Carey KA, Martin J, Koyner JL, Edelson DP, Gilbert ER, Mayampurath A, Afshar M, Churpek MM. Development and external validation of deep learning clinical prediction models using variable-length time series data. J Am Med Inf Assoc. 2024;31(6):1322–30.
33. Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: International Conference on Machine Learning. PMLR; 2018. pp. 1861–1870.
34. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347; 2017.
35. Tamboli D, Chen J, Jotheeswaran K P, et al. Reinforced sequential decision-making for sepsis treatment: the posnegdm framework

with mortality classifier and transformer[J]. IEEE J Biomed Health Inf. 2024.
36. Peng X, Ding Y, Wihl D, Gottesman O, Komorowski M, Li-wei HL, Ross A, Faisal A, Doshi-Velez F. Improving sepsis treatment strategies by combining deep and kernel-based reinforcement learning. In: AMIA Annual Symposium Proceedings, vol. 2018. American Medical Informatics Association; 2018. p. 887.
37. Sinha S, Mandlekar A, Garg A. S4rl: surprisingly simple self-supervision for offline reinforcement learning in robotics. In: Conference on Robot Learning. PMLR; 2022. pp. 907–917.
38. Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, Pérez P. Deep reinforcement learning for autonomous driving: a survey. IEEE Trans Intell Transp Syst. 2021;23(6):4909–26.
39. Deng Y, Bao F, Kong Y, Ren Z, Dai Q. Deep direct reinforcement learning for financial signal representation and trading. IEEE Trans Neural Netw Learn Syst. 2016;28(3):653–64.
40. Emerson H, Guy M, McConville R. Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes. J Biomed Inf. 2023;142:104376.
41. Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D. Concrete problems in ai safety. arXiv preprint arXiv:1606.06565; 2016.
42. Altman E. Constrained Markov decision processes. Boca Raton: Routledge; 2021.
43. Beutler FJ, Ross KW. Optimal policies for controlled Markov chains with a constraint. J Math Anal Appl. 1985;112(1):236–52.
44. Beutler FJ, Ross KW. Time-average optimal constrained semi-Markov decision processes. Adv Appl Probab. 1986;18(2):341–59.
45. Kallenberg LC. Linear programming and finite Markovian control problems. Amsterdam: Mathematical Centre; 1983.
46. Clouse JA, Utgoff PE. A teaching method for reinforcement learning. In: Machine Learning Proceedings 1992, pp. 92–101. Elsevier, Amsterdam; 1992
47. Moldovan TM, Abbeel P. Safe exploration in markov decision processes. arXiv preprint arXiv:1205.4810; 2012.
48. Khattar V, Ding Y, Sel B, Lavaei J, Jin M. A cmdp-within-online framework for meta-safe reinforcement learning. arXiv preprint arXiv:2405.16601; 2024.
49. Kalagarla KC, Jain R, Nuzzo P. A sample-efficient algorithm for episodic finite-horizon MDP with constraints. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 35; 2021. pp. 8030–8037.
50. Xiong X, Wang J, Zhang F, Li K. Combining deep reinforcement learning and safety based control for autonomous driving. arXiv preprint arXiv:1612.00147; 2016.
51. Gu Z, Gao L, Ma H, Li SE, Zheng S, Jing W, Chen J. Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning. IEEE Trans Intell Transp Syst. 2023;24(9):9966–83.
52. Yoo SJ, Gu YH, et al. Safety AARL: weight adjustment for reinforcement-learning-based safety dynamic asset allocation strategies. Expert Syst Appl. 2023;227:120297.
53. Yuan Y, Shi J, Yang J, Li C, Cai Y, Tang B. Conservative q-learning for mechanical ventilation treatment using diagnose transformer-encoder. In: 2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). IEEE; 2023. pp. 2346–2351.
54. Kondrup F, Jiralerspong T, Lau E, et al. Deep conservative reinforcement learning for personalization of mechanical ventilation treatment[J].
55. Kaushik P, Kummetha S, Moodley P, Bapi RS. A conservative q-learning approach for handling distribution shift in sepsis treatment strategies. arXiv preprint arXiv:2203.13884; 2022.
56. Cai X, Chen J, Zhu Y, et al. Towards real-world applications of personalized anesthesia using policy Constraint Q Learning for Propofol Infusion Control[J]. IEEE J Biomed Health Inf. 2023;28(1):459–69.
57. Mataric MJ. Reward functions for accelerated learning. In: Machine Learning Proceedings 1994. Elsevier, Amsterdam; 1994. pp. 181–189.
58. Lin R, Stanley MD, Ghassemi MM, Nemati SA, deep deterministic policy gradient approach to medication dosing and surveillance in the ICU. In: 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC). IEEE; 2018. pp 4927–31.
59. Eghbali N, Alhanai T, Ghassemi MM. Patient-specific sedation management via deep reinforcement learning. Front Dig Health. 2021;3:608893.
60. Singer M, Deutschman CS, Seymour CW, Shankar-Hari M, Annane D, Bauer M, Bellomo R, Bernard GR, Chiche J-D, Coopersmith CM, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). JAMA. 2016;315(8):801–10.
61. Elixhauser A, Steiner C, Harris DR, Coffey RM. Comorbidity measures for use with administrative data. Med Care. 1998;36(1):8–27.
62. Hug C. Detecting hazardous intensive care patient episodes using real-time mortality models. PhD thesis; 2009.
63. Blu T, Thévenaz P, Unser M. Linear interpolation revitalized. IEEE Trans Image Process. 2004;13(5):710–9.
64. Tutz G, Ramzan S. Improved methods for the imputation of missing data by nearest neighbor methods. Comput Stat Data Anal. 2015;90:84–99.
65. Hendrycks D, Gimpel K. A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136; 2016.
66. Nitsch J, Itkina M, Senanayake R, Nieto J, Schmidt M, Siegwart R, Kochenderfer MJ, Cadena C, Out-of-distribution detection for automotive perception. In: 2021 IEEE International Intelligent Transportation Systems Conference (ITSC). IEEE; 2021. p. 2938–43.
67. Watkins CJ, Dayan P. Q-learning. Mach Learn. 1992;8:279–92.
68. Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, Riedmiller M, Fidjeland AK, Ostrovski G, et al. Human-level control through deep reinforcement learning. Nature. 2015;518(7540):529–33.
69. Uehara M, Shi C, Kallus N. A review of off-policy evaluation in reinforcement learning. arXiv preprint arXiv:2212.06355; 2022.
70. Jia Y, Burden J, Lawton T, Habli I. Safe reinforcement learning for sepsis treatment. In: 2020 IEEE International Conference on Healthcare Informatics (ICHI). IEEE; 2020. pp. 1–7.
71. Tang S, Wiens J. Model selection for offline reinforcement learning: practical considerations for healthcare settings. In: Machine Learning for Healthcare Conference. PMLR; 2021. pp. 2–35
72. Le H, Voloshin C, Yue Y. Batch policy learning under constraints. In: International conference on machine learning. PMLR; 2019. pp. 3703–3712.
73. Rivers E, Nguyen B, Havstad S, Ressler J, Muzzin A, Knoblich B, Peterson E, Tomlanovich M. Early goal-directed therapy in the treatment of severe sepsis and septic shock. N Engl J Med. 2001;345(19):1368–77.
74. Asfar P, Meziani F, Hamel J-F, Grelon F, Megarbane B, Anguel N, Mira J-P, Dequin P-F, Gergaud S, Weiss N, et al. High versus low blood-pressure target in patients with septic shock. N Engl J Med. 2014;370(17):1583–93.