Extended Abstract

Motivation Generalizing reinforcement learning methods across hospital systems presents a critical challenge for clinical decision support, as patient populations, treatment protocols, and data collection practices vary substantially between institutions. While RL has demonstrated potential for healthcare applications, most existing research focuses on single-site datasets, providing limited insight into real-world policy transferability. This work addresses the gap by benchmarking deep reinforcement learning algorithms across distinct hospital systems to evaluate cross-institutional generalization for the clinically relevant task of potassium repletion in ICU patients.

Method We construct standardized patient cohorts from MIMIC-IV (Beth Israel Deaconess Medical Center) and STARR-OMOP (Stanford Health Care) datasets, focusing on ICU patients who received intravenous potassium chloride and had recorded potassium measurements. We formulate potassium repletion as a Markov Decision Process with 4-hour time windows, discrete action space of KCl doses (0, 10, 20, 40 mEq), and state features including patient demographics, vital signs, and laboratory values. The reward function uses a Gaussian centered at the clinical target potassium level (4.25 mmol/L). We evaluate three algorithms: Behavior Cloning (BC) for clinician policy modeling, Fitted Q-Iteration (FQI) for value-based learning, and Conservative Q-Learning (CQL) for safe offline RL with distributional constraints.

Implementation All methods utilize identical two-layer MLP architectures with 256 ReLU units, trained with Adam optimization and inverse-frequency weighting to address severe class imbalance (82-85% of timesteps involve no KCl administration). Missing laboratory values are imputed using bin-based interpolation across adjacent 4-hour windows. Training employs soft target network updates and validation-based early stopping.

Results Off-policy evaluation using weighted importance sampling reveals distinct algorithmic behaviors across institutions. CQL demonstrates closer alignment with clinician practice, learning policies with mean doses similar to observed clinical behavior. In contrast, FQI exhibits substantial deviation toward higher dosing strategies, with learned policies favoring doses significantly above clinician norms. BC shows intermediate performance with dataset-dependent alignment. Importance sampling weight distributions indicate that CQL maintains better distributional support, while FQI generates policies that frequently select actions rarely observed in clinical practice, leading to clipped importance weights and potentially unreliable value estimates.

Discussion The observed differences in learned policies highlight important considerations for cross-institutional RL deployment. CQL's conservative regularization successfully prevents value overestimation for out-of-distribution actions, maintaining clinical alignment across both datasets. FQI's tendency toward aggressive dosing likely stems from Q-value overestimation amplified by class imbalance weighting, demonstrating the risks of unconstrained offline RL in clinical settings. The variation in baseline patient characteristics between institutions suggests that hospital-specific factors influence policy learning and may require domain adaptation techniques for successful generalization.

Conclusion This work demonstrates that algorithm choice significantly influences both safety and generalizability in clinical RL applications. While conservative approaches like CQL show promise for maintaining clinical alignment across institutions, the observed performance differences indicate that further research is necessary to develop robust evaluation frameworks for multi-institutional healthcare data. Our findings suggest that successful clinical RL deployment is a question worth investigating if we want to use RL for healthcare, requiring continued development of domain adaptation techniques and standardized benchmarking protocols that balance optimization objectives with practical medical constraints.

Cross-Institution RL Benchmarking for Non-Synthetic Clinical Settings

Kalyani Limaye
Institute of Computational and Mathematical Engineering
Stanford University
limayk@stanford.edu

Abstract

This study presents a comprehensive benchmarking of deep reinforcement learning algorithms for clinical decision support across distinct hospital systems. We specifically examine potassium repletion using real-world data from MIMIC-IV and STARR-OMOP. We construct standardized patient cohorts and evaluate three algorithms—Behavior Cloning (BC), Fitted Q-Iteration (FQI), and Conservative Q-Learning (CQL)—using weighted importance sampling for off-policy evaluation. Our results reveal some algorithmic differences in policies learned across institutions: CQL demonstrates closer alignment with clinician behavior across both datasets, while FQI exhibits notable deviation toward higher dosing strategies. These findings highlight important considerations for generalizing reinforcement learning policies across healthcare institutions and suggest the value of conservative approaches for clinical applications where distributional shift and safety considerations are relevant.

1 Introduction

The challenge of cross-institutional generalization in healthcare RL extends beyond simple domain adaptation, encompassing fundamental questions about the transferability of learned medical decision-making policies. Hospital systems differ not only in their patient demographics and clinical protocols but also in their electronic health record systems, laboratory measurement frequencies, and institutional treatment philosophies. These variations create a complex landscape where a policy optimized for one institution may perform poorly or even unsafely when applied to another. Furthermore, the high-stakes nature of clinical decision-making amplifies the consequences of poor generalization, as suboptimal policies could directly impact patient outcomes. Understanding these generalization challenges is crucial for the practical deployment of RL-based clinical decision support systems, as any real-world implementation must demonstrate robustness across diverse healthcare environments rather than excellence within a single, controlled setting. While RL has demonstrated potential for clinical decision support, much of the existing research focuses on optimizing policies within single-site datasets, providing limited insight into real-world variability.

This project proposes benchmarking DRL algorithms across distinct hospital systems — specifically, MIMIC-IV (Beth Israel Deaconess Medical Center) and STARR-OMOP (Stanford Health Care) — by constructing standardized patient cohorts across datasets. By controlling for MDP structure, we aim to enable consistent evaluation across datasets. We further seek to analyze how algorithm performance shifts between institutions, identifying critical barriers to robust policy generalization.

Stanford CS224R 2025 Final Report

2 Related Work

Prior work in healthcare reinforcement learning has primarily focused either on single-institution real-world datasets or on synthetic environments. Raghu et al. (2018) applied model-based RL to optimize sepsis treatment strategies using real-world ICU data from MIMIC-III. However, evaluations were restricted to a single hospital system, leaving questions about policy transferability across institutions unexplored.

While some early efforts toward standardization exist, such as the evaluation guidelines proposed by Gottesman et al. (2019), broader benchmarking for reinforcement learning in healthcare remains limited. More recently, Hargrave et al. (2024) introduced the EpiCare benchmark to standardize offline RL evaluation for clinical decision-making, but their framework operates on simulated patient trajectories rather than real clinical records, limiting insights into real-world variability.

Our project aims to address both of these gaps by benchmarking RL methods on real-world clinical datasets and explicitly evaluating policy generalization across distinct hospital systems.

More recently, Prasad et al. (2022) introduced an RL-based framework for guiding electrolyte replacement in critical care, centering on efficient, effective, and patient-oriented policies. We adopt their task formulation and a similarly simplified MDP structure for potassium repletion, but whereas Prasad et al. focus on cost-efficiency and resource optimization within a single institution, our work emphasizes benchmarking and explicitly generalizing learned policies across distinct hospital systems.

3 Experimental Setup

3.1 Markov Decision Process

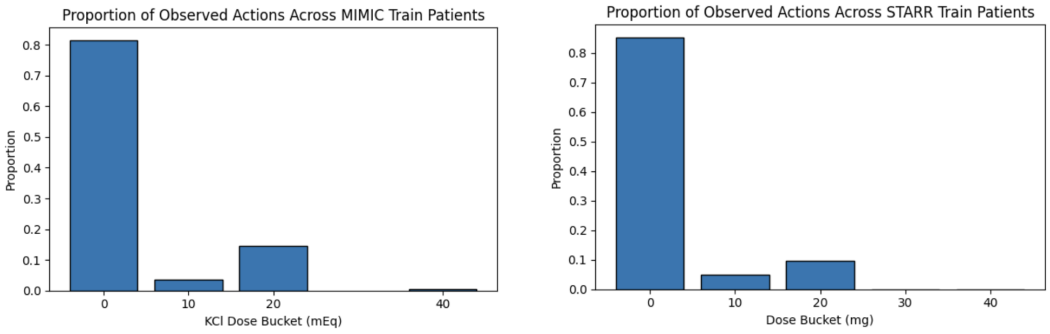
We study the task of potassium repletion Prasad et al. (2022) using MIMIC IV Johnson et al. (2024) (Beth Israel Deaconess Medical Center) and STARR-OMOP (Stanford Health Care) datasets. For each dataset, we construct a cohort of hospital stays in which patients (i) received at least one intravenous potassium-chloride (KCl) dose and (ii) had at least one recorded potassium lab (serum or whole-blood). For this study, we sample 200 patients from each cohort.

Time discretisation. Every hospital stay that passes the cohort filter is sliced into fixed, non-overlapping 4-hour windows. To avoid trajectories that are almost entirely missing signal, we discard stays with (i) fewer than six valid windows, or (ii) a usable potassium measurement in $< 60\%$ of windows **and** a non-zero-dose frequency below $\frac{1}{12}$ (i.e. less than one intervention per 48 h).

Action space Our action is intravenous administration of potassium chloride (KCl), which we discretize into the clinically representative bins

$$\mathcal{A} = \{0,\ 10,\ 20,\ 40\}\ \text{mEq}.$$

Timesteps with no KCl administration action default to the 0 mEq bin.



(a) MIMIC cohort (n=3498) (b) STARR cohort (n=7192)

Figure 1: Empirical proportions of observed KCl dose buckets in the training data.

As seen in the plots above, both cohorts exhibit a pronounced skew toward the zero-dose bin (here n = number of stays). In MIMIC, 82% of timesteps receive no KCl, about 3% receive 10 mEq, 14% receive 20 mEq, and under 1% receive 40 mEq. In STARR, 85% of timesteps are zero-dose, 5% are 10 mEq, 9% are 20 mEq, and almost none are 40 mEq. This extreme imbalance arises because every interval without a recorded drug is encoded as the 0 mEq "do nothing" action, making zero-dose by far the most common choice. Such skew forces naïve learners to default to the majority class and motivates our use of oversampling or weighted-loss techniques to ensure that rarer but clinically important dosing actions are properly learned.

State space The feature vector for window t is $s_t = [\text{age}, \text{sex}, \text{weight}, K_{\text{pre}}, \text{extra covariates}]$, where K_{pre} is the latest potassium lab before the action. Extra covariates include heart rate, systolic blood pressure, diastolic blood pressure, respiratory rate, serum creatinine, and serum anion gap.

Missing lab values within the primary time interval $[t,\, t + \Delta t)$ are imputed using a bin-based interpolation procedure. When no observation occurs in the target window, we iteratively examine up to three adjacent bins of width Δt,

$$[t + k\,\Delta t,\ t + (k+1)\,\Delta t) \quad k \in \{0, \pm 1, \pm 2\}$$

This search spans ± 3 bins (i.e. ± 12 h) to accommodate the typical once-daily frequency of potassium measurements. Within each bin, we compute the arithmetic mean of all recorded values, adopting the first non-empty bin mean as the imputed value. If all evaluated bins are empty, the value is set to 0.0 to denote absence. This approach preserves temporal proximity, leverages local measurement density to mitigate interpolation bias, and explicitly flags missing daily measurements.

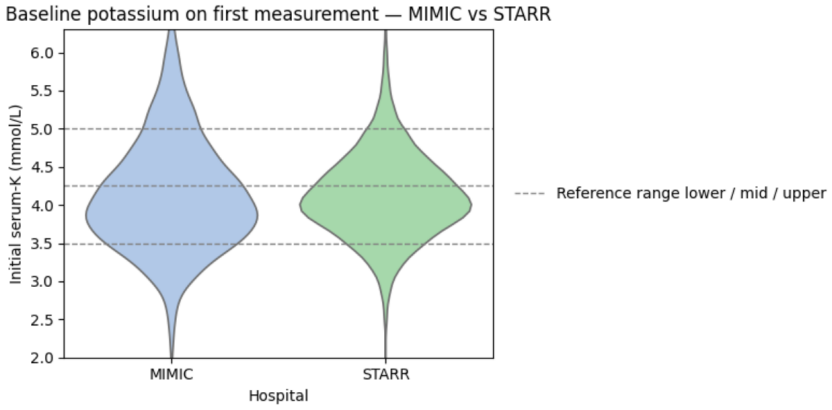


Figure 2: Violin plots of the first measured serum potassium value per stay in MIMIC and STARR, with dashed lines indicating the laboratory reference range (3.5, 4.25, and 5.0 mmol/L).

The plot above shows initial distribution of measured potassium accross patient stays. Both cohorts exhibit a central tendency near 4.0 mmol/L, but the MIMIC distribution is noticeably wider, with more patients below 3.5 mmol/L and above 5.0 mmol/L. In contrast, STARR values are more tightly concentrated within the reference range, suggesting fewer extreme hypo- or hyperkalemic cases. This population difference indicates that our algorithms are more likely to encounter rarer states in the MIMIC cohort.

Reward function. Let K_{t+1}^{post} be the first potassium lab obtained after dose a_t. We assign

$$r_t = \exp\Big(-\frac{(K_{t+1}^{\text{post}} - 4.25)^2}{2\sigma^2}\Big), \quad \sigma = 0.25,$$

a Gaussian that peaks at the mid-point of the clinical reference range for normal potassium $(3.5 - 5.0)$ mmol/L. Rewards are clipped to $[0, 1]$ and paired with the next state to form (s_t, a_t, r_t, s_{t+1}) tuples; episode termination occurs at discharge.

4 Method

4.1 Algorithms

Behavior Cloning (BC) To capture the clinician's decision-making directly from data, we first fit a *behavior-cloning* model that learns to mimic observed dosing choices. This approach requires no reward signal and provides a convenient estimate of the behavior policy π_b for subsequent OPE. Concretely, we parameterize

$$\texttt{BC\_Policy}: \; s \in \mathbb{R}^d \;\mapsto\; \big[z_1, \ldots, z_n\big], \qquad \pi_b(a_k \mid s) \;=\; \frac{\exp(z_k)}{\sum_{j=1}^{n} \exp(z_j)},$$

where each a_k is one of the n discrete dose buckets. We train by minimizing a weighted categorical cross-entropy,

$$\mathcal{L}_{\mathrm{BC}}(\phi) = -\frac{1}{N} \sum_{i=1}^{N} w_{a_i} \, \log \pi_b(a_i \mid s_i),$$

with inverse-frequency weights $w_k \propto (\#\{a_i = k\})^{-\alpha}$ ($\alpha \in (0, 1)$) to correct for severe class imbalance. After training, π_b can be queried deterministically or stochastically via the network's `get_action` and `get_action_probs` methods.

Fitted Q–Iteration (FQI) To learn a dosing policy that maximizes expected cumulative reward, we employ *Fitted Q–Iteration*, a batch RL algorithm that approximates the Bellman operator by repeated regression. Starting from a fixed dataset of transitions $\{(s_t, a_t, r_t, s_{t+1}, a_{t+1})\}$, we fit a Q-network

$$Q_\theta(s) \;=\; \big[Q_\theta(s, a_1), \, \ldots, \, Q_\theta(s, a_n)\big]$$

by minimizing the per-sample weighted Bellman-residual

$$y_t = r_t + \gamma \, (1 - d_t) \, Q_{\bar{\theta}}(s_{t+1}, a_{t+1}), \quad \mathcal{L}_{\mathrm{FQI}}(\theta) = \frac{1}{B} \sum_{t \in \text{batch}} w_{a_t} \, \big(Q_\theta(s_t, a_t) - y_t\big)^2.$$

Here d_t marks terminal steps, $\bar{\theta}$ is a *target network* slowly updated via $\bar{\theta} \leftarrow (1-\tau)\bar{\theta} + \tau\theta$ for stability, and weights w_{a_t} again compensate for under-represented dose actions. After convergence, the greedy policy $\pi_e(s) = \arg\max_a Q_\theta(s, a)$ is used for evaluation.

Conservative Q-Learning (CQL) Offline value estimates can be overly optimistic on out-of-distribution actions, so we adopt *Conservative Q-Learning* Kumar et al. (2020) to bias the learned policy toward safer, well-supported actions. CQL augments the FQI objective with a log-sum-exp penalty on the Q-values,

$$\mathcal{L}_{\mathrm{CQL}}(\theta) = \mathcal{L}_{\mathrm{FQI}}(\theta) + \alpha \Big(\mathbb{E}_s \big[\log \sum_a e^{Q_\theta(s,a)} \big] - \mathbb{E}_{(s,a)\sim\mathcal{D}} [Q_\theta(s, a)] \Big),$$

where $\alpha > 0$ controls conservatism. We implement this via d3rlpy's `DiscreteCQLConfig` (v2.8.1) with a `MeanQFunctionFactory` and `DefaultEncoderFactory`. The resulting greedy policy $\pi_{\mathrm{CQL}}(s) = \arg\max_a Q_\theta(s, a)$ balances return maximization against the risk of unsupported dosing choices.

4.2 Off-Policy Evaluation

Since deploying our learned policies in a live clinical setting is infeasible, we estimate their expected return using *off-policy evaluation* (OPE). In this framework, we distinguish between the *behavior policy* π_b, which generated the observed trajectories, and the *target policy* π_e, whose value we wish to estimate. We approximate π_b by fitting a behaviour-cloning (BC) model to the clinician's logged data Gottesman et al. (2018), and take the greedy policies learned by FQI or CQL as our π_e.

Among OPE estimators, we choose **weighted importance sampling** (WIS) for three reasons: (i) it remains fully *model-free*, avoiding any bias from a simulator; (ii) it is *consistent* and asymptotically unbiased; and (iii) by self-normalizing the importance weights it dramatically reduces variance compared to ordinary importance sampling.

$$\hat{V}_{\mathrm{WIS}} = \frac{\sum_{i=1}^{N} w_i\, G_i}{\sum_{i=1}^{N} w_i}, \quad w_i = \prod_{t=0}^{T_i-1} \frac{\pi_e(a_{i,t} \mid s_{i,t})}{\pi_b(a_{i,t} \mid s_{i,t})}, \quad G_i = \sum_{t=0}^{T_i-1} \gamma^t\, r_{i,t}.$$

In contrast, the ordinary IS estimator $\hat{V}_{\mathrm{IS}} = \frac{1}{N}\sum_i w_i G_i$ can suffer from extremely high variance when likelihood ratios w_i vary widely. WIS's normalization trades a small bias for a substantial variance reduction, making it more reliable on finite clinical datasets.

5 Results & Discussion

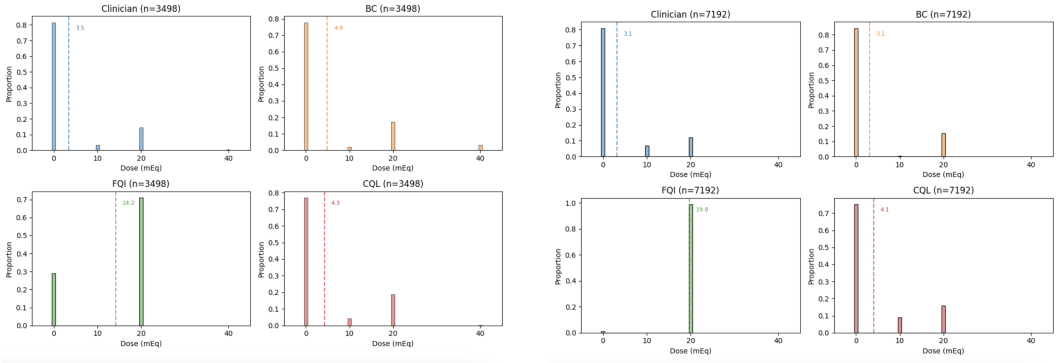
We evaluate the greedy policies learned by FQI and CQL with the weighted–importance–sampling (WIS) estimator introduced in Section 4.2. Table 1 summarises the point estimates and clipped weight statistics; Figure 3 shows the corresponding dose distributions.

Dataset	Method	WIS Value Estimate	Clipped mean	Clipped std
MIMIC	FQI	1414.7	5.89	4.80
MIMIC	CQL	1319.2	10.00	0.00
STARR	FQI	1696.5	5.02	4.94
STARR	CQL	1473.5	9.27	2.52

Table 1: WIS estimates and moments of the *clipped* importance-sampling weights for greedy FQI and CQL policies.

On both hospitals the FQI policy attains the higher WIS (1414.7 on MIMIC, 1696.5 on STARR), reflecting its willingness to deviate from clinician practice when the estimated Q–function predicts a gain. Its clipped weights have lower means ($\approx$ 5–6) but higher dispersion (std $\approx$ 4.8–4.9), signalling moderate yet uneven divergence from the behaviour policy.

CQL's conservative penalty pushes many weights to the clipping ceiling (mean ≈ 10). This tightens the variance (std 0 on MIMIC, 2.5 on STARR) and keeps the action distribution closer to clinicians, but costs 7–13 % of estimated return. For MIMIC + CQL every importance weight was truncated at the upperbound, giving a clipped mean of 10 and a standard deviation of 0. This extreme occurs when the learned greedy policy selects actions the clinician almost never used, so the raw ratios explode and are uniformly clipped.



(a) MIMIC cohort (b) STARR cohort

Figure 3: Dose–bucket proportions (0–40 mEq) for clinicians and learned policies. Vertical dashed lines mark the mean prescribed dose in each panel.

The plots above show the learned action distributions of our algorithms compared to clinician behavior on each dataset. We notice in the action distributions that CQL demonstrates better alignment with clinician behavior across both datasets. This likely stems from CQL's regularization term that

penalizes Q-values for out-of-distribution actions, constraining the learned policy to remain within the behavioral distribution observed in training data and preventing value overestimation for rarely observed high-dose actions. This stability could suggest that CQL's conservative regularization adapts to institutional differences by maintaining distributional alignment with local clinical practice, rather than optimizing toward a fixed dosing strategy.

In contrast, we observe that FQI most drastic departure from clinician distribution, learning policies that favor substantially higher doses with means of 14.2 mEq (MIMIC) and 19.8 mEq (STARR). This bias toward out-of-distribution actions occurs because FQI optimizes Q-values through iterative bootstrapping without distributional constraints, while class imbalance weighting ($\alpha = 0.6$) inadvertently encourages exploration of infrequent high-dose actions. The algorithm's target Q-network updates create a feedback loop where high-dose actions receive inflated values despite being rarely prescribed clinically. The tighter initial potassium distribution observed in STARR may paradoxically worsen FQI's overestimation problem, as the algorithm encounters fewer examples of high-dose interventions during training, leading to more pronounced Q-value inflation for out-of-distribution actions.

BC shows intermediate performance with dataset-dependent behavior: close alignment in STARR (3.1 mEq) but more aggressive dosing in MIMIC (4.9 vs 3.5 mEq). Algorithm performance varies systematically between datasets, with BC benefiting from STARR's concentrated clinician distribution that provides clearer imitation signals, while FQI becomes increasingly aggressive in STARR due to sparser high-dose examples amplifying Q-value overestimation. CQL maintains consistent performance across datasets, demonstrating robustness that reinforces the value of conservative approaches for clinical reinforcement learning applications.

6 Conclusion

Our investigation into cross-institutional reinforcement learning for potassium repletion reveals important considerations for deploying RL algorithms across diverse healthcare settings. Notably, although FQI achieved higher off-policy evaluation (OPE) estimates than CQL in both MIMIC-IV and STARR-OMOP, its action-selection distribution deviated substantially from clinician practice, prescribing high doses far more often. CQL, by contrast, produced dosing histograms much closer to the clinician baseline, suggesting greater "realism" and safer behavior despite slightly lower OPE scores. We also observed differences in the initial potassium distributions between the two cohorts, which may help explain why each algorithm performs differently across sites. The variation in results indicates both the choice of RL algorithm and the underlying patient population characteristics shape policy behavior and must be considered in clinical decision support.

However, these findings are not conclusive- the tradeoff between return-maximization and clinical alignment emphasizes that no single algorithm clearly dominates across institutions. Instead, our study serves as a commentary on a few factors—algorithmic bias, state-distribution shifts, and cohort heterogeneity—that underlie cross-institutional challenges in healthcare RL. Deeper investigation is needed to understand these effects and guide safe, generalizable policy deployment.

7 Team Contributions

This project was completed individually. I am currently a research assistant in a Biomedical Data Science lab, where RL generalization in non-synthetic settings is an active area of interest. I accessed the MIMIC-IV and STARR-OMOP datasets through my lab for this project. While my lab mentors were not directly involved in the course project, they provided guidance and helpful recommendations when necessary. I intend to extend and build on this work beyond the class to explore deeper insights into cross-institution generalization.

Changes from Proposal Our original proposal included evaluating model-based RL methods alongside model-free baselines, but this study focused only on model-free approaches (FQI and CQL). In future work, we could build a simple simulator of how potassium levels respond to different doses—using historical lab and treatment data—and then test planning-based RL algorithms within that simulator to see if they can improve on our current policies.

8 Limitations & Future Work

To address limitations in this study and provide more comprehensive commentary, we plan to pursue the following open questions in future work:

- **Improved interpolation for sparsity.** Refine our bin-based lab imputation—e.g. via learned time-series models—to reduce bias from missing measurements.
- **Reward function design.** Explore alternative reward schemes that penalize both under- and over-correction of potassium more explicitly, to sharpen the value gap between dosing actions.
- **Larger cohort evaluation.** Expand beyond the initial 200-patient sample to test our methods on a substantially larger dataset, improving statistical power and robustness.
- **Confidence intervals.** Compute and report uncertainty estimates (e.g. bootstrap confidence intervals) for all OPE metrics to quantify estimation variability.
- **Cross-institutional policy testing.** Evaluate each learned policy by applying it to the other dataset (e.g. train on MIMIC, test on STARR and vice versa) to assess generalizability across hospital systems.

References

Omer Gottesman et al. 2019. Guidelines for reinforcement learning in healthcare. *Nature Medicine* 25, 1 (2019), 16–18.

Omer Gottesman, Fredrik D. Johansson, Joshua Meier, Jack Dent, Donghun Lee, Srivatsan Srinivasan, Linying Zhang, Yi Ding, David Wihl, Xuefeng Peng, Jiayu Yao, Isaac Lage, Christopher Mosch, Li-Wei H. Lehman, Matthieu Komorowski, Aldo Faisal, Leo Anthony Celi, David A. Sontag, and Finale Doshi-Velez. 2018. Evaluating Reinforcement Learning Algorithms in Observational Health Settings. *CoRR* abs/1805.12298 (2018). arXiv:1805.12298 `http://arxiv.org/abs/1805.12298`

Ryan Hargrave et al. 2024. EpiCare: A Benchmark for Offline Reinforcement Learning in Healthcare. In *Proceedings of the Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*.

Alistair E. W. Johnson, Luca Bulgarelli, Tom J. Pollard, Benjamin Gow, Benjamin Moody, Steven Horng, Leo A. Celi, and Roger Mark. 2024. MIMIC-IV (version 3.1). `https://physionet.org/content/mimiciv/3.1/`. PhysioNet. `https://doi.org/10.13026/kpb9-mt58`.

Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. 2020. Conservative Q-Learning for Offline Reinforcement Learning. *CoRR* abs/2006.04779 (2020). arXiv:2006.04779 `https://arxiv.org/abs/2006.04779`

Niranjani Prasad, Aishwarya Mandyam, Corey Chivers, Michael Draugelis, C William Hanson, Barbara E Engelhardt, and Krzysztof Laudanski. 2022. Guiding Efficient, Effective, and Patient-Oriented Electrolyte Replacement in Critical Care: An Artificial Intelligence Reinforcement Learning Approach. *Journal of Personalized Medicine* 12, 5 (2022), 661. `https://doi.org/10.3390/jpm12050661`

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, et al. 2018. Model-based reinforcement learning for sepsis treatment. *arXiv preprint arXiv:1811.09602* (2018).

A Implementation Details

All methods share the same training regimen for a fair comparison. For BC and FQI we use a two-layer MLP with 64 ReLU units per layer; CQL employs d3rlpy's `DefaultEncoderFactory` (two-layer MLP with 256 units). All networks are optimized with Adam (learning rate 1×10^{-5}, weight decay 1×10^{-4}) on mini-batches of 512 transitions sampled from the complete clinician replay buffer. We set the discount factor to $\gamma = 0.99$ and perform a soft update of each target network

after every gradient step using $\tau = 0.1$. Training runs for 100 epochs (approx 100 k gradient steps), and the checkpoint with the lowest validation TD-loss is chosen for evaluation.

Input features are standardized (zero mean, unit variance) via a `StandardScaler` fit on the training trajectories. BC is trained with weighted cross-entropy using inverse-frequency class weights; FQI minimizes a weighted Bellman-residual mean-squared error; CQL adds the log-sum-exp conservatism penalty on top of FQI's loss. All experiments were implemented in Python 3.8 with PyTorch 1.12, d3rlpy 2.8.1, NumPy 1.21, and scikit-learn 1.0.