Reinforcement Learning for Individual Optimal Policy from Heterogeneous Data

Annie Qu

UC Santa Barbara

Rui Miao
UT Dallas

Babak Shahbaba
UC Irvine

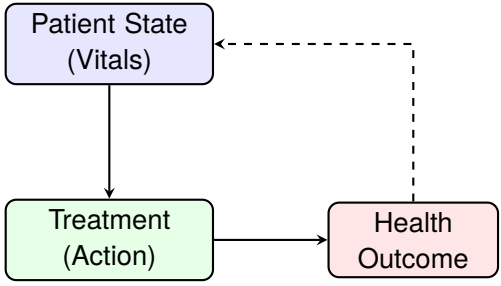
Precision Medicine: From Averages to Individuals

Traditional Medicine:

- One-size-fits-all treatments.
- Static guidelines based on population averages.

The Individualized Dynamic Regimens:

- Continuously monitor patient states (e.g., ICU vitals, Smartwatches).
- Dynamically adjust treatments.
- Maximize long-term health outcomes.



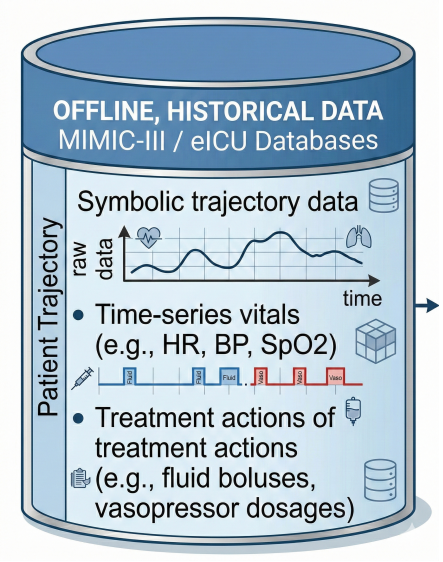
Motivating Example: Sepsis in the ICU

Sepsis is a life-threatening response to infection.

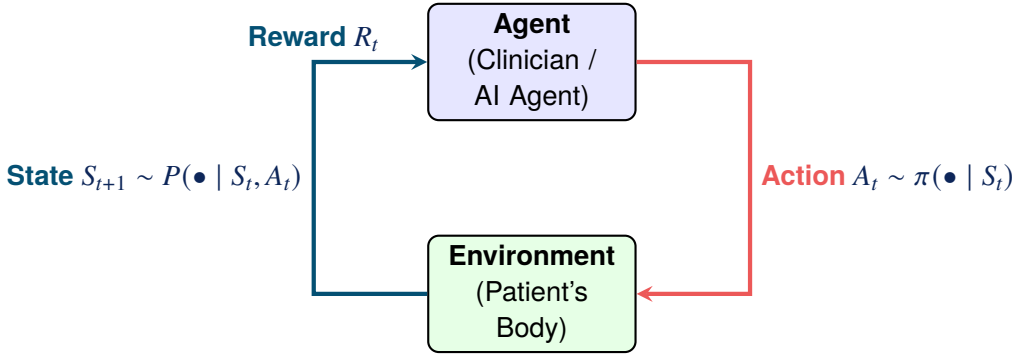
- Clinicians must balance giving **IV Fluids** and **Vasopressors**.
- Too much fluid? Lungs fill with water. Too much vasopressor? Tissue death.

The Fatal Roadblock in Healthcare

- We cannot run algorithms in the ICU to "trial-and-error" treatments on patients.
- We must learn optimal strategies from **offline, historical data** (e.g., MIMIC-III database).



Reinforcement Learning (RL)



- **Policy π:** A conditional distribution $\pi(a \mid s)$, action $a \in \mathcal{A}$, state $s \in \mathcal{S}$.
- **Q-functionl $Q_\pi(s,a)$:** The expected *long-term cumulative reward* if you are in State s, take Action a, and then follow Policy π. With some *discount factor* $0 < \gamma < 1$,

$$Q_\pi(s,a) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t R_t \mid S_0 = s, A_0 = a, S_{t+1} \sim P(\bullet \mid S_t, A_t), A_t \sim \pi(\bullet \mid S_t)\right]$$

- **Goal:** Find the optimal policy π^* that maximizes the expected value.

Value Functions and Policy Objective

State Value Function

$$V^i_{\pi^i}(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i_t \,\middle|\, S^i_0 = s,\ A^i_t \sim \pi^i\right]$$

Expected total reward starting from state s.

V–Q Relationship

$$V^i_{\pi^i}(s) = \sum_{a \in \mathcal{A}} Q^i_{\pi^i}(s,a)\, \pi^i(a \mid s)$$

Policy Value (Objective)

$$J^i(\pi^i) \triangleq (1-\gamma)\, \mathbb{E}_{S^i_0 \sim \nu}\left[V^i_{\pi^i}(S^i_0)\right]$$

- $(1-\gamma)$: normalizes to $[0, R_{\max}]$
- ν: initial state distribution

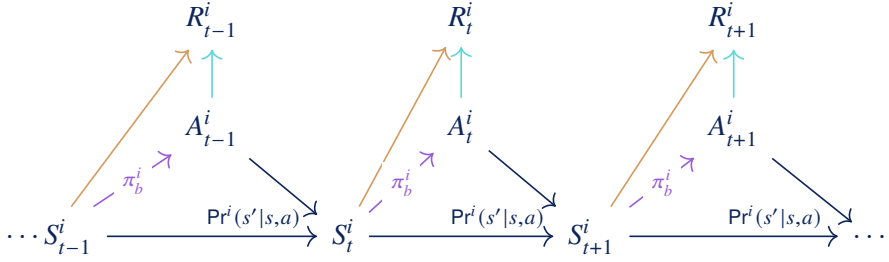
Individualized Goal:

$$\max_{\pi^1,\ldots,\pi^N} \sum_{i=1}^{N} J^i(\pi^i)$$

Find the best policy *for each patient*.

Markov decision processes (MDPs) framework

For individual $i \in [N] = \{1, \ldots, n\}$, the episode $\left\{S_t^i, A_t^i, R_t^i\right\}_{t \geq 0}$ follows MDP $\mathcal{M}^i = \left\{\mathcal{S}, \mathcal{A}, \mathsf{Pr}^i, r^i, \gamma\right\}$.



At each decision time $t \geq 0$:

- The action $\mathcal{A} \ni A_t^i \sim \pi_b^i(\bullet \mid S_t^i)$: the *behavior policy*.
- Individual i's status transits to $\mathcal{S} \ni S_{t+1}^i \sim \mathsf{Pr}^i(\bullet \mid S_t^i, A_t^i)$
- Receives an immediate reward R_t^i, with $\mathbb{E}[R_t^i \mid \text{past}] = r^i(S_t^i, A_t^i)$

Assume **Markovianity** and $\mathsf{Pr}^i(\bullet \mid s, a)$ and $\pi_b^i(\bullet \mid s)$ are **time-stationary**.

Why Learning Individually Fails

Challenge 1: Low sample efficiency

- Each subject i has only T^i transitions — typically *short* in healthcare.
- Cannot borrow information from other subjects.

Challenge 2: Coverage fails for each individual

Discounted Visitation Prob.

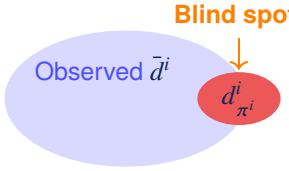
$$d^i_{\pi^i}(s,a) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t p^i_{\pi^i,t}(s,a)$$

(s,a) pairs visited when following *target* π^i.

Empirical Visitation Prob.

$$\bar{d}^i(s,a) = \frac{1}{T^i}\sum_{t=0}^{T^i-1} p^i_{\pi^i_b,t}(s,a)$$

(s,a) pairs actually *observed* in data.



Require $d^i_{\pi^i} \ll \bar{d}^i$
(absolutely continuous)
$\Rightarrow$ infeasible for 1 subject!

Existing approaches Munos and Szepesvári [2008], Antos et al. [2008], Xie et al. [2021], Zhan et al. [2022] require $d^i_{\pi^i}$ (or $d^i_{\pi^i_*}$) $\ll \bar{d}^i$.

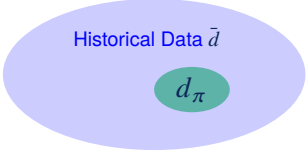
The Challenge of Offline RL: Coverage

Only have a static dataset $\mathcal{D}$ generated by historical doctors (the **Behavior Policy**, π_b).

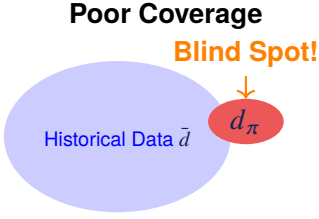
The Coverage Assumption

To safely evaluate a new policy π, the historical data must have "covered" the actions the new policy wants to take.
New policy π explored state-actions visitation distribution d_π is *absolutely continuous* ($\ll$) w.r.t. empirical distribution $\bar{d}$ discovered by data ($d_\pi \ll \bar{d}$)

Good Coverage



Poor Coverage



The Challenge: Heterogeneity vs. Coverage

Patients are highly **heterogeneous**. We want to learn **individualized** policies π^i.

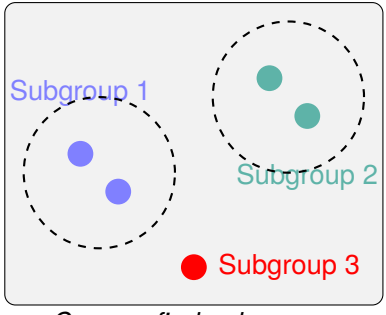
Approach 1: Purely Individual RL

- Learn strictly from subject i's history.
- **Fails Coverage:** One subject rarely experiences all treatments. The coverage assumption is impossibly strict.
- Low sample efficiency.

Approach 2: Homogeneous RL

- Pool all patients together.
- **Fails Heterogeneity:** Averages out crucial differences.
- Bias.

The Missing Middle



Can we find subgroups and borrow strength?

Our Solution: Weak Partial Coverage

Instead of relying on a single individual's sparse data, we leverage the **entire population's** data to satisfy coverage for the individual.

Strict Individual Coverage (Impossible)

Requires target policy π^i to be covered solely by Patient i's history.

$$d^i_{\pi^i} \ll \bar{d}^i$$

Our Weak Partial Coverage (Realistic)

Requires target policy π^i to be covered by the **grand average** of the population's history.

$$d^i_{\pi^i} \ll \bar{d} = \frac{1}{N}\sum_{j=1}^{N} \bar{d}^j$$

To make this work, we need a robust way to link individuals together.

Our Framework: P4L

Penalized **P**essimistic **P**ersonalized **P**olicy **L**earning

To solve the heterogeneity and coverage problems simultaneously, our method relies on **Three Key Pillars**:

1. **Latent Shared Structure:** Borrowing statistical strength across the population.
2. **Pessimism:** Safe exploration when data is missing.
3. **Multi-Centroid Penalization:** Automatically discovering subject's characteristics.

Borrowing Strength via Shared Structure

Instead of learning N separate models, we learn **one global model** that adapts based on a subject-specific "latent vector" $\mathbf{u}^i$.

The Mixed-Effects Analogy:

$$Y^i = \underbrace{X^i\alpha}_{\text{Population}} + \underbrace{Z^i\mathbf{u}^i}_{\text{Individual}}$$

In RL (Q-Functions):

$$Q^i_\pi(\text{state}, \text{action}) = Q_\pi(\text{state}, \text{action}, \mathbf{u}^i)$$

How this helps:

- Q is a global neural network trained on **everyone's** data.
- $\mathbf{u}^i$ tweaks the network for Patient i.
- If Patient i never received Action A, the network infers the outcome based on similar patients who did!

Key Lemma: Why the Min-Max Objective?

Lemma (Value Estimation Error)

For any candidate $\widetilde{Q}^i$ and policy π^i, the value estimation error satisfies:

$$J^i(\pi^i) - \widetilde{J}^i(\pi^i) = \mathbb{E}_{(s,a)\sim d^i_{\pi^i}}\Big[\underbrace{R^i + \gamma \widetilde{Q}^i(S^i_+, \pi^i(S^i_+)) - \widetilde{Q}^i(s,a)}_{\text{Bellman Error}(Z^i; \widetilde{Q}^i, \pi^i)}\Big]$$

- $d^i_{\pi^i}$ is **unknown** — cannot compute directly.
- Under **weak partial coverage** $d^i_{\pi^i} \ll \bar{d} = \frac{1}{N}\sum_{j=1}^N \bar{d}^j$, we upper-bound:
$$\left|J^i(\pi^i) - \widetilde{J}^i(\pi^i)\right| \le \sup_{f\in\mathcal{F}} \bar{\mathbb{E}}\big[f(s,a;u^i)\cdot \text{BellmanErr}(Z^i, \widetilde{Q}^i, \pi^i)\big]$$
- $f \in \mathcal{F}$: test function class capturing the visitation ratio $d^i_{\pi^i}/\bar{d}$.
- **Minimize Bellman Error** over population data $\Rightarrow$ bound value estimation error.

This motivates the following min-max formulation.

Shared Structure Enables Population Coverage

Summing over **all individuals** under shared latent variables u^i:

Suppose $\pi^i(\cdot) = \pi(\cdot; u^i)$, $Q^i_{\pi^i}(\cdot) = Q_\pi(\cdot; u^i)$, and the maximizing $f^i = f(\cdot; u^i)$. Then:

$$\sum_{i=1}^{N} \left| J^i - \widetilde{J}^i \right| \leq \underbrace{\frac{1}{N} \sum_{i=1}^{N} \frac{1}{T^i} \sum_t f(S_t^i, A_t^i; u^i) \cdot \text{BellmanErr}_t\left(Z^i, \widetilde{Q}(\cdot; u^i), \pi(\cdot; u^i)\right)}_{\widehat{\Phi}(\widetilde{Q}, f, \pi, \mathbf{u})}$$

- Grand average $\bar{d}$ pools coverage from *all* individuals — much richer.
- Individual i never tried action a? **Other subjects** may have, satisfying $d^i_{\pi^i} \ll \bar{d}$.
- Shared Q and π networks trained on **everyone's data simultaneously**.

To minimize $\widehat{\Phi}$, we solve:

Formally

Empirical we solve a min-max problem for given π and $\mathbf{u} = \left\{u^i\right\}_{i\in[N]}$

$$\underbrace{\min_{\widetilde{Q}\in\mathcal{Q}} \max_{f\in\mathcal{F}} \widehat{\Phi}(\widetilde{Q}, f, \pi, \mathbf{u})}_{\text{Policy Evaluation Err}},$$

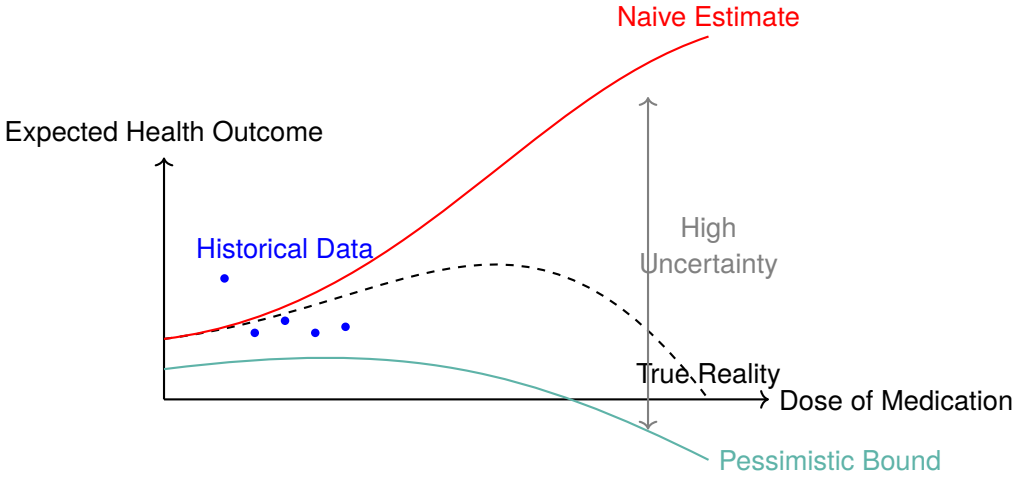
where

$$\widehat{\Phi}(\widetilde{Q}, f, \pi, \mathbf{u}) \triangleq \frac{1}{N}\sum_{i=1}^{N}\frac{1}{T^i}\sum_{t=0}^{T^i-1} f(S_t^i, A_t^i; u^i) \times \text{BellmanErr}_t\left(Z^i, \widetilde{Q}(\bullet; u^i), \pi(\bullet; u^i)\right)$$

- $\text{BellmanErr}_t(Z^i; \widetilde{Q}^i, \pi^i) := R_t^i + \gamma\widetilde{Q}^i(S_{t+1}^i, \pi^i(S_{t+1}^i)) - \widetilde{Q}^i(S_t^i, A_t^i)$;
- f is some test function in space $\mathcal{F}$ covering visitation probability ratio;
- $\mathcal{Q}$ is the space of candidate Q-functions;
- T^i is the length of episode for subject i.

Pessimism (Safety First)

Even with shared data, some state-action pairs will be rarely explored globally. Neural networks will often wildly overestimate rewards in these "blind spots."



Solution: We construct an uncertainty bound around our estimate, and optimize the policy assuming the **worst-case scenario** within that bound.

Pessimistic policy optimization

Uncertainty set

Given **uncertainty level** $\alpha_{NT} > 0$, define

$$\Omega(\pi, \mathbf{u}, \alpha_{NT}) \triangleq \Big\{ Q \in \mathcal{Q} : \underbrace{\max_{f \in \mathcal{F}} \widehat{\Phi}(Q, f, \pi, \mathbf{u})}_{\text{Policy Evaluation Err}} \leq \alpha_{NT} \Big\}$$

- Consider all candidates Q from $\Omega(\pi, \mathbf{u}, \alpha_{NT})$ [They are close enough to the truth!]
- Maximize the value of the most pessimistic Q in uncertainty set.

Find the in-class optimal policy with latent variables

$$(\widehat{\pi}, \widehat{\mathbf{u}}) \in \arg \max_{\pi \in \Pi, \mathbf{u}} \min_{Q \in \Omega(\pi, \mathbf{u}, \alpha_{NT})} J(\pi, \mathbf{u}), \quad \text{where}$$

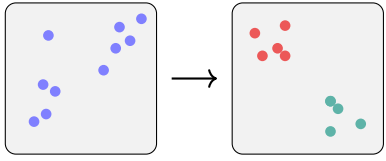
$$\text{Value of the policy } J(\pi, \mathbf{u}) \triangleq (1-\gamma) \sum_{i=1}^{N} \underbrace{\mathbb{E}_{S_0^i \sim \nu} Q(S_0^i, \pi(S_0^i; u^i); u^i)}_{\text{expected value for } i} .$$

Clustering Phenotyping via Penalization

We want our latent variables $\mathbf{u}^i$ to cluster naturally into subgroups (e.g., “Septic Shock” vs. “Mild Infection”).

Multi-Centroid Penalty:

$$\text{Penalty} = \sum_{i=1}^{N} \min_{k \in [K]} \|\mathbf{u}^i - \mathbf{v}^k\|_2^2$$

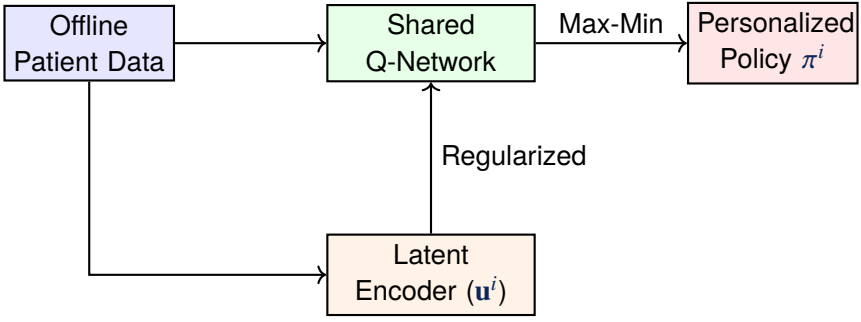


- $\mathbf{u}^i$: Individual patient feature.
- $\mathbf{v}^k$: Center of subgroup k.

Instead of learning N noisy policies, we effectively learn K highly robust policies, fine-tuned for individuals ($K < N$).

Model Architecture

We implement this using an Alternating Direction Method of Multipliers (ADMM) embedded in our workflow.



- **Step 1:** Update the Q-Network to evaluate the current policy.
- **Step 2:** Update individual embeddings ($\mathbf{u}^i$) and cluster them.
- **Step 3:** Update the Policy Network to maximize pessimistic Q-values.

Theoretical Guarantees

"Borrowing strength" works theoretically.

Theorem (Informal Regret Bound)

Assuming trajectory length $T \gg N$, and true subgroups are sufficiently separated, our Penalized Pessimistic estimator achieves an average regret bounded by:

$$J(\pi_*, \mathbf{u}_*) - J(\widehat{\pi}, \widehat{\mathbf{u}}) = O_{\mathrm{p}}\left(\sqrt{\frac{\text{Complexity}}{N \times T}}\right),$$

where

$$J(\pi, \mathbf{u}) = (1-\gamma) \sum_{i=1}^{N} \mathbb{E}_{S_0^i \sim \nu} Q(S_0^i, \pi(S_0^i; u^i); u^i)$$

The Takeaway: If you learn strictly individually, your error shrinks at a rate of $1/\sqrt{T}$. By clustering and sharing the Q-function, our error shrinks at $1/\sqrt{NT}$.

We achieve the Oracle rate as if we knew the true patient subgroups in advance.

Simulation I: Synthetic Environment

Data-generating mechanism ($\gamma = 0.8$):

- $S_0^i \sim \mathcal{N}(0, \mathbf{I}_2)$; $A_t^i \sim \text{Bernoulli}(\frac{1}{2})$
- State transition ($\epsilon_t^i \sim \mathcal{N}(0, 0.25\,\mathbf{I}_2)$):

$$[S_{t+1}^i]_1 = 0.8(2A_t^i - 1)[S_t^i]_1 + c_1^i [S_t^i]_2 + \epsilon_1$$
$$[S_{t+1}^i]_2 = c_2^i [S_t^i]_1 + 0.8(1 - 2A_t^i)[S_t^i]_2 + \epsilon_2$$

- Reward:
$R_t^i = 0.9\,\text{expit}\big((2A_t^i - 1)([S_t]_1 - 2[S_t]_2)\big) + \text{Unif}[-0.1, 0.1]$

Heterogeneity via 3 groups:

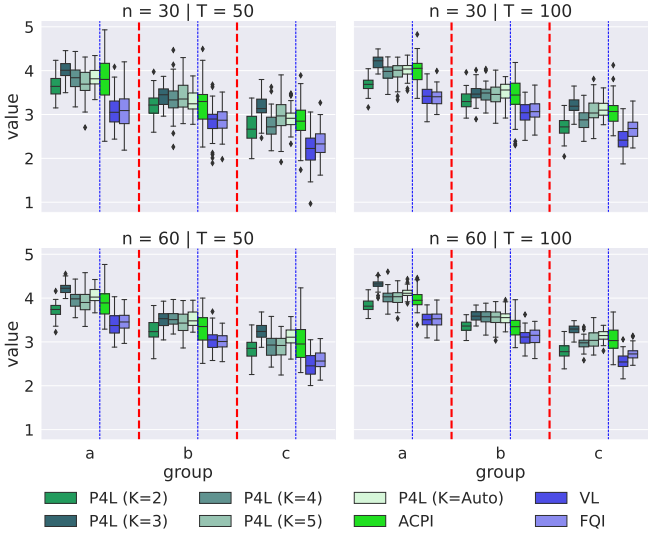
Param.	(a)	(b)	(c)
c_1^i	0	0.6	−0.7
c_2^i	−0.6	0.4	0.5
Prop.	0.4	0.3	0.3

Design:

- $N \in \{30, 60\}$, $T \in \{50, 100, 200\}$
- 50 replications
- Evaluate: 1000 MC trajectories/group

Competitors: FQI, VL, ACPI

Simulation I: Results



Heterogeneous methods:

- **P4L (ours)**
- ACPI [Chen et al., 2022]

Homogeneous methods:

- FQI [Munos, 2003, Le et al., 2019]
- VL [Luckett et al., 2019]

Take-away:

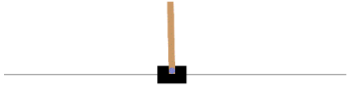
- P4L with oracle $K=3$ consistently best.
- VL/FQI biased by heterogeneity.
- ACPI: larger variance (no cross-group information).

Simulation II: OpenAI CartPole

- **State:** cart position/velocity; pole angle/angular velocity
- **Action:** push right ($+1$) or left (-1)
- **Behavior policy:** $\text{sign}(\theta_t)$
- **Reward:** 1 per step; terminate if $|\theta| > 12$ or $|x| > 2.4$; max 300 steps

Heterogeneous environments:

- Pole length $\in [2, 10]$; push force $\in [0.15, 0.85]$
- Three settings of (length/force) per experiment
- 100 offline trajectories per setting (300 total)
- 50 replications



Three experimental designs:

- **Setting A:** vary length, fix force at 0.85
- **Setting B:** fix length at 5, vary force

Simulation II: CartPole Results

	CartPole — playing steps (mean ± SE)		
Setting A: fix force = 0.85	(2/0.85)	(5/0.85)	(10/0.85)
P4L ($K = 3$)	92.4±17.6	**182.8**±4.3	**226.0**±12.3
P4L (K = Auto)	**98.2**±12.1	174.2±6.9	213.0±19.0
ACPI	89.4±26.5	168.2±24.3	200.1±15.0
FQI	63.4±17.7	65.9±13.7	100.2±11.1
VL	32.7±3.8	71.5±10.1	95.1±10.7
Setting B: fix length = 5	(5/0.15)	(5/0.5)	(5/0.85)
P4L ($K = 3$)	**192.4**±8.1	204.6±8.7	**164.0**±10.2
P4L (K = Auto)	174.2±15.7	206.0±9.1	154.8±8.9
ACPI	168.4±26.5	**216.4**±19.5	151.1±13.9
FQI	121.4±17.7	154.9±24.1	122.0±19.1
VL	146.7±3.8	181.5±10.5	110.1±14.2

P4L (K=3) leads in most columns; *Auto* selection reliably near-oracle.

Real Data Application: Sepsis Treatment (MIMIC-III)

Multi-parameter Intelligent Monitoring in Intensive Care (MIMIC-III).

- $N = 17,621$ Sepsis patients
- From 5 ICUs at a Boston teaching hospital.

In pre-processed data

1. **State**: Top 10 principal components (96% of the total variation) + SOFA score.

   SOFA = Sequential Organ Failure Assessment

2. **Actions:** Vasopressor Injection (3 levels) $\times$ Intravenous fluid Injection (3 levels).

$$|\mathcal{A}| = 3 \times 3 = 9$$

3. **Timestep:** 4 hour.
4. **Rewards:**

$$r(s_t, a, s_{t+1}) = -s_{t+1}^{\text{SOFA}}.$$

   Discount factor $\gamma = 0.8$.

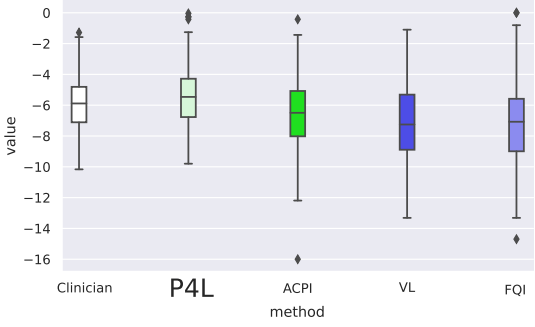
Real Data Application: Sepsis Treatment (MIMIC-III)

Evaluation via Digital Twin

Since we cannot test on patients, we trained a deep generative simulator (PerSim) on held-out data to safely estimate outcomes.

Results:

- Standard AI (Homogeneous) fails to beat the human clinician baseline.
- **Ours (P4L)** achieves the highest simulated survival reward by discovering superior *individualized* treatment regimes.



Summary & Future Work

The P4L Framework:

- **Heterogeneity:** Personalized latent variables replace one-size-fits-all.
- **Offline Coverage:** Weak partial coverage via population-level borrowing.
- **Safety:** Pessimism avoids dangerous recommendations in unexplored regions.
- **Theory:** Oracle regret rate $O_{\mathrm{p}}(1/\sqrt{NT})$ without knowing subgroups.
- **Practice:** Outperforms clinicians in simulated Sepsis treatment.

Future Directions:

- **New individuals:** Classify a new subject to a subgroup and transfer its optimal policy.
- **Non-stationary MDPs:** Policy learning for time-varying heterogeneous environments.
- **Unmeasured confounding:** Individualized RL with unmeasured confounders or time-to-event rewards.

Thank You!

Questions?

Rui Miao, Babak Shahbaba, Annie Qu (2025) Reinforcement learning for individual optimal policy from heterogeneous data. *Ann. Statist.* 53(4): 1513-1534. DOI: 10.1214/25-AOS2512

References I

A. Antos, C. Szepesvári, and R. Munos. Learning near-optimal policies with bellman-residual minimization based fitted policy iteration and a single sample path. *Machine Learning*, 71(1):89–129, 2008.

E. Y. Chen, R. Song, and M. I. Jordan. Reinforcement learning with heterogeneous data: Estimation and inference. *arXiv preprint arXiv:2202.00088*, 2022.

H. Le, C. Voloshin, and Y. Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712. PMLR, 2019.

D. J. Luckett, E. B. Laber, A. R. Kahkoska, D. M. Maahs, E. Mayer-Davis, and M. R. Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.

R. Munos. Error bounds for approximate policy iteration. In *ICML*, volume 3, pages 560–567. Citeseer, 2003.

R. Munos and C. Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9(5), 2008.

T. Xie, C.-A. Cheng, N. Jiang, P. Mineiro, and A. Agarwal. Bellman-consistent pessimism for offline reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 15702–15716, 2021.

W. Zhan, L. Shi, J. D. Lee, and Y. Chi. Offline reinforcement learning with realizability and single-policy concentrability. In *Advances in Neural Information Processing Systems*, volume 35, pages 2730–2742, 2022.