Exploring Time-Step Size in Reinforcement Learning for Sepsis Treatment

Yingchuan Sun[1], **Shengpu Tang**[1]
{yingchuan.sun,shengpu.tang}@emory.edu

[1]**Department of Computer Science, Emory University, USA**

Abstract

Existing studies on reinforcement learning (RL) for sepsis management have mostly followed an established problem setup, in which patient data were aggregated into 4-hour time steps. Although concerns have been raised regarding the coarseness of this time-step size, which might distort patient dynamics and lead to suboptimal treatment policies, the extent to which this happens in practice remains unexplored. In this work, we conducted empirical experiments for a controlled comparison of four time-step sizes ($\Delta t = 1, 2, 4, 8$ h) on this task, following a consistent offline RL pipeline. Our goal was to quantify how time-step size influences state representation learning, model selection, and off-policy evaluation. Our results show that smaller time-step sizes (1 h and 2 h) yielded higher estimated returns than the canonical 4 h setting without reducing the effective sample size (ESS), however this is influenced by how importance ratios are truncated during evaluation. In addition, we found that tailoring the action space definition to the distribution treatments under each time-step size led to improved policy performance. Our work highlights that time-step size and action-space definition are core design choices that shape policy learning for sepsis treatment.

1 Introduction

Reinforcement learning (RL) has shown great promise for sequential decision-making in healthcare, enabling data-driven treatment policies for complex conditions such as sepsis (Komorowski et al., 2018). Unlike typical RL problems in which states and actions are implicitly assumed to occur at regular intervals, electronic health record (EHR) data are collected at irregular intervals, e.g., vital signs and laboratory measurements occur only when patients interact with the healthcare system. This irregularity poses significant challenges for the direct application of RL on such data.

A common workaround is to discretize irregularly sampled data into fixed-length time steps. For example, aggregating measurements into 4-hour time steps as in the landmark AI Clinician work (Komorowski et al., 2018). However, studies have demonstrated that such discretization can introduce biases and obscure rapid physiological changes, negatively impacting downstream policy learning and evaluation (Schulam & Saria, 2018). So far this bias has been studied only in theory, with no empirical comparison across different time-step sizes. To date, almost all RL-based sepsis management studies (including the AI Clinician) adhered to the 4-hour time step and have not systematically studied the impact of other time-step sizes on the entire policy learning pipeline (see Appendix A.1).

In this work, we explore the impact of using 1-, 2-, 4-, or 8-hour time steps in the MIMIC-III sepsis treatment task. While this may seem to be a simple modification for preprocessing, we note that this has important consequences on the study cohort and action space definition, which poses challenges for establishing a "fair" comparison. With these considerations in mind, we learned and evaluated treatment policies at the four different time-step sizes separately following an identical offline RL

pipeline, which includes latent state representation learning, behavior cloning, batch-constrained Q-learning, hyperparameter selection, and off-policy evaluation (OPE). We found that finer time-step sizes (1 h and 2 h) improved the OPE performance over the conventional 4 h setting, whereas the coarse 8 h time-step size degraded performance. Using a clinically relevant action space further improved behavior cloning performance compared with a quantile-based action space, which corresponds to potentially more reliable OPE metrics. While we observed smaller time-step sizes to lead to comparable or even higher effective sample sizes (ESS) than larger time-step sizes, we caution that this result could be an artifact of how importance ratios were truncated during evaluation, pointing to yet another challenge for establishing a fair comparison. Our findings highlight that both time-step size and action space discretization are core design choices that shape the learned policy.

2 Background and Related Work

2.1 Time Step Discretization in RL for Healthcare

Definitions.

- *Time step* k. For each ICU admission we discretize the timeline into T consecutive windows of fixed length Δt. The windows start at an anchor time t_0. For the sepsis task, t_0 is 28 h before the estimated sepsis onset, and the windows end at most 52 h after onset, yielding a trajectory of up to 80 h. We define the boundaries

$$t_k = t_0 + k\Delta t,\ k = 0, \ldots, T.$$

  The k-th time step is the half-open interval $[t_k, t_{k+1})$ for $k = 0, \ldots, T-1$.
- *State* s_k. All vitals and labs **recorded inside** $[t_k, t_{k+1})$ are aggregated into a raw feature vector o_k (together with static demographics); we then embed it as $s_k = f(o_k) \in \mathcal{S}$.
- *Action* a_k. Treatments administered **during the next window** $[t_{k+1}, t_{k+2})$, binned into a discrete pair (IV fluid dosage, vasopressor dosage). Thus a_k is chosen after observing s_k and affects the transition to s_{k+1}.

With rewards r_k and a terminal flag $done_k$, the trajectory is

$$\tau = (s_0, a_0, r_0, \ldots, s_{T-1}, a_{T-1}, r_{T-1}, s_T).$$

Design Choices. When applying reinforcement learning to ICU sepsis management, most studies discretize each patient's EHR into 4-hour time steps ($\Delta t = 4$ h), treating every time step as a single Markov decision step. All measurements and treatments within the step are collapsed into the corresponding state-action pair. Once this aggregation is set, the critical design question is how to construct those states and actions from raw data. A widely adopted choice, first popularized by the *AI Clinician* study (Komorowski et al., 2018), is the *interval-end* representation: s_k consists of the vitals and labs measured at the end of the k-th 4-hour time step, on the assumption that these values best capture the patient's post-treatment physiology.

Prior work. In Appendix A.1 we summarize recent RL-for-sepsis studies. Nearly all adopted $\Delta t = 4$ h, inherited from the AI Clinician Work (Komorowski et al., 2018). Indeed, Lu et al. (2020) found that using 1-hour time steps substantially altered the learned policy, suggesting that a 4-hour discretization might obscure important decision timing. However, no controlled study has been done to compare different Δt values in otherwise identical setups.

3 Experimental Setup

In our experiments, we applied an identical offline RL pipeline to patient data discretized at four time-step sizes: 1 h, 2 h, 4 h, and 8 h. We describe the details below.

3.1 Dataset & Cohort Construction

We conducted our study on the MIMIC-III v1.4 critical care database (Johnson et al., 2016), focusing on adult ICU patients who developed sepsis. Following the work of Subramanian & Killian (2020),

we extracted each ICU admission's time series from 28 hours before the first sepsis onset to up to 52 hours after onset, yielding an episode trajectory with maximum 80 hours for each sepsis case. Patients younger than 18 were excluded, as were those with implausible data entries (e.g., physiologically impossible vital signs). For each Δt, this resulted in a final cohort of approximately $19,000$ trajectories.

3.2 Offline RL Pipeline

Our pipeline for offline RL comprises the following stages: *Pre-processing of raw EHR data* $\rightarrow$ *Approximate Information State (AIS)*$\rightarrow$ *Behavior Cloning (BC)* $\rightarrow$ *Batch-Constrained Q-learning (BCQ)* $\rightarrow$ *Weighted Importance Sampling (WIS)* for off-policy evaluation (OPE).

Data Pre-processing. All trajectories were discretized into fixed-length time steps, separately for each time-step size. The data extraction process produced 33 time-varying continuous features per time step, in addition to 5 static demographic and contextual features. The complete list of these features is provided in Appendix A.2. We considered two versions of discrete action space of size 25: a QUANTILE-5 grid following Komorowski et al. (2018) and a CLINICAL-THRESHOLD grid with hand-picked dose cut-offs following Tang et al. (2020). In both versions of the action space, intravenous fluid (IV fluid) and vasopressor doses were each divided into 5 levels, yielding $5 \times 5 = 25$ possible actions per Δt. The bin boundaries are summarized in Table 1. As mentioned in Section 3.1, each ICU admission was represented as a sequence of state–action–next state transitions with up to an 80-hour horizon, forming a trajectory. A time step was retained only if it contained at least one chart, lab, or intervention entry; otherwise it is skipped. Recording typically stopped when the patient died or was discharged, so the last non-empty time step naturally marked the end of the trajectory. If no such event occurred within the 80-hour horizon, the sequence was truncated at the last non-empty step within that window. Raw time-series data (vitals, labs, etc.) were cleaned by removing implausible outliers and then normalized (per-feature z-scoring using training-set statistics). We performed feature imputation for missing values to obtain complete state vectors at each time step. Finally, we split the cohort into training, validation, and test sets (70/15/15% of episodes) using a fixed random seed.

Table 1: Binning strategies for discretizing intravenous (IV) fluids and vasopressors.

Level	**Quantile-5** (Komorowski et al., 2018)		**Clinical-Threshold** (Tang et al., 2020)	
	IV fluids (mL/Δt)	Vasopressor (μg kg^{-1} min^{-1})	IV fluids (mL/Δt)	Vasopressor (μg kg^{-1} min^{-1})
0	$= 0$	$= 0$	$= 0$	$= 0$
1	$(0, q_{25})$	$(0, q_{25})$	$(0, 500)$	$(0, 0.08)$
2	$[q_{25}, q_{50})$	$[q_{25}, q_{50})$	$[500, 1000)$	$[0.08, 0.20)$
3	$[q_{50}, q_{75})$	$[q_{50}, q_{75})$	$[1000, 2000)$	$[0.20, 0.45)$
4	$\geq q_{75}$	$\geq q_{75}$	≥ 2000	≥ 0.45

q_{25}, q_{50}, q_{75} denote the 25th, 50th, and 75th empirical percentiles of the non-zero dose distributions for IV fluids and vasopressors, computed separately.

Approximate Information State. To address the partial observability in patient trajectories, we learned a compact latent state representation using a recurrent neural network, using the approximate information state (AIS) in Subramanian et al. (2021) and Killian et al. (2020). Specifically, we trained a gated recurrent unit (GRU) encoder (Cho et al., 2014) that, at each time t, maps the concatenated 33-dimensional observation vector, 5-dimensional demographic context, and the action a_{t-1} taken at time t, to a D-dimensional latent state z_t. The GRU encoder was optimized via a dual-head objective: one decoder head reconstructs the current observation vector, while another head predicts the next observation x_{t+1} given the current latent state and action a_t in the form of a parameterized distribution $P(x_{t+1}|z_t, a_t)$. We trained the representation model on the training set trajectories, monitoring the negative log-likelihood (NLL) of the reconstructions/predictions on the validation set. For each time-step size, we ran an identical grid search over 5 different latent dimensions and 6 different learning rates (see Appendix A.4). The checkpoint with lowest validation NLL was selected to extract the latent states at each time step. We treat the D-dimensional latent state z_t as the AIS summarizing the patient's history up to time t in a Markovian fashion.

Behavior Cloning. We learned an estimated behavior policy $\hat{\pi}_B$ to mimic the clinicians' treatment decisions, which we use for OPE. The model takes the patient's state representation s_t as input and predicts $\hat{\pi}_B(a|s)$, a probability distribution of actions that clinicians would take. We implemented and compared two behavior cloning approaches: a k-nearest-neighbors (kNN) classifier with $k = 100$ (Raghu et al., 2018) and a 3-layer feed-forward neural network. Both models were trained on the training set to classify the clinician's chosen action at each Δt. We evaluated their predictive performance using micro-average area under the receiver operating characteristic curve (AUROC) on the validation set, and selected the better-performing model as our clinician policy estimate.

Batch-Constrained Q-learning. We adopted the Batch-Constrained Q-learning (BCQ) algorithm (Fujimoto et al., 2019) for offline policy optimization. In our implementation, the agent's state input was the AIS latent z_t described above. We defined a sparse reward signal reflecting patient outcomes, similar to Tang et al. (2020): the episode terminal reward was $+100$ if the patient survived to hospital discharge (or was alive at 52 h post-onset) and 0 otherwise. The BCQ implementation used in our experiments employs a single two-layer feed-forward Q-network. We also trained a separate behavior-cloning head within BCQ that proposed actions constrained to the behavior dataset's support. To guard against out-of-distribution actions, we performed offline filtering: at each decision point, any action whose estimated behavior probability $\hat{\pi}_B(a|s)$ fell below a threshold ε was disallowed. We trained the BCQ agent on the batch of training trajectories for a fixed number of epochs, using five different random seeds and eight values of ε (see Appendix A.4), and selected the final policy which we denote π_{eval}.

Off-policy Evaluation. We evaluated the performance of the learned policy using off-policy evaluation (OPE), specifically weighted importance sampling (WIS). The WIS estimator used importance weights to re-weight the returns of test trajectories under the assumption that test data were generated by the behavior policy π_B; using our learned $\hat{\pi}_B$ model, we computed per-step importance ratios $\rho_t = \frac{\pi_{\text{eval}}(a_t|s_t)}{\hat{\pi}_B(a_t|s_t)}$ for each action a_t taken by clinicians, and then took a weighted average of the observed returns and normalizing by the sum of the importance weights across all evaluation trajectories (Liu & Brunskill, 2022). To control the estimator's variance, we truncated the cumulative importance ratios $W = \prod_{t=1}^{H} \rho_t$ at a maximum of $W \leq 10^3$ (Ionides, 2008). For each policy, we estimated the expected return and its standard error via bootstrap resampling (1000 bootstrap samples from the test set trajectories). We also recorded the effective sample size (ESS) of the WIS estimator (Elvira et al., 2022), which reflects how many trajectories contribute meaningfully after weighting. In Section 4.4, we report the WIS estimated performance for each policy along with the ESS. All results are reported separately for the different discretization experiments. To complement the quantitative metrics, we also use heat maps to visualize how the learned BCQ policies redistribute treatment probabilities relative to clinicians.

4 Results

We applied our experimental pipeline to **four time-step sizes** ($\Delta t = 1, 2, 4, 8$ h) under **two action-space designs**—QUANTILE-5 and CLINICAL-THRESHOLD. For every Δt / action-space pair we report the following: (i) cohort size and episode length, (ii) BC performance via AUROC, (iii) AIS reconstruction error across latent dimensions, and (iv) performance of the policy learned by BCQ, measured by WIS and action frequency heatmaps. These results allow us to isolate the individual and joint effects of time-step size and action discretization on every stage of the learning pipeline.

4.1 Cohort Statistics

Table 2 reports the number of ICU admissions and time-step counts for each time-step size Δt. Across all time-step sizes, the cohort size remained around 19 000 admissions but dropped slightly at larger Δt because stays shorter than one step were excluded. The total number of time steps decreased by a factor of approximately 0.53 each time Δt doubles (rather than the ideal 0.5). This is because we retained any partial step at the end of each trajectory (i.e. we used $\lceil L/\Delta t \rceil$ rather than $\lfloor L/\Delta t \rfloor$), so even a small remainder became a full extra step. Minor cohort drift (a few trajectories drop out at coarser time-step sizes) and slight changes in missing-data filtering further nudged the

ratio up to ≈ 0.53. The bottom two rows show the average steps per admission and the step count as a percentage of the 1 h case. Note that these small cohort mismatches across time-step sizes complicate direct comparisons, while our analysis proceeds using the available cohorts as extracted.

Table 2: Cohort size and time-step counts for different time-step sizes Δt, plus average steps per ICU admission and scaling relative to the 1 h case.

	1 h	2 h	4 h	8 h
Number of ICU Admissions	18 995	18 987	18 906	18 783
Number of Time Steps	889 227	468 984	247 713	132 038
Avg. Steps per Admission	46.8	24.7	13.1	7.0
Steps (% of 1 h)	100%	52.7%	27.9%	14.8%

4.2 State Representation Models

Following Section 3.2, we first evaluated the quality of our AIS encoder across four time-step sizes, $\Delta t \in \{1, 2, 4, 8\}$ h, and two action-space discretizations (QUANTILE-5 vs. CLINICAL-THRESHOLD). Table 3 reports both the chosen latent size and its resulting minimum validation mean square error (MSE). QUANTILE-5 required a reduced latent dimension (64) at coarser time-step sizes (4 h and 8 h) to stabilize training, whereas the fixed-threshold encoder consistently used 128 dimensions. Across all time-step sizes, the AIS encoder reaches virtually identical validation performance for both discretization schemes (Δ MSE < 0.001), indicating that any policy-level differences we observe subsequently are not driven by differences in representation quality. We also observe that validation MSE tends to increase as Δt grows. This is likely because the AIS encoder is trained to predict future observations with a prediction horizon of Δt (e.g., the average heart rate over the next step), so forecasting 1 h ahead is inherently easier than 8 h ahead.

Table 3: AIS encoder results across time-step sizes: selected latent dimension and corresponding minimum validation MSE with 95 % bootstrap confidence intervals from 1000 bootstrap samples.

Δt (h)	QUANTILE-5		CLINICAL-THRESHOLD	
	latent dim	MSE [95% CI]	latent dim	MSE [95% CI]
1	128	0.2456 [0.2293, 0.2638]	128	0.2454 [0.2286, 0.2655]
2	128	0.3073 [0.2823, 0.3395]	128	0.3074 [0.2825, 0.3397]
4	64	0.4676 [0.3961, 0.5852]	128	0.4669 [0.3944, 0.5854]
8	64	0.4340 [0.4239, 0.4512]	128	0.4348 [0.4250, 0.4521]

4.3 Behavior-Cloning Models

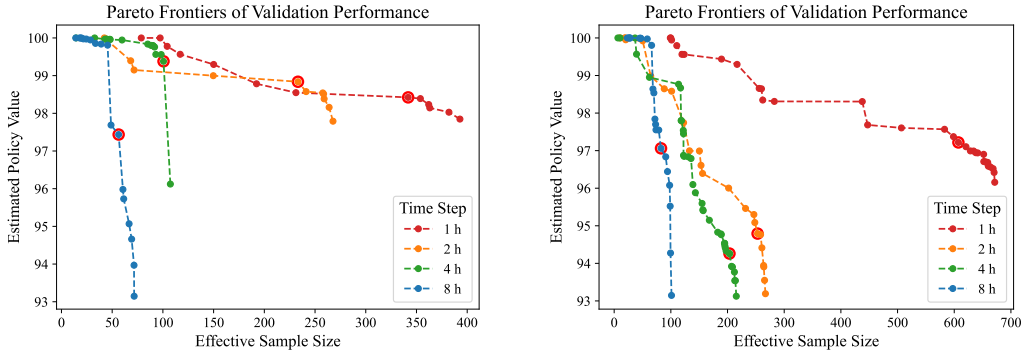
To ensure reliable off-policy evaluation via weighted-importance sampling (WIS), we trained behavior-cloning (BC) models that estimate clinician action probabilities. Table 4 shows micro-AUROC of BC across four time steps ($\Delta t = 1, 2, 4, 8$ h) for both QUANTILE-5 and CLINICAL-THRESHOLD action spaces. In every case we observed AUROC is close to or above 0.90, satisfying common WIS-variance guidelines (Jeong et al., 2024). CLINICAL-THRESHOLD consistently outperforms QUANTILE-5. While performance decreases slightly as Δt increases (due to lost temporal detail), the performance gap remains stable. Overall, the high BC fidelity confirms that our models supply accurate, well-calibrated behavior probabilities, ensuring that downstream WIS estimates will be unbiased and low-variance.

4.4 Policy Performance

Section 4.3 shows the ESS-WIS Pareto frontiers obtained during validation. Rather than maximizing WIS alone, we chose the checkpoint on each frontier that met a practical trade-off: WIS had to be

Table 4: BC performance (micro-AUROC) across time-step sizes with 95% confidence intervals estimated from 1000 bootstrap samples.

Action space	Δt (h)	AUROC [95% CI]
QUANTILE-5	1	0.8975 [0.8965, 0.8984]
	2	0.9063 [0.9049, 0.9076]
	4	0.9074 [0.9057, 0.9091]
	8	0.9007 [0.8983, 0.9031]
CLINICAL-THRESHOLD	1	0.9572 [0.9565, 0.9579]
	2	0.9533 [0.9523, 0.9544]
	4	0.9372 [0.9356, 0.9389]
	8	0.9140 [0.9115, 0.9165]



(a) QUANTILE-5 action space

(b) CLINICAL-THRESHOLD action space

Figure 1: Pareto frontiers of validation WIS versus ESS for each time step Δt. Dashed lines trace the non-dominated points; hollow markers denote the model selected for test-time evaluation.

near frontier-optimal and ESS could not fall below the stability threshold used in tuning ($\sim$40 samples). This guards against high-value but high-variance models. The performance of the resulting checkpoints are reported in Table 5.

Action-space differences. The QUANTILE-5 discretization achieves the highest test WIS for every Δt but at the cost of much smaller ESS, most pronounced at 8 h (43 vs. 83). This suggests that QUANTILE-5 estimates, while numerically superior, may be less reliable than their CLINICAL-THRESHOLD counterparts.

Time-step effect. Finer time-step sizes (1–2 h) consistently dominate coarser ones in WIS without a drastic ESS penalty. The 1 h QUANTILE-5 policy, for example, exceeds the clinician baseline by 5.3 ± 0.6 WIS points while still retaining 331 ± 16 effective samples. In contrast, at 8 h both BCQ variants collapse onto near-zero vasopressor doses (Section 5), implying a conservative strategy that may limit potential gains.

ESS trends. In principle, a finer time-step size (smaller Δt) yields more decision points, giving the evaluation policy more chances to diverge from the behavior policy and thereby inflating the variance of importance weights—so one would expect ESS to shrink as Δt becomes smaller. Empirically we observe the opposite: ESS is highest at $\Delta t = 1$–2 h and lowest at 8 h.

The explanation lies in our variance-control scheme (see Section 3.2): we truncated each trajectory weight W at a fixed threshold $W_{\max} = 10^3$. Because smaller Δt yields larger H, more trajectories now meet the ceiling, as confirmed by the right–hand mass in Appendix A.3. Clipping a large fraction of weights at the same value compresses the heavy tail and reduces the variance, thereby inflating the ESS. In other words, the observed improvement of ESS is actually an artifact of time-step size independent clipping of importance ratios, rather than evidence of improved OPE.

Table 5: Test-set WIS value and ESS for BCQ and clinician policies across Δt.

Δt (h)	Policy	Threshold ε	$\widehat{V}_{\text{test}}$ (WIS)	ESS_{test}
1	Quantile-5	0.50	99.07 ± 0.48	331 ± 16
	Clinical-threshold	0.30	95.14 ± 0.79	642 ± 20
	Observed π_b	–	93.81 ± 0.43	2795
2	Quantile-5	0.9999	97.11 ± 1.25	229 ± 13
	Clinical-threshold	0.50	92.55 ± 2.04	222 ± 14
	Observed π_b	–	93.79 ± 0.41	2785
4	Quantile-5	0.50	96.56 ± 2.03	94 ± 9
	Clinical-threshold	0.9999	94.50 ± 1.44	210 ± 12
	Observed π_b	–	93.87 ± 0.58	2791
8	Quantile-5	0.00	98.27 ± 1.20	43 ± 5
	Clinical-threshold	0.30	95.17 ± 2.60	83 ± 7
	Observed π_b	–	94.03 ± 0.51	2764

To obtain a fair comparison across resolutions, in future work we will replace the fixed ceiling with an adaptive rule, e.g. truncating at the 95^{th} percentile of $\{W_i\}$ for each Δt.

5 Discussion & Conclusion

What we contribute. Prior RL-for-sepsis studies almost universally employ a fixed 4 h time grid. We present, to our knowledge, the first systematic comparison of four time-step sizes (1, 2, 4, 8 h) within a single, controlled pipeline: identical cohort-extraction code, the same AIS encoder architecture, identical BC and BCQ hyper-parameter grids, and a common WIS evaluator. Our analysis covers cohort drift, representation learning, BC calibration, policy quality (WIS+ESS), and action redistribution, thereby isolating the specific impact of time discretization and action-space design.

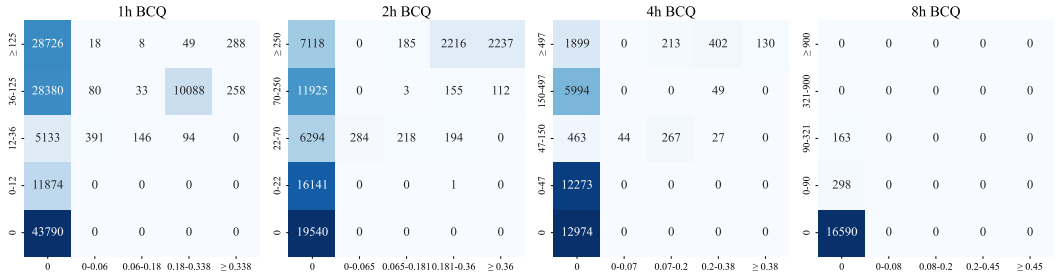
Why a fair comparison is hard. Changing Δt inevitably alters (i) the cohort (short stays drop out at coarser time steps), (ii) the number of decision steps per episode, (iii) the dose that each discrete action represents, and (iv) the evaluation horizon. Hence policies trained at different time-step sizes cannot be judged on exactly the same trajectories or action supports. Our strategy is to keep every component other than Δt fixed, then interpret results jointly through WIS and ESS so that variance differences are explicit.

Action-space design remains open. Equal-frequency (QUANTILE-5) bins align with the data distribution but lack clinical meaning; clinically chosen cut-offs (CLINICAL-THRESHOLD) map directly to practice yet were devised for 4 h time steps. Simply scaling doses by time-step sizes (e.g. halving thresholds when moving from 4 h to 2 h) could retain semantics, but requires careful validation—a promising direction for future work.

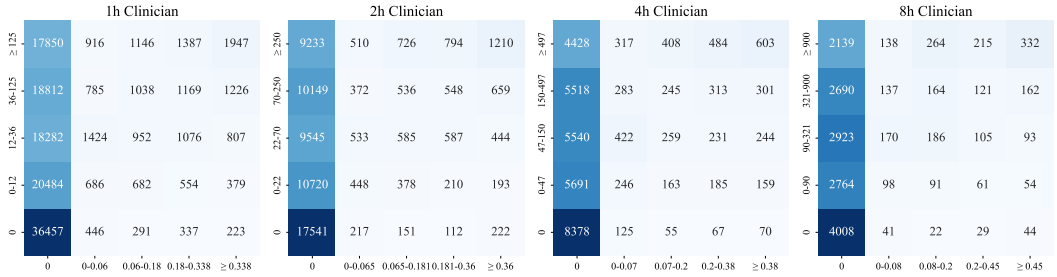
Cross-granularity evaluation. Evaluating a 4 h policy on 2 h data (or vice-versa) would demand hierarchical decision models or marginalizing over unobserved mid-step. We leave such multi-rate OPE as an important extension.

Key empirical insight. Across both action spaces, finer grids (1–2 h) deliver higher WIS without an excessive loss of ESS, supporting calls to move below the conventional 4 h time step. Nevertheless, ever-smaller steps are not automatically better: clinical events rarely unfold minute-by-minute, and excessively fine time-step sizes would inflate horizon length, variance, and computational cost. Finally, our evaluation still relies on a fixed threshold truncation of trajectory weights, and developing an adaptive truncation rule remains an open problem for fair comparison across time-step sizes.

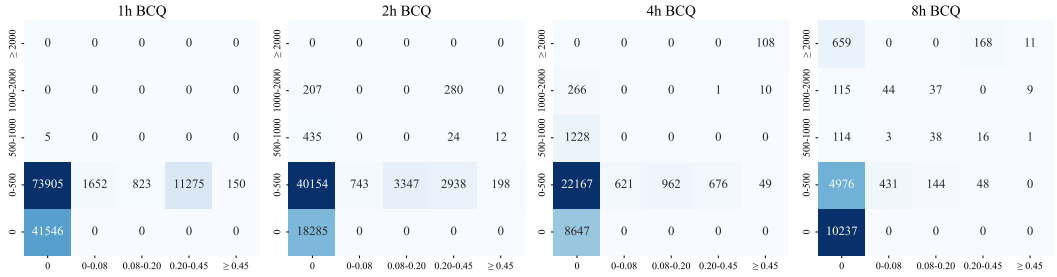
Take-home message. Time-step size is a critical design knob in the RL-for-sepsis task. Our results advocate using finer time-step sizes in sepsis management, which outperform the standard 4 h grid. These results call for principled choices of every component—time-step size, action-space discretization, and OPE—rather than relying on inherited defaults.



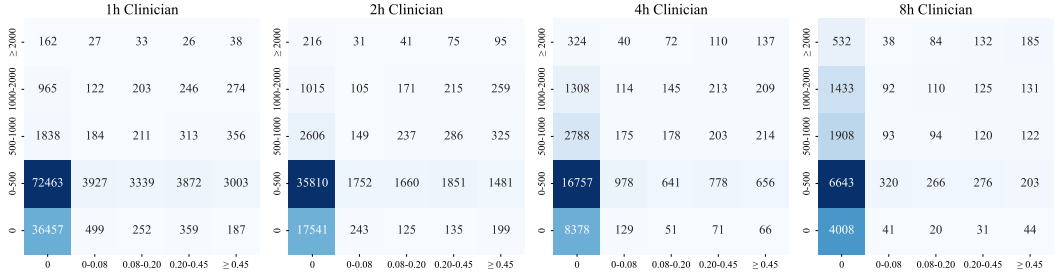
(a) BCQ, QUANTILE-5



(b) Observed clinicians, QUANTILE-5



(c) BCQ, CLINICAL-THRESHOLD



(d) Observed clinicians, CLINICAL-THRESHOLD

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

References

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.

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.

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

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.

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

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.

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.

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.

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.

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.

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.

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.

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.

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

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.

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.

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.

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.

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

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

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.

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 (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9387–9396. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/tang20c.html.

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.

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

A Appendix

A.1 Time-Step Size Selections in RL Research on Sepsis Care

Table 6: RL studies for sepsis care, summarizing time-step choices and key design aspects.

Paper	Δt	Algorithm	Dataset	Cohort	Notes
Raghu et al. (2017)	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. (2018)	4 h	Batch Q-learning	MIMIC-III (+eRI*)	17.1k	AI Clinician; 750 states, 25 actions; external validation.
Jeter et al. (2019)	4 h	Reproduction study	MIMIC-III	5.4k	Finds no-action policy often rivals AI Clinician; urges caution.
Yu et al. (2019)	1 h	Deep IRL	MIMIC-III	14.0k	Learns reward; highlights mortality factors (e.g. PaO_2).
Tang et al. (2020)	4 h	Set-valued DQN	MIMIC-III	20.9k	Returns top-k near-optimal dose sets for clinician choice.
Killian et al. (2020)	4 h	Offline DQN	MIMIC-III	17.9k	Sequential latent encodings outperform raw features.
Lu et al. (2020)	1–4 h	Dueling DDQN	MIMIC-III	17k+	Sensitivity study on features, reward, time discretization.
Fatemi et al. (2021)	4 h	Dead-end discovery	MIMIC-III	17k+	Identifies high-risk states; secures policy to avoid them.
Satija et al. (2021)	4 h	MO-SPIBB	MIMIC-III	17k+	Safe policy improvement under performance constraints.
Ji et al. (2021)	4 h	Trajectory inspection	MIMIC-III	17k+	Clinician "what-if" review reveals policy flaws; validation tool.
Liang et al. (2023)	4 h	Episodic-memory DQN	MIMIC-III	17.9k	Memory module boosts sample efficiency, lowers est. mortality.
Choudhary et al. (2024)	4 h	Tabular MDP	MIMIC-III	∼18k	ICU-Sepsis benchmark: 715 states, 25 actions.
Tu et al. (2025)	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 7: 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 Histograms of Trajectory Weights

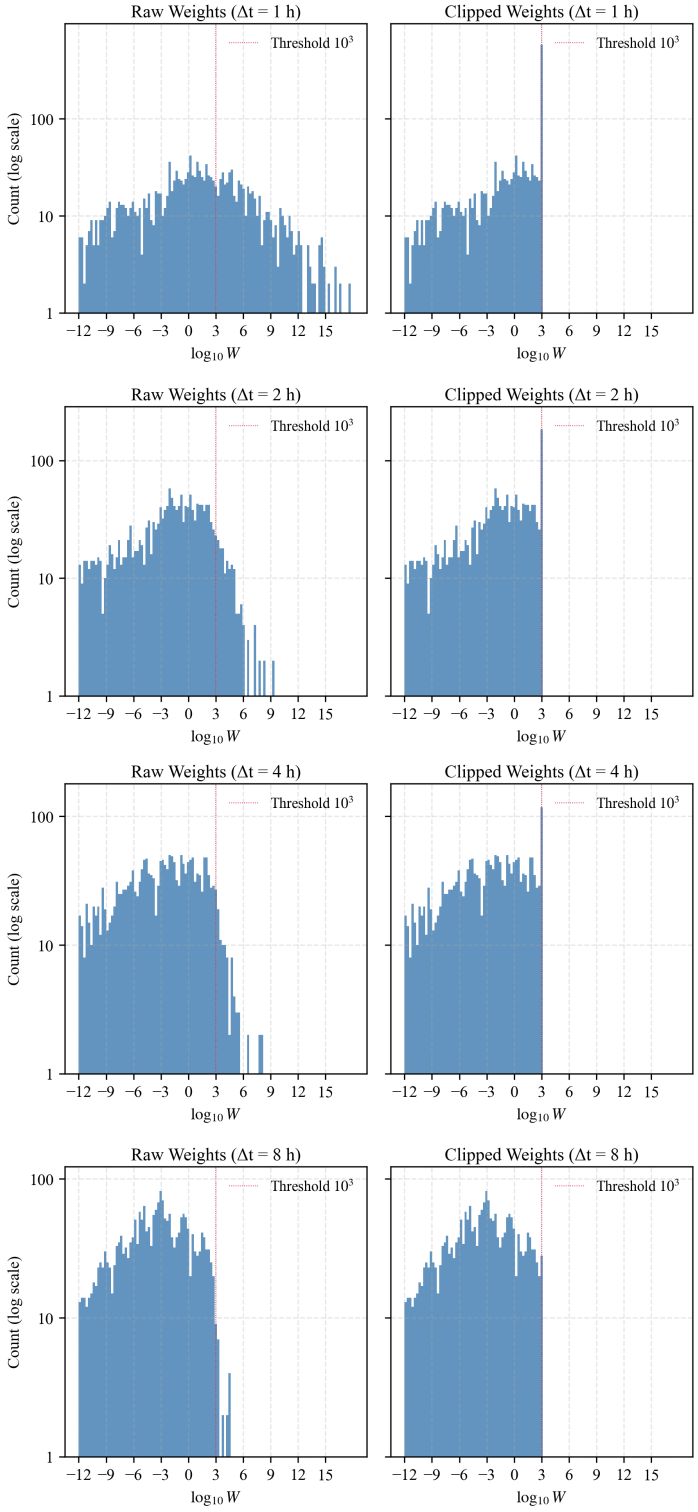


Figure 3: Log–histograms of clipped importance–sampling trajectory weights for different time–step sizes with threshold $\varepsilon = 0.5, iteration = 10000$. All panels share the same y-axis (log–scale) and clipping threshold $W_{\max} = 10^3$.

A.4 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	{1e-5, 3e-5, 1e-4, 3e-4, 1e-3, 5e-4}
BCQ (with 5 random restarts):	
– Threshold, ε	$\{0, 0.01, 0.05, 0.1, 0.3, 0.5, 0.75, 0.999\}$
– Learning rate	3e-4
– Weight decay	1e-3
– Hidden layer size	256