# Exploring Time-Step Size in Reinforcement Learning for Sepsis Treatment

**Anonymous Author(s)**
Affiliation
Address
email

## Abstract

Existing studies on reinforcement learning (RL) for sepsis management have mostly aggregated patient data into 4-hour time steps. Although this coarseness may distort patient dynamics and lead to suboptimal policies, the extent to which this is a problem in practice remains unexplored. In this work, we conducted controlled experiments of four time-step sizes ($\Delta t = 1, 2, 4, 8$ h), following an identical offline RL pipeline to quantify effects on state representation learning, behavior cloning, policy training, and off-policy evaluation. Under our model-selection criteria, 1 h time-step size yielded the highest estimated returns; however, we caution that this naive comparison is not "fair" because the evaluation makes different assumptions about the underlying problem. Our work highlights that time-step size is a core design choice in offline RL for healthcare and emphasizes the importance of thoughtful evaluation.

## 1 Introduction

Reinforcement learning (RL) has shown promise for sequential decision-making in healthcare, enabling data-driven sepsis treatment policies [1]. Unlike typical RL problems with discrete steps, electronic health record (EHR) data are often recorded at irregular time intervals. This irregularity poses challenges for the direct application of RL to such data. A common workaround is a fixed-length discretization of time. For example, the landmark work by Komorowski et al. [1] used 4 hour time-steps. However, such discretization may introduce biases and obscure physiological changes, negatively impacting policy learning and evaluation [2]. So far this bias has been studied only in theory; almost all empirical work on this domain has adhered to the 4-hour time step and has not systematically explored the impact of other time-step sizes on the entire RL pipeline.

In this work, we applied RL to the MIMIC-III sepsis management domain with four time-step sizes ($\Delta t = 1, 2, 4, 8$ h). While this may seem like a simple change in preprocessing, it also alters the problem formulation, the cohort, and action space definition, posing challenges for a "fair" comparison. To facilitate analysis across different time step sizes, we used the same cohort, designed normalized action spaces, and learned and evaluated policies separately for each $\Delta t$ following an identical offline RL pipeline. Our results show that the 1-h time-step size yielded the highest estimated returns. However, because changing $\Delta t$ induces a different MDP used during evaluation, such a naive comparison is not fair. Our work highlights that time-step size is a core design choice for healthcare RL that affects both learning and evaluation.

## 2 Related Works

Table 4 in Appendix A.1 summarizes recent studies on RL for sepsis. Nearly all adopted $\Delta t = 4$ h, inherited from the seminal work by Komorowski et al. [1]. Jeter et al. [3] criticizes the coarse $\Delta t$ for potentially failing to capture rapid physiological changes. Lu et al. [4] found that using 1 h time steps altered the learned policy, suggesting that a 4 h step might obscure decision timing. To our knowledge, no controlled study has been done to compare different $\Delta t$ values.

Under review. Do not distribute.

# 3 Formulations

Suppose the timeline starts at an anchor time $t_0$ and ends at an ending time $T$. We discretize the continuous timeline into non-overlapping windows of size $\Delta t$. We define the boundaries between consecutive windows $t_k = t_0 + k\Delta t,\ k = 0, \ldots, L$, where $L = \lceil (T - t_0)/\Delta t \rceil$ is the total number of time steps. The $k$-th time step is the half-open interval $[t_k, t_{k+1})$ for $k = 0, \ldots, T-1$.

Typically, we model healthcare RL problems as partially observable Markov decision processes (POMDPs). All information recorded **within the window** $[t_k, t_{k+1})$ is aggregated into a (raw) observation vector $o_k$. A learned encoder $f(\cdot)$ is used to derive the state at step $k$ from the history, $s_k = f(o_{0:k})$. Based on $s_k$, the action selected and executed **within the subsequent window** $[t_{k+1}, t_{k+2})$ is denoted as $a_k$, yielding a reward $r_k$ [5]. This action affects the transition to the next state $s_{k+1}$. When a terminal state $s_T$ is reached, the process terminates, generating a trajectory $\tau = (s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T)$.

# 4 Experimental Setup

We applied an identical offline RL pipeline (Fig. 1) to data discretized at $\Delta t \in 1, 2, 4, 8$ h, including the following stages: *Pre-processing* → *Approximate Information State (AIS)* → *Behavior Cloning (BC)* → *Batch-Constrained Q-learning (BCQ)* → *Weighted Importance Sampling (WIS)* (OPE).

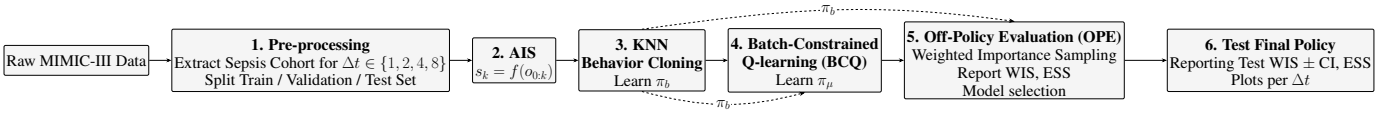



Figure 1: Overview of the offline RL pipeline.

## 4.1 Dataset & Cohort Construction

We used MIMIC-III v1.4 database [6], focusing on the adult ICU patients with sepsis following [7]. For each hospitalization, we kept the first ICU stay and extracted demographic data and time-series data. We then estimated the sepsis onset using the Sepsis-3 criteria [8]. For each stay we assembled trajectories from 28 h pre-onset to 52 h post-onset (up to 80 h). We handled outliers, missing values and implausible data following [7], and built a separate cohort for each $\Delta t$. Since trajectories shorter than $\Delta t$ are excluded, the cohort sizes differ across $\Delta t$. To ensure fair comparison, we conducted all experiments on a unified cohort defined as the intersection of the cohorts for all $\Delta t$. We then split the cohort into 70/15/15% (train/validation/test).

## 4.2 Offline RL Pipeline

**Data Preprocessing.** Trajectories were discretized separately for each $\Delta t$. Each step has 33 time-varying features plus 5 demographic features (Appendix A.2), forming the observations of a POMDP. Following [1], we defined the action space using total volume of intravenous (IV) fluids and the maximum dose of vasopressors within a time step, each binned into 5 levels (25 actions total). Following [9], we used clinically motivated bins and designed NORMALIZED-THRESHOLD boundaries to enable cross–$\Delta t$ comparison (Table 1). A sparse reward of $+100$ was given for survival (at discharge or at end of trajectory).

Table 1: Normalized-Threshold action space for discretizing intravenous (IV) fluids and vasopressors.

| **Level** | IV fluids (mL/$\Delta t$) | Vasopressor ($\mu$g kg$^{-1}$ min$^{-1}$) |
|---|---|---|
| 0 | $= 0$ | $= 0$ |
| 1 | $(0,\ 125\Delta t)$ | $(0,\ 0.08)$ |
| 2 | $[125\Delta t,\ 250\Delta t)$ | $[0.08,\ 0.20)$ |
| 3 | $[250\Delta t,\ 500\Delta t)$ | $[0.20,\ 0.45)$ |
| 4 | $\geq 500\Delta t$ | $\geq 0.45$ |

**Approximate Information State.** We learned latent states from the history of observations with a GRU encoder [10], following [11] and [12]. At each time step $k$, the encoder maps the 33-dimensional observation, 5-dimensional demographic context, and action $a_{k-1}$ to a $D$-dimensional latent state $s_k$.

A dual-head objective reconstructs $o_k$ and predicts $o_{k+1}$ via $P(o_{k+1}|s_k, a_k)$. We conducted a grid search, selecting models with the lowest validation negative log-likelihood (NLL).

**Behavior Cloning.** We learned $k$-nearest-neighbors (kNN) classifiers for behavior cloning of clinicians' policy $\pi_b(a|s)$ [13]. We performed a hyperparameter search over $k$ and the distance metric separately for the train/validation/test set. Best classifiers were selected based on their macro and micro averaged area under the receiver operating characteristic curve (AUROC) via 5-fold cross validation, and were used as $\pi_b$ for BCQ and OPE.

**Batch-Constrained Q-learning.** To avoid extrapolation error beyond $\pi_b$ [14], we used discrete batch-constrained Q-learning (BCQ) [15]. Our Q-network is a 3-layer network that estimates $Q(s, a)$, together with a target network that was updated by Polyak averaging. At each update, Q-network selected the next action from a set generated by $\pi_b$ from BC, in which actions whose estimated probability fell below a threshold $\varepsilon$ were masked out. We trained the network with a Huber loss using 5 seeds and 8 values of $\varepsilon$ (see Appendix A.6), and applied OPE on validation set to select the final policy $\pi_\mu$.

**Off-policy Evaluation.** We evaluated $\pi_\mu$ using weighted importance sampling (WIS) for off-policy evaluation (OPE). In WIS, we first computed per-step importance ratios $\rho_k = \frac{\pi_\mu(a_k|s_k)}{\pi_b(a_k|s_k)}$, and then took a weighted average of the observed returns, normalizing by the sum of the importance weights [16]. To control the estimator variance, we truncated the cumulative importance ratios $W = \prod_{k=1}^{H} \rho_k$ at $W \leq 1.438^H$ [17]. We recorded the effective sample size (ESS) [18], which reflects how many trajectories contribute meaningfully after weighting. In Section 5, we present the validation ESS–WIS Pareto frontier for candidate policies, which consists of the set of candidate policies for which no other policy simultaneously achieves both higher WIS and higher ESS. We also report WIS and ESS with standard errors estimated via bootstrapping for each policy, with results shown separately for each $\Delta t$. To complement these metrics, we further include heatmaps showing how the BCQ policy redistributes action probabilities relative to the clinician policy.

# 5 Results

**Cohort Statistics.** In Appendix A.7, we compare the cohort sizes across $\Delta t$. The cohort sizes decrease with coarser $\Delta t$, reflecting the exclusion of trajectories shorter than one step. For all experiments, we report results on a unified cohort that includes trajectories present under all $\Delta t$, which contains 18,377 admissions with a mortality rate of 5.9%.

**AIS Encoder.** In Table 2, we report the final selected hyperparameters and validation performance of the AIS encoder for each $\Delta t$. $\Delta t = 8$ required a smaller latent size of 32, whereas the other $\Delta t$ all used 128. We observe that validation MSE increases with larger $\Delta t$, which is expected given the longer prediction horizons.

Table 2: AIS encoder (GRU) results across time-step sizes: selected latent dimension, learning rate, and minimum validation MSE with 95% confidence intervals from 1000 bootstrap samples.

| $\Delta t$ (h) | Latent Dim | Learning Rate | MSE [95% CI] |
|---|---|---|---|
| 1 | 128 | 0.001 | 0.2288 [0.2181, 0.2424] |
| 2 | 128 | 0.001 | 0.2678 [0.2655, 0.2702] |
| 4 | 128 | 0.001 | 0.4011 [0.3940, 0.4110] |
| 8 | 32 | 0.001 | 0.4351 [0.4286, 0.4420] |

**Behavior Cloning.** We summarize the hyperparameter grid and results in Appendix A.6 and Table 7. Across all datasets, as $k$ increases from 21 to $5\sqrt{n}$, validation macro and micro AUROC generally improves. Based on the validation performance, we selected KNN classifiers with Euclidean distance and $k = 5\sqrt{n}$ as $\pi_b$, yielding macro AUROC $> 0.75$ and micro AUROC $\approx 0.95$. While class imbalance can reduce macro AUROC and inflate micro AUROC, the overall performance is acceptable [19].

**Off-Policy Evaluation.** Section 5 shows validation ESS-WIS Pareto frontiers for candidate policies. Across $\Delta t$, WIS is high ($\approx 100$) when ESS is small, then declines as ESS increases, reflecting the bias-variance trade-off. Coarser $\Delta t$ (4-8 h) achieves higher ESS without reducing WIS, whereas finer $\Delta t$ (1-2 h) produce lower ESS overall because they introduce more decision points, inflating variance

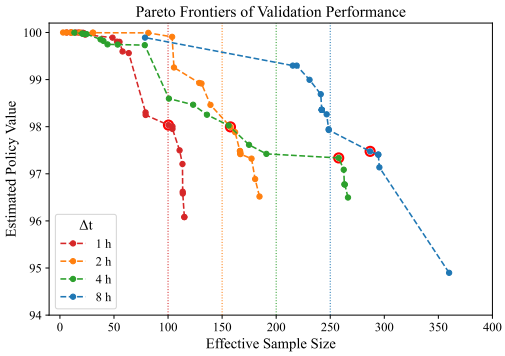



Figure 2: Pareto frontiers of validation WIS versus ESS for each time step $\Delta t$. Dashed lines trace the non-dominated points; hollow markers denote the model selected for testing; dotted lines with different colors represent the thresholds used as the boundary for model selection across $\Delta t$.

and thus reducing ESS. For each $\Delta t$, we selected a different ESS cutoff and chose the policy with ESS $\geq$ the cutoff that achieved the highest WIS. More details are provided in Appendix A.6.

**Test Performance.** Table 3 summarizes test performance. When compared against the observed rewards induced by the behavior policy $\pi_b$, BCQ $\pi_\mu$ achieves a higher WIS at 1 h, while its performance is comparable to $\pi_b$ at coarser $\Delta t$ (2–8 h). For all $\Delta t$, the test ESS exceeds the corresponding validation ESS, with coarser $\Delta t$ (4–8 h) yielding higher ESS, consistent with trends on the validation set. Coarser $\Delta t$ (4–8 h) adopt a higher BCQ threshold ($\varepsilon = 0.5$) than finer $\Delta t$ (1–2 h; $\varepsilon = 0.1$), indicating that policies are more conservative at coarser time scales. Appendix A.5 compares action frequencies between the selected $\pi_\mu$ and $\pi_b$. Both policies most frequently select zero vasopressor and low IV-fluid doses. Compared with $\pi_b$, $\pi_\mu$ assigns more probability mass to zero or low IV-fluid doses under zero vasopressor, producing a more skewed action distribution over a small set of actions.

Table 3: Test-set WIS value and ESS for BCQ and clinician (Observed $\pi_b$) policies across $\Delta t$. Observed $\pi_b$ results are identical across $\Delta t$. Values are reported with 100 bootstrap mean $\pm$ std.

| $\Delta t$ (h) | Policy | Threshold $\varepsilon$ | $\widehat{V}_{\text{test}}$ (WIS) | $\text{ESS}_{\text{test}}$ |
|---|---|---|---|---|
| 1 | BCQ $\pi_\mu$ | 0.10 | $97.88 \pm 1.01$ | $175.93 \pm 12.77$ |
| 2 | BCQ $\pi_\mu$ | 0.10 | $94.03 \pm 1.50$ | $196.29 \pm 12.54$ |
| 4 | BCQ $\pi_\mu$ | 0.50 | $94.25 \pm 1.27$ | $292.76 \pm 14.89$ |
| 8 | BCQ $\pi_\mu$ | 0.50 | $94.66 \pm 1.23$ | $280.03 \pm 15.71$ |
| All $\Delta t$ | Observed $\pi_b$ | – | $94.09 \pm 0.44$ | 2757.00 |

## 6 Conclusion & Discussion

While most prior work on RL for sepsis followed the AI Clinician [1] with 4 h time steps, we provide, to our knowledge, the first systematic comparison across 1, 2, 4, and 8 h using an identical offline RL pipeline. To enable a fairer comparison, we extracted the same cohort and train/val/test splits across $\Delta t$, designed a normalized action space for each $\Delta t$, conducted AIS and kNN grid searches per $\Delta t$, and clipped importance ratios by horizon in WIS. These choices offer a robust reference for future fair comparisons in similar tasks. Still, a fully fair evaluation remains challenging: in our task, policies are evaluated across different MDPs (induced by different $\Delta t$), so the WIS/ESS are not directly comparable across $\Delta t$. In future work, we will evaluate policies on a common resolution (e.g., evaluate policies learned under 1 h and 8 h both on an 8 h dataset) and our pipeline design makes this feasible. Our results show that time-step size is a crucial design choice that can substantially shape the learned policies in RL for sepsis task. Our results advocate for careful reconsideration from the community of different time-step sizes in sepsis management beyond the conventional 4 h setup, in order to learn better policies and make fairer evaluation across time-step sizes.

# References


[1] Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*, 24(11):1716–1720, 2018. URL https://doi.org/10.1038/s41591-018-0213-5.

[2] Peter Schulam and Suchi Saria. Discretizing logged interaction data biases learning for decision-making, 2018. URL https://arxiv.org/abs/1810.03025.

[3] Russell Jeter, Christopher Josef, Supreeth Shashikumar, and Shamim Nemati. Does the "Artificial Intelligence Clinician" learn optimal treatment strategies for sepsis in intensive care? *arXiv preprint arXiv:1902.03271*, 2019. URL https://arxiv.org/abs/1902.03271.

[4] MingYu Lu, Zachary Shahn, Daby Sow, Finale Doshi-Velez, and Li wei H. Lehman. Is deep reinforcement learning ready for practical applications in healthcare? a sensitivity analysis of duel-ddqn for hemodynamic management in sepsis patients, 2020. URL https://arxiv.org/abs/2005.04301.

[5] Shengpu Tang, Jiayu Yao, Jenna Wiens, and Sonali Parbhoo. Off by a beat: Temporal misalignment in offline RL for healthcare. In *RLC 2025 Workshop on Practical Insights into Reinforcement Learning for Real Systems*, 2025. URL https://openreview.net/forum?id=yRMY2a1rjR.

[6] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

[7] Jayakumar Subramanian and Taylor Killian. Sepsis cohort from mimic dataset. https://github.com/microsoft/mimic_sepsis, 2020. Accessed: 2025-05-22.

[8] Mervyn Singer, Clifford S Deutschman, Christopher W Seymour, Manu Shankar-Hari, Djillali Annane, Michael Bauer, Rinaldo Bellomo, Gordon R Bernard, Jean-Daniel Chiche, Craig M Coopersmith, et al. The third international consensus definitions for sepsis and septic shock (sepsis-3). *JAMA*, 315(8):801–810, 2016. doi: 10.1001/jama.2016.0287.

[9] Shengpu Tang, Aditya Modi, Michael Sjoding, and Jenna Wiens. Clinician-in-the-loop decision making: Reinforcement learning with near-optimal set-valued policies. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9387–9396. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/tang20c.html.

[10] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation, 2014. URL https://arxiv.org/abs/1406.1078.

[11] Jayakumar Subramanian, Amit Sinha, Raihan Seraj, and Aditya Mahajan. Approximate information state for approximate planning and reinforcement learning in partially observed systems, 2021. URL https://arxiv.org/abs/2010.08843.

[12] Taylor W. Killian, Haoran Zhang, Jayakumar Subramanian, Mehdi Fatemi, and Marzyeh Ghassemi. An empirical study of representation learning for reinforcement learning in healthcare, 2020. URL https://arxiv.org/abs/2011.11235.

[13] Aniruddh Raghu, Omer Gottesman, Yao Liu, Matthieu Komorowski, Aldo Faisal, Finale Doshi-Velez, and Emma Brunskill. Behaviour policy estimation in off-policy policy evaluation: Calibration matters. *arXiv preprint arXiv:1807.01066*, 2018.

[14] Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi velez, and Leo Celi. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25, 01 2019. doi: 10.1038/s41591-018-0310-5.

[15] Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration, 2019. URL https://arxiv.org/abs/1812.02900.

[16] Yao Liu and Emma Brunskill. Avoiding overfitting to the importance weights in offline policy optimization, 2022. URL https://openreview.net/forum?id=dLTXoSIcrik.

[17] Edward L. Ionides. Truncated importance sampling. *Journal of Computational and Graphical Statistics*, 17(2):295–311, 2008. ISSN 10618600. URL http://www.jstor.org/stable/27594308.

[18] Víctor Elvira, Luca Martino, and Christian P Robert. Rethinking the effective sample size. *International Statistical Review*, 90(3):525–550, 2022.

[19] Hyewon Jeong, Siddharth Nayak, Taylor Killian, and Sanjat Kanjilal. Identifying differential patient care through inverse intent inference, 2024. URL https://arxiv.org/abs/2411.07372.

[20] Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment - a deep reinforcement learning approach, 2017. URL https://arxiv.org/abs/1705.08422.

[21] Chao Yu, Guoqi Ren, and Jiming Liu. Deep inverse reinforcement learning for sepsis treatment. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 1–3, 2019. doi: 10.1109/ICHI.2019.8904645.

[22] Mehdi Fatemi, Taylor W. Killian, Jayakumar Subramanian, and Marzyeh Ghassemi. Medical dead-ends and learning to identify high-risk states and treatments. In *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=4CRpaV4pYp.

[23] Harsh Satija, Philip S Thomas, Joelle Pineau, and Romain Laroche. Multi-objective spibb: Seldonian offline policy improvement with safety constraints in finite mdps. *Advances in Neural Information Processing Systems*, 34:2004–2017, 2021.

[24] Christina X Ji, Michael Oberst, Sanjat Kanjilal, and David Sontag. Trajectory inspection: A method for iterative clinician-driven design of reinforcement learning studies. *AMIA Summits on Translational Science Proceedings*, 2021:305, 2021.

[25] Dayang Liang, Huiyi Deng, and Yunlong Liu. The treatment of sepsis: an episodic memory-assisted deep reinforcement learning approach. *Applied Intelligence*, 53(9):11034–11044, 2023.

[26] Kartik Choudhary, Dhawal Gupta, and Philip S. Thomas. ICU-Sepsis: A benchmark MDP built from real medical data. *Reinforcement Learning Journal*, 4:1546–1566, 2024.

[27] Rui Tu, Zhipeng Luo, Chuanliang Pan, Zhong Wang, Jie Su, Yu Zhang, and Yifan Wang. Offline safe reinforcement learning for sepsis treatment: Tackling variable-length episodes with sparse rewards. *Human-Centric Intelligent Systems*, 5(1):63–76, 2025.

# A appendix

## A.1 Related Works

Table 4: RL studies for sepsis care, summarizing time-step choices and key design aspects.

| Paper | $\Delta t$ | Algorithm | Dataset | Cohort | Notes |
|---|---|---|---|---|---|
| Raghu et al. [20] | 4 h | Dueling DDQN | MIMIC-III | 17.9k | Continuous state; 5×5 IV/vaso bins; first DL-RL policy (–3.6 % mortality). |
| Komorowski et al. [1] | 4 h | Batch Q-learning | MIMIC-III (+eRI*) | 17.1k | AI Clinician; 750 states, 25 actions; external validation. |
| Jeter et al. [3] | 4 h | Reproduction study | MIMIC-III | 5.4k | Finds no-action policy often rivals AI Clinician; urges caution. |
| Yu et al. [21] | 1 h | Deep IRL | MIMIC-III | 14.0k | Learns reward; highlights mortality factors (e.g. $PaO_2$). |
| Tang et al. [9] | 4 h | Set-valued DQN | MIMIC-III | 20.9k | Returns top-$k$ near-optimal dose sets for clinician choice. |
| Killian et al. [12] | 4 h | Offline DQN | MIMIC-III | 17.9k | Sequential latent encodings outperform raw features. |
| Lu et al. [4] | 1–4 h | Dueling DDQN | MIMIC-III | 17k+ | Sensitivity study on features, reward, time discretization. |
| Fatemi et al. [22] | 4 h | Dead-end discovery | MIMIC-III | 17k+ | Identifies high-risk states; secures policy to avoid them. |
| Satija et al. [23] | 4 h | MO-SPIBB | MIMIC-III | 17k+ | Safe policy improvement under performance constraints. |
| Ji et al. [24] | 4 h | Trajectory inspection | MIMIC-III | 17k+ | Clinician "what-if" review reveals policy flaws; validation tool. |
| Liang et al. [25] | 4 h | Episodic-memory DQN | MIMIC-III | 17.9k | Memory module boosts sample efficiency, lowers est. mortality. |
| Choudhary et al. [26] | 4 h | Tabular MDP | MIMIC-III | ∼18k | ICU-Sepsis benchmark: 715 states, 25 actions. |
| Tu et al. [27] | 1 h | CQL (offline) | MIMIC-III | 14.0k | Safety-aware CQL with dense rewards for variable-length stays. |

*eRI: Philips *e*ICU Research Institute cohort for external validation; DDQN: Double Deep Q-Network; DQN: Deep Q-Network; IRL: Inverse Reinforcement Learning; CQL: Conservative Q-Learning; MO-SPIBB: Multi-Objective Safe Policy Improvement with Baseline Bootstrapping.

## A.2 Extracted Features for State Representation

Table 5: Observed features extracted from the MIMIC-III database. The upper panel lists the 33-dimensional time-varying continuous variables fed to the GRU encoder, following the default code configuration. The lower panel lists the 5 static demographic / contextual variables appended to each trajectory.

**33-d Time-varying continuous features**

| | | |
|---|---|---|
| Glasgow Coma Scale | Heart Rate | Sys. BP |
| Dia. BP | Mean BP | Respiratory Rate |
| Body Temp (°C) | $FiO_2$ | Potassium |
| Sodium | Chloride | Glucose |
| INR | Magnesium | Calcium |
| Hemoglobin | White Blood Cells | Platelets |
| PTT | PT | Arterial pH |
| Lactate | $PaO_2$ | $PaCO_2$ |
| $PaO_2/FiO_2$ | Bicarbonate ($HCO_3$) | $SpO_2$ |
| BUN | Creatinine | SGOT |
| SGPT | Bilirubin | Base Excess |

**5-d Demographic and contextual features**

| | | | | |
|---|---|---|---|---|
| Age | Gender | Weight | Ventilation Status | Re-admission Status |

## A.3 Extracted Cohort Sizes

Table 6: Extracted cohort size of MIMIC-Sepsis at different time steps.

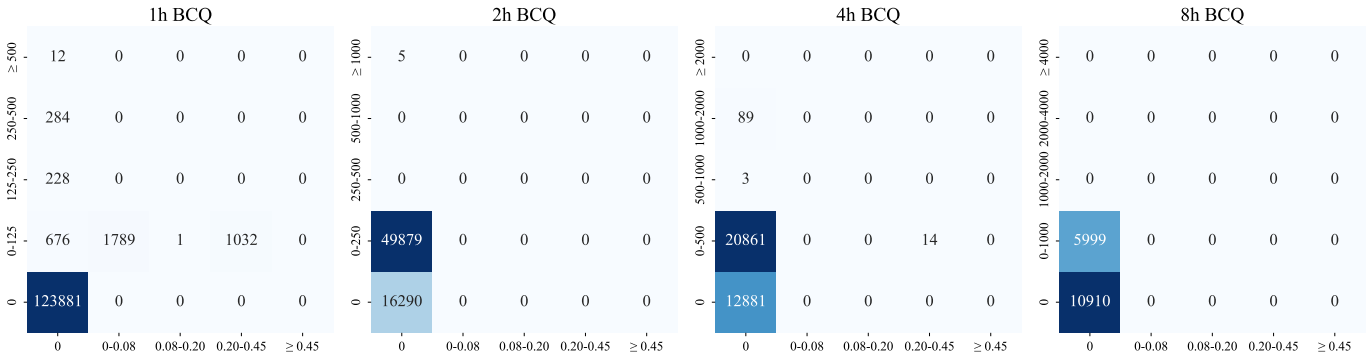
| $\Delta t$ (h) | Cohort Size |
|---|---|
| 1 | 18,995 |
| 2 | 18,987 |
| 4 | 18,906 |
| 8 | 18,783 |

## A.4 BC Performance

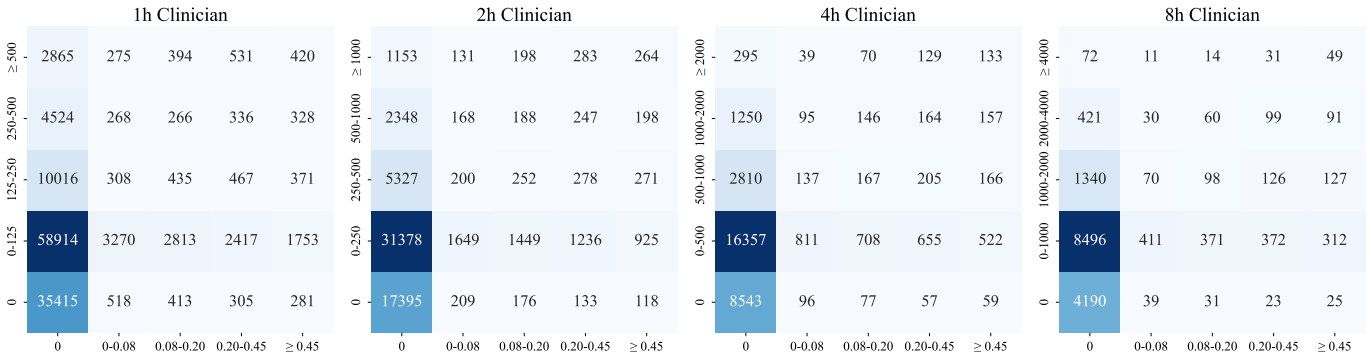
Table 7: Estimated behavior policy performance (Macro and Micro AUROC) on the validation sets across time-step sizes, with 95% confidence intervals from 1000 bootstrap samples.

| $\Delta t$ (h) | Macro AUROC [95% CI] | Micro AUROC [95% CI] |
|---|---|---|
| 1 | 0.7715 [0.7678, 0.7753] | 0.9449 [0.9443, 0.9456] |
| 2 | 0.8047 [0.7998, 0.8095] | 0.9491 [0.9482, 0.9500] |
| 4 | 0.8143 [0.8071, 0.8211] | 0.9507 [0.9496, 0.9518] |
| 8 | 0.7576 [0.7429, 0.7720] | 0.9454 [0.9435, 0.9472] |

## A.5 Action Heatmaps




(a) BCQ




(b) Observed clinicians

Figure 3: Frequency heatmap of IV-fluid (y-axis; mL) and vasopressor (x-axis; $\mu$g kg$^{-1}$ min$^{-1}$) doses for $\Delta t \in \{1, 2, 4, 8\}$. BCQ policies are compared with the empirical clinician distribution. Darker cells indicate more frequent selections.

### A.6 Additional Hyperparameter Details

Table 8: Hyperparameter values used for training GRU encoder and BCQ models.

| Hyperparameter | Searched Settings |
|---|---|
| **RNN:** | |
| – Embedding dimension, $d_S$ | $\{8, 16, 32, 64, 128\}$ |
| – Learning rate | $\{1\times10^{-5}, 3\times10^{-5}, 1\times10^{-4}, 3\times10^{-4}, 5\times10^{-4}, 1\times10^{-3}\}$ |
| **kNN:** | |
| – Number of neighbors, $k$ | $k_i = \exp\left(\ln 21 + \frac{i}{7}(\ln(5\sqrt{n}) - \ln 21)\right)$[a] |
| – Distance metric | {Euclidean, Manhattan} |
| **BCQ (with 5 random restarts):** | |
| – Threshold, $\varepsilon$ | $\{0, 0.01, 0.05, 0.1, 0.3, 0.5, 0.75, 0.999\}$ |
| – Learning rate | $3\times10^{-4}$ |
| – Weight decay | $1\times10^{-3}$ |
| – Hidden layer size | 256 |

[a] $i = 0, 1, \ldots, 7$. $n$ denotes the size of the flattened dataset.

Table 9: ESS cutoffs for model selection.

| $\Delta t$ (h) | ESS Cutoff |
|---|---|
| 1 | 100 |
| 2 | 150 |
| 4 | 200 |
| 8 | 250 |

### A.7 Extracted Cohort Sizes

Table 10: Extracted cohort size of MIMIC-Sepsis at different time steps.

| $\Delta t$ (h) | Cohort Size |
|---|---|
| 1 | 18,995 |
| 2 | 18,987 |
| 4 | 18,906 |
| 8 | 18,783 |