RC-DQN: A RISK-AWARE DUAL-CRITIC CONFORMAL DEEP Q-NETWORK FOR DEXMEDETOMIDINE DOSE TITRATION IN POSTOPERATIVE ICU PATIENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Dexmedetomidine dose titration in postoperative intensive care unit (ICU) patients is a sequential decision-making problem requiring a balance among prespecified treatment objectives, physiological tolerance, and action support. Applying offline reinforcement learning to this setting is complicated by the ordinal dose space, the scalarization of treatment utility and physiological cost, and extrapolation to actions that are poorly supported by retrospective data. We developed a Risk-Aware Dual-Critic Conformal Deep Q-Network (RC-DQN) using retrospective trajectories from two critical-care databases. Patient records were organized into six-hour decision windows, and dexmedetomidine exposure was represented by 16 ordered actions. RC-DQN combines an ordinal behavior-policy model, separate utility and physiological-cost-score critics, risk-modulated conformal calibration, an absolute cost-score shield, and rule-based fallback. The primary offline evaluation used terminal-outcome fitted Q evaluation (FQE) to estimate the mean initial-state value for the alive-without-delirium terminal outcome. An independent FQE evaluated returns under the prespecified composite reward, while action matching, ablation analysis, and bidirectional cross-database sensitivity analysis provided complementary assessments. Among the evaluated policies, RC-DQN produced the numerically highest within-database point estimates for both the alive-without-delirium endpoint and composite-reward return across both databases. Its endpoint estimates were higher than the corresponding observed test-cohort proportions, while action-matching analysis showed partial agreement with documented clinical decisions. Cross-database evaluation produced asymmetric results when source-trained policies and evaluators were applied to target-database states, indicating sensitivity to database-specific distributions. Overall, RC-DQN provides a structured approach for retrospectively learning and evaluating ordinal medication-titration policies under behavioral-support and physiological-cost constraints. The source code and datasets are submitted as Supplementary Materials for Reproducibility.

1 INTRODUCTION

Dexmedetomidine is a highly selective α_2-adrenoceptor agonist with sedative, anxiolytic, sympatholytic, and opioid-sparing properties and is commonly used to provide sedation while minimizing respiratory depression in critically ill and postoperative patients (DailyMed, 2026; Lewis et al., 2025). Evidence comparing dexmedetomidine with other ICU sedatives has reported heterogeneous, context-dependent outcomes in terms of sedation quality, delirium, ventilation-related recovery, and ICU resource utilization (Lewis et al., 2022). Recent perioperative evidence indicates that dexmedetomidine can reduce postoperative delirium in selected older surgical populations, while hypotension and bradycardia remain relevant trade-offs (Li et al., 2026; Luney et al., 2026).

Dexmedetomidine dosing in postoperative ICU patients can therefore be formulated as a sequential, risk-sensitive titration problem. In clinical practice, bedside titration is guided by target sedation levels, concurrent organ-support requirements, hemodynamic tolerance, and clinicians' integrated judgment. However, current guidelines and prescribing information specify general sedation targets and recommended infusion ranges but do not provide a patient-state-dependent decision rule for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

mapping multimodal physiological trajectories to discrete dosing actions (van den Boogaard et al., 2026; DailyMed, 2026).

Reinforcement learning (RL) provides a computational framework for learning sequential decision policies from fixed trajectories of states, actions, and rewards (Gottesman et al., 2019; Zhang et al., 2026). Because online exploration is generally unethical or impractical in clinical settings, offline RL is more appropriate for learning from retrospective electronic health record data (Gottesman et al., 2019; Prudencio et al., 2024). A major methodological challenge is extrapolation error, whereby a learned policy may assign spuriously high values to actions that are poorly represented in the historical data. Conservative and behavior-regularized offline RL methods seek to mitigate this problem by limiting value overestimation for actions with insufficient support in the behavior dataset (Kumar et al., 2020; Fujimoto et al., 2019; Wang et al., 2026).

Uncertainty-aware action selection is particularly consequential for medication titration in critical care, where unsupported decisions may expose patients to avoidable harm. Under exchangeability, conformal prediction provides finite-sample, distribution-free marginal coverage guarantees for prediction sets (Angelopoulos & Bates, 2023; Vovk, 2026). ConformalDQN extended this approach to reinforcement-learning-based mechanical ventilation control (Eghbali et al., 2025). Dexmedetomidine titration, however, presents additional methodological challenges, including ordinal dose modeling, control of short-term physiological risk, and explicit restrictions on unsupported high-dose actions.

To address these methodological gaps, we introduce the Risk-Aware Dual-Critic Conformal Deep Q-Network (RC-DQN) for dexmedetomidine dose titration. Our contributions are threefold. First, we introduce a truncated-Poisson behavior-policy head that uses the ordinal structure of 16 dexmedetomidine treatment levels to estimate action support in the retrospective data. Second, we separate the standardized composite-reward utility critic from a threshold-regularized critic that produces a model-estimated discounted physiological-cost score; this score is used for constraint and ranking purposes rather than interpreted as a calibrated adverse-event probability. Third, we combine risk-modulated conformal action admissibility with an absolute cost-score threshold and prespecified high-dose veto and fallback rules to construct a constrained retrospective decision policy. The conformal component addresses marginal inclusion of observed behavior actions under its assumptions, the cost critic supplies a model-defined constraint score, and the deterministic rules impose additional prespecified restrictions; none of these components constitutes a guarantee of clinical safety or effectiveness. We evaluate these components using delirium-related terminal-outcome FQE, an independent composite-reward FQE, action-pattern analyses, ablation comparisons, and bidirectional cross-database sensitivity analyses. These analyses provide a retrospective comparison of policy behavior and model-estimated outcomes across the two databases.

2 Related Work

2.1 Dexmedetomidine Titration in Postoperative Critical Care

Evidence syntheses comparing dexmedetomidine with other sedative or perioperative regimens report heterogeneous effects on delirium, mechanical ventilation, ICU length of stay, and hemodynamic events across populations and protocols (Lewis et al., 2022; Li et al., 2026; Luney et al., 2026). This evidence characterizes population-level efficacy and adverse-event profiles but does not provide a state-dependent rule for sequential dose adjustment during an ICU stay.

2.2 Offline Reinforcement Learning for Clinical Sequential Decisions

Offline reinforcement learning has been studied for sepsis management, mechanical ventilation, glucose control, and medication dosing, where policies must be learned from observational trajectories without online exploration (Gottesman et al., 2019; Qiu et al., 2026; Paul & Greiner, 2026). Approaches to extrapolation error include batch-constrained action support, conservative value estimation, behavior regularization, and implicit or in-sample value learning (Fujimoto et al., 2019; Kumar et al., 2020; Zhang et al., 2026; Prudencio et al., 2024; Kostrikov et al., 2022; Chen et al., 2025). Empirical action frequencies and policy agreement can characterize behavioral support but do not establish positivity over the full patient-state space (Gottesman et al., 2019; Amortila et al., 2026). FQE enables target-policy evaluation without deployment, although its reliability depends on

evaluator specification, function approximation, effective horizon, and target-policy occupancy coverage (Zhang et al., 2022; Wang et al., 2024). General offline RL formulations also need not exploit ordered dose structure or separate long-horizon utility from short-horizon physiological cost.

2.3 Uncertainty-Aware and Physiologically Constrained Action Selection

Mechanical ventilation is a relevant analogue because it involves recurrent ICU decisions, delayed outcomes, and state-dependent physiological constraints. UBER-CQL investigated reinforcement-learning-based ventilator management, while ConformalDQN added conformal action filtering to Double DQN; recent offline safe RL work has also learned persistent safety sets from fixed data (Liu et al., 2026; Eghbali et al., 2025; Choudhury et al., 2026). Under exchangeability, conformal prediction can provide finite-sample marginal coverage without a parametric error model (Angelopoulos & Bates, 2023; Vovk, 2026). Behavioral support, conformal admissibility, and physiological acceptability nevertheless address different questions: a frequently observed action may have elevated physiological cost, whereas a low predicted cost does not establish sufficient support for reliable value extrapolation. Conformal inclusion is also not a guarantee of clinical safety or optimality. Existing conformal action filtering was not designed to jointly handle ordered medication doses, a learned cost score, and deterministic high-dose or empty-set management.

2.4 Positioning and Technical Contributions of RC-DQN

Offline RL research specifically addressing dexmedetomidine dosing remains limited. Lee et al. (2024) studied delirium prevention using conservative value learning. RC-DQN instead formulates postoperative ICU titration as a coupled problem of ordinal behavioral-support estimation, utility–cost separation, and constrained action selection.

The contribution is not the independent addition of an ordinal head, a cost critic, and a conformal filter. The truncated-Poisson behavior head shares information across ordered dose levels and supplies probabilities for both support estimation and the nonconformity score. Separate critics preserve distinct utility and physiological-cost targets; the cost score influences threshold regularization, Bellman action selection, and inference-time constraints but is not interpreted as a calibrated adverse-event probability. Risk modulation incorporates this score into conformal admissibility, after which an absolute cost threshold, cost-adjusted ranking, high-dose vetoes, and deterministic fallback address cases in which probability-based support is insufficient or no candidate remains admissible. RC-DQN thus integrates structured support estimation and physiological constraints throughout the policy pipeline rather than applying them as independent post-hoc filters.

3 Materials and Methods

As shown in Figure 1, RC-DQN follows five stages: offline trajectory construction, joint learning of efficacy Q-values, physiological-cost scores, and ordinal behavior probabilities, risk-modulated conformal calibration, constrained action selection, and offline policy evaluation. Calibration combines behavior probabilities and cost scores to estimate an admissibility threshold; action selection then applies this threshold together with an absolute cost shield, cost-adjusted ranking, and prespecified veto-and-fallback rules. The complete sequence is summarized in Algorithm 1.

3.1 Data Construction

Postoperative ICU stays with documented dexmedetomidine exposure were divided into 6-hour decision windows. Each transition contained an 18-dimensional clinical state, one of 16 ordered infusion-rate actions (including action 0 for no infusion), a composite utility reward, and a physiological side-effect cost. Episodes were split by patient into training, validation, and test sets; database-specific state and total-reward standardizers were fitted on the training split and reused for its validation and test splits. Cohort construction, preprocessing, state and action definitions, and the complete expert-prespecified reward specification are reported in Appendix A.1 and the subsequent data-preparation and MDP subsections.

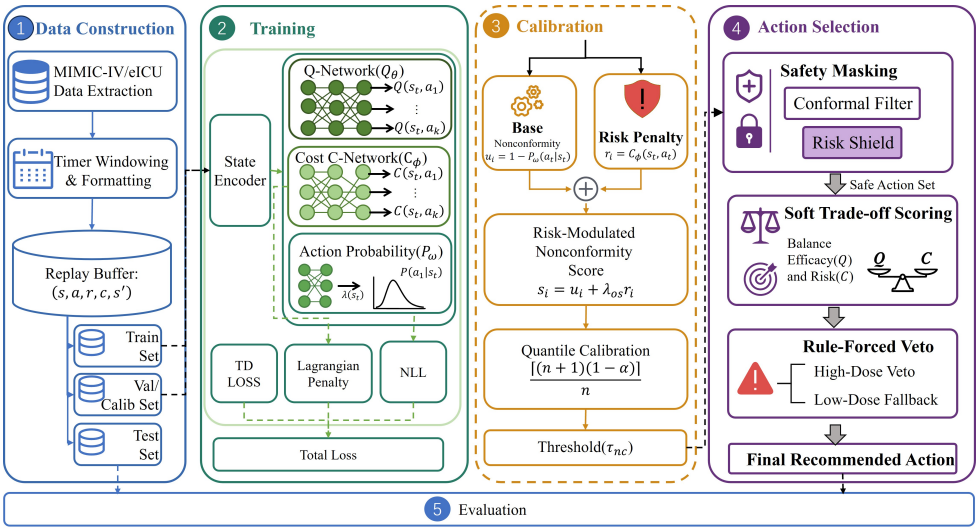

Figure 1: Overview of the RC-DQN framework.

3.2 Training

3.2.1 Trainable Components and Output Parameterization

RC-DQN uses a shared two-layer state encoder with three output heads: an efficacy critic Q_θ, a nonnegative physiological-cost-score critic C_ϕ, and an ordinal behavior-policy model P_ω. This separation preserves distinct utility and cost signals, while the behavior model supports constrained selection over ordered, sparsely observed dose levels. The architecture addresses three coupled limitations of offline medication dosing: independent categorical logits ignore dose ordering, a single scalar critic can obscure infrequent cost signals with positive utility, and filtering solely by historical action probability can retain behaviorally common but high-cost-score actions. RC-DQN therefore links ordinal support estimation, dual utility–cost evaluation, and risk-modulated conformal selection. The state-dependent truncated-Poisson head parameterizes the finite action support $\mathcal{A} = \{0, \ldots, 15\}$ using a learned rate λ_s:

$$P_\omega(a = k \mid s) = \frac{\exp(k \log \lambda_s - \log k!)}{\sum_{j \in \mathcal{A}} \exp(j \log \lambda_s - \log j!)} \tag{1}$$

The probabilities are normalized over the finite action set and used for negative-log-likelihood training and conformal scoring; this is an ordinal inductive bias rather than a generative claim about clinician dosing. C_ϕ uses a Softplus output and is interpreted as a threshold-regularized discounted physiological-cost score, not a calibrated adverse-event probability. The conformal selector combines P_ω and C_ϕ, so the cost critic and action-support model provide complementary constraints. Additional interpretation is provided in Appendix A.4.

3.2.2 Dual-Critic and Joint Training Objectives

RC-DQN uses separate discount factors for efficacy and physiological cost. For each sampled transition $(s_t, a_t, r_t, c_t, s_{t+1}, d_t)$, the target network computes next-step Q-values and costs. The bootstrap action is selected by maximizing a risk-adjusted target-network score:

$$a^* = \arg\max_{a \in \mathcal{A}} \left[Q_{\theta^-}(s_{t+1}, a) - \lambda_{\text{safe}} C_{\phi^-}(s_{t+1}, a) \right]. \tag{2}$$

The efficacy and cost targets are then defined as:

$$y_t^Q = r_t + \gamma_Q (1 - d_t) Q_{\theta^-}(s_{t+1}, a^*) \tag{3}$$

$$y_t^C = c_t + \gamma_C (1 - d_t) C_{\phi^-}(s_{t+1}, a^*) \tag{4}$$

Here, γ_Q determines the discounting of utility rewards, whereas γ_C determines the discounting of physiological costs. Policy training used $\gamma_Q = 0.99$; γ_C was treated as a tunable hyperparameter and specified independently of γ_Q.

The ordinal behavior-policy head is trained by minimizing the negative log-likelihood of the observed clinician actions, $\mathcal{L}_{\mathrm{NLL}} = -\mathbb{E}_{\mathcal{D}} \log P_\omega(a_t \mid s_t)$. This term fits the behavior-policy model whose probabilities are subsequently used to construct conformal nonconformity scores. The implemented objective also includes an auxiliary predicted-cost threshold penalty, $\mathcal{L}_{\mathrm{Risk}} = \mathbb{E}_{\mathcal{D}} \max(C_\phi(s_t, a_t) - d_{\mathrm{safe}}, 0)$. The joint training objective is:

$$\mathcal{L}_{\mathrm{TD},Q} = \mathbb{E}_{\mathcal{D}} \left[\mathrm{Huber} \left(Q_\theta(s_t, a_t) - y_t^Q \right) \right] \tag{5}$$

$$\mathcal{L}_{\mathrm{Asym},C} = \mathbb{E}_{\mathcal{D}} \left[w_t \cdot \mathrm{Huber} \left(C_\phi(s_t, a_t) - y_t^C \right) \right], \quad w_t = \begin{cases} \tau, & y_t^C > C_\phi(s_t, a_t), \\ 1 - \tau, & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathcal{L}_{\mathrm{Risk}} = \mathbb{E}_{\mathcal{D}} \left[\max \left(C_\phi(s_t, a_t) - d_{\mathrm{safe}}, 0 \right) \right] \tag{7}$$

$$\mathcal{L} = \mathcal{L}_{\mathrm{TD},Q} + \mathcal{L}_{\mathrm{Asym},C} + \lambda_{\mathrm{nll}} \mathcal{L}_{\mathrm{NLL}} + \lambda_{\mathrm{safe}} \mathcal{L}_{\mathrm{Risk}} \tag{8}$$

The asymmetric cost loss penalizes underestimation more strongly; with $\tau = 0.95$, equal-magnitude underestimation receives 19 times the weight of overestimation. The threshold regularizer enters the joint objective but is not an additive penalty inside the utility TD loss or a learned Lagrange multiplier. Because the output heads share an encoder, this regularizer can nevertheless affect the learned representation. The cost critic also affects the next action supplying the target Q-value through a^*, although target construction has no gradient propagation, and it controls conformal modulation, shielding, and cost-adjusted ranking at inference. Further interpretation is provided in Appendix A.4.

3.3 Calibration

After model training, RC-DQN uses a held-out validation/calibration set to estimate the threshold for its conformal action-admissibility criterion. For each observed state-action pair (s_i, a_i) in the calibration set, the trained model computes the estimated behavior-policy probability $P_\omega(a_i \mid s_i)$ and the threshold-regularized discounted physiological-cost score $C_\phi(s_i, a_i)$. The risk-modulated nonconformity score is defined as

$$u_i = 1 - P_\omega(a_i \mid s_i) + \lambda_{\mathrm{os}} C_\phi(s_i, a_i). \tag{9}$$

For n calibration examples, the threshold is set to the finite-sample-adjusted empirical quantile

$$\tau_{\mathrm{nc}} = \mathrm{Quantile} \left(\{u_i\}_{i=1}^n, \frac{\lceil (n+1)(1-\alpha) \rceil}{n} \right), \tag{10}$$

where $\alpha \in (0, 1)$ is the miscoverage level. Low behavior probability and high estimated cost both increase nonconformity, reducing the admissibility of historically common actions when the cost critic assigns an elevated score. The resulting threshold concerns marginal inclusion of observed behavior actions under the calibration assumptions; it is not a calibrated adverse-event probability or a statement of policy optimality. This action-set construction differs from conformal OPE intervals for target-policy returns (Zhang et al., 2023); repeated windows within trajectories and cross-database shift also affect the applicable exchangeability assumptions (Barber et al., 2023). Further statistical interpretation is provided in Appendix A.4.

3.4 Constrained Action Selection

3.4.1 Admissibility Mask and Cost-Adjusted Scoring

At inference, RC-DQN first computes the risk-modulated nonconformity score for each action,

$$u(s, a) = 1 - P_\omega(a \mid s) + \lambda_{\mathrm{os}} C_\phi(s, a). \tag{11}$$

and retains actions satisfying both the calibrated nonconformity threshold and the prespecified absolute cost-score threshold:

$$\mathcal{A}_{\mathrm{adm}}(s) = \{a \in \mathcal{A} : u(s, a) \le \tau_{\mathrm{nc}} \text{ and } C_\phi(s, a) \le d_{\mathrm{safe}}\}. \tag{12}$$

When $\mathcal{A}_{\mathrm{adm}}(s)$ is nonempty, the candidate maximizes cost-adjusted utility over admissible actions:

$$a_{\mathrm{cand}} = \arg \max_{a \in \mathcal{A}_{\mathrm{adm}}(s_t)} \left[Q_\theta(s_t, a) - \lambda_{\mathrm{safe}} C_\phi(s_t, a) \right]. \tag{13}$$

3.4.2 HIGH-DOSE VETO AND ZERO-DOSE FALLBACK

RC-DQN redirects the two highest-dose candidates (actions 14 and 15) to action 0 when estimated behavior probability is below 0.05 or estimated cost exceeds $0.5d_{\text{safe}}$; an empty admissible set also falls back to action 0. Thus, $a_{\text{high}} = 14$ in

$$a_t = \begin{cases} 0, & \mathcal{A}_{\text{adm}}(s_t) = \emptyset, \\ 0, & a_{\text{cand}} \geq a_{\text{high}} \text{ and } \left(P_\omega(a_{\text{cand}} \mid s_t) < 0.05 \text{ or } C_\phi(s_t, a_{\text{cand}}) > 0.5d_{\text{safe}}\right), \\ a_{\text{cand}}, & \text{otherwise.} \end{cases} \quad (14)$$

Action 0 denotes no infusion in the current window. This fallback is a prespecified constraint on unsupported or high-cost-score recommendations and does not imply that action 0 is clinically safe or optimal for every state; it does not replace individualized assessment of whether withholding or discontinuing dexmedetomidine is appropriate.

3.5 EVALUATION

Fitted Q evaluation (FQE) estimated each fixed target policy's value from retrospective trajectories by averaging initial-state values over held-out episodes,

$$\hat{V}^\pi = \frac{1}{N} \sum_{i=1}^{N} \hat{Q}^\pi(s_0^{(i)}, \pi(s_0^{(i)})). \quad (15)$$

The primary binary terminal endpoint was alive without delirium; first-window delirium was excluded, later first delirium terminated the trajectory, and the remaining terminal windows represented death before delirium or alive without delirium. Death before recorded delirium was encoded as a mutually exclusive terminal event rather than analyzed with cause-specific hazards or cumulative-incidence functions (Austin et al., 2016). MIMIC-IV excluded 180 of 3,389 training-plus-validation episodes and 27 of 599 test episodes, leaving 3,209 and 572 eligible episodes; eICU-CRD excluded 9 of 640 and 1 of 114, leaving 631 and 113. Outcome-specific FQE assigned reward 1 to the selected terminal outcome and 0 otherwise, clipped initial Q values to $[0, 1]$, and averaged them over eligible initial states.

An independent FQE evaluated the unstandardized composite reward over complete trajectories. We compared RC-DQN with clinician-policy FQE, DDQN, CQL, HUBL, RIQL, BCQ, ACTIVE-IQL, and ConformalDQN (van Hasselt et al., 2016; Kumar et al., 2020; Geng et al., 2024; Yang et al., 2024; Fujimoto et al., 2019; Chen et al., 2025; Eghbali et al., 2025). The four reported summaries were the mean initial-state endpoint FQE estimate with a fixed-model episode-bootstrap 95% interval, its descriptive difference from the observed test-cohort proportion, total-reward FQE, and global action match. Episode resampling with replacement produced the 2.5th and 97.5th percentiles while holding the policy and FQE fixed; unlike procedures that refit FQE, these intervals characterize test-episode sampling variation rather than evaluator-estimation uncertainty (Hao et al., 2021).

In-database FQE was fitted on the corresponding database's training and validation transitions. Cross-database sensitivity transported each fixed source-trained policy and source-fitted FQE to target-database test states, thereby assessing their joint sensitivity to distribution shift rather than separating policy from evaluator transport. Target observed outcomes and action matches were descriptive only; complete directional results are reported in Appendix A.6.

4 RESULTS

4.1 EXPERIMENTAL SETUP

Experiments were conducted independently on MIMIC-IV and eICU-CRD using patient-level training, validation, and test partitions. Policy learning used the training partition, conformal calibration used the validation partition, and evaluation used held-out test trajectories. After excluding episodes with delirium recorded in the first decision window, the primary terminal-outcome analysis included 572 MIMIC-IV and 113 eICU-CRD test episodes. RC-DQN was compared with clinician-policy FQE and seven offline-RL baselines. We reported the mean initial-state FQE estimate for alive

without delirium, its descriptive difference from the observed test-cohort proportion, total-reward FQE, and global action match.

The experiments addressed three questions: whether RC-DQN produced different in-database FQE estimates from the comparison policies; how its recommendations differed in action distribution, physiologically vulnerable states, and demographic subgroups; and how model configurations and cross-database distribution shift affected the reported estimates.

4.2 Main In-Database Results

Table 1 summarizes the in-database evaluation on 572 eligible MIMIC-IV and 113 eligible eICU-CRD test episodes, whose observed alive-without-delirium proportions were 47.03% and 76.11%, respectively. RC-DQN achieved the highest endpoint and total-reward FQE point estimates on both datasets, whereas the highest action match was obtained by different comparison policies.

Table 1: In-database offline evaluation. Endpoint is the mean initial-state FQE estimate for alive without delirium; Diff. is endpoint estimate minus the observed cohort proportion; Match is global decision-window action match.

Data	Policy	Endpoint, % (95% CI)	Total FQE	Diff. (pp)	Match (%)
MIMIC	Clinician	47.98 (45.79–50.09)	7.1081	–	–
	DDQN	47.07 (45.12–49.03)	4.2460	0.04	49.06
	CQL	48.88 (46.96–50.69)	8.9669	1.85	57.38
	BCQ	49.00 (47.02–50.84)	8.2866	1.97	**61.00**
	HUBL	49.77 (47.81–51.64)	9.1171	2.74	57.01
	RIQL	49.26 (47.32–51.14)	9.1570	2.23	58.93
	ACTIVE-IQL	49.46 (47.47–51.34)	9.0681	2.44	60.90
	ConformalDQN	50.23 (47.99–52.35)	6.7720	3.20	3.75
	RC-DQN	**50.49 (48.50–52.45)**	**10.2554**	**3.47**	50.17
eICU	Clinician	70.01 (67.27–72.71)	22.5793	–	–
	DDQN	76.25 (72.57–80.29)	19.4691	0.15	42.51
	CQL	77.27 (75.40–79.36)	25.1931	1.17	56.41
	BCQ	78.26 (76.24–80.38)	19.8372	2.16	54.39
	HUBL	77.23 (75.22–79.31)	25.1906	1.12	**56.68**
	RIQL	77.51 (75.36–79.88)	21.7446	1.40	51.34
	ACTIVE-IQL	77.48 (73.82–80.90)	7.2614	1.37	29.07
	ConformalDQN	78.15 (76.23–80.21)	22.7521	2.04	1.08
	RC-DQN	**78.43 (76.46–80.56)**	**25.6019**	**2.33**	49.67

RC-DQN exceeded the next-highest endpoint point estimate by 0.26 percentage points on MIMIC-IV (50.49% versus 50.23% for ConformalDQN) and by 0.17 points on eICU-CRD (78.43% versus 78.26% for BCQ); these margins were small, and the conditional episode-bootstrap intervals overlapped. For total-reward FQE, RC-DQN improved on the strongest non-RC-DQN baseline from 9.1570 to 10.2554 on MIMIC-IV and from 25.1931 to 25.6019 on eICU-CRD. In contrast, BCQ obtained the highest action match on MIMIC-IV (61.00% versus 50.17% for RC-DQN), and HUBL obtained the highest action match on eICU-CRD (56.68% versus 49.67%). The difference in rankings indicates that the higher FQE point estimates did not result simply from reproducing recorded clinician actions.

4.3 Cross-Database Stress Test

Cross-database transfer produced strongly direction-dependent estimates across the evaluated models. MIMIC-IV-fitted policy–FQE pairs yielded endpoint estimates of 43.12–44.96% on eICU-CRD test states, whereas the reverse transfer yielded 84.04–85.88%. RC-DQN was at the upper end of the model range in both directions, with endpoint/total-reward estimates of 44.96%/9.1454 for MIMIC-IV-to-eICU-CRD transfer and 85.88%/28.8888 in the reverse direction. Its action match decreased to 39.30% and 36.59%, respectively. The dominant change between transfer directions demonstrates substantial source dependence of the jointly transported policy and evaluator; accordingly, this ex-

periment was treated as a distribution-shift stress test rather than a comparison with target-database clinician outcomes (Appendix A.6).

4.4 Policy Behavior and Physiological Constraint Response

RC-DQN shifted probability mass toward action 0 on MIMIC-IV: compared with the recorded clinician distribution, action 0 increased from 62.05% to 78.29%, while high-index actions (7–15) decreased from 13.40% to 4.79%. On eICU-CRD, RC-DQN preserved a similar split between action 0 and actions 1–6 to the clinician data but selected no high-index actions. The comparison policies exhibited distinct behaviors: ConformalDQN assigned 11.92% and 9.52% of decisions to high-index actions on MIMIC-IV and eICU-CRD, whereas BCQ assigned 0.97% and 0.00%. ACTIVE-IQL was particularly database-dependent, with its high-index action frequency increasing from 0.94% on MIMIC-IV to 45.80% on eICU-CRD.

Within the prespecified low-heart-rate or low-systolic-pressure strata, RC-DQN further shifted decisions toward action 0 and away from high-index actions relative to the clinician distribution. On MIMIC-IV, action 0 increased from 52.60% to 72.92% and high-index actions decreased from 11.98% to 7.29%; on eICU-CRD, action 0 increased from 37.29% to 55.93%, while neither policy selected high-index actions in this stratum. ACTIVE-IQL showed a qualitatively different cross-database response, assigning 61.02% of eICU-CRD risk-state windows to high-index actions. In MIMIC-IV windows with heart rate < 50 bpm, RC-DQN also had a lower mean action level than the clinician and ConformalDQN policies (1.58 versus 3.79 and 2.95). The corresponding eICU-CRD heart-rate stratum contained only two windows and was therefore not used for comparative interpretation. These results characterize constraint-response patterns rather than physiological outcomes (Appendix A.7).

4.5 Robustness Across Demographic Subgroups

The subgroup analysis did not reveal a consistent reversal of the RC-DQN endpoint-estimate-minus-observed difference across sex or age strata. On MIMIC-IV, subgroup point differences ranged from 0.5 to 4.8 percentage points, compared with an overall difference of 3.5 points. On eICU-CRD, the corresponding subgroup differences ranged from 0.7 to 5.6 points, compared with 2.3 points overall. Uncertainty increased in the smaller strata: most intervals included zero, and the eICU-CRD subgroup intervals were particularly wide. The analysis therefore provides a descriptive robustness diagnostic but does not support a precise performance ranking across demographic groups (Appendix A.8).

4.6 Configuration-Level Ablation Analysis

The dependency-aware ablation defined A as ordinal behavior modeling; B as the cost critic, absolute mask, cost-adjusted ranking, and fallback; and C as the cost term in conformal nonconformity. Because C operates on the cost score supplied by B, the evaluated configurations followed this dependency structure. Behavior-only conformal thresholds were recalibrated on validation data before evaluation. Table 2 compares the resulting configurations across endpoint FQE, total-reward FQE, predicted-minus-observed difference, and action match.

Table 2: Dependency-aware ablation results; OC, SC, and OCSC are defined in the text.

Data	Configuration	Endpoint, % (95% CI)	Total FQE	Diff. (pp)	Match (%)
MIMIC	OC	43.33 (41.87–44.84)	11.0802	−3.69	27.99
	SC	49.57 (47.60–51.51)	8.9867	2.54	34.97
	OCSC	49.97 (48.21–51.59)	**11.7565**	2.94	24.02
	Full RC-DQN	**50.49 (48.50–52.45)**	10.2554	**3.47**	**50.17**
eICU	OC	77.82 (75.74–80.05)	24.6607	1.71	48.08
	SC	76.60 (74.67–78.69)	19.9013	0.49	**52.92**
	OCSC	78.21 (76.31–80.13)	25.1069	2.10	48.51
	Full RC-DQN	**78.43 (76.46–80.56)**	**25.6019**	**2.33**	49.67

Metric rankings varied across configurations. On MIMIC-IV, cost-shielded variants raised the endpoint estimate from 43.33% for Ordinal Conformal to 49.57–50.49%; Ordinal Cost-Shielded Conformal with the risk term disabled had the highest total-reward FQE (11.7565), whereas full RC-DQN had the highest endpoint estimate (50.49%) and action match (50.17%). On eICU-CRD, full RC-DQN had the highest endpoint and total-reward estimates (78.43% and 25.6019), while Softmax Cost-Shielded Conformal had the highest action match (52.92%). Thus, endpoint value, composite return, and behavioral agreement captured different configuration properties.

5 Discussion and Future Work

RC-DQN combines ordinal action-support estimation, dual utility–cost critics, risk-modulated conformal admissibility, and deterministic constraints. It yielded the highest in-database endpoint and composite-return FQE point estimates among evaluated policies, but margins were small and other policies had higher action match.

The cost critic and conformal selector address complementary failure modes: C_ϕ scores state–action constraint cost, whereas P_ω estimates retrospective action support. Their combination filters unsupported or elevated-cost-score actions before utility ranking, and the deterministic fallback resolves high-dose vetoes or empty admissible sets, helping explain the ablation trade-offs.

Interpretation remains dependent on offline support and FQE specification, including target-policy occupancy, function approximation, reward definition, and horizon (Rhee, 2026). The reported intervals condition on fixed policies and evaluators; cross-database analysis transports both together; and conformal calibration concerns marginal inclusion of observed actions rather than trajectory-level coverage or policy optimality (Barber et al., 2023). Future work should combine complementary off-policy estimators, state–action coverage diagnostics, repeated-seed evaluation, and preference-conditioned robust offline RL.

6 Conclusion

RC-DQN is a risk-aware, dual-critic conformal offline reinforcement learning method for dexmedetomidine dose titration in postoperative ICU patients with documented dexmedetomidine exposure. RC-DQN combines a truncated-Poisson ordinal behavior model, an efficacy critic, a threshold-regularized physiological-cost-score critic, risk-modulated conformal calibration, absolute cost-score filtering, and a rule-based fallback mechanism to limit unsupported or elevated-cost-score action selection in retrospective critical care data.

In offline evaluations using MIMIC-IV and eICU-CRD trajectories, RC-DQN produced numerically higher mean initial-state FQE estimates for the alive-without-delirium terminal outcome and numerically higher returns under the prespecified composite reward than the evaluated comparators. RC-DQN also frequently selected action 0 or lower action indices, including in states satisfying prespecified bradycardia or hypotension criteria. Bidirectional cross-database sensitivity analyses and demographic subgroup analyses provided additional but descriptive assessments under distribution shift and across demographic strata. These findings support the technical feasibility of training and retrospectively evaluating an uncertainty-aware, constrained offline RL policy for dexmedetomidine titration.

AI USE STATEMENT

During preparation of this manuscript, the authors used OpenAI ChatGPT to assist with language editing and grammatical review, organization of literature-search leads, code review and debugging, and refinement of figure and table presentation. The authors reviewed, verified, and revised all AI-assisted text, code suggestions, citations, calculations, and visual outputs against the source code, original results, and cited literature.

REPRODUCIBILITY STATEMENT

To support the reproducibility of our work, we have submitted the source code, datasets, implementation details, and necessary instructions as the Supplementary Materials. We believe that these resources provide sufficient information for reproducing the main experimental results presented in this paper.

ETHICS STATEMENT

This study used de-identified, credentialed-access data from MIMIC-IV and eICU-CRD in accordance with their respective data-use requirements. No new patient recruitment or clinical intervention was conducted. The learned policies and retrospective estimates are intended for research evaluation and not for direct clinical deployment.

REFERENCES

Philip Amortila, Audrey Huang, Akshay Krishnamurthy, and Nan Jiang. A unifying view of coverage in linear off-policy evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/858e773f5c99a4186b548f3f531fb07e-Abstract-Conference.html.

A. N. Angelopoulos and S. Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Peter C. Austin, Douglas S. Lee, and Jason P. Fine. Introduction to the analysis of survival data in the presence of competing risks. *Circulation*, 133(6):601–609, 2016. doi: 10.1161/CIRCULATIONAHA.115.017719.

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023. doi: 10.1214/23-AOS2276.

T. Chen, R. Cai, F. Wu, and X. Zhang. ACTIVE: Offline reinforcement learning via adaptive imitation and in-sample V-ensemble. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c06f788963f0ce069f5b2dbf83fe7822-Abstract-Conference.html.

Ayan Choudhury, Janaka Brahmanage, Akshat Kumar, and Praveen Paruchuri. Persistent safety set guided offline safe reinforcement learning. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 4224–4232, 2026. doi: 10.24963/ijcai.2026/470. URL https://www.ijcai.org/proceedings/2026/470.

DailyMed. Dexmedetomidine hydrochloride in 0.9% sodium chloride injection, prescribing information. National Library of Medicine, https://dailymed.nlm.nih.gov/dailymed/drugInfo.cfm?setid=f140de5a-f13d-46b4-a742-049c63e474f5, 2026. Accessed June 7, 2026.

N. Eghbali, T. Alhanai, and M. M. Ghassemi. Distribution-free uncertainty quantification in mechanical ventilation treatment: a conformal deep Q-learning framework. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 27960–27968, 2025. doi: 10.1609/aaai.v39i27.35013.

E. W. Ely, S. K. Inouye, G. R. Bernard, et al. Delirium in mechanically ventilated patients: validity and reliability of the confusion assessment method for the intensive care unit (CAM-ICU). *JAMA*, 286(21):2703–2710, 2001. doi: 10.1001/jama.286.21.2703.

S. Fujimoto, D. Meger, and D. Precup. Off-policy deep reinforcement learning without exploration. In *Proceedings of the International Conference on Machine Learning*, volume 97, pp. 2052–2062, 2019.

S. Geng, A. Pacchiano, A. Kolobov, and C.-A. Cheng. Improving offline RL by blending heuristics. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/b40cd8bb212ad8ad0a8c8a43c4da5f0b-Abstract-Conference.html.

O. Gottesman, F. Johansson, M. Komorowski, et al. Guidelines for reinforcement learning in healthcare. *Nature Medicine*, 25(1):16–18, 2019. doi: 10.1038/s41591-018-0310-5.

Botao Hao, Xiang Ji, Yaqi Duan, Hao Lu, Csaba Szepesvari, and Mengdi Wang. Bootstrapping fitted Q-evaluation for off-policy inference. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 4074–4084. PMLR, 2021. URL https://proceedings.mlr.press/v139/hao21b.html.

A. E. W. Johnson, L. Bulgarelli, L. Shen, et al. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10:1, 2023. doi: 10.1038/s41597-022-01899-x.

I. Kostrikov, A. Nair, and S. Levine. Offline reinforcement learning with implicit Q-learning. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=68n2s9ZJWF8.

A. Kumar, A. Zhou, G. Tucker, and S. Levine. Conservative Q-learning for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 1179–1191, 2020.

H. Y. Lee, S. Chung, J. Lee, et al. Reinforcement learning model for optimizing dexmedetomidine dosing to prevent delirium in critically ill patients. *npj Digital Medicine*, 7:325, 2024. doi: 10.1038/s41746-024-01335-x.

K. Lewis, F. Alshamsi, K. L. Carayannopoulos, et al. Dexmedetomidine vs other sedatives in critically ill mechanically ventilated adults: a systematic review and meta-analysis of randomized trials. *Intensive Care Medicine*, 48(7):811–840, 2022. doi: 10.1007/s00134-022-06712-2.

K. Lewis, M. C. Balas, J. L. Stollings, et al. A focused update to the clinical practice guidelines for the prevention and management of pain, anxiety, agitation/sedation, delirium, immobility, and sleep disruption in adult patients in the ICU. *Critical Care Medicine*, 53(3):e711–e727, 2025. doi: 10.1097/CCM.0000000000006574.

Anan Li, Haopeng Wu, Kai Zhang, and Jianzhi Ying. Efficacy and hemodynamic safety of perioperative dexmedetomidine for the prevention of postoperative delirium in elderly patients undergoing cardiac surgery: an updated systematic review and meta-analysis of randomized controlled trials. *Frontiers in Neurology*, 17:1895312, 2026. doi: 10.3389/fneur.2026.1895312. URL https://www.frontiersin.org/journals/neurology/articles/10.3389/fneur.2026.1895312/full.

Huidong Liu, Hang Yu, Qiyang Zhang, Jiarui Dou, Xianlei Long, Jiantao Shi, and Fuqiang Gu. Uncertainty-guided adaptive conservative offline reinforcement learning for safer mechanical ventilation. In *Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence*, pp. 6788–6796, 2026. doi: 10.24963/ijcai.2026/755. URL https://www.ijcai.org/proceedings/2026/755.

M. Luney, L. Holdsworth, A. Hagana, V. Kaminskaite, L. Qiu, J. Berner, K. Pattinson, J. Nanchahal, G. S. Collins, and M. Costa. Effectiveness of drug interventions to prevent delirium after surgery for older adults: systematic review and network meta-analysis of randomised controlled trials. *BMJ*, 392:e085539, 2026. doi: 10.1136/bmj-2025-085539.

Animesh Kumar Paul and Russell Greiner. Learning dynamic binary treatment policies under treatment selection bias: A conservative Q-learning approach with representation balancing. *Journal of Biomedical Informatics*, 176:104994, 2026. doi: 10.1016/j.jbi.2026.104994. URL https://doi.org/10.1016/j.jbi.2026.104994.

T. J. Pollard, A. E. W. Johnson, J. D. Raffa, L. A. Celi, R. G. Mark, and O. Badawi. The eICU collaborative research database, a freely available multi-center database for critical care research. *Scientific Data*, 5:180178, 2018. doi: 10.1038/sdata.2018.178.

R. F. Prudencio, M. R. O. A. Maximo, and E. L. Colombini. A survey on offline reinforcement learning: taxonomy, review, and open problems. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8):10237–10257, 2024. doi: 10.1109/TNNLS.2023.3250269.

Feier Qiu, Ying Chen, Xiuxian Wang, Na Geng, and Zhitao Yang. Clinician-informed offline reinforcement learning for vasopressor administration optimization in shock management. *Engineering Applications of Artificial Intelligence*, 167:113818, 2026. doi: 10.1016/j.engappai.2026.113818. URL https://doi.org/10.1016/j.engappai.2026.113818.

Kihun Rhee. Confounding masquerading as improvement: A systematic evaluation of offline reinforcement learning for stroke antithrombotic treatment in a 129,000-patient registry. In *Proceedings of the 11th Machine Learning for Healthcare Conference*, volume 340 of *Proceedings of Machine Learning Research*, pp. 1689–1725. PMLR, 2026. URL https://proceedings.mlr.press/v340/rhee26a.html.

C. N. Sessler, M. S. Gosnell, M. J. Grap, et al. The richmond agitation-sedation scale: validity and reliability in adult intensive care unit patients. *American Journal of Respiratory and Critical Care Medicine*, 166(10):1338–1344, 2002. doi: 10.1164/rccm.2107138.

Mark van den Boogaard, Florien Ham, Ingeborg van Dusseldorp, et al. The dutch clinical practice guideline: Pain, analgesia/sedation, delirium, immobility and sleep disturbances; PADIS. *Intensive and Critical Care Nursing*, 95:104438, 2026. doi: 10.1016/j.iccn.2026.104438. URL https://pubmed.ncbi.nlm.nih.gov/42090788/.

H. van Hasselt, A. Guez, and D. Silver. Deep reinforcement learning with double Q-learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016. doi: 10.1609/aaai.v30i1.10295. URL https://ojs.aaai.org/index.php/AAAI/article/view/10295.

Vladimir Vovk. Universality of conformal prediction under the assumption of randomness. In *Proceedings of the 37th International Conference on Algorithmic Learning Theory*, volume 313 of *Proceedings of Machine Learning Research*, pp. 1–18. PMLR, 2026. URL https://proceedings.mlr.press/v313/vovk26a.html.

Jiayi Wang, Zhengling Qi, and Raymond K. W. Wong. A fine-grained analysis of fitted Q-evaluation: Beyond parametric models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 51273–51302. PMLR, 2024. URL https://proceedings.mlr.press/v235/wang24be.html.

Qingjun Wang, Hongtu Zhou, Hang Yu, Junqiao Zhao, Yanping Zhao, Chen Ye, Ziqiao Wang, and Guang Chen. Beyond penalization: Diffusion-based out-of-distribution detection and selective regularization in offline reinforcement learning. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/3a2dbdf0aca2006f4751abf0804620c0-Abstract-Conference.html.

R. Yang, H. Zhong, J. Xu, A. Zhang, C. Zhang, L. Han, and T. Zhang. Towards robust offline reinforcement learning under diverse data corruption. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/43d7bc009cf5171e7af77a91ee4bb890-Abstract-Conference.html.

Ruiqi Zhang, Xuezhou Zhang, Chengzhuo Ni, and Mengdi Wang. Off-policy fitted Q-evaluation with differentiable function approximators: Z-estimation and inference theory. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26713–26749. PMLR, 2022. URL https://proceedings.mlr.press/v162/zhang22al.html.

Yingying Zhang, Chengchun Shi, and Shikai Luo. Conformal off-policy prediction. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 2751–2768. PMLR, 2023. URL https://proceedings.mlr.press/v206/zhang23c.html.

Yujia Zhang, Lin Li, Wei Wei, Jianguo Wu, Yi Ma, and Jiye Liang. Efficient offline reinforcement learning via peer-influenced constraint. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/6f95a050a7eedab83ae7c27bae38e2e0-Abstract-Conference.html.

A SUPPLEMENTARY MATERIAL

A.1 DATA

This study utilized two high-quality, publicly available intensive care databases: MIMIC-IV (Medical Information Mart for Intensive Care IV) and eICU-CRD (eICU Collaborative Research Database)(Johnson et al., 2023; Pollard et al., 2018). Both databases have undergone rigorous de-identification and peer-reviewed validation. The target cohort consisted of postoperative critically ill patients with documented dexmedetomidine exposure during the ICU stay. Specifically, the inclusion criteria applied during the data extraction phase were: (1) adult patients aged $\geq$ 18 years; (2) patients who underwent surgical treatment and were admitted to the ICU within 24 hours after surgery; and (3) patients with at least one documented dexmedetomidine administration record during the ICU stay. The resulting processed RL cohorts contained 3,988 ICU-stay episodes and 68,401 six-hour decision windows from MIMIC-IV and 754 episodes and 16,866 decision windows from eICU-CRD.

A.2 DATA PREPARATION

Data preparation was performed separately for MIMIC-IV and eICU-CRD using a common pipeline. The process involved temporal aggregation, construction of the prespecified state and action spaces, patient-level splitting, missing-data handling, and training-split-based standardization.

1. Temporal Segmentation. We divided each patient's ICU stay into consecutive 6-hour windows anchored to the time of ICU admission. For point-based records (e.g., laboratory tests, vital signs), values within each window were aggregated using mean or sum. For interval-based records (e.g., mechanical ventilation, vasopressor infusion, dexmedetomidine administration), recorded time spans were distributed across overlapping windows with values allocated proportionally by duration of overlap. All windowed feature tables were merged on patient identifiers and window indices.

2. State and Outcome Variables. The final 18-dimensional state vector comprised sex, age, age-adjusted Charlson Comorbidity Index (aCCI), estimated mortality risk, postoperative delirium risk, Acute Physiology Score (APS), Richmond Agitation-Sedation Scale (RASS) (Sessler et al., 2002), CAM-ICU (Ely et al., 2001), Sequential Organ Failure Assessment (SOFA), Glasgow Coma Scale (GCS), systolic blood pressure, mean arterial pressure, body mass index, heart rate, urine output in ml/kg/h, mechanical-ventilation status, renal-replacement-therapy type, and vasoactive-drug use. ICU mortality, 30- and 90-day survival information, ventilation duration, delirium outcome, ICU length of stay, and post-ICU hospital stay were retained separately as endpoint variables and were not included in the model state.

3. Splitting, Missing Data, and Standardization. ICU stays were randomly partitioned at the patient level into training, validation, and test sets in proportions of 70%, 15%, and 15%, respectively, so that windows from one stay could not occur in multiple splits. MIMIC-IV contained 2,791/598/599 episodes and 47,765/10,182/10,454 decision windows in the training/validation/test splits; the corresponding eICU-CRD counts were 527/113/114 episodes and 11,800/2,483/2,583 windows. Columns with at least 95% missingness were removed. Variables with less than 30% missingness were imputed using K-nearest neighbors ($k = 3$). Variables with 30%–95% missingness were forward-filled within each patient; when the first value in a trajectory was missing, it was initialized using the global column mean. Any remaining missing values were replaced by the corresponding column mean. State and total-reward standardizers were fitted only on the training split of each database and then reused for that database's validation and test splits. Reward construction itself used the unstandardized aggregated clinical measurements.

A.3 RL PROBLEM FORMULATION

We formulated the decision-making process for postoperative dexmedetomidine administration as a Markov Decision Process (MDP), defined by the tuple$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$:

State: The state $s_t \in \mathcal{S}$ is the standardized 18-dimensional vector defined above for a 6-hour decision window. It contains baseline risk and severity information together with contemporaneous se-

dation, neurological, hemodynamic, renal, respiratory-support, and vasoactive-treatment variables. Trajectory-level outcomes were stored separately and were not used as state inputs.

Action: The action $a_t \in \mathcal{A}$ denotes the dexmedetomidine infusion-rate category. The action space contains 16 ordered nonnegative treatment levels: action 0 represents no recorded dexmedetomidine infusion in the current 6-hour window, whereas actions 1–15 represent increasing nonzero dose bins obtained by clipping the infusion rate to [0.1, 1.5] μg·kg^{-1}·h^{-1} and discretizing it into 15 ordered intervals. Action 0 was therefore modeled as the lowest level of the same ordered treatment variable rather than as an unrelated categorical intervention. Reward: The revised prespecified composite reward separates nonterminal clinical-state rewards from a mutually exclusive terminal outcome reward. For a nonterminal transition (s_t, a_t, s_{t+1}), the raw composite reward is

$$r_t^{\text{raw}} = r_t^{\text{CAM}} + r_t^{\text{RASS}} + r_t^{\text{organ}} + r_t^{\text{side}},$$

where all components except the SOFA-change term are evaluated from the unstandardized measurements in s_{t+1}. A negative CAM-ICU screen receives $+1$, a positive screen receives -5, and values other than 0 or 1 receive 0. RASS in the target interval $[-2, 1]$ receives $+0.5$ and RASS outside this interval receives -0.5; missing CAM-ICU or RASS values contribute 0. The organ/support component applies additive penalties of -0.1 per point of SOFA increase from s_t to s_{t+1}, -0.5 for GCS ≤ 8, -0.3 for urine output < 0.5 ml/kg/h, -0.1 for active mechanical ventilation, -0.2 for active renal replacement therapy, and -0.1 for active vasoactive-drug treatment. The side-effect component applies -2 for heart rate < 50 beats/min and one additional -2 penalty when systolic blood pressure is < 90 mmHg or mean arterial pressure is < 65 mmHg.

At the final transition, no nonterminal component is added. The terminal reward depends only on the trajectory-level delirium and ICU-death indicators: it is -30 for ICU death, -20 for delirium without prior ICU death, and $+20$ for being alive without delirium; death takes precedence when both terminal indicators are present. The numerical reward weights were prespecified by clinical experts to encode the intended relative priorities among delirium, sedation, organ-support, hemodynamic-tolerance, and terminal outcomes. They should be interpreted as an expert-defined reward design rather than a validated cardinal clinical-utility scale. The resulting raw total reward is standardized using the training-split reward distribution before utility-critic training, and the same transformation is applied to validation and test rewards. The safety cost used to train the physiological-cost-score critic is not the standardized total reward. It is derived only from the unstandardized side-effect component as $c_t = \max(-r_t^{\text{side}}, 0)$, so terminal, sedation, delirium, and organ/support rewards do not enter the cost-score target.

A.4 Additional Model-design and Implementation Details

The architecture was designed to address three limitations that are especially problematic for medication dosing in offline critical-care data. First, treating ordered dose levels as unrelated softmax classes does not encode their ordinal structure and permits irregular probability changes across neighboring, sparsely observed dose bins. Second, collapsing efficacy and adverse-event penalties into a single scalar reward can allow positive long-term utility to mask infrequent but safety-related cost signals. Third, a conformal filter based only on historical action probability may inherit unsafe historical practice when a frequently observed action is risky in a specific physiological state. RC-DQN therefore organizes the model around three linked modules: a truncated-Poisson ordinal behavior model that introduces parameter sharing across the ordered action support, a dual-critic efficacy-cost evaluator for risk separation, and a risk-modulated conformal selector with deterministic veto-and-fallback rules for offline action selection.

The truncated-Poisson head introduces parameter sharing and a unimodal ordinal inductive bias across the 16 treatment levels. Its probabilities are used for negative log-likelihood training and subsequent conformal scoring. This is a discriminative behavior-policy parameterization for structured action-support estimation, not a claim that clinician dosing decisions follow a Poisson data-generating process or that the head is a validated generative model of clinical dosing behavior.

The dual-critic evaluator separates efficacy and physiological cost at the network-output level while retaining a shared state encoder. The two critics are trained with different targets and discount factors, and the cost TD error is weighted asymmetrically. Both critic losses and the cost-threshold regularizer enter the joint training objective; therefore, although the output heads are distinct, their gradients can interact through the shared representation. Consequently, C_ϕ is used as a constraint

and ranking score rather than interpreted as a calibrated adverse-event probability or an unbiased clinical-risk estimate.

The risk-modulated conformal layer first constructs a set of actions satisfying both the nonconformity threshold and an absolute cost-score threshold. It applies risk-adjusted utility scoring only within this set and then enforces prespecified conservative constraints: suspicious high-dose candidates are vetoed, and states with no admissible action fall back to action 0. This makes the learned policy a constrained offline decision rule rather than a raw neural-network argmax.

The auxiliary term $\mathcal{L}_{\mathrm{Risk}}$ is a predicted-cost threshold regularizer included in the joint loss with coefficient λ_{safe}; minimizing it pushes observed-action cost scores above d_{safe} downward toward the threshold. This term is not an additive penalty on $\mathcal{L}_{\mathrm{TD},Q}$ and is not a learned Lagrange multiplier. Nevertheless, it can indirectly affect the efficacy critic because the Q and cost heads share an encoder. The cost critic also influences efficacy learning through the bootstrap action a^*: C_{ϕ^-} helps determine which next action supplies the target Q-value, although target construction is performed without gradient propagation. Thus, physiological-cost constraints operate during training through the joint objective, shared representation, and cost-adjusted Bellman action selection, and during inference through risk-modulated conformal calibration, the absolute cost-score shield, cost-penalized ranking, and veto-and-fallback rules.

Observed clinician actions provide information about behavioral support but do not establish the counterfactual optimality or safety of those actions in every patient state. Under exchangeability, the conformal statement concerns marginal inclusion of observed behavior actions according to the specified nonconformity score. The present component constructs action-admissibility sets and is distinct from conformal off-policy methods designed to form prediction intervals for target-policy returns (Zhang et al., 2023). Conventional conformal validity depends on the exchangeability structure used for calibration; repeated decision windows within an ICU trajectory and cross-database distribution shift therefore require separate consideration (Barber et al., 2023).

A.5 RC-DQN ALGORITHM

Algorithm 1 summarizes the three operational phases of RC-DQN: joint network training on offline transitions, conformal calibration on held-out state–action pairs, and constrained action selection for a new state. Data preprocessing and FQE are excluded because they are respectively upstream data construction and downstream policy evaluation procedures.

Algorithm 1 RC-DQN pseudocode

Require: Training transitions $\mathcal{D}_{\text{tr}}$; calibration pairs $\mathcal{D}_{\text{cal}}$; action set $\mathcal{A} = \{0, \dots, 15\}$
Require: $\gamma_Q, \gamma_C, \lambda_{\text{nll}}, \lambda_{\text{safe}}, \lambda_{\text{os}}, d_{\text{safe}}, \tau, \alpha$
Ensure: Trained $Q_\theta, C_\phi, P_\omega$; conformal threshold τ_{nc}; constrained policy π

function TRAIN($\mathcal{D}_{\text{tr}}$)
  Initialize online networks $Q_\theta, C_\phi, P_\omega$ and target networks Q_{θ^-}, C_{ϕ^-}
  for each optimization step **do**
    Sample minibatch $\mathcal{B} = \{(s_i, a_i, r_i, c_i, s'_i, d_i)\}$ from $\mathcal{D}_{\text{tr}}$
    for each transition $i \in \mathcal{B}$ **do**
      $a_i^* \leftarrow \arg\max_{a \in \mathcal{A}} [Q_{\theta^-}(s'_i, a) - \lambda_{\text{safe}} C_{\phi^-}(s'_i, a)]$
      $y_i^Q \leftarrow r_i + \gamma_Q (1 - d_i) Q_{\theta^-}(s'_i, a_i^*)$
      $y_i^C \leftarrow c_i + \gamma_C (1 - d_i) C_{\phi^-}(s'_i, a_i^*)$
      $w_i \leftarrow \tau$ if $y_i^C > C_\phi(s_i, a_i)$, and $1 - \tau$ otherwise
    end for
    $\mathcal{L}_Q \leftarrow |\mathcal{B}|^{-1} \sum_i \text{Huber}(Q_\theta(s_i, a_i) - y_i^Q)$
    $\mathcal{L}_C \leftarrow |\mathcal{B}|^{-1} \sum_i w_i \, \text{Huber}(C_\phi(s_i, a_i) - y_i^C)$
    $\mathcal{L}_{\text{NLL}} \leftarrow -|\mathcal{B}|^{-1} \sum_i \log P_\omega(a_i \mid s_i)$
    $\mathcal{L}_{\text{Risk}} \leftarrow |\mathcal{B}|^{-1} \sum_i \max(C_\phi(s_i, a_i) - d_{\text{safe}}, 0)$
    $\mathcal{L} \leftarrow \mathcal{L}_Q + \mathcal{L}_C + \lambda_{\text{nll}} \mathcal{L}_{\text{NLL}} + \lambda_{\text{safe}} \mathcal{L}_{\text{Risk}}$
    Update all trainable parameters by descending $\nabla \mathcal{L}$
    Update $(\theta^-, \phi^-) \leftarrow$ TARGETUPDATE$(\theta^-, \phi^-, \theta, \phi)$
  end for
  return $Q_\theta, C_\phi, P_\omega$
end function

function CALIBRATE($\mathcal{D}_{\text{cal}}$)
  Freeze $Q_\theta, C_\phi, P_\omega$ and set $n \leftarrow |\mathcal{D}_{\text{cal}}|$
  for each observed pair $(s_i, a_i) \in \mathcal{D}_{\text{cal}}$ **do**
    $u_i \leftarrow 1 - P_\omega(a_i \mid s_i) + \lambda_{\text{os}} C_\phi(s_i, a_i)$
  end for
  $q \leftarrow \lceil (n + 1)(1 - \alpha) \rceil / n$
  $\tau_{\text{nc}} \leftarrow$ QUANTILE$(\{u_i\}_{i=1}^n, q)$
  return τ_{nc}
end function

function SELECTACTION(s, τ_{nc})
  for each $a \in \mathcal{A}$ **do**
    $q_a \leftarrow Q_\theta(s, a)$; $c_a \leftarrow C_\phi(s, a)$; $p_a \leftarrow P_\omega(a \mid s)$
    $u_a \leftarrow 1 - p_a + \lambda_{\text{os}} c_a$
  end for
  $\mathcal{A}_{\text{adm}} \leftarrow \{a \in \mathcal{A} : u_a \le \tau_{\text{nc}} \text{ and } c_a \le d_{\text{safe}}\}$
  if $\mathcal{A}_{\text{adm}} = \emptyset$ **then**
    return 0
  end if
  $a_{\text{cand}} \leftarrow \arg\max_{a \in \mathcal{A}_{\text{adm}}} (q_a - \lambda_{\text{safe}} c_a)$
  if $a_{\text{cand}} \ge 14$ **and** $(p_{a_{\text{cand}}} < 0.05$ **or** $c_{a_{\text{cand}}} > 0.5 d_{\text{safe}})$ **then**
    return 0
  end if
  return a_{cand}
end function

$\tau_{\text{nc}} \leftarrow$ CALIBRATE$(\mathcal{D}_{\text{cal}})$
$\pi(s) \leftarrow$ SELECTACTION(s, τ_{nc})

A.6 Complete Cross-database Sensitivity Results

Table 3: Source MIMIC-IV policy–FQE pairs applied to eICU-CRD test states.

Policy	Endpoint, % (95% CI)	Total FQE	Match (%)
DDQN	33.27 (29.20–37.88)	3.2470	41.50
CQL	43.29 (39.48–47.10)	7.9533	47.43
BCQ	43.16 (39.37–46.94)	7.7404	46.67
HUBL	43.23 (39.49–47.13)	8.5356	43.52
RIQL	43.12 (39.02–47.29)	8.2197	41.62
ACTIVE-IQL	44.22 (39.88–48.50)	8.1623	43.94
ConformalDQN	43.63 (39.32–48.10)	5.6772	11.65
RC-DQN	44.96 (40.90–49.18)	9.1454	39.30

Table 4: Source eICU-CRD policy–FQE pairs applied to MIMIC-IV test states.

Policy	Endpoint, % (95% CI)	Total FQE	Match (%)
DDQN	79.25 (77.67–80.92)	22.9854	51.87
CQL	85.20 (84.27–86.09)	28.3088	37.10
BCQ	83.56 (82.48–84.60)	22.3901	36.30
HUBL	84.04 (83.12–84.95)	28.3927	35.10
RIQL	84.04 (83.08–84.96)	24.3462	38.44
ACTIVE-IQL	77.01 (75.21–78.81)	6.1692	19.10
ConformalDQN	84.79 (83.91–85.71)	25.5927	38.37
RC-DQN	85.88 (84.88–86.84)	28.8888	36.59

A.7 Action Distributions and Physiological Strata

Table 5: Action distributions in prespecified physiological-risk states.

Data	Policy	Action 0	Actions 1–6	Actions ≥ 7
MIMIC	Clinician	52.60%	35.42%	11.98%
	DDQN	59.90%	35.42%	4.69%
	CQL	79.17%	20.83%	0.00%
	BCQ	86.46%	10.94%	2.60%
	HUBL	69.79%	27.60%	2.60%
	RIQL	75.00%	20.83%	4.17%
	ACTIVE-IQL	90.10%	8.85%	1.04%
	ConformalDQN	0.00%	90.10%	9.90%
	RC-DQN	72.92%	19.79%	7.29%
eICU	Clinician	37.29%	62.71%	0.00%
	DDQN	76.27%	15.25%	8.47%
	CQL	27.12%	72.88%	0.00%
	BCQ	38.98%	61.02%	0.00%
	HUBL	27.12%	72.88%	0.00%
	RIQL	44.07%	55.93%	0.00%
	ACTIVE-IQL	30.51%	8.47%	61.02%
	ConformalDQN	0.00%	93.22%	6.78%
	RC-DQN	55.93%	44.07%	0.00%

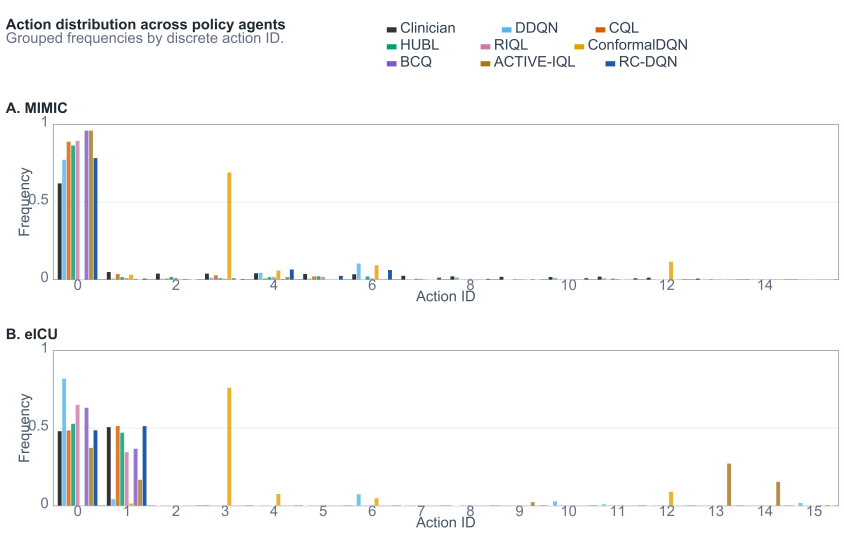

Figure 2: Decision-window action distributions for clinician and learned policies on MIMIC-IV and eICU-CRD.

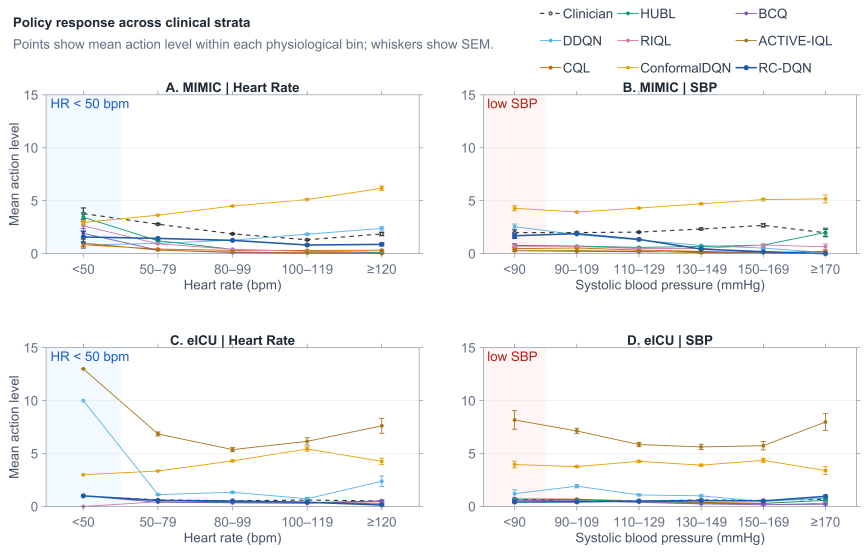

Figure 3: Mean action level across heart-rate and systolic-blood-pressure strata. Points indicate means, whiskers indicate SEM, and shaded regions identify prespecified strata.

A.8 Demographic Subgroup Analysis

Within sex and age strata, RC-DQN endpoint FQE minus the observed proportion was 3.5 percentage points (95% CI -0.2 to 7.0; $N = 572$) overall on MIMIC-IV, 4.7 (0.3 to 8.8) for males, 0.5 (-7.0 to 7.4) for females, 4.8 (-0.3 to 10.1) below age 65, and 1.8 (-3.8 to 7.0) at age 65 or older. On eICU-CRD, the overall difference was 2.3 (-5.6 to 10.1; $N = 113$), with corresponding subgroup estimates of 0.7 (-7.9 to 10.4), 5.6 (-7.9 to 20.6), 3.0 (-7.1 to 15.1), and 1.7 (-8.9 to 12.4). Most intervals included zero; these contrasts do not establish fairness, subgroup benefit, or effect homogeneity.

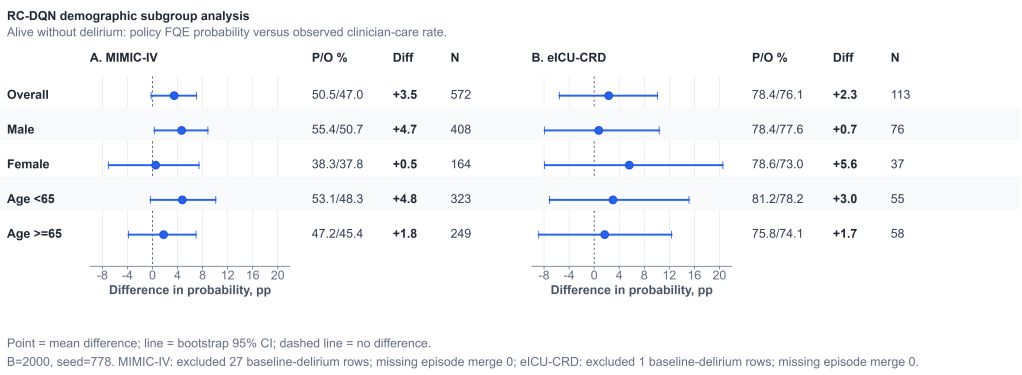

Figure 4: Demographic subgroup analysis. Points show endpoint-FQE minus observed proportions; whiskers are episode-bootstrap 95% intervals conditional on fixed policy and FQE.

A.9 Expanded Ablation Interpretation

Ordinal Conformal retained A but omitted B and C. Softmax Cost-Shielded Conformal replaced A with categorical logits, retained B, and used behavior-only conformal filtering. Ordinal Cost-Shielded Conformal retained A and B but disabled C; it reused full-model weights and recalibrated the behavior-only threshold from $1 - P(a_i \mid s_i)$ on validation data at the stored confidence level. Full RC-DQN included A, B, and C. Because these configurations change linked components and lack pairwise confidence intervals, they do not identify independent module effects.