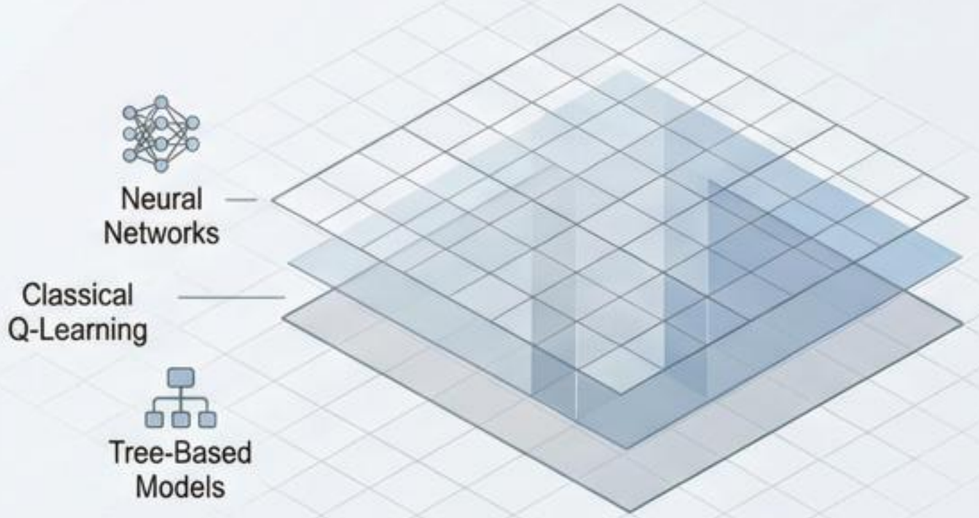
Offline Reinforcement Learning for Sepsis Treatment: A Comparative Study of Q-Learning Approaches

Evaluating Neural, Classical, and Tree-Based Methods on MIMIC-IV Data



Authors:
[Your Name(s)]

Affiliation:
[Your Institution]

Date:
[Presentation Date]

Conference/Course:
[If applicable]

Why Reinforcement Learning for Sepsis Treatment?

Evaluating Neural, Classical, and Tree-Based Methods on MIMIC-IV Data

Sepsis: A Critical Healthcare Challenge

- Leading cause of death in ICUs worldwide
- 30-50% mortality rate for severe sepsis
- Complex treatment decisions: IV fluids, vasopressors, antibiotics
- High variability in clinician treatment strategies

Why Reinforcement Learning?

- Sequential decision-making under uncertainty
- Learn optimal treatment policies from historical data
- Potential to reduce mortality through personalized treatment
- Can identify patterns beyond human clinical intuition

The Challenge: Offline RL

- Cannot interact with patients (ethical constraints)
- Must learn from fixed historical datasets
- Risk of learning unsafe policies from limited data
- Need robust evaluation without deployment

Key Questions:

Can RL learn effective sepsis treatment policies from observational data?

Which RL architectures are most suitable for offline medical settings?

How do we ensure learned policies are clinically valid and safe?

PROBLEM STATEMENT:

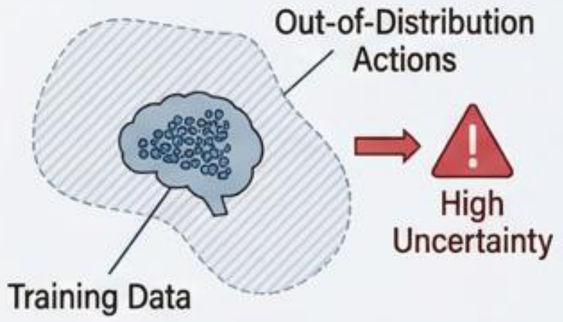
Offline Reinforcement Learning Challenges in Healthcare

The Core Problem: Learn optimal treatment policies from fixed observational data without environment interaction

1. Distributional Shift

Training data reflects clinician behavior (behavior policy π_b)

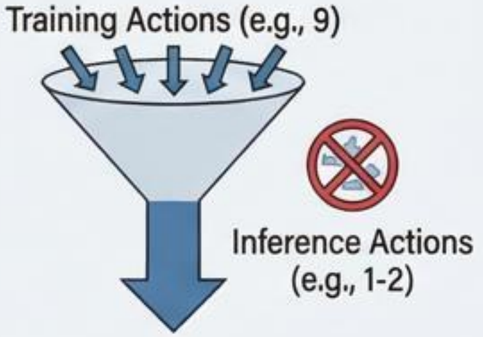
- Learned policy may recommend actions outside training distribution
- Out-of-distribution actions have unreliable Q-value estimates
- Risk: Overestimating value of rarely-taken actions



2. Action Space Collapse

Models may learn accurate Q-values but output deterministic policies

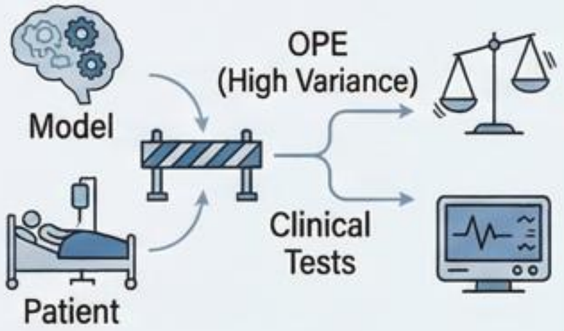
- Collapses to 1-2 actions at inference despite using all actions in training
- Dangerous in medical settings: different patients need different treatments
- Challenge: Maintaining diversity while maximizing expected outcome



3. Evaluation Without Deployment

Cannot deploy untested policies on real patients (ethical constraints)

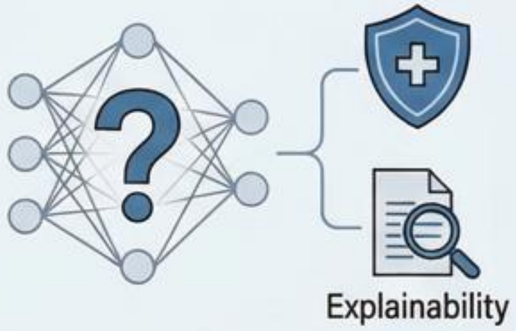
- Off-policy evaluation (OPE) has high variance
- Clinical sensitivity tests needed beyond numerical metrics
- Must validate physiological relationships (BP ↓ → vasopressor ↑)



4. Safety and Interpretability

Black-box neural networks lack clinical transparency

- Need to verify learned policies respect medical knowledge
- Must prevent policies that ignore critical vital signs
- Regulatory and ethical requirements for explainability



Research Objective: Compare 4 diverse RL architectures to identify which methods are most promising for offline medical RL and understand fundamental limitations

- Double DQN (Neural)
- Linear Q-Learning (Classical)
- GBM Q-Learning (Tree-Based)
- IQL Multi-Task (Modern Offline RL)

MIMIC-IV Dataset and Markov Decision Process Design

Dataset: MIMIC-IV | MDP State Space | Action Space | Reward Function

MIMIC-IV Dataset

MIMIC-IV (Version IV)

54,926 Adult ICU Patients
Beth Israel Deaconess Medical Center (2008-2019)
1,613,734 Total Transitions

Preprocessing

Data Split

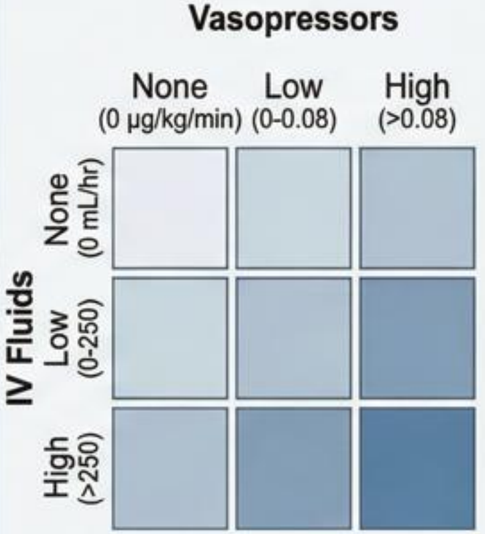
- **Train:** 70% (38,025 Patients)
- **Validation:** 15% (8,234 Patients)
- **Test:** 15% (8,667 Patients)

MDP State & Action Space

State Space (54 Features)

- **Vital Signs:** HR, BP, Temp, SpO2, RR
- **Laboratory Values:** Lactate, pH, Glucose, Creatinine, etc.
- **Blood Gas:** Arterial pH, PaCO2, Bicarbonate
- **Organ Function:** SOFA Score
- **Demographics:** Age, Gender, Weight
- **Ventilation:** Mechanical Status
- **Time:** Hours since ICU Admission

Action Space



Simplified 9 Discrete Actions

Original: 25 Discrete Actions (5x5 grid)

Reward & Trajectory

Reward Function

- **+15** (Survival)
- **-15** (Death at 90 days)
- **Intermediate Rewards:** ±0.1 to ±0.25
  SOFA Score Changes
  SOFA Decrease → Positive
  SOFA Increase → Negative
- **Sparse Reward Signal:** Most timesteps small intermediate rewards

Trajectory

- **Discount Factor:** $\gamma = 0.99$ (Long-term Outcomes)
- **Trajectory Length:** Variable, 4-hour Timesteps, Average ~30 Steps/Patient

Authors: [Your Name(s)] **Affiliation:** [Your Institution]

Date: [Presentation Date] **Conference/Course:** [If applicable]

Prior Work in RL for Sepsis Treatment

Early Research & The AI Clinician

Komorowski et al. (2018) - The AI Clinician

- First major work applying RL to sepsis using MIMIC-III
- Used discretized state space (750 clusters) + 25 actions
- Fitted Q-Iteration with ensemble of models

Key Finding

- AI policy suggested different treatments than clinicians

Limitation

- Validation on same dataset, no prospective trials

Modern Offline RL & Current Challenges

Offline RL Methods

Conservative Q-Learning (CQL) - Kumar et al. (2020)

- Adds penalty to Q-values for out-of-distribution actions
- $L_{CQL} = \text{logsumexp}(Q(s, \cdot)) - Q(s, a_{data})$
- Used in our Dueling DQN implementation

Batch-Constrained Q-Learning (BCQ) - Fujimoto et al. (2019)

- Restricts policy to actions likely under behavior policy
- Only select actions with $\pi_b(a|s) >$ threshold
- Used in our Double DQN implementation

Implicit Q-Learning (IQL) - Kostrikov et al. (2021)

- Avoids max operator through expectile regression
- Learns separate V-function and Q-function
- Our novel contribution: IQL + Multi-Task Learning

Challenges Identified in Prior Work

- Q-value overestimation in neural networks
- Lack of action diversity at deployment
- Difficulty validating without clinical trials
- Need for better evaluation metrics beyond WDR

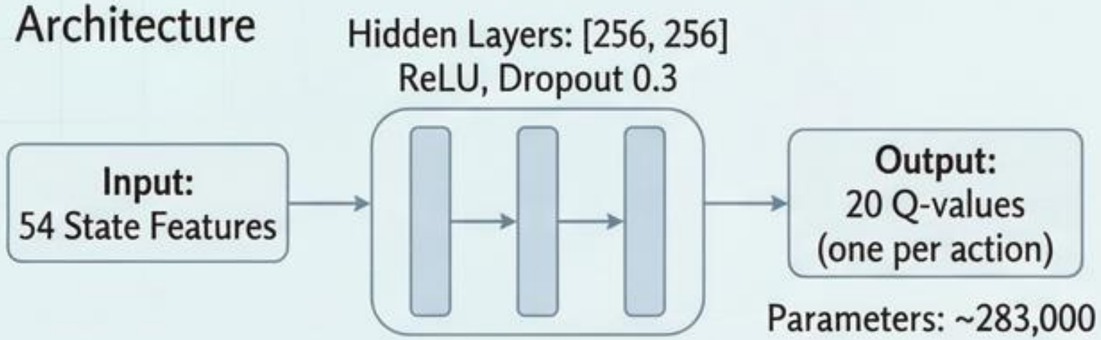
Our Contribution

Systematic comparison of 4 diverse architectures (neural, classical, tree-based, modern offline) with comprehensive clinical validation

METHODOLOGY - PART 1: Neural and Classical Approaches

Model 1: Double Deep Q-Network (Double DQN)

Architecture



Key Components & Training

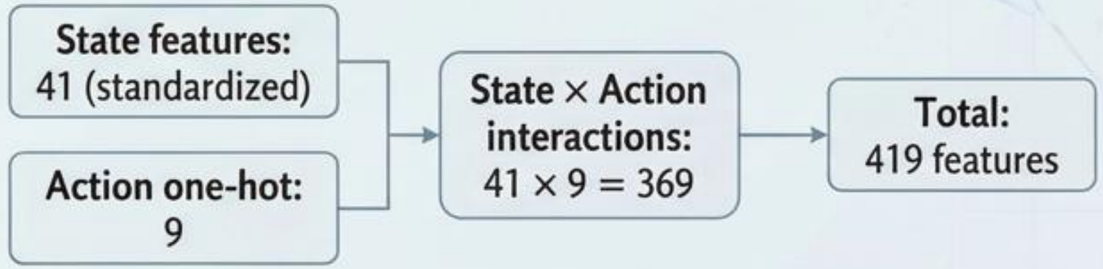
- **Online Network:** Selects actions, actively trained
- **Target Network:** Evaluates values, updated every 1,000 steps
- **TD Target:** $r + \gamma \cdot Q_{target}(s', \text{argmax}_a\, Q_{online}(s', a'))$
- **BCQ Constraint:** Only allow actions with $\pi_b(a|s) > 0.1\%$
- **Loss:** Smooth L1 (Huber) loss
- **Optimizer:** Adam, lr=1×10^{-4}

Purpose

Baseline neural approach with standard overestimation mitigation.

Model 2: Linear Q-Learning with Feature Engineering

Architecture: Feature Construction



Training Procedure & Form

- **Mathematical Form:** $Q(s,a) = w^T \phi(s,a)$
  where $\phi(s,a) = [s_{scaled}; 1_a; s_{scaled} \otimes 1_a]$
- **Algorithm:** Fitted Q-Iteration with SGD
- **Update Rule:** $w \leftarrow w + \alpha(\nabla Q \cdot TD_{error} - \lambda w)$
- **Hyperparameters:** Learning rate $\alpha = 0.001$, L2 regularization $\lambda = 0.01$
- Converged at 335 epochs

Purpose

Classical interpretable approach with explicit feature engineering.

Tree-Based and Modern Offline RL Approaches

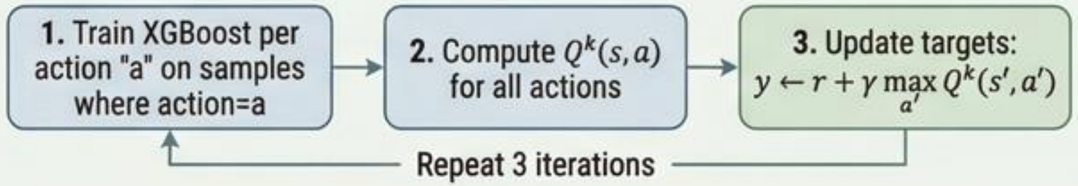
Evaluating Neural, Classical, and Tree-Based Methods on MIMIC-IV Data

Model 3: Gradient Boosting Q-Learning (GBM)

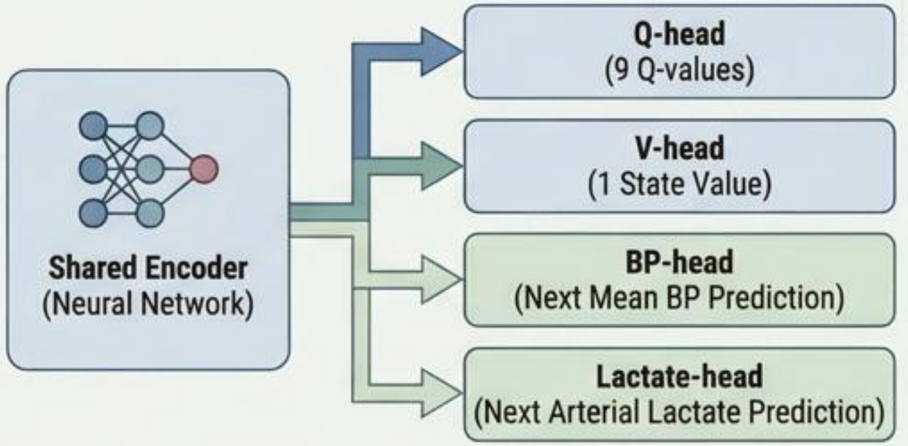
Key Components & Advantages

- **9 Independent XGBoost Regressors** (one per action)
- Captures non-linear relationships without manual feature engineering
- Separate models prevent Q-value interference across actions
- Trees handle mixed continuous/categorical features naturally
- No gradient vanishing issues

Iterative Q-Learning Training



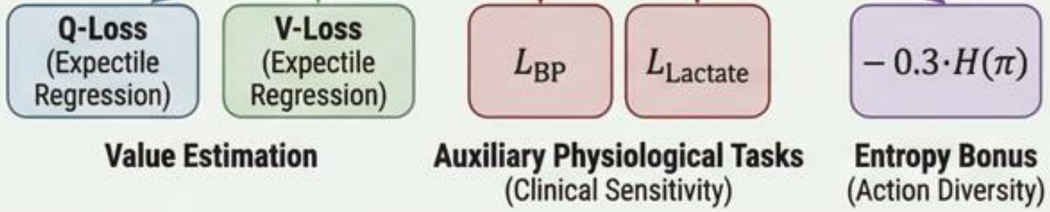
Model 4: Implicit Q-Learning with Multi-Task Learning (IQL)



Key Mechanisms & Multi-Task Loss

- **Shared Encoder & Multi-Task Learning** for representation
- No "max" operator: Uses expectile regression for value estimation
- Explicit diversity mechanisms through auxiliary tasks and entropy bonus

$$L_{\text{total}} = L_{\text{Q}} + L_{\text{V}} + 0.1 \cdot L_{\text{BP}} + 0.1 \cdot L_{\text{Lactate}} - 0.3 \cdot H(\pi)$$



Comprehensive Evaluation Framework

1. Weighted Doubly Robust (WDR) Estimator

Purpose: Estimate policy value from offline data

$\text{WDR} = \mathbb{E}_\tau[V^{WDR}(\pi_e)]$ where

$$V^{WDR} = Q(s_0, \pi_e(s_0)) + \sum_t \gamma^t \rho_{0:t}\left(r_t + \gamma Q(s_{t+1}, \pi_e) - Q(s_t, a_t)\right)$$

$\rho_t = \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}$ [importance weight]

Key Features:

- Combines Q-function estimates with importance sampling
- Lower variance than pure IS, less bias than pure model-based
- Bootstrap confidence intervals (1,000 samples)
- Importance weights clipped to [0, 10] for stability

Interpretation:

- **WDR > 7.56**: Policy outperforms clinicians (✓)
- **WDR ≈ 7.56**: Comparable to clinicians (~)
- **WDR < 7.56**: Policy underperforms (✗)

2. Clinical Sensitivity Tests

Purpose: Verify learned policies respond correctly to physiological changes

Test 1: Blood Pressure Sensitivity

- Vary mean BP from -2σ to +2σ (standardized)
- Measure correlation: BP ↔ Vasopressor recommendation
- **Pass criteria:** Spearman r < -0.2 (negative correlation)
- **Clinical reasoning:** Low BP should increase vasopressors (BP ↓ → vasopressor ↑)

Test 2: Lactate Sensitivity

- Vary arterial lactate from -1σ to +3σ
- Measure correlation: Lactate ↔ Treatment intensity (IV +Vaso)
- **Pass criteria:** Spearman r > 0.2 (positive correlation)
- **Clinical reasoning:** High lactate (hypoperfusion) needs aggressive treatment

Test 3: Action Diversity

- Sample 2,000 random states, count unique actions
- Compute Shannon entropy: $H = -\sum p(a) \log p(a)$
- **Pass criteria:** ≥6 unique actions AND H > 1.2
- **Clinical reasoning:** Different patients need different treatments

Overall Clinical Validation: Pass ≥2/3 tests

3. Additional Metrics

Agreement with Clinicians

- Percentage of test states where $\pi_{\text{learned}}(s) = \pi_{\text{clinician}}(s)$
- High agreement (>90%) suggests behavior cloning, not improvement
- Low agreement (<20%) with high WDR suggests novel effective policy

Effective Sample Size (ESS)

$$\text{ESS} = \frac{(\sum w_t)^2}{\sum w_t^2}$$ where w_t are importance weights

- Measures quality of importance sampling
- Low ESS (<0.01) indicates high variance in WDR estimate

Action Distribution

- Number of unique actions used at inference
- Shannon entropy of action distribution
- Identifies deterministic collapse

Evaluating Neural, Classical, and Tree-Based Methods on MIMIC-IV Data

Comparative Performance Across All Models

Model	WDR Value	95% CI	Agreement	Clinical Tests	Action Diversity	Status
Clinician Baseline	7.560	[7.370, 7.751]	100%	N/A	N/A	Baseline 😐
Double DQN	149.269	[147.661, 151.011]	98.2%	0/3 ❌	6/20	Failed ❌
Linear Q-Learning	5.639	± 0.859	24.6%	0/3 ❌	1/9	Failed ❌
GBM Q-Learning	6.704	± 0.167	9.2%	0/3 ❌	2/9	Best WDR ✅
IQL Multi-Task	14.520	± 0.091	9.8%	1/3 ⚠️	9/9	Best Diversity ⚠️

Key Observations

✅ **Best Value Estimation**: GBM (6.704, only 11% error vs clinician)

❌ **Worst Estimation**: Double DQN (1,875% overestimation)

✅ **Most Realistic Neural**: IQL (14.52, still 92% overestimation but 10× better than Double DQN)

✅ **Best Diversity**: IQL (9/9 actions, H=1.48)

❌ **Worst Diversity**: Linear (1/9 actions, completely deterministic)

⚠️ **No Model Succeeded Overall**: All failed majority of clinical tests

📊 **Accuracy vs Diversity Tradeoff**: Accurate models (Linear, GBM): Low diversity. Diverse models (IQL): Overestimation persists.

Model Performance Analysis: Successes, Failures, and Key Lessons

Double DQN

Catastrophic Overestimation

- **WDR:** 149.27 (1,875% error) ❌
- **Agreement:** 98.2% (Behavior Cloning) ⚠️
- **Clinical:** 0/3 Passed ❌
- **Actions:** 6/20 ❌

Why it failed: Neural TD learning compounds errors through bootstrapping, leading to massive overestimation.

Lesson: Standard neural DQN is unsuitable for offline medical RL due to severe overestimation and behavior mimicking.

Linear Q-Learning

Interpretable but Inflexible

- **WDR:** 5.64 (25% below clinician) 📉
- **Agreement:** 24.6% (Highest Non-Neural) ✅
- **Clinical:** 0/3 Passed ❌
- **Actions:** 1/9 (Collapse) ❌

Why it failed: Linear approximation cannot capture non-linear decision boundaries, leading to a single action output.

Lesson: Sepsis treatment requires non-linear function approximation; interpretability alone is insufficient. ⚠️

GBM Q-Learning

Accurate but Deterministic

- **WDR:** 6.70 (89% of clinician, most realistic) ✅
- **Agreement:** 9.2% (Low agreement) 📉
- **Clinical:** 0/3 Passed ❌
- **Actions:** 2/9 (Collapse) ❌

Why it failed: Greedy argmax over learned Q-functions defaults to globally best action, despite accurate Q-values, causing action collapse.

Lesson: Accurate Q-learning does not guarantee diverse policies in offline RL; explicit diversity mechanisms are needed. ⚠️

IQL Multi-Task

Diversity Success, Overestimation Remains

- **WDR:** 14.52 (92% overestimation, improved vs DQN) 📉
- **Agreement:** 9.8% 📉
- **Clinical:** 1/3 Passed (Diversity only) ⚠️
- **Actions:** 9/9 (Full Diversity) ✅

Why auxiliary failed: Auxiliary task weights (0.1) were too low, allowing the RL objective to dominate.

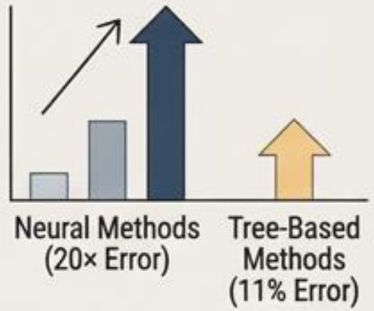
Lesson: Multi-task learning is promising for enforcing diversity, but requires stronger auxiliary losses (0.5+). ⚠️

Overall Lessons & Implications

1. **Non-parametric methods (trees)** provide more realistic value estimates than neural networks in offline settings.
2. **Explicit mechanisms for action diversity** (e.g., multi-task, entropy) are crucial to prevent deterministic collapse.
3. **Multi-task learning** shows promise for clinical sensitivity and diversity but needs careful tuning of auxiliary losses.

KEY FINDINGS: FUNDAMENTAL CHALLENGES IN OFFLINE MEDICAL RL

1. THE Q-OVERESTIMATION PROBLEM



Observation: Neural methods (Double DQN, IQL) massively overestimate.

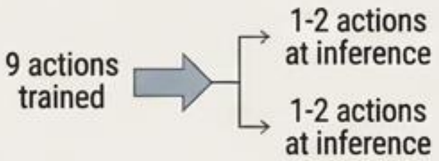
Tree-based methods (GBM) stay realistic.

Root Cause: Neural networks + TD learning + limited data → compound bootstrapping errors.

max operator in Bellman update selects highest (possibly spurious) Q-value.
Trees bounded by training data

Implication: Non-parametric methods (trees, kernels) preferred for offline medical RL.

2. THE ACTION COLLAPSE PHENOMENON



Observation: Models using all actions in training collapse to 1-2 actions at inference.

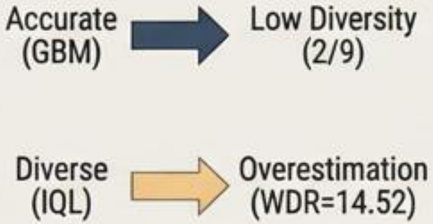
Linear: 9 trained → 1 test;
GBM: 9 trained → 2 test;
IQL: Only model maintaining 9/9

Root Cause: Offline RL learns Q-values only for state-action pairs similar to training.

Greedy policy argmax Q(s,a) selects globally highest-valued action.

Implication: Explicit diversity mechanisms (entropy, multi-task) required, not just accurate Q-learning.

3. THE DIVERSITY-PERFORMANCE TRADEOFF



Observation:

Accurate models → Low diversity

Diverse models → Overestimation

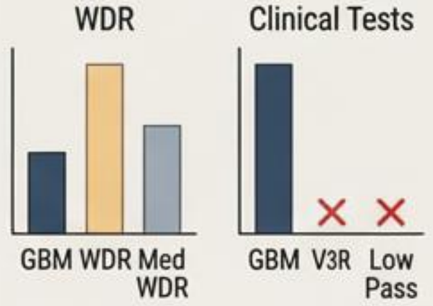
Adding stochasticity degrades performance.

Root Cause: Q-functions learned assuming deterministic deployment.

Introducing randomness violates Bellman optimality assumptions.

Implication: Diversity must be incorporated during training (IQL approach), not at inference.

4. CLINICAL SENSITIVITY INDEPENDENCE FROM WDR



Observation: Best WDR (GBM) fails all clinical tests (0/3).

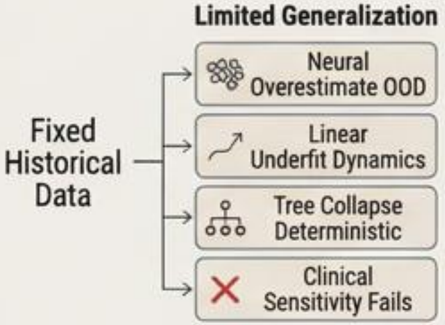
Medium WDR (IQL) passes diversity test (1/3).

No correlation between numerical performance and physiological validity.

Root Cause: WDR measures expected return under learned policy, doesn't verify policy respects medical knowledge.

Implication: Multi-metric evaluation essential; WDR alone insufficient for medical AI.

5. THE OFFLINE RL EXTRAPOLATION PROBLEM



Summary: All models struggle to generalize beyond training distribution.

Manifestations:
Neural models overestimate OOD actions;
Linear models underfit complex dynamics;
Tree models collapse to deterministic policies;
Even with diversity, clinical sensitivity fails.

Fundamental Limitation: Cannot validate new policies without online interaction (ethical barrier in medicine).

Challenges, Limitations, and Future Directions

Why Offline RL is Hard for Sepsis & Study Limitations

Fundamental Challenges
Data, Safety, Evaluation

1. Data Limitations: Coverage problem in continuous space; rare actions (<1%) & confounding.

2. Safety vs. Performance Tension: No middle ground found between conservative (safe, low perf.) & aggressive (high est., unsafe) methods.

3. Evaluation Without Ground Truth: Ethical constraints prevent online testing; high variance in off-policy evaluation (OPE).

Study Limitations
Model, Data, Evaluation, Generalizability

Model Selection: Unexplored models (Model-Based, Causal); limited tuning/seeds.

Data Quality: Missing data, errors, coarse 4-hour timesteps.

Evaluation: Wide WDR confidence intervals; heuristic clinical tests.

Generalizability: Single hospital/system; dated 2008-2019 data.

What Worked? What Didn't?

Successes

Realistic Value Estimates (GBM 6.70 vs 7.56).

Multi-Task Learning (IQL) enforces diversity.

Tree-Based Methods avoid catastrophic overestimation.

Identified fundamental offline RL limitations.

Failures

No model achieved all goals (value + diversity + sensitivity).

Action collapse persists across architectures.

Weak auxiliary losses for physiological relationships.

Cannot recommend deployment of any policy.

Future Directions & Key Takeaways

1. Stronger Multi-Task Learning: Increase weights (0.5+), add physiological predictions, multi-objective optimization.

2. Causal Reinforcement Learning: Explicitly model causality, use instrumental variables, combine with causal frameworks.

3. Hybrid Approaches: Rule-based safety constraints, RL for exploration with clinician decisions, offline pre-training + online fine-tuning.

4. Better Evaluation: Develop validated clinical benchmarks, counterfactual estimation, prospective observational studies.

5. Model-Based RL: Learn patient dynamics model, plan actions (sample efficient), quantify uncertainty.

Key Takeaway: Current offline RL is promising but requires significant methodological advancements in safety, causality, and evaluation to be viable for high-stakes medical applications.

Conclusion and Key Takeaways

Summary of Work: Comparative Study

MIMIC-IV Data & Evaluation

Double DQN: Neural network baseline

Linear Q-Learning: Classical interpretable approach

Gradient Boosting Q-Learning: Non-parametric tree ensemble

IQL Multi-Task: Modern offline RL with auxiliary learning

Comprehensive Evaluation: Value estimation (WDR), clinical sensitivity, action diversity, agreement.

Key Contributions: Evidence & Challenges

1. Empirical Evidence on Method Suitability

- Tree-based (GBM) achieve most realistic value estimates
- Neural TD methods suffer catastrophic overestimation
- Multi-task learning is effective diversity mechanism

2. Identified Fundamental Challenges

- **Action collapse:** Universal across architectures (1-2 actions)
- **Diversity-performance tradeoff:** Cannot optimize simultaneously
- **Offline extrapolation problem:** Limited data prevents safe learning

Established Evaluation Framework

3. Multi-Metric Validation is Critical

WDR Alone Insufficient

- High variance estimates
- Lacks physiological validation

Clinical Sensitivity Essential

- Verifies medical knowledge
- Crucial for safety

Framework established highlighting gap between numerical performance and clinical validity.