Benchmarking Reinforcement Learning Algorithms for ICU Ventilator Settings: An Interpretable and Probabilistic Patient Environment for Doctor Agents

Ya-Hsi Chang[1], **Po-Chih Kuo**[1].

[1]Department of Computer Science, National Tsing Hua University, Hsinchu, Taiwan
kuopc@cs.nthu.edu.tw

Abstract

This supplementary material complements the main paper by providing additional details and results to support the benchmarking of RL algorithms for ICU ventilator settings. It includes cohort descriptions for the MIMIC-IV and eICU datasets, preprocessing rules, hyperparameter settings for RL agents, training curves, visualizations of agent behaviors and workflows, and detailed reward design comparisons across multiple algorithms (BC, NFQ, DQN, DDQN, SAC, BCQ, CQL). These elements aim to enhance reproducibility and offer deeper insights into the experimental setup and findings.

Data

Cohort Selection Flow

We selected adult ICU stays ($\geq$ 20 years old) with at least 24 hours of invasive MV (excluding cases where ventilation was initiated for surgical procedures), excluding DNR/DNI cases. From each dataset, we extracted patient demographics (age, gender) and hourly measurements of relevant clinical variables. Figure S1 presents the workflow used for selecting the study cohort and relevant variables.

Preprocessing Aggregation Rules

- HR: Prefer abnormal values (<60 or >130 bpm); else median.
- RR: Prefer abnormal values (<12 or >30 bpm); else median.
- SpO_2: Select minimum nonzero value.
- FiO_2, RR_{set}: Select maximum observed value.

Variable Discretization Rules

To ensure interpretability and alignment with clinical practice, all continuous physiological and ventilator variables were discretized into clinically meaningful bins before model training. The bin thresholds were based on commonly used early warning systems and ventilator management guidelines. The discretization scheme is summarized below:

- HR: [0, 40, 50, 60, 70, 80, 90, 100, 110, 120, 130, 180]
- RR: [0, 4, 8, 11, 14, 17, 20, 24, 27, 30, 33, 40]
- SpO_2: [0, 80, 90, 91, 93, 95, 98, 100]
- FiO_2: [21, 30, 35, 40, 50, 60, 70, 80, 90, 100]
- RR_{set}: [0, 4, 8, 11, 14, 17, 20, 24, 27, 30, 33, 40]

Characteristic	**MIMIC-IV**	**eICU**
Total Stays	10,633	3,877
Race Distribution		
Asian	291 (2.74%)	45 (1.16%)
Black	1,178 (11.08%)	414 (10.68%)
Hispanic	403 (3.79%)	256 (6.60%)
Others	1,957 (18.40%)	275 (7.09%)
White	6,804 (63.99%)	2,887 (74.46%)
Gender Distribution		
Female	4,441 (41.77%)	1,639 (42.27%)
Male	6,192 (58.23%)	2,237 (57.70%)
Age (Mean ± SD)	64.05 ± 16.15	59.78 ± 15.78
Sepsis	8,949 (84.16%)	1,588 (40.96%)
ARDS	328 (3.08%)	370 (9.54%)

Table S1: Cohort Description for MIMIC-IV and eICU.

Each interval represents a categorical bin used to encode the corresponding continuous measurement. For example, a heart rate of 95 bpm falls into the $[90, 100)$ bin. The same thresholds were used consistently across MIMIC-IV and eICU datasets to preserve clinical interpretability and ensure comparability between datasets.

Experiments

For the Patient Environment development and RL agent training, we use fixed seeds—`random.seed(42)` and `np.random.seed(42)`—to ensure reproducibility.

RL Algorithms

The following algorithms were evaluated in this study:

- **BC (Behavior Cloning)**: A supervised learning method that mimics expert demonstrations to learn a policy.

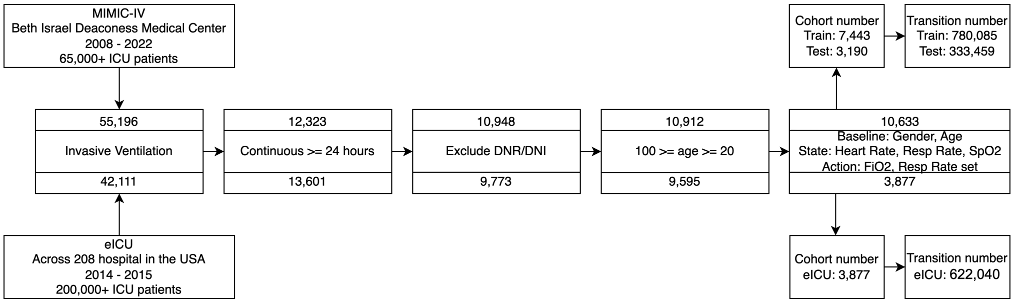

Figure S1: Cohort and Variable Selection Workflow.

- **NFQ (Neural Fitted Q-iteration)**: An off-policy approach using neural networks to iteratively fit the Q-function.
- **DQN (Deep Q-Network)**: A value-based method employing deep learning to approximate the Q-function for discrete actions.
- **DDQN (Double Deep Q-Network)**: An enhancement of DQN that reduces overestimation bias in Q-value updates.
- **SAC (Soft Actor-Critic)**: An off-policy actor-critic algorithm that optimizes both reward and policy entropy.
- **BCQ (Batch-Constrained Q-learning)**: A batch RL method that constrains the policy to align with the behavior policy in offline settings.
- **CQL (Conservative Q-Learning)**: An offline RL technique that learns a conservative Q-function estimate to mitigate overestimation.

RL Hyperparameter Settings

The hyperparameters in Table S2 were selected based on preliminary experiments and established practices in discrete offline RL. For instance, the learning rate of 1×10^{-5} ensures stable convergence (except Behavior Cloning), while the batch size of 1,024 balances computational efficiency and training performance. The use of batch normalization (`use_batch_norm = True`) and a dropout rate of 0.5 helps mitigate overfitting given the complexity of ICU data.

We experimented with a range of hyperparameters including `batch_size` $\in$ {32, 64, 128, 256, 512, 1024, 2048}, `learning_rate` $\in$ {3e-4, 1e-5, 1e-6}, `gamma` $\in$ $\{0.75, 0.98, 0.99\}$, `use_batch_norm` $\in$ {True, False}, and `dropout_rate` $\in$ {0, 0.5}. We identified a configuration under which all methods consistently converge on the training set.

MMD illustrates

Figure S2 illustrates that for each observed (O, A), the simulator's predicted next-state distribution to the empirical real distribution.

Computing Infrastructure

For related works, ehrMGAN was run on a machine with an NVIDIA GeForce RTX 3060 GPU and 24 GB RAM. All other experiments were conducted on an Apple M3 Mac with 16 GB RAM. We used Python 3.9 and d3rlpy version 2.8.1 for reinforcement learning implementation. The operating system for the M3 Mac experiments was macOS Sonoma 14.6.

Hyperparameter	Value
n_steps	20,000
target_update_interval	1,000
learning_rate	1×10^{-5}
gamma	0.98
batch_size	1,024
use_batch_norm	True
dropout_rate	0.5

Table S2: Offline RL Hyperparameter Settings in d3rlpy.

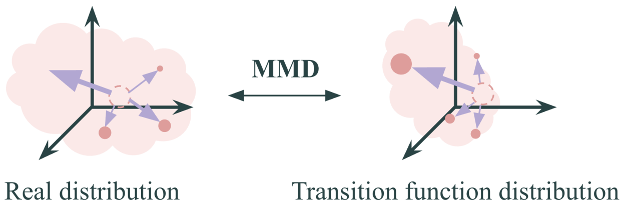

Figure S2: MMD for Empirical Transition vs. Simulated Distributions.

Experimental Results

Effect of K in Action-Based KNN

Figure S3 demonstrates how the choice of K in Action-Based KNN affects MMD. While a small K better preserves fidelity for seen (O, A) pairs, larger K improves generalization on unseen pairs. We select $K = 100$ as a compromise.

Transition Level MMD Evaluation

Table S3 show the MMD score of transition level evalution across different method and different initial categories.

To evaluate whether the Action-based KNN method significantly improves upon the baseline, we conducted a Wilcoxon signed-rank test on the paired MMD values. The Wilcoxon signed-rank test is a non-parametric statistical test

Method	Internal (Training)	Internal (Seen Keys)	Internal (Unseen Keys)	External (Seen Keys)	External (Unseen Keys)
Keep Current State	0.1655	0.1601	0.1639	0.1653	0.1980
Random Walk	0.2478	0.2460	0.2462	0.2555	0.2323
Bayesian	0.3347	0.2474	0.3567	0.2600	0.3706
Action-Based KNN	**0.0351**	**0.0411**	**0.0585**	**0.0561**	0.0835
GRU	0.0589	0.0485	**0.0585**	0.0688	**0.0814**
Transformer	0.0603	0.0458	0.0588	0.0568	0.0847

Table S3: Transition-Level MMD for Internal and External Evaluation across Seen and Unseen Keys. Lower is better.

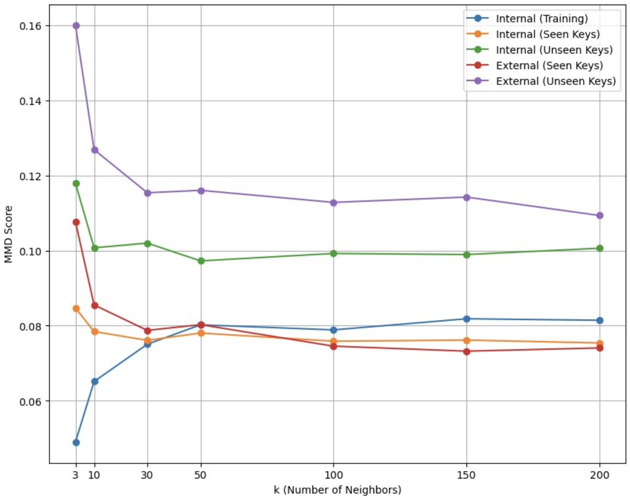

Figure S3: MMD for Varying K in KNN.

suitable for small sample sizes and paired data, as it does not require the assumption of normality. In this case, it assesses whether the MMD values for action based KNN are significantly lower than those for Keep Current State.

The test yielded a test statistic of 15.0 and a one-tailed p-value of 0.03125. Since $p = 0.03125 < 0.05$, we reject the null hypothesis and conclude that the action based KNN method significantly improves the transition-level MMD compared to the Keep Current State baseline.

RL Benchmark Improvement

As shown in Tables S25, S26, and S27, we report results for high-severity patients across the three datasets. The improvements are statistically significant for all models, except SAC on the OOD test set dataset.

Training and Anomaly Curves

As shown in Figure S4, most RL algorithms exhibit increasing rewards with training epochs. However, the performance of BC (blue) declines after few epoch, likely due to overfitting. When trained with a lower learning rate (1e-6), as shown in Figure S5, BC performance improves. Notably, most RL agents outperform the physician policy baseline. An exception is SAC (purple) trained on the OOD test set, which performs poorly across patient groups of varying severity.

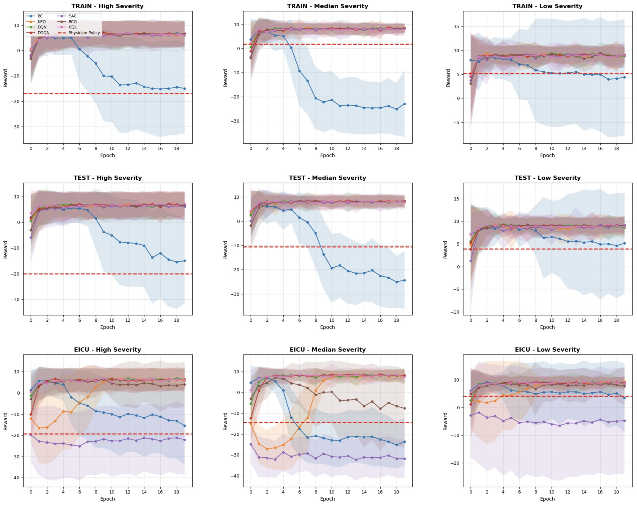

Figure S4: Training Curve (Learning Rate $1e^{-5}$).

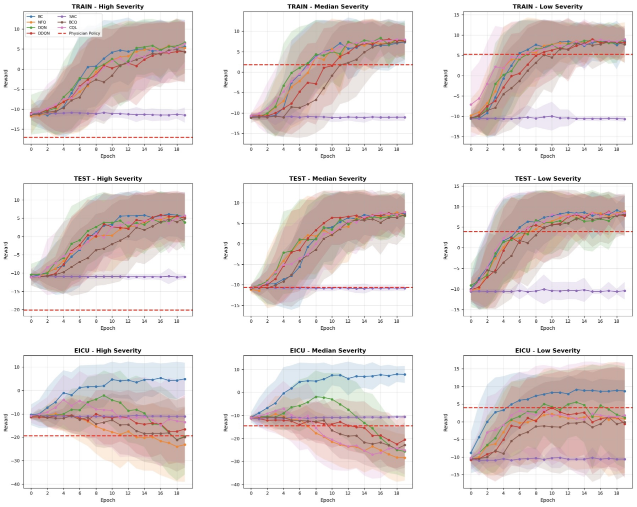

Figure S5: Training Curve (Learning Rate $1e^{-6}$).

Agent Demonstrations

As shown in Figure S6, the topmost trajectory under the Physician Policy exhibits behavior similar to that of the BC agent, where FiO_2 (red) is maintained at a high level (90%–100%) to ensure stable SpO_2 before being gradually reduced—reflecting a relatively conservative strategy. In contrast, other RL-based agents adopt a more proactive approach: after observing a clear improvement in SpO_2 (rising to around 93–95%) within the first two hours, they reduce FiO_2 more quickly, with most meeting the extubation criteria within 15 hours.

Effect of Reward Design Variants

Tables S4–S24 present the effects of different reward design variants across algorithms. This includes the default setting, as well as ablations without the Action-Stability penalty (NoActPen), without intermediate rewards (NoIntermRew), with increased extubation reward (HighExtubRew), and without the extubation reward (NoExtubRew). Bold font indicates the best performance within the same severity level (high, medium, or low) across different reward design settings.

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	5.7	7.83	8.89	6.13	6.74	8.74	6.3	7.25	**9.22**	6.33	7.24	8.95	**6.49**	**8.01**	8.86
Extubation Meet Rate (%)	93.0	**98.0**	99.0	93.0	95.0	99.0	95.0	95.0	**100.0**	**97.0**	96.0	**100.0**	96.0	**98.0**	99.0
Avg. Trajectory Length (hrs)	20.03	15.9	**13.08**	**17.35**	17.94	14.68	19.36	16.59	13.47	22.95	18.09	14.77	20.41	**15.88**	14.03
Avg. Time to Meet (hrs)	20.75	**15.76**	**13.04**	**17.46**	18.19	14.79	19.99	16.92	13.47	23.36	17.95	14.77	20.92	15.99	14.07
Action Diversity	**33**	25	**19**	29	**31**	14	22	24	14	27	23	16	26	20	17
Anomalous Actions (%)	7.0	**2.0**	1.0	7.0	5.0	1.0	5.0	5.0	**0.0**	**3.0**	4.0	**0.0**	4.0	**2.0**	1.0

Table S4: Reward Design Comparison Across All Severity Categories for Training Set (BC).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.87	8.58	9.29	6.79	8.57	9.16	6.69	8.51	**9.44**	**6.94**	**8.63**	9.19	6.87	8.46	9.26
Extubation Meet Rate (%)	95.0	**100.0**	**100.0**	**97.0**	**100.0**	**100.0**	94.0	**100.0**	**100.0**	95.0	**100.0**	**100.0**	94.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	17.41	15.02	13.13	18.54	15.05	13.66	**17.13**	15.41	**12.26**	17.34	**14.58**	13.01	17.33	16.32	12.98
Avg. Time to Meet (hrs)	18.12	15.02	13.13	18.99	15.05	13.66	**17.97**	15.41	**12.26**	18.04	**14.58**	13.01	18.18	16.44	12.98
Action Diversity	13	11	13	15	**12**	**15**	14	11	13	11	9	12	**19**	10	**15**
Anomalous Actions (%)	5.0	**0.0**	**0.0**	**3.0**	**0.0**	**0.0**	6.0	**0.0**	**0.0**	5.0	**0.0**	**0.0**	6.0	1.0	**0.0**

Table S5: Reward Design Comparison Across All Severity Categories for Training Set (NFQ).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.82	8.48	**9.4**	**6.87**	8.36	9.27	6.58	8.28	9.22	6.84	**8.51**	9.1	6.72	8.38	9.33
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	95.0	**100.0**	**100.0**	94.0	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	94.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	**15.99**	**15.28**	12.58	16.48	15.35	**12.4**	17.24	15.83	13.26	17.52	15.71	13.99	18.67	17.35	13.5
Avg. Time to Meet (hrs)	**16.73**	**15.28**	12.58	17.13	15.35	**12.4**	18.09	15.83	13.26	18.08	15.71	13.99	19.61	17.48	13.5
Action Diversity	17	**11**	12	15	9	11	18	10	13	18	**11**	12	**24**	**11**	**17**
Anomalous Actions (%)	6.0	**0.0**	**0.0**	5.0	**0.0**	**0.0**	6.0	**0.0**	**0.0**	**4.0**	**0.0**	**0.0**	6.0	1.0	**0.0**

Table S6: Reward Design Comparison Across All Severity Categories for Training Set (DQN).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	**7.1**	**8.84**	9.22	6.64	8.49	9.21	6.93	8.71	9.26	7.08	8.42	**9.29**	6.61	8.66	9.12
Extubation Meet Rate (%)	95.0	**100.0**	**100.0**	93.0	**100.0**	**100.0**	95.0	**100.0**	**100.0**	**96.0**	**100.0**	**100.0**	94.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	16.86	**14.13**	13.26	**15.79**	15.06	13.3	16.89	14.2	**12.7**	16.64	15.26	13.24	18.76	17.59	14.6
Avg. Time to Meet (hrs)	17.54	**14.13**	13.26	**16.68**	15.06	13.3	17.57	14.2	**12.7**	17.17	15.26	13.24	19.7	17.73	14.6
Action Diversity	12	7	13	14	8	12	17	11	14	18	7	**15**	**19**	**12**	**15**
Anomalous Actions (%)	5.0	**0.0**	**0.0**	7.0	**0.0**	**0.0**	5.0	**0.0**	**0.0**	**4.0**	**0.0**	**0.0**	6.0	1.0	**0.0**

Table S7: Reward Design Comparison Across All Severity Categories for Training Set (DDQN).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	7.05	8.18	9.25	**7.15**	**8.46**	9.24	6.89	8.26	9.12	6.6	8.24	**9.31**	6.37	8.38	9.27
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	94.0	**100.0**	**100.0**	93.0	**100.0**	**100.0**
Avg. Trajectory Length (hrs)	**14.9**	17.06	**12.33**	15.06	**14.89**	13.35	17.06	15.61	13.84	16.25	16.25	12.8	17.6	15.6	13.26
Avg. Time to Meet (hrs)	15.6	17.06	**12.33**	**15.59**	**14.89**	13.35	17.75	15.61	13.84	17.03	16.25	12.8	18.62	15.6	13.26
Action Diversity	15	9	10	**23**	**12**	**13**	11	9	11	19	**12**	12	19	10	**13**
Anomalous Actions (%)	6.0	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	6.0	**0.0**	**0.0**	7.0	**0.0**	**0.0**

Table S8: Reward Design Comparison Across All Severity Categories for Training Set (SAC).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.64	8.55	**9.44**	6.59	**8.67**	9.15	6.68	8.61	9.06	**7.01**	8.42	9.27	6.99	8.48	9.24
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	93.0	**100.0**	**100.0**	92.0	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	99.0	**100.0**
Avg. Trajectory Length (hrs)	16.72	14.5	**12.39**	16.11	**14.27**	12.79	15.58	15.52	13.6	**15.41**	16.43	13.51	19.8	16.5	13.23
Avg. Time to Meet (hrs)	17.53	14.5	**12.39**	17.0	**14.27**	12.79	16.57	15.52	13.6	**15.99**	16.43	13.51	20.63	16.63	13.23
Action Diversity	16	9	13	14	10	13	15	**14**	**17**	17	10	14	**30**	**14**	14
Anomalous Actions (%)	6.0	**0.0**	**0.0**	7.0	**0.0**	**0.0**	8.0	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	**5.0**	1.0	**0.0**

Table S9: Reward Design Comparison Across All Severity Categories for Training Set (BCQ).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.87	8.37	9.11	6.71	8.57	9.24	**7.23**	**8.61**	9.18	6.86	8.54	**9.34**	6.35	8.54	9.16
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	93.0	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	94.0	**100.0**	**100.0**	93.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	**15.88**	16.29	13.76	15.96	15.98	13.1	15.89	15.69	13.29	16.18	**15.59**	**12.72**	17.9	16.93	13.65
Avg. Time to Meet (hrs)	16.64	16.29	13.76	16.86	15.98	13.1	**16.52**	15.69	13.29	16.95	**15.59**	**12.72**	18.95	17.06	13.65
Action Diversity	14	11	16	20	10	16	21	**15**	**18**	21	14	16	**23**	14	17
Anomalous Actions (%)	6.0	**0.0**	**0.0**	7.0	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	6.0	**0.0**	**0.0**	7.0	1.0	**0.0**

Table S10: Reward Design Comparison Across All Severity Categories for Training Set (CQL).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	5.8	6.93	9.16	6.33	6.89	**9.17**	**6.69**	**8.02**	8.68	6.09	7.24	9.02	5.85	7.52	9.04
Extubation Meet Rate (%)	92.0	95.0	**100.0**	**97.0**	95.0	99.0	96.0	**99.0**	99.0	91.0	**99.0**	**100.0**	91.0	96.0	**100.0**
Avg. Trajectory Length (hrs)	19.19	17.25	14.05	20.96	17.61	**12.4**	18.61	17.3	14.29	**16.9**	20.84	13.96	17.96	**15.79**	14.45
Avg. Time to Meet (hrs)	19.72	17.35	14.05	21.39	17.13	**12.29**	19.09	17.36	14.26	**17.69**	20.9	13.96	18.52	**15.86**	14.45
Action Diversity	**29**	**30**	15	25	28	**18**	28	20	16	28	22	17	28	20	14
Anomalous Actions (%)	8.0	5.0	**0.0**	**3.0**	5.0	1.0	4.0	**1.0**	1.0	9.0	**1.0**	**0.0**	9.0	4.0	**0.0**

Table S11: Reward Design Comparison Across All Severity Categories for Test Set (BC).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	**7.06**	8.49	9.25	6.96	**8.64**	9.14	7.02	8.43	**9.39**	6.95	8.54	9.16	6.96	8.46	9.19
Extubation Meet Rate (%)	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**
Avg. Trajectory Length (hrs)	15.84	15.46	12.82	**15.68**	**15.18**	13.85	15.83	15.66	**12.64**	16.18	15.23	14.09	16.54	16.86	13.6
Avg. Time to Meet (hrs)	16.46	15.46	12.82	**16.29**	**15.18**	13.85	16.45	15.66	**12.64**	16.82	15.23	14.09	17.2	16.86	13.6
Action Diversity	11	12	12	15	9	13	15	**13**	**14**	14	11	13	**18**	10	**14**
Anomalous Actions (%)	**5.0**	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**

Table S12: Reward Design Comparison Across All Severity Categories for Test Set (NFQ).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	7.09	8.32	**9.34**	6.89	8.54	9.19	**7.36**	8.51	9.33	7.22	8.62	9.14	7.03	**8.68**	9.08
Extubation Meet Rate (%)	95.0	**100.0**	**100.0**	94.0	**100.0**	**100.0**	**97.0**	**100.0**	**100.0**	95.0	**100.0**	**100.0**	94.0	**100.0**	**100.0**
Avg. Trajectory Length (hrs)	16.11	16.41	**12.82**	15.98	15.5	12.97	16.76	15.58	12.94	**15.39**	**15.24**	13.3	17.37	16.16	13.54
Avg. Time to Meet (hrs)	16.75	16.41	**12.82**	16.74	15.5	12.97	17.15	15.58	12.94	**15.99**	**15.24**	13.3	18.22	16.16	13.54
Action Diversity	14	**17**	13	16	10	12	16	13	**15**	13	11	12	**21**	16	14
Anomalous Actions (%)	5.0	**0.0**	**0.0**	6.0	**0.0**	**0.0**	**3.0**	**0.0**	**0.0**	5.0	**0.0**	**0.0**	6.0	**0.0**	**0.0**

Table S13: Reward Design Comparison Across All Severity Categories for Test Set (DQN).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	7.19	8.48	9.19	6.97	8.5	9.2	**7.32**	**8.54**	9.22	6.47	8.52	**9.33**	6.96	8.48	9.24
Extubation Meet Rate (%)	95.0	**100.0**	**100.0**	95.0	**100.0**	**100.0**	**98.0**	**100.0**	**100.0**	93.0	**100.0**	**100.0**	94.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	**15.42**	15.19	13.19	16.74	15.55	13.21	17.41	15.28	13.38	16.35	**15.13**	**12.82**	17.87	16.19	12.88
Avg. Time to Meet (hrs)	**16.02**	15.19	13.19	17.41	15.55	13.21	17.68	15.28	13.38	17.28	**15.13**	**12.82**	18.76	16.31	12.88
Action Diversity	14	8	13	14	8	13	18	**14**	16	13	9	12	**19**	13	**17**
Anomalous Actions (%)	5.0	**0.0**	**0.0**	5.0	**0.0**	**0.0**	**2.0**	**0.0**	**0.0**	7.0	**0.0**	**0.0**	6.0	1.0	**0.0**

Table S14: Reward Design Comparison Across All Severity Categories for Test Set (DDQN).

Metric	`default`			`NoActPen`			`NoIntermRew`			`HighExtubRew`			`NoExtubRew`		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.68	8.38	9.24	**6.72**	**8.53**	**9.34**	6.68	8.43	9.07	6.42	8.31	9.23	6.55	8.4	9.03
Extubation Meet Rate (%)	**94.0**	**100.0**	**100.0**	93.0	**100.0**	**100.0**	**94.0**	**100.0**	**100.0**	**94.0**	99.0	**100.0**	**94.0**	**100.0**	**100.0**
Avg. Trajectory Length (hrs)	15.98	15.47	13.63	**15.49**	**14.79**	**12.73**	16.78	15.64	13.67	17.55	15.33	13.59	17.25	15.75	13.93
Avg. Time to Meet (hrs)	16.74	15.47	13.63	**16.35**	**14.79**	**12.73**	17.6	15.64	13.67	18.41	15.38	13.59	18.1	15.75	13.93
Action Diversity	15	**17**	12	13	10	**13**	15	11	12	14	**17**	**13**	**23**	10	12
Anomalous Actions (%)	**6.0**	**0.0**	**0.0**	7.0	**0.0**	**0.0**	**6.0**	**0.0**	**0.0**	**6.0**	1.0	**0.0**	**6.0**	**0.0**	**0.0**

Table S15: Reward Design Comparison Across All Severity Categories for Test Set (SAC).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.95	8.39	9.23	6.74	8.45	9.25	**7.28**	8.6	9.28	6.78	8.56	9.25	6.64	**8.61**	**9.32**
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	**97.0**	**100.0**	**100.0**	96.0	**100.0**	**100.0**	94.0	**100.0**	**100.0**	92.0	**100.0**	**100.0**
Avg. Trajectory Length (hrs)	**15.42**	15.95	**12.97**	18.06	14.87	13.21	16.63	**14.85**	13.11	16.36	15.54	13.21	18.01	16.8	13.27
Avg. Time to Meet (hrs)	**16.13**	15.95	**12.97**	18.49	14.87	13.21	17.14	**14.85**	13.11	17.15	15.54	13.21	19.23	16.8	13.27
Action Diversity	16	12	15	15	**14**	15	19	9	16	16	9	15	**22**	12	**18**
Anomalous Actions (%)	6.0	**0.0**	**0.0**	**3.0**	**0.0**	**0.0**	4.0	**0.0**	**0.0**	6.0	**0.0**	**0.0**	8.0	**0.0**	**0.0**

Table S16: Reward Design Comparison Across All Severity Categories for Test Set (BCQ).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	**7.16**	8.23	9.29	6.93	**8.68**	9.23	6.97	8.57	9.25	6.94	8.41	9.3	5.93	8.47	**9.33**
Extubation Meet Rate (%)	**95.0**	99.0	**100.0**	**95.0**	**100.0**	**100.0**	94.0	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	90.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	16.54	15.89	13.12	17.07	**14.72**	13.44	**15.46**	15.34	13.03	17.32	16.21	13.25	18.33	18.42	**12.86**
Avg. Time to Meet (hrs)	17.2	16.01	13.12	17.74	**14.72**	13.44	**16.19**	15.34	13.03	18.02	16.21	13.25	19.92	18.57	**12.86**
Action Diversity	21	13	16	22	14	16	21	13	**18**	19	11	12	**24**	**18**	16
Anomalous Actions (%)	**5.0**	1.0	**0.0**	**5.0**	**0.0**	**0.0**	6.0	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	10.0	1.0	**0.0**

Table S17: Reward Design Comparison Across All Severity Categories for Test Set (CQL).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	5.72	7.06	9.01	**6.94**	7.01	9.02	5.74	**7.56**	8.53	6.55	6.91	**9.26**	5.41	7.45	8.81
Extubation Meet Rate (%)	92.0	98.0	**100.0**	**98.0**	96.0	99.0	91.0	98.0	98.0	94.0	96.0	**100.0**	90.0	**99.0**	**100.0**
Avg. Trajectory Length (hrs)	19.54	20.64	14.27	19.24	17.39	13.43	18.03	**17.38**	14.44	**17.04**	19.08	**13.09**	18.31	19.51	14.65
Avg. Time to Meet (hrs)	20.12	20.87	14.27	19.55	**17.4**	13.36	18.95	17.45	14.42	**17.78**	18.99	**13.09**	19.31	19.45	14.65
Action Diversity	28	21	15	28	**28**	17	**29**	22	**20**	21	19	15	27	26	14
Anomalous Actions (%)	8.0	2.0	**0.0**	**2.0**	4.0	1.0	9.0	2.0	2.0	6.0	4.0	**0.0**	10.0	**1.0**	**0.0**

Table S18: Reward Design Comparison Across All Severity Categories for OOD Test Set (BC).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.65	8.12	9.32	6.46	8.14	9.23	**7.14**	**8.33**	**9.33**	6.66	8.13	9.3	6.08	7.5	8.91
Extubation Meet Rate (%)	**94.0**	**100.0**	**100.0**	93.0	**100.0**	**100.0**	**94.0**	99.0	**100.0**	**94.0**	**100.0**	**100.0**	92.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	18.15	18.91	**12.66**	17.21	19.31	13.78	**15.47**	**18.13**	13.2	17.45	19.22	12.84	20.47	24.46	15.65
Avg. Time to Meet (hrs)	19.05	18.91	**12.66**	18.2	19.31	13.78	**16.2**	**18.27**	13.2	18.31	19.22	12.84	21.9	24.67	15.65
Action Diversity	20	16	17	20	16	17	18	16	16	20	14	17	**25**	**18**	**21**
Anomalous Actions (%)	**6.0**	**0.0**	**0.0**	7.0	**0.0**	**0.0**	**6.0**	1.0	**0.0**	**6.0**	**0.0**	**0.0**	8.0	1.0	**0.0**

Table S19: Reward Design Comparison Across All Severity Categories for OOD Test Set (NFQ).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.86	8.46	9.18	6.84	8.48	9.24	6.86	8.7	9.32	**6.9**	**8.92**	**9.41**	6.33	8.06	9.22
Extubation Meet Rate (%)	94.0	**100.0**	**100.0**	**95.0**	**100.0**	**100.0**	94.0	**100.0**	**100.0**	93.0	**100.0**	**100.0**	91.0	99.0	**100.0**
Avg. Trajectory Length (hrs)	15.94	17.63	13.76	17.4	16.38	13.12	16.51	15.85	**12.52**	**15.4**	**15.08**	12.59	17.71	21.89	13.84
Avg. Time to Meet (hrs)	16.7	17.63	13.76	18.11	16.38	13.12	17.31	15.85	**12.52**	**16.25**	**15.08**	12.59	19.03	22.07	13.84
Action Diversity	17	**18**	15	18	17	16	18	14	15	18	15	16	**25**	15	**21**
Anomalous Actions (%)	6.0	**0.0**	**0.0**	**5.0**	**0.0**	**0.0**	6.0	**0.0**	**0.0**	7.0	**0.0**	**0.0**	9.0	1.0	**0.0**

Table S20: Reward Design Comparison Across All Severity Categories for OOD Test Set (DQN).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.77	8.61	9.19	6.77	8.4	**9.22**	6.57	**8.67**	9.18	**6.99**	8.58	9.18	6.27	7.98	9.19
Extubation Meet Rate (%)	93.0	**100.0**	**100.0**	**94.0**	99.0	**100.0**	**94.0**	**100.0**	**100.0**	**94.0**	**100.0**	**100.0**	92.0	98.0	**100.0**
Avg. Trajectory Length (hrs)	**16.26**	16.75	**13.36**	17.61	**16.5**	13.47	17.38	16.74	13.81	16.5	16.76	13.88	17.45	19.98	14.35
Avg. Time to Meet (hrs)	**17.18**	16.75	**13.36**	18.48	**16.63**	13.47	18.23	16.74	13.81	17.3	16.76	13.88	18.62	20.31	14.35
Action Diversity	22	15	16	17	**16**	15	17	14	15	17	13	16	**25**	**16**	**17**
Anomalous Actions (%)	7.0	**0.0**	**0.0**	**6.0**	1.0	**0.0**	**6.0**	**0.0**	**0.0**	**6.0**	**0.0**	**0.0**	8.0	2.0	**0.0**

Table S21: Reward Design Comparison Across All Severity Categories for OOD Test Set (DDQN).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	-19.78	-24.91	-1.8	-20.28	-25.62	-2.53	**-12.35**	**-15.44**	**0.61**	-17.02	-21.65	-2.4	-13.97	-18.93	-0.55
Extubation Meet Rate (%)	7.0	1.0	73.0	19.0	1.0	70.0	15.0	4.0	**76.0**	14.0	4.0	71.0	**27.0**	**11.0**	75.0
Avg. Trajectory Length (hrs)	54.52	77.66	40.8	91.26	87.35	45.32	**32.64**	**41.27**	**33.64**	52.92	69.69	44.07	49.68	65.59	36.42
Avg. Time to Meet (hrs)	**13.57**	**12.0**	**13.3**	16.37	13.0	13.73	27.07	40.0	16.75	30.86	18.75	16.01	33.48	53.36	14.24
Action Diversity	37	42	26	24	28	17	53	50	**31**	**58**	**54**	28	48	45	24
Anomalous Actions (%)	69.0	67.0	3.0	**12.0**	**57.0**	**1.0**	83.0	91.0	10.0	68.0	75.0	4.0	59.0	67.0	6.0

Table S22: Reward Design Comparison Across All Severity Categories for OOD Test Set (SAC).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	**6.46**	7.12	8.66	5.91	**7.58**	8.94	6.3	7.43	**9.23**	5.48	6.46	9.13	4.55	1.06	8.7
Extubation Meet Rate (%)	**94.0**	99.0	**100.0**	91.0	**100.0**	**100.0**	92.0	98.0	**100.0**	89.0	99.0	**100.0**	92.0	83.0	99.0
Avg. Trajectory Length (hrs)	18.8	25.75	16.44	19.73	22.45	14.84	**16.84**	**19.51**	**13.5**	18.62	25.43	14.19	27.27	47.95	15.17
Avg. Time to Meet (hrs)	19.74	25.91	16.44	21.29	22.45	14.84	**17.88**	**19.83**	**13.5**	20.43	25.65	14.19	27.11	42.73	14.08
Action Diversity	25	20	19	23	20	17	22	19	**22**	27	17	19	**28**	**30**	21
Anomalous Actions (%)	**6.0**	1.0	**0.0**	9.0	**0.0**	**0.0**	8.0	2.0	**0.0**	11.0	1.0	**0.0**	7.0	10.0	**0.0**

Table S23: Reward Design Comparison Across All Severity Categories for OOD Test Set (BCQ).

Metric	default			NoActPen			NoIntermRew			HighExtubRew			NoExtubRew		
	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low	High	Med	Low
Total Cumulative Reward	6.83	8.5	**9.2**	6.36	**8.61**	9.11	6.69	8.53	9.02	**6.88**	8.14	**9.2**	5.8	7.13	8.87
Extubation Meet Rate (%)	**93.0**	**100.0**	**100.0**	**93.0**	**100.0**	**100.0**	**93.0**	99.0	**100.0**	**93.0**	98.0	**100.0**	**93.0**	99.0	**100.0**
Avg. Trajectory Length (hrs)	17.03	18.62	**13.4**	18.29	17.24	13.54	17.58	**16.44**	14.04	**16.45**	18.29	14.07	22.32	29.78	15.19
Avg. Time to Meet (hrs)	18.01	18.62	**13.4**	19.23	17.24	13.54	18.6	**16.57**	14.04	**17.39**	18.58	14.07	23.41	30.04	15.19
Action Diversity	25	18	19	23	20	19	24	16	18	25	20	**20**	**26**	**22**	**20**
Anomalous Actions (%)	**7.0**	**0.0**	**0.0**	**7.0**	**0.0**	**0.0**	**7.0**	1.0	**0.0**	**7.0**	2.0	**0.0**	**7.0**	1.0	**0.0**

Table S24: Reward Design Comparison Across All Severity Categories for OOD Test Set (CQL).

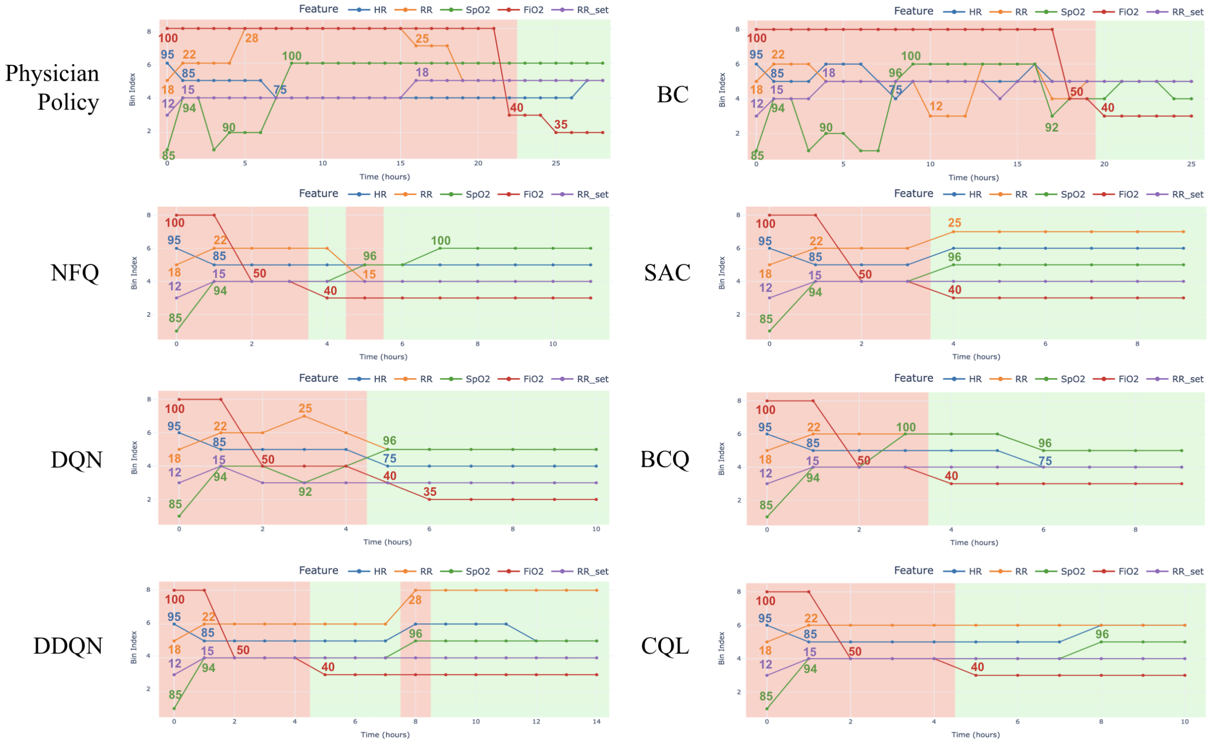

Figure S6: Comparison of Agent Behaviors from Same Initial State. This figure compares the trajectories of different RL agents (e.g., BC, DQN) starting from identical initial patient states, showcasing how various reward designs influence decision-making and ventilator setting adjustments over time.

Algorithm	Metric	RL Mean	Physician Mean	p-value
BC	Total Cumulative Reward	5.70	-16.98	**2.52e-17**
BC	Extubation Meet Rate (%)	93.00	50.00	**1.63e-11**
BC	Avg. Trajectory Length (hrs)	20.03	87.76	**9.36e-31**
BC	Avg. Time to Meet (hrs)	20.75	54.64	**6.80e-18**
NFQ	Total Cumulative Reward	6.87	-16.98	**1.23e-18**
NFQ	Extubation Meet Rate (%)	95.00	50.00	**1.03e-12**
NFQ	Avg. Trajectory Length (hrs)	17.41	87.76	**7.54e-32**
NFQ	Avg. Time to Meet (hrs)	18.12	54.64	**5.03e-20**
DQN	Total Cumulative Reward	6.82	-16.98	**1.46e-18**
DQN	Extubation Meet Rate (%)	94.00	50.00	**4.23e-12**
DQN	Avg. Trajectory Length (hrs)	15.99	87.76	**1.71e-32**
DQN	Avg. Time to Meet (hrs)	16.73	54.64	**5.13e-21**
DDQN	Total Cumulative Reward	7.11	-16.98	**7.12e-19**
DDQN	Extubation Meet Rate (%)	95.00	50.00	**1.03e-12**
DDQN	Avg. Trajectory Length (hrs)	16.86	87.76	**4.70e-32**
DDQN	Avg. Time to Meet (hrs)	17.54	54.64	**1.82e-20**
SAC	Total Cumulative Reward	7.05	-16.98	**8.50e-19**
SAC	Extubation Meet Rate (%)	94.00	50.00	**4.23e-12**
SAC	Avg. Trajectory Length (hrs)	14.90	87.76	**7.21e-33**
SAC	Avg. Time to Meet (hrs)	15.60	54.64	**7.52e-22**
BCQ	Total Cumulative Reward	6.64	-16.98	**2.20e-18**
BCQ	Extubation Meet Rate (%)	94.00	50.00	**4.23e-12**
BCQ	Avg. Trajectory Length (hrs)	16.72	87.76	**3.74e-32**
BCQ	Avg. Time to Meet (hrs)	17.53	54.64	**1.87e-20**
CQL	Total Cumulative Reward	6.87	-16.98	**1.28e-18**
CQL	Extubation Meet Rate (%)	94.00	50.00	**4.23e-12**
CQL	Avg. Trajectory Length (hrs)	15.88	87.76	**1.92e-32**
CQL	Avg. Time to Meet (hrs)	16.64	54.64	**4.01e-21**

Table S25: Statistical test results on High Severity patients from the Training set. Bolded p-values indicate significance ($p < 0.05$).

Algorithm	Metric	RL Mean	Physician Mean	p-value
BC	Total Cumulative Reward	5.80	-20.07	**5.15e-22**
BC	Extubation Meet Rate (%)	92.00	39.00	**3.18e-15**
BC	Avg. Trajectory Length (hrs)	19.19	94.67	**4.54e-35**
BC	Avg. Time to Meet (hrs)	19.72	55.41	**1.03e-16**
NFQ	Total Cumulative Reward	7.06	-20.07	**2.18e-23**
NFQ	Extubation Meet Rate (%)	95.00	39.00	**3.72e-17**
NFQ	Avg. Trajectory Length (hrs)	15.84	94.67	**2.69e-36**
NFQ	Avg. Time to Meet (hrs)	16.46	55.41	**6.88e-19**
DQN	Total Cumulative Reward	7.09	-20.07	**2.02e-23**
DQN	Extubation Meet Rate (%)	95.00	39.00	**3.72e-17**
DQN	Avg. Trajectory Length (hrs)	16.11	94.67	**4.51e-36**
DQN	Avg. Time to Meet (hrs)	16.75	55.41	**9.63e-19**
DDQN	Total Cumulative Reward	7.19	-20.07	**1.59e-23**
DDQN	Extubation Meet Rate (%)	95.00	39.00	**3.72e-17**
DDQN	Avg. Trajectory Length (hrs)	15.42	94.67	**3.18e-36**
DDQN	Avg. Time to Meet (hrs)	16.02	55.41	**3.13e-19**
SAC	Total Cumulative Reward	6.68	-20.07	**5.43e-23**
SAC	Extubation Meet Rate (%)	94.00	39.00	**1.73e-16**
SAC	Avg. Trajectory Length (hrs)	15.98	94.67	**3.95e-36**
SAC	Avg. Time to Meet (hrs)	16.74	55.41	**9.55e-19**
BCQ	Total Cumulative Reward	6.95	-20.07	**2.82e-23**
BCQ	Extubation Meet Rate (%)	94.00	39.00	**1.73e-16**
BCQ	Avg. Trajectory Length (hrs)	15.42	94.67	**2.19e-36**
BCQ	Avg. Time to Meet (hrs)	16.13	55.41	**3.96e-19**
CQL	Total Cumulative Reward	7.16	-20.07	**1.71e-23**
CQL	Extubation Meet Rate (%)	95.00	39.00	**3.72e-17**
CQL	Avg. Trajectory Length (hrs)	16.54	94.67	**6.84e-36**
CQL	Avg. Time to Meet (hrs)	17.20	55.41	**1.86e-18**

Table S26: Statistical test results on High Severity patients from the Test set. Bolded p-values indicate significance ($p < 0.05$).

Algorithm	Metric	RL Mean	Physician Mean	p-value
BC	Total Cumulative Reward	5.72	-19.38	**6.15e-22**
BC	Extubation Meet Rate (%)	92.00	37.00	**4.38e-16**
BC	Avg. Trajectory Length (hrs)	19.54	94.29	**8.23e-34**
BC	Avg. Time to Meet (hrs)	20.12	50.30	**1.28e-13**
NFQ	Total Cumulative Reward	6.65	-19.38	**5.33e-23**
NFQ	Extubation Meet Rate (%)	94.00	37.00	**2.28e-17**
NFQ	Avg. Trajectory Length (hrs)	18.15	94.29	**1.81e-34**
NFQ	Avg. Time to Meet (hrs)	19.05	50.30	**2.58e-14**
DQN	Total Cumulative Reward	6.86	-19.38	**3.19e-23**
DQN	Extubation Meet Rate (%)	94.00	37.00	**2.28e-17**
DQN	Avg. Trajectory Length (hrs)	15.94	94.29	**2.67e-35**
DQN	Avg. Time to Meet (hrs)	16.70	50.30	**6.26e-16**
DDQN	Total Cumulative Reward	6.77	-19.38	**3.98e-23**
DDQN	Extubation Meet Rate (%)	93.00	37.00	**1.02e-16**
DDQN	Avg. Trajectory Length (hrs)	16.26	94.29	**5.66e-35**
DDQN	Avg. Time to Meet (hrs)	17.18	50.30	**1.01e-15**
SAC	Total Cumulative Reward	-19.78	-19.38	5.66e-01
SAC	Extubation Meet Rate (%)	7.00	37.00	**3.04e-07**
SAC	Avg. Trajectory Length (hrs)	54.52	94.29	**2.49e-09**
SAC	Avg. Time to Meet (hrs)	13.57	50.30	**4.41e-18**
BCQ	Total Cumulative Reward	6.46	-19.38	**8.66e-23**
BCQ	Extubation Meet Rate (%)	94.00	37.00	**2.28e-17**
BCQ	Avg. Trajectory Length (hrs)	18.80	94.29	**2.17e-34**
BCQ	Avg. Time to Meet (hrs)	19.74	50.30	**1.21e-13**
CQL	Total Cumulative Reward	6.83	-19.38	**3.45e-23**
CQL	Extubation Meet Rate (%)	93.00	37.00	**1.02e-16**
CQL	Avg. Trajectory Length (hrs)	17.03	94.29	**7.06e-35**
CQL	Avg. Time to Meet (hrs)	18.01	50.30	**4.65e-15**

Table S27: Statistical test results on High Severity patients from the OOD Test set. Bolded p-values indicate significance ($p < 0.05$).