Reinforcement Learning Beyond Classical Assumptions

Zeyu Bian
University of Miami

Joint work with
Chengchun Shi, Zhengling Qi, and Lan Wang

Reinforcement Learning Applications

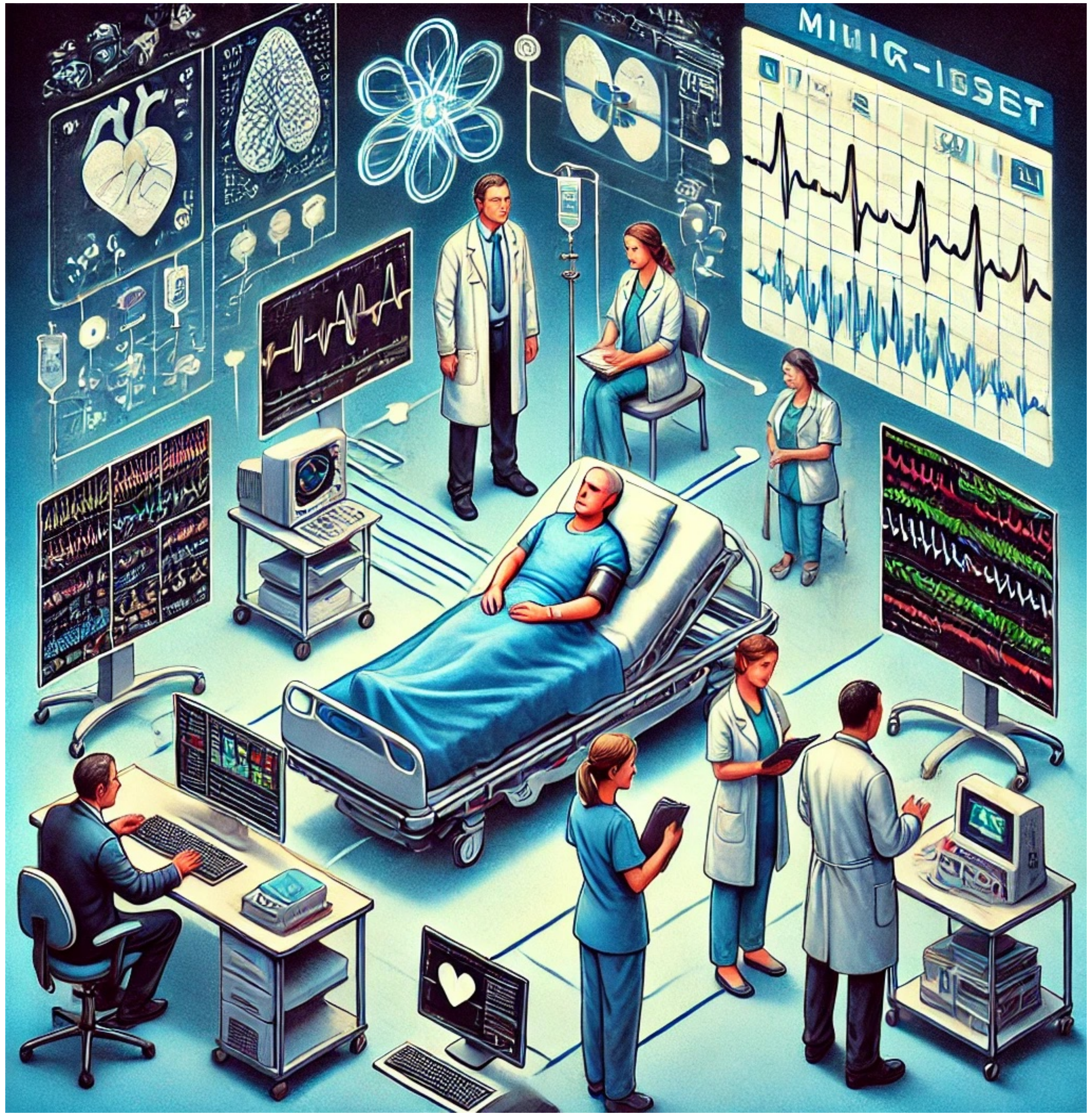
Mobile Health

Self Driving

Ridesharing

Game

Reinforcement Learning in Healthcare



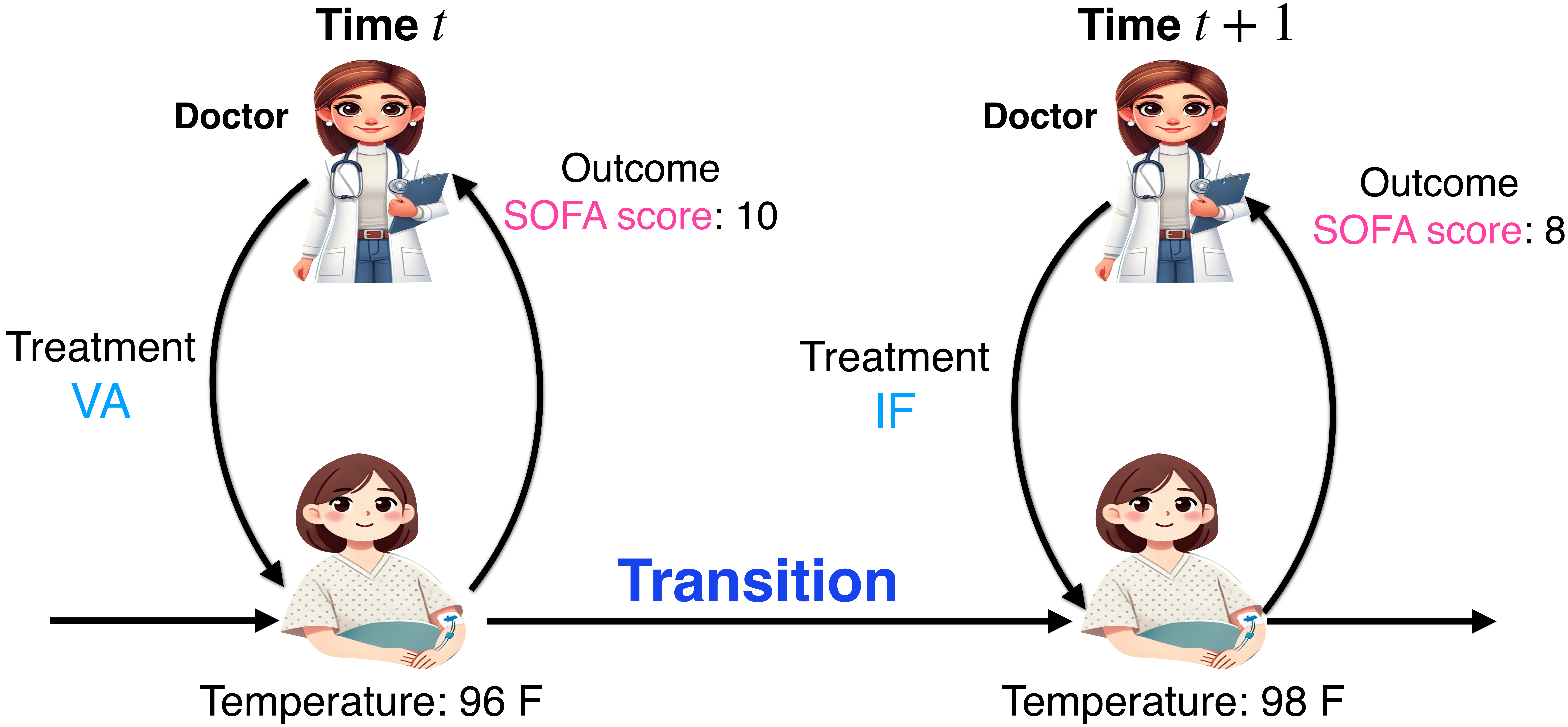
Sepsis

Longitudinal data of sepsis patients from MIMIC-III.

- **Objective**: evaluate patients' long-term outcomes under different treatment strategies.
- **Treatment**: intravenous fluids (IF) vs. vasopressors (VA).

- **Outcome**: SOFA score: measures organ failure.
- **Covariates**: gender, weight, etc.

Sequential Decision Making (Healthcare)



SOFA score measures organ failure, lower scores indicate better outcomes.

Sequential Decision Making

Policy

Policy $\pi \equiv \{\pi_t\}_t$: observation $\mapsto$ probability distribution over the actions.

- One size fits all: $\pi_t^O(IF \mid o) = 1, \forall o.$
- Tailored, stochastic: $\pi_t^T(VA \mid \text{female}) = 0.7.$

Question: how can we measure the effectiveness of a policy?

Policy Evaluation

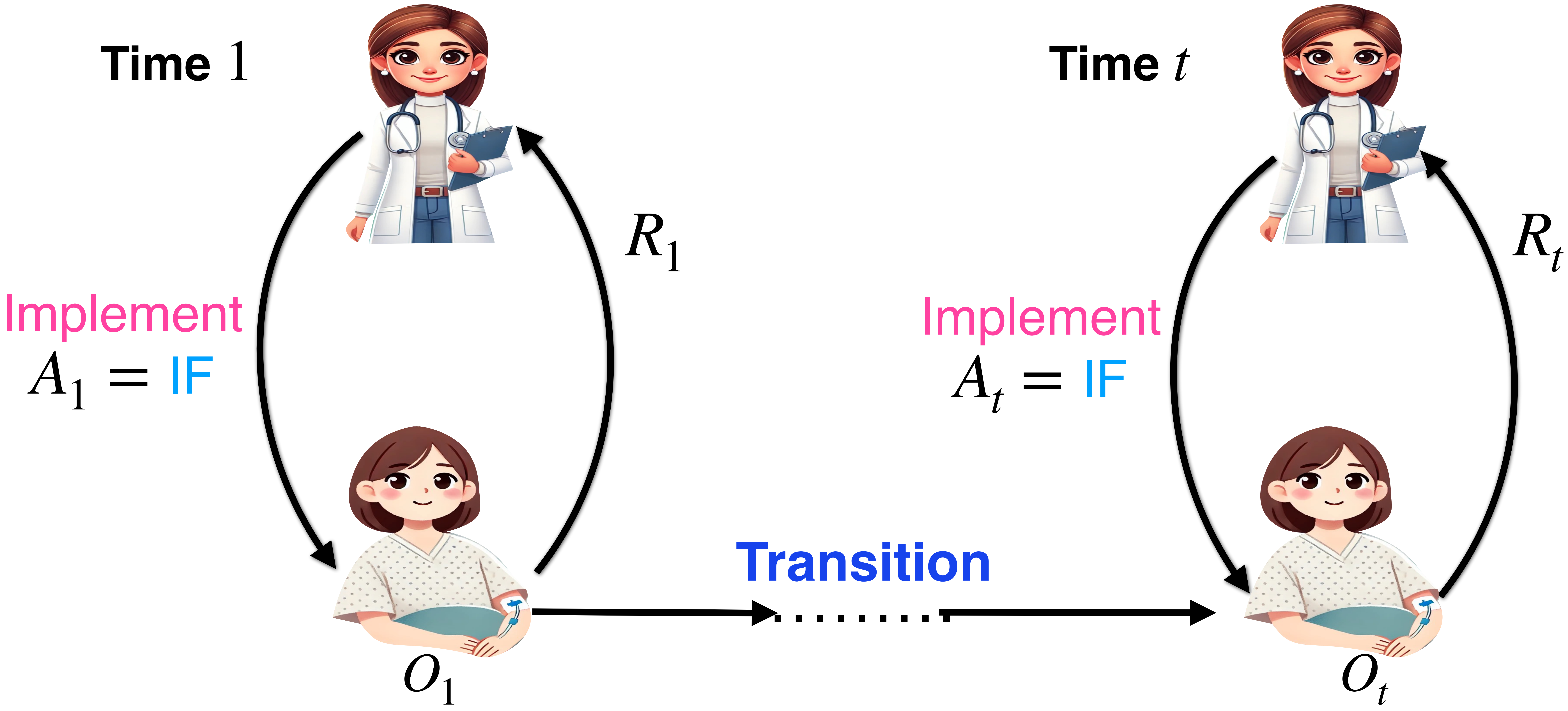
Aim: evaluate the target value $\mathbb{E}^{\pi}(R_t \mid O_1)$ under policy π.

- π is an intervention, and $\mathbb{E}^{\pi}(R_t \mid O_1)$ is analogous to the potential outcome.

Sepsis example:

- $\pi_t^O(IF \mid o) = 1, \forall t, \forall o.$
- $\mathbb{E}^{\pi}(R_t \mid \text{female})$: expected SOFA score at time t, for a female patient if we had applied IF.

Policy Evaluation by Direct Implementation (On-Policy)



$\mathbb{E}^{\pi}(R_t \mid O_1)$ can be approximated using sample average.

Limitation of On-Policy Evaluation

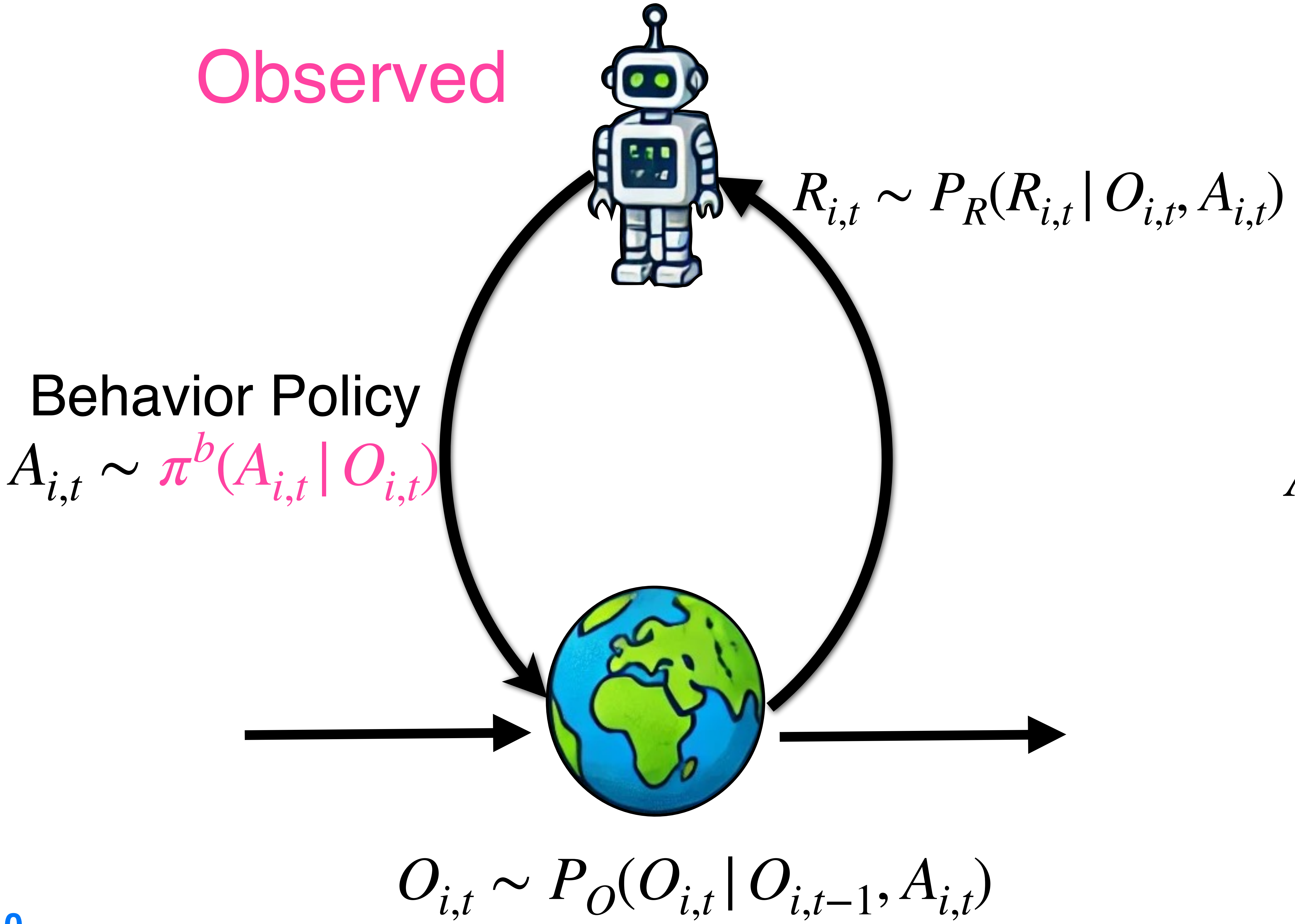
Directly implementing a policy involves potential risks and high costs.

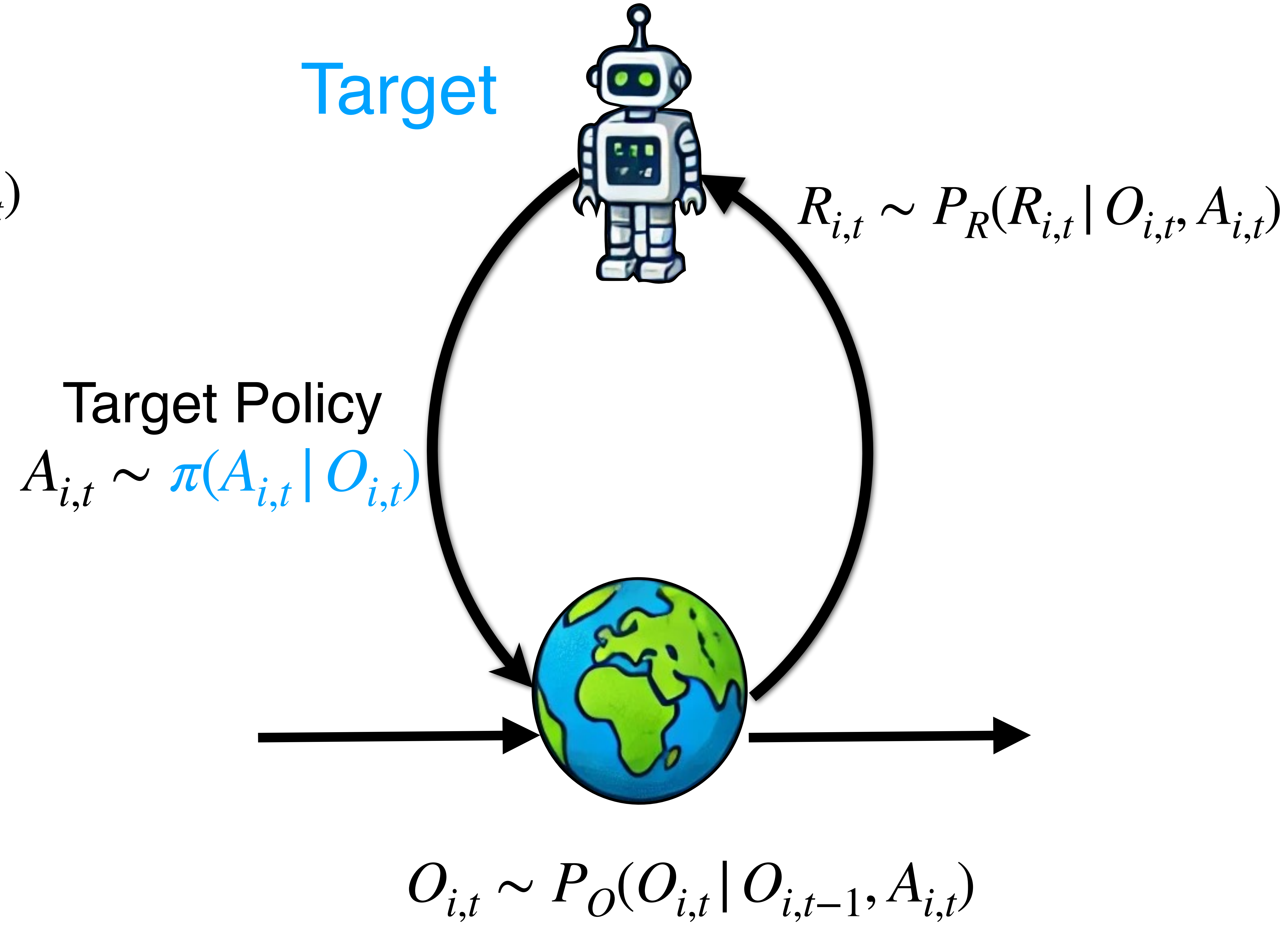
Ridesharing

Self Driving

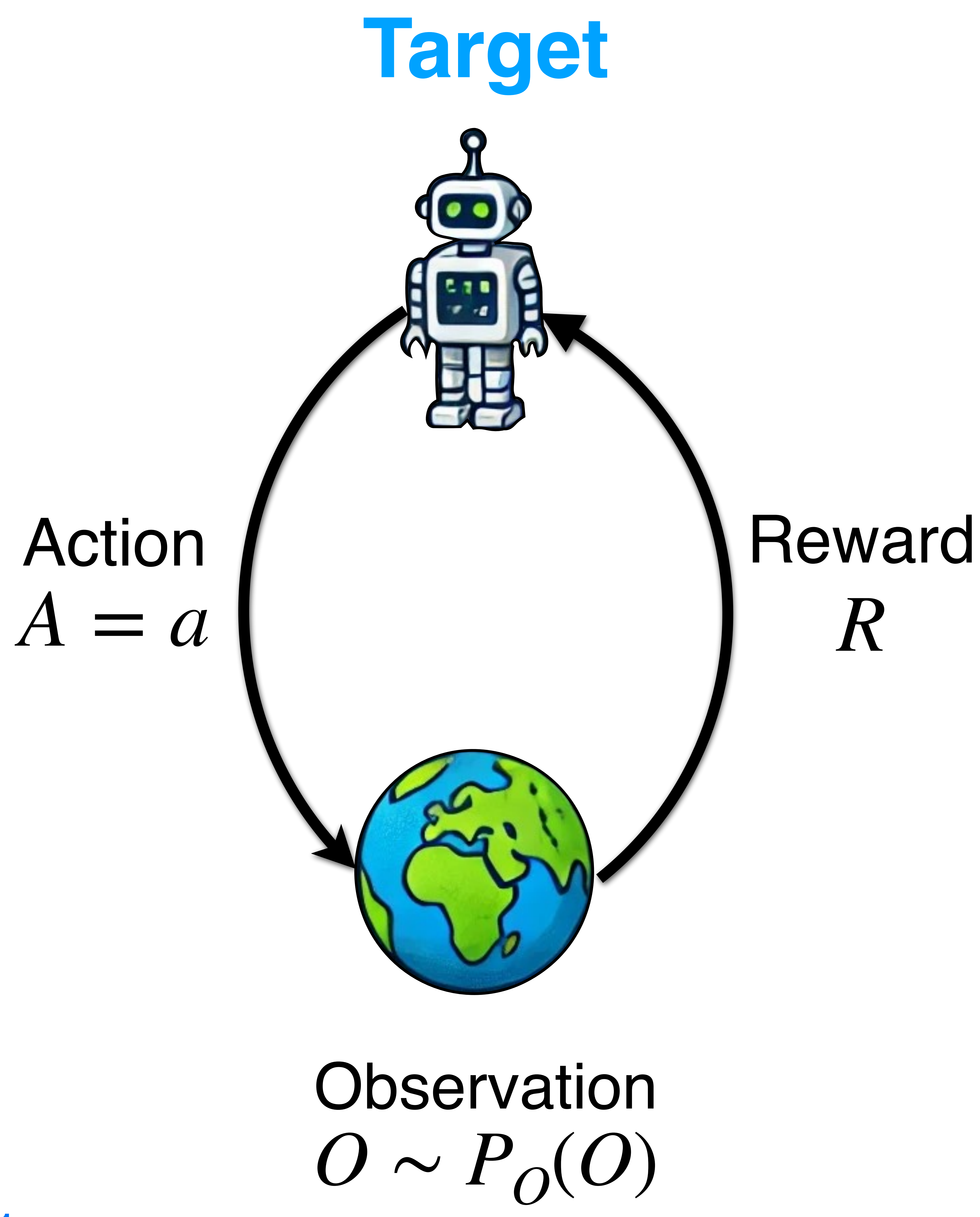
Off-policy Evaluation (OPE)

OPE: evaluate $\mathbb{E}^{\pi}(R_t \mid O_1)$ using offline data (observed) $\{(O_{i,t}, A_{i,t}, R_{i,t}) : 1 \leq i \leq N, 1 \leq t \leq T\}$ generated by π^b.





Connection to Causal Inference



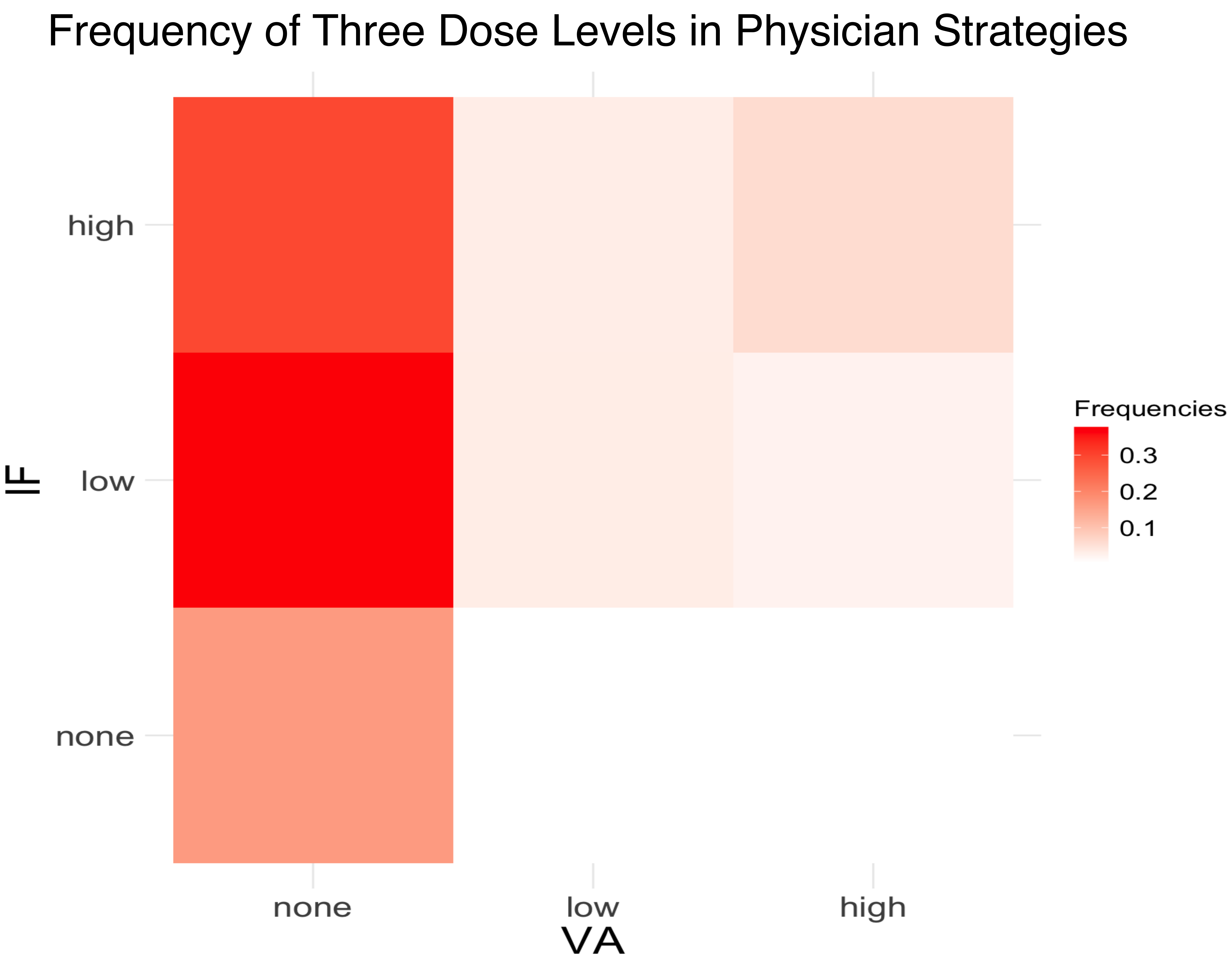
$T = 1$; binary action: $a = 0, 1$.

- $\forall o, \pi(1 \mid o) = 1;$ and $\pi'(0 \mid o) = 1$.
- CATE: $\mathbb{E}^{\pi}(R \mid o) - \mathbb{E}^{\pi'}(R \mid o)$.
- ATE: $\mathbb{E}_O \left[\mathbb{E}^{\pi}(R \mid O) - \mathbb{E}^{\pi'}(R \mid O) \right]$.

Physicians' Preferences in Sepsis Data

Three dosing levels: none, low, and high.

Limited impact of VA (Zhou et al., 2022)

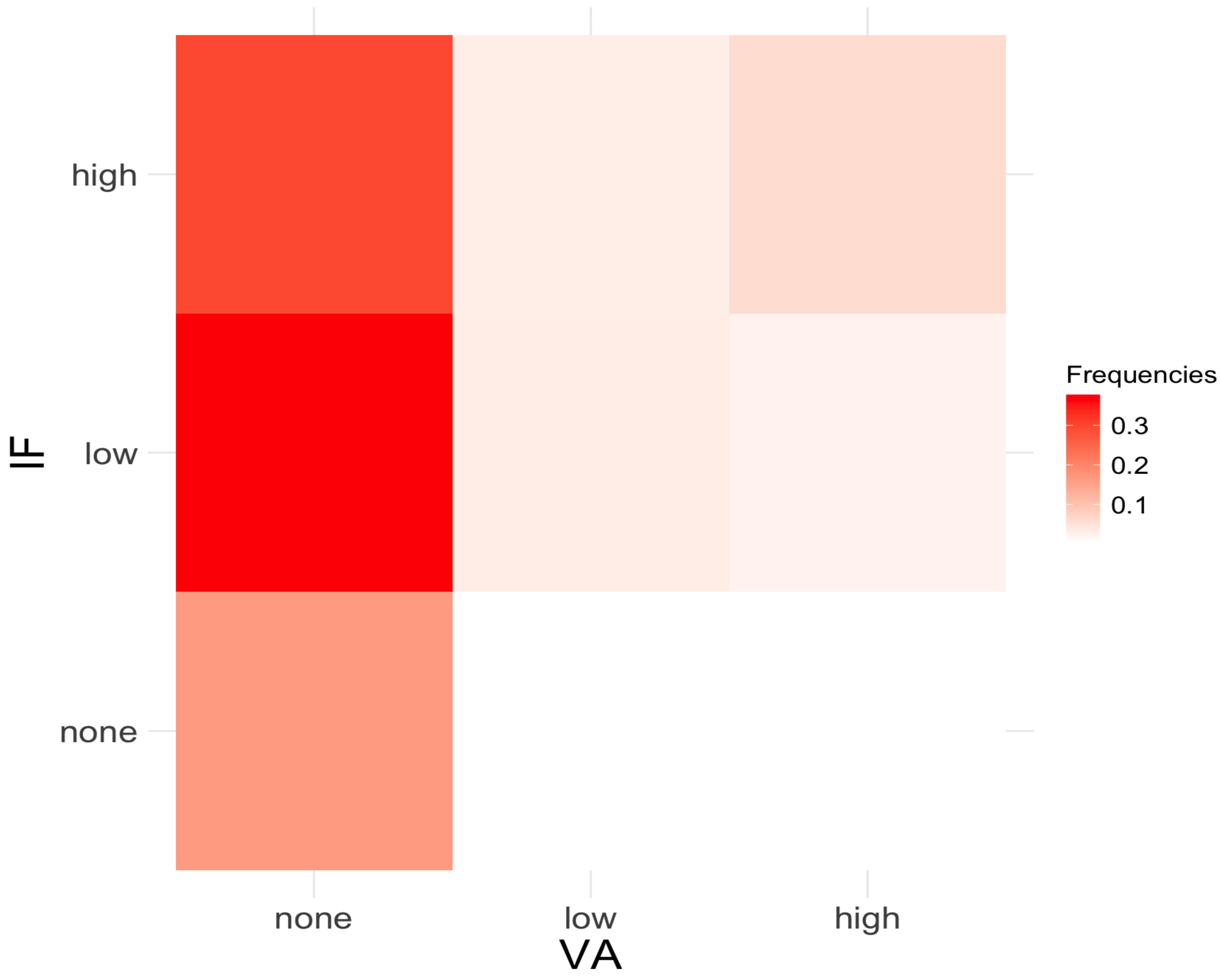


Target Policies

Compare two policies:

- One size fits all policy π^{O} : always low IF.
- Tailored policy π^{T}: a low IF if SOFA < 11; a high IF dose otherwise.

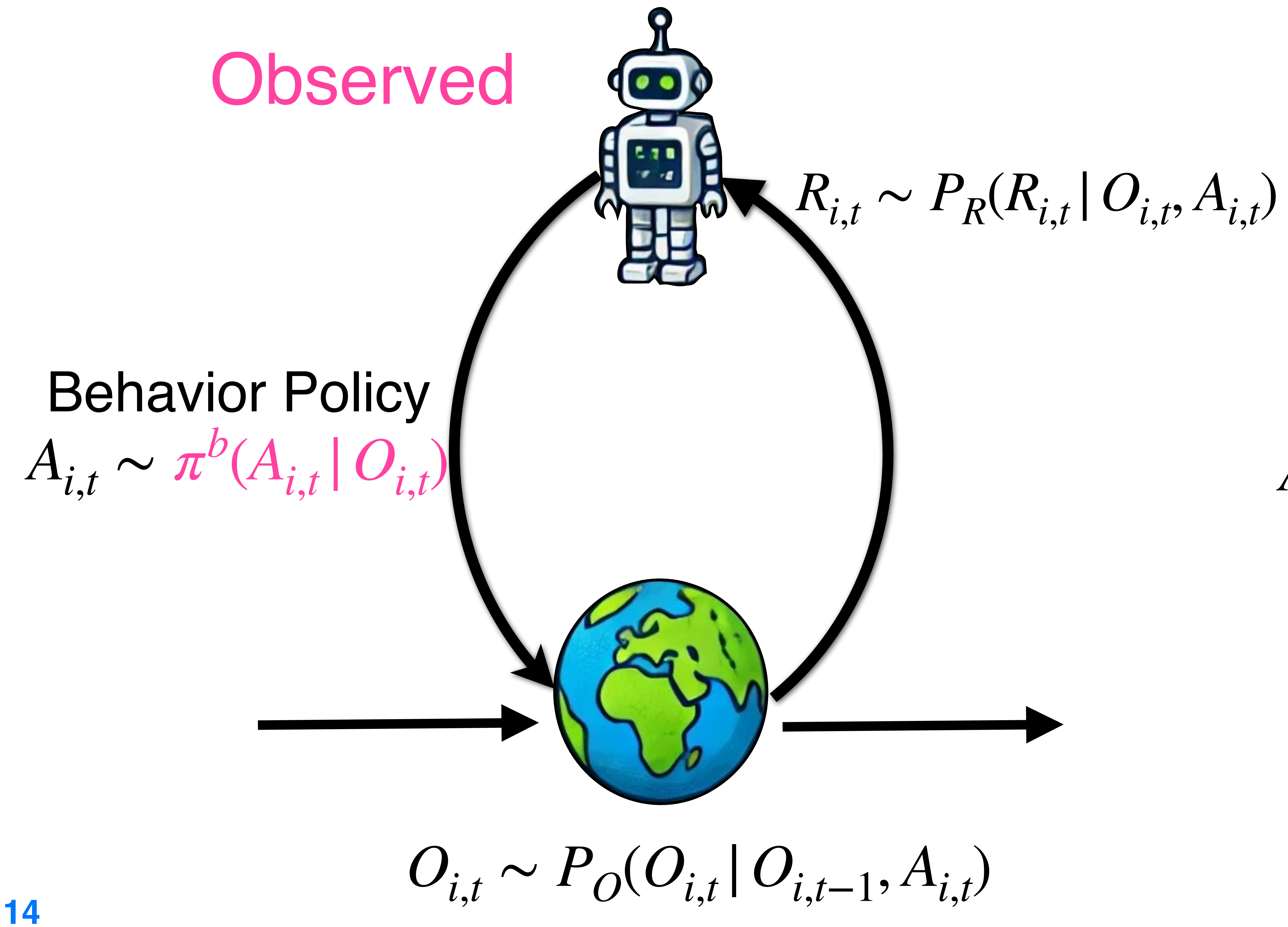
Frequency of Three Dose Levels in Physician Strategies

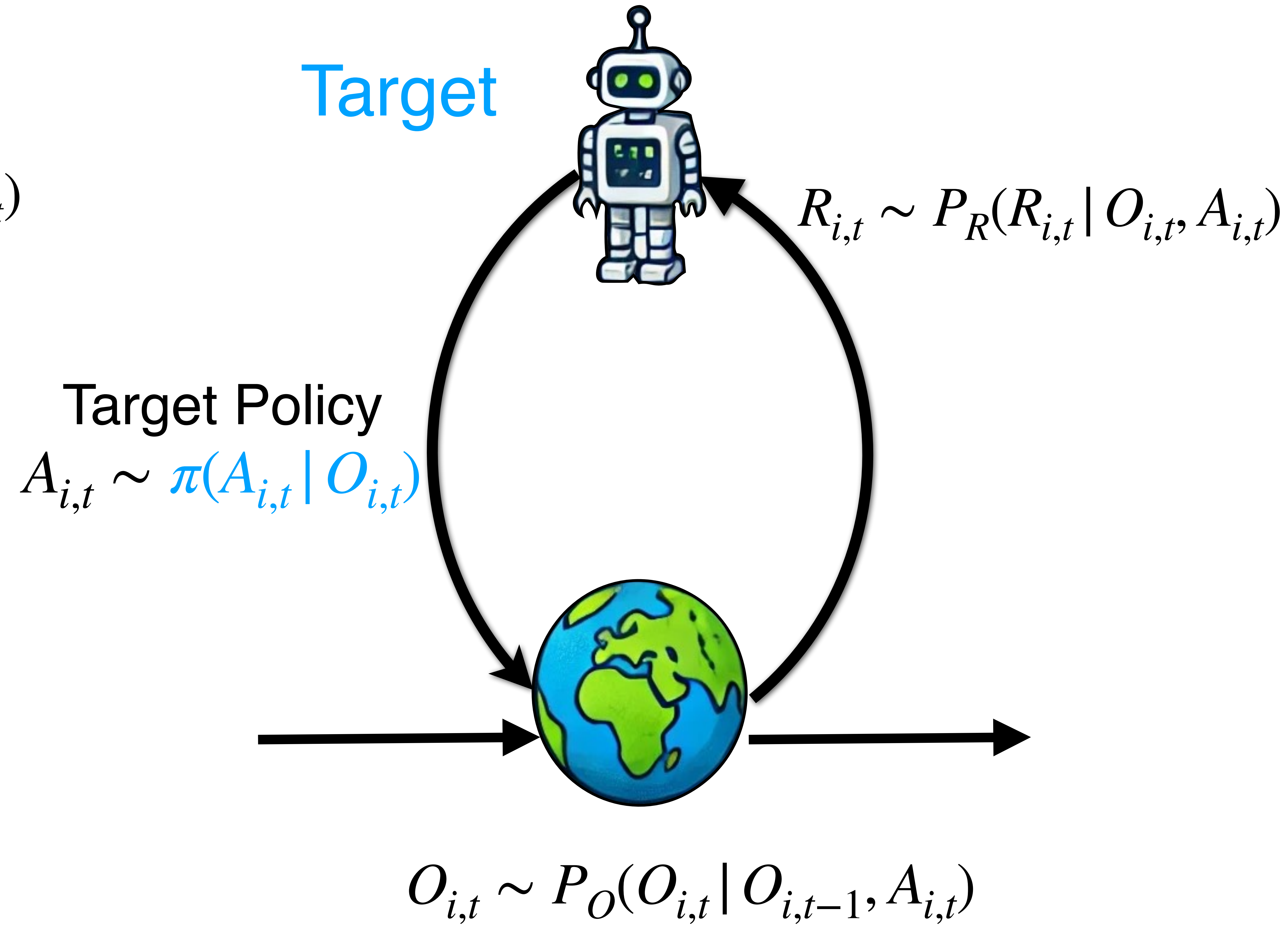


SOFA score > 11 has a 90% mortality rate (Jones et al., 2009).

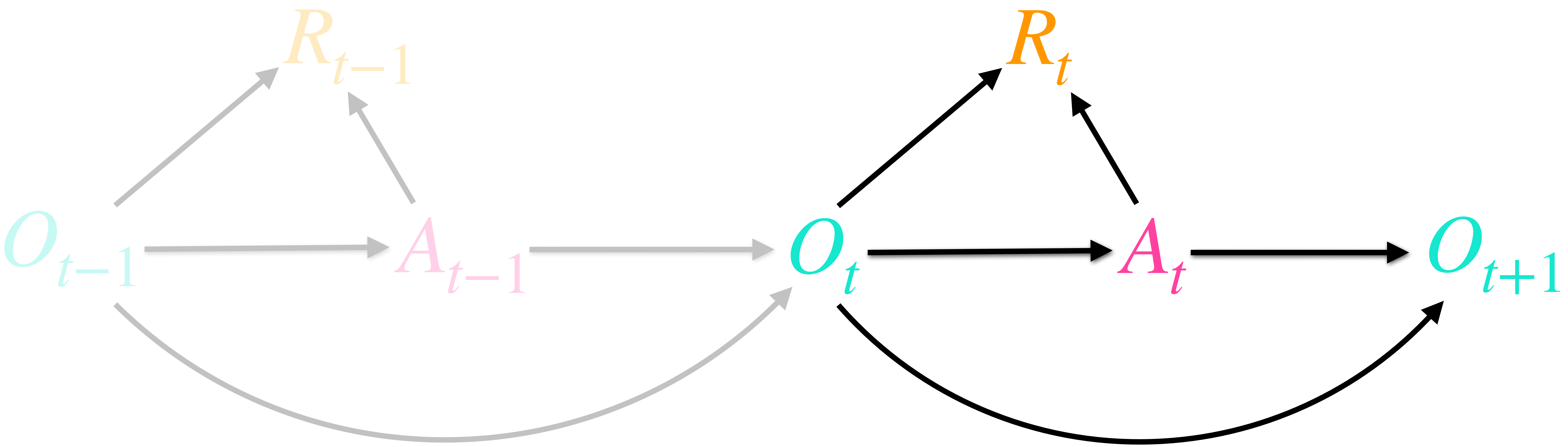
Off-policy Evaluation (OPE)

OPE: evaluate $\mathbb{E}^{\pi}(R_t \mid O_1)$ using offline data generated by π^b.





Classical Key Assumptions in RL



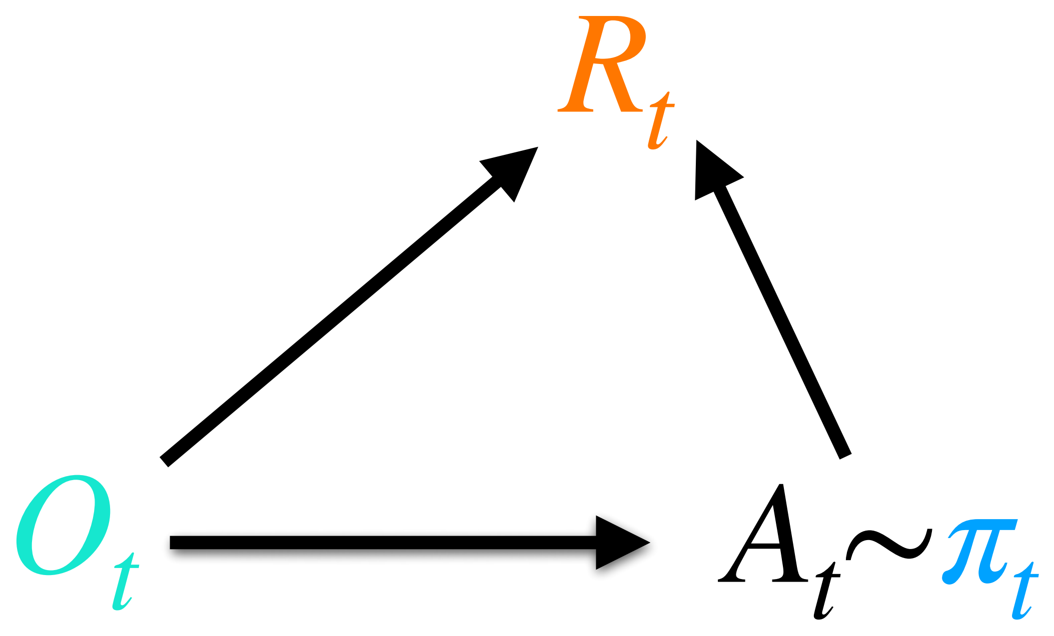
(i) **Markov**: $P_t(O_{t+1}, R_t \mid O_t, A_t, \{O_j, A_j, R_j\}_{1 \le j < t}) = P_t(O_{t+1}, R_t \mid O_t, A_t)$.

(ii) **Stationarity**: the transition $P(\,\cdot\,,\,\cdot \mid \cdot\,,\,\cdot\,)$ does not depend on t.

(iii) **Homogeneity**: the subjects are i.i.d.

OPE Method: Backward Induction under Classical Assumptio

Step 1



$$\mu_t^\pi \equiv \mathbb{E}^\pi(R_t \mid O_t) = \sum_a \underbrace{\mathbb{E}(R_t \mid O_t, a)}_{Q_t^\pi(O_t,a)} \pi_t(a \mid O_t)$$

Step 2

$$\mu_{t-1}^\pi \equiv \mathbb{E}^\pi(R_t \mid O_{t-1}) = \sum_a \underbrace{\mathbb{E}(\mu_t^\pi \mid O_{t-1}, a)}_{Q_{t-1}^\pi(O_{t-1},a)} \pi_{t-1}(a \mid O_{t-1})$$

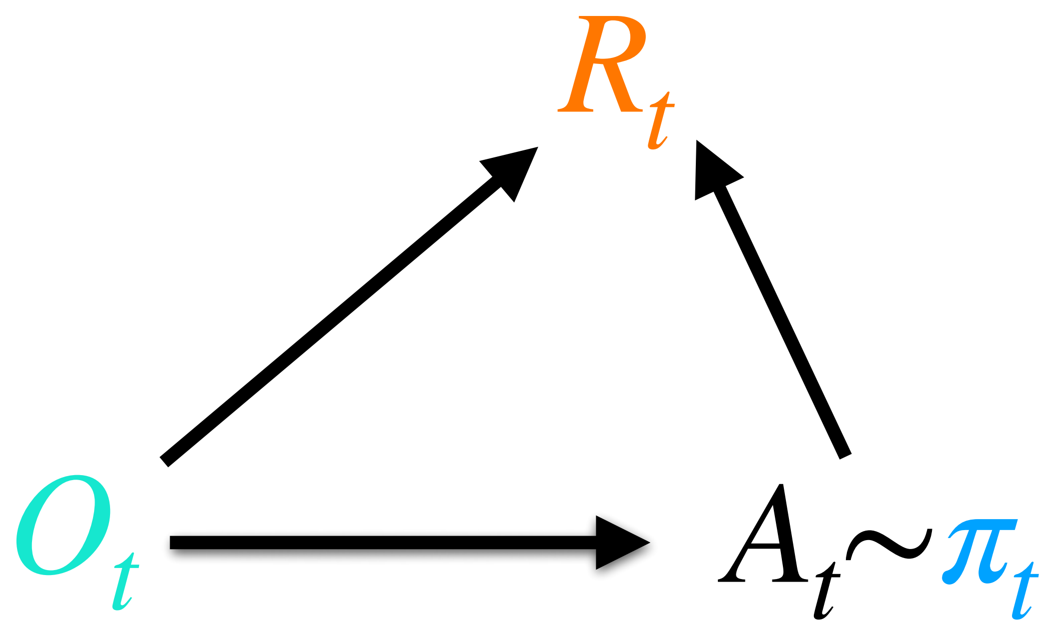
$\vdots$

Step $t+1$ (Target)

$$\mu_1^\pi \equiv \mathbb{E}^\pi(R_t \mid O_1) = \sum_a \underbrace{\mathbb{E}(\mu_2^\pi \mid O_1, a)}_{Q_1^\pi(O_1,a)} \pi_1(a \mid O_1)$$

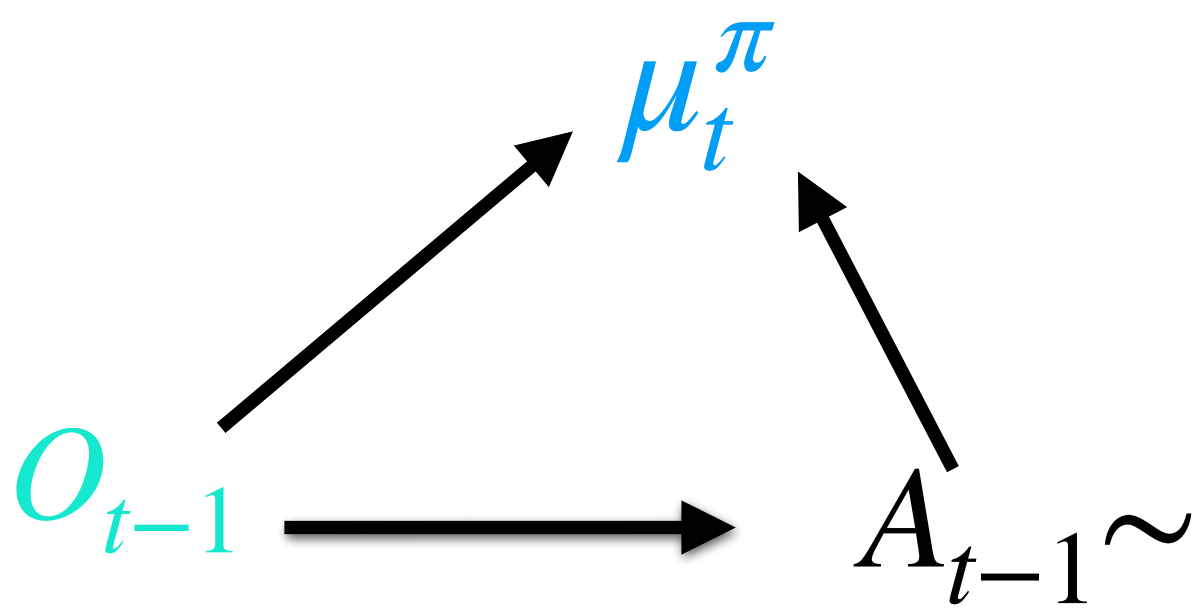
OPE Method: Backward Induction under Classical Assumption

Step 1



$$\mu_t^\pi \equiv \mathbb{E}^\pi(R_t \mid O_t) = \sum_a \underbrace{\mathbb{E}(R_t \mid O_t, a)}_{Q_t^\pi(O_t, a)} \pi_t(a \mid O_t)$$

Step 2



$$\mu_{t-1}^\pi \equiv \mathbb{E}^\pi(R_t \mid O_{t-1}) = \sum_a \underbrace{\mathbb{E}(\mu_t^\pi \mid O_{t-1}, a)}_{Q_{t-1}^\pi(O_{t-1}, a)} \pi_{t-1}(a \mid O_{t-1})$$

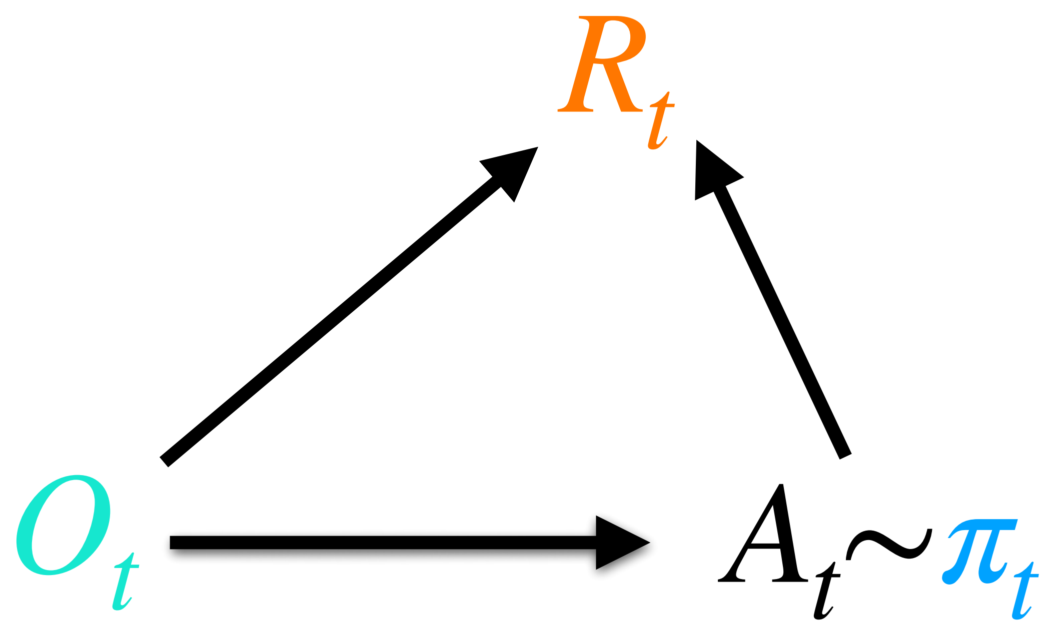
⋮

Step $t+1$
(Target)

$$\mu_1^\pi \equiv \mathbb{E}^\pi(R_t \mid O_1) = \sum_a \underbrace{\mathbb{E}(\mu_2^\pi \mid O_1, a)}_{Q_1^\pi(O_1, a)} \pi_1(a \mid O_1)$$

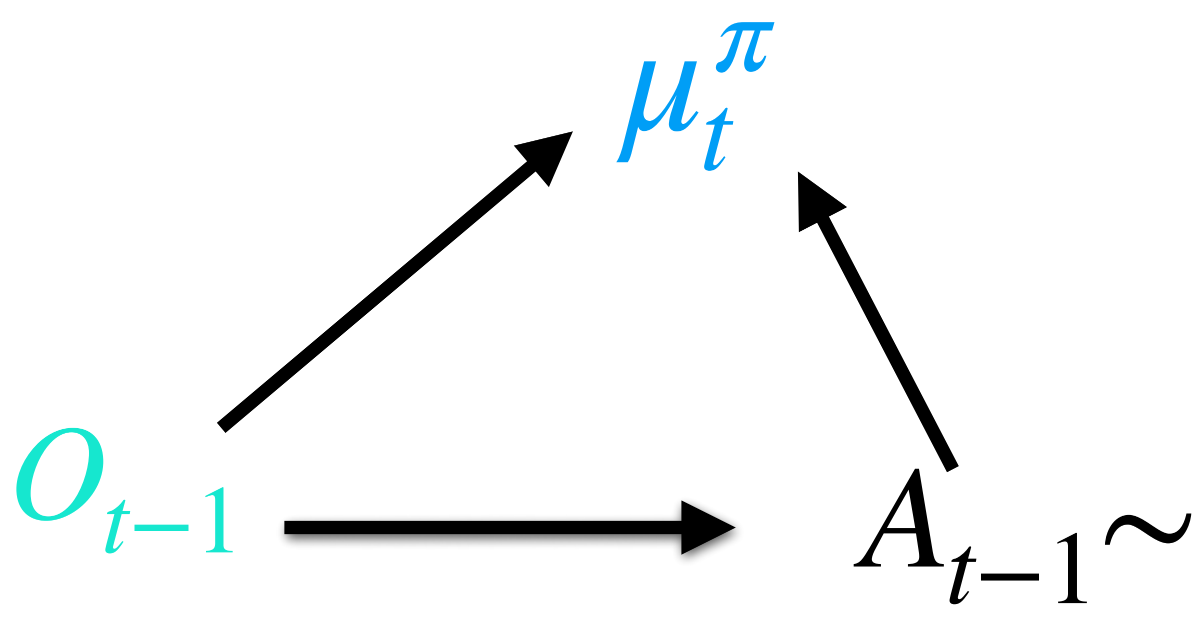
OPE Method: Backward Induction under Classical Assumptio

Step 1



$$\mu_t^\pi \equiv \mathbb{E}^\pi(R_t \mid O_t) = \sum_a \underbrace{\mathbb{E}(R_t \mid O_t, a)}_{Q_t^\pi(O_t,a)} \pi_t(a \mid O_t)$$

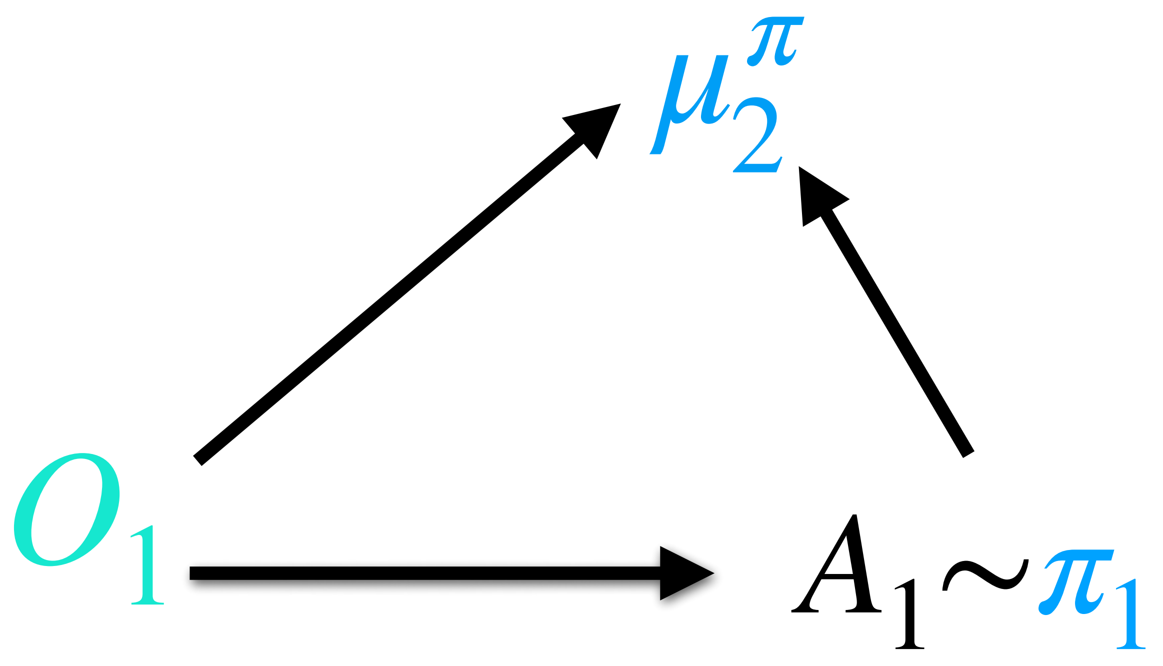
Step 2



$$\mu_{t-1}^\pi \equiv \mathbb{E}^\pi(R_t \mid O_{t-1}) = \sum_a \underbrace{\mathbb{E}(\mu_t^\pi \mid O_{t-1}, a)}_{Q_{t-1}^\pi(O_{t-1},a)} \pi_{t-1}(a \mid O_{t-1})$$

⋮ ⋮ ⋮ ⋮

Step $t+1$ (Target)



$$\mu_1^\pi \equiv \mathbb{E}^\pi(R_t \mid O_1) = \sum_a \underbrace{\mathbb{E}(\mu_2^\pi \mid O_1, a)}_{Q_1^\pi(O_1,a)} \pi_1(a \mid O_1)$$

Impact of Three Key Assumptions

Consider backward induction,

(i) **Markov**: Q_t^π and π_t only depends on current $\Longrightarrow$ simplify decision process.

(ii) **Stationarity**: Q_t^π can be learned using all T time points.

(iii) **Homogeneity**: Q_t^π can be learned using all N subjects.

- (ii) and (iii) allow us to use the data effectively.

Possible Violation of Assumptions in Practice

Sepsis

i. **Heterogeneous** treatment responses (Evans et al. 2021).

ii. Information over 10 years $\Longrightarrow$ **non-stationary**.

iii. Questionnaire responses may only partially reflect the patient's state $\Longrightarrow$ **non-Markov**.

No method addresses all three challenges simultaneously.

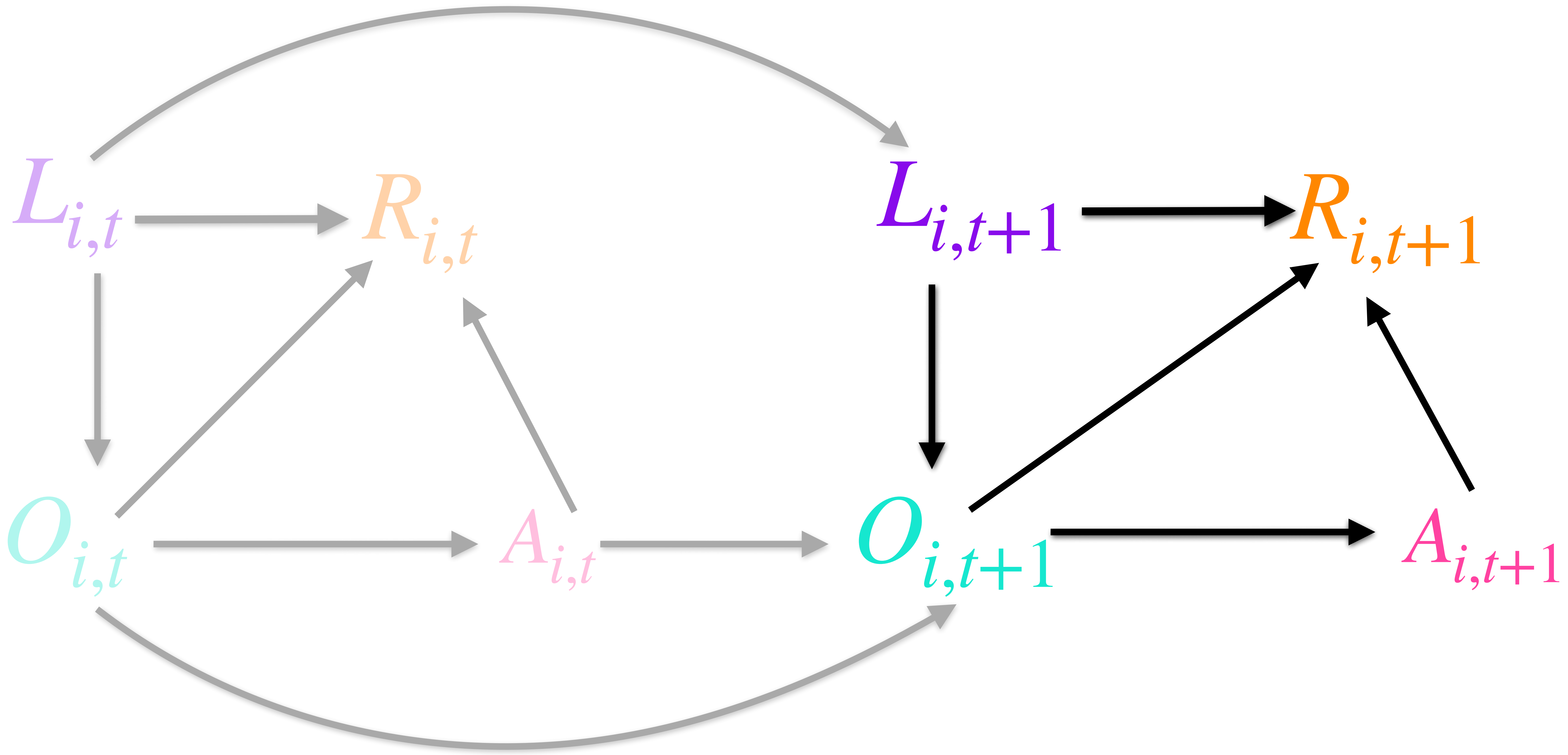
Literature Review

	Heterogeneous	Non-stationary	Non-Markov	$t \to \infty$
Fitted-Q		✓	✓	$t \ll N$
Importance Sampling		✓	✓	fixed t
Double RL	✓			✓

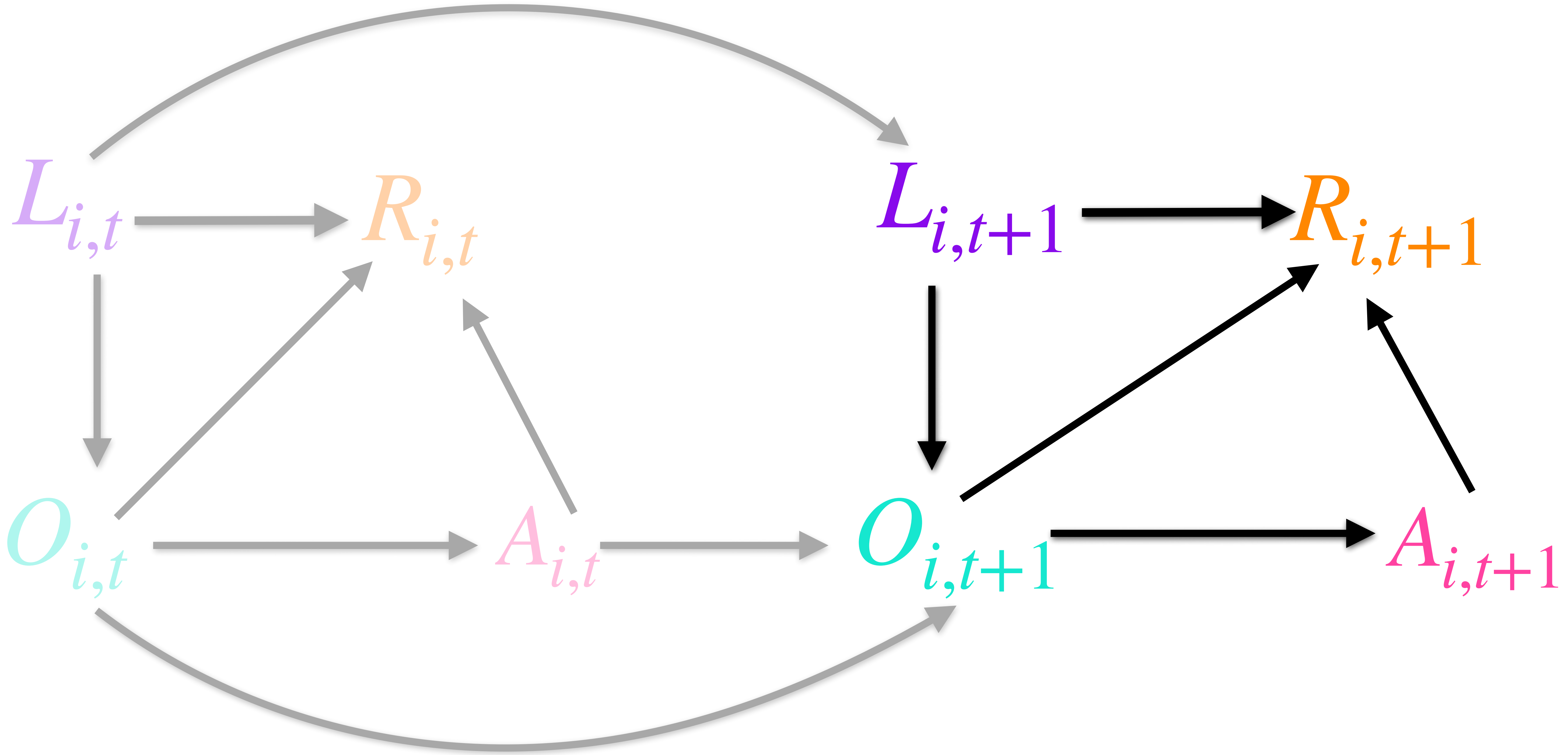
Our General Framework

Markov holds only when conditioned on individual- and time-specific **latent** factors $\{U_i\}_{i=1}^N$ and $\{V_t\}_{t=1}^T$.

$L_{i,t} = (U_i, V_t)$



Our General Framework



- Heterogeneity: $\{U_i\}$, e.g., genetic information.
- Non-stationary: $\{V_t\}$, e.g., disease progression.
- Non-Markov: $(O_{i,t+1}, R_{i,t}) \not\perp\!\!\!\perp \{O_{i,j}, A_{i,j}, R_{i,j}\}_{1 \le j < t} \mid (O_{i,t}, A_{i,t})$.

Adjust for Unobserved Latent Factors

Inspired by the **two-way fixed effects** model:

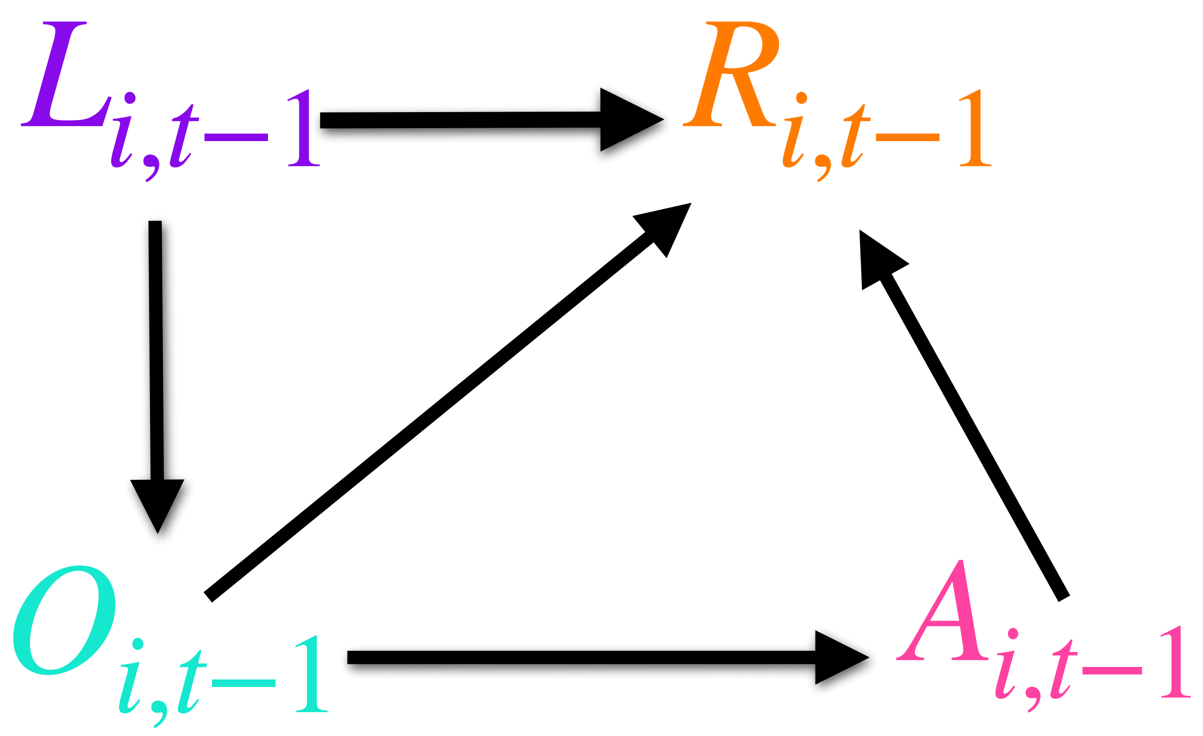
- Practical model to account for unobserved variables.
- Model

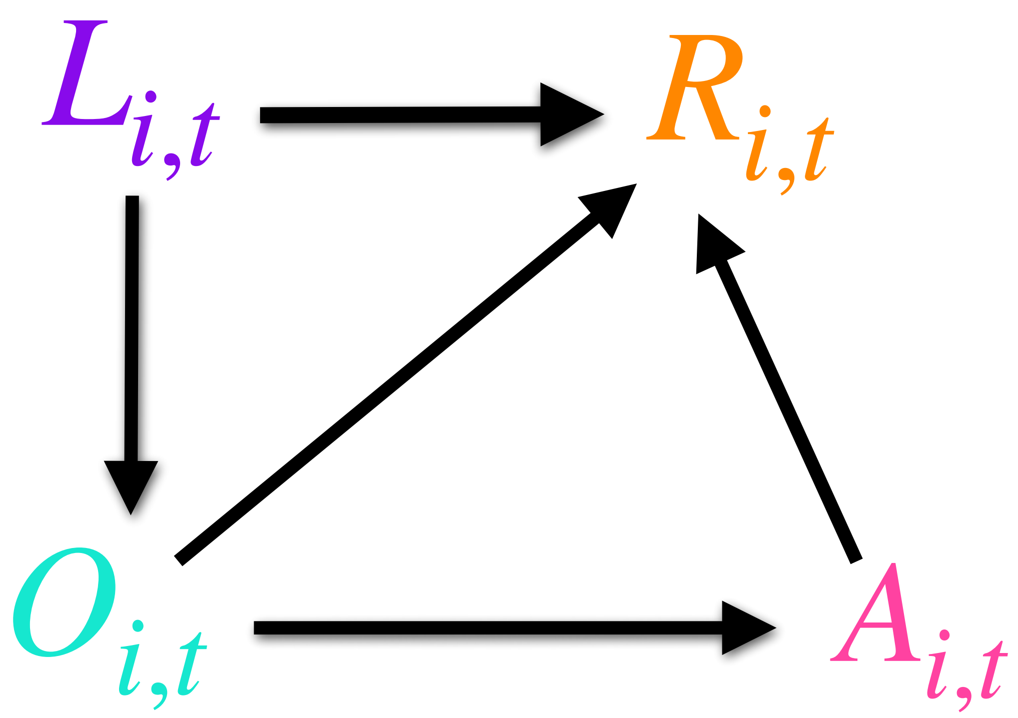
$$R_{i,t} = \underbrace{\theta_i}_{\text{subject effect}} + \underbrace{\lambda_t}_{\text{time effect}} + \underbrace{r(O_{i,t}, A_{i,t})}_{\text{main effect}} + \varepsilon_{i,t}.$$

- Solving

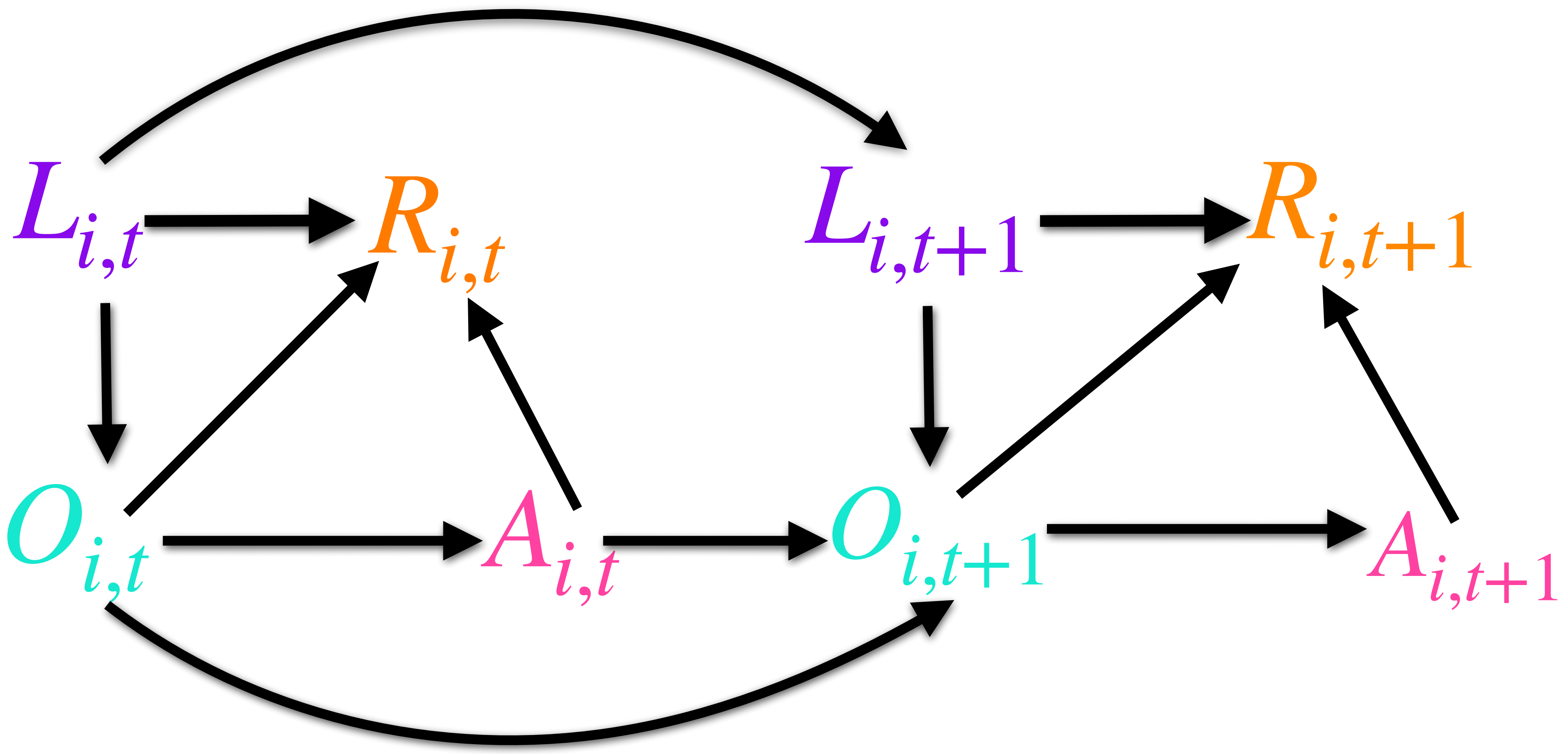
$$\hat{r} = \text{argmin}_{r \in \mathscr{R}} \frac{1}{NT} \sum_{i,t} \left[R_{i,t} - \theta_i - \lambda_t - r(O_{i,t}, A_{i,t}) \right]^2.$$

Comparison with the 2WFE Model





Classical 2WFE Model



Our Model

Additive Assumption

The transition is additive w.r.t. u_i, v_t and (o, a):

$$p(O_{i,t+1} \mid u_i, v_t, o, a)$$

$$= \omega_u p_{u_i}(O_{i,t+1} \mid u_i) + \omega_v p_{v_t}(O_{i,t+1} \mid v_t) + \omega_0 p_0(O_{i,t+1} \mid o, a),$$

with $\omega_u + \omega_v + \omega_0 = 1$.

$\omega_0 = 1 \implies$ Markov Assumption.

Two-way Structure of the Q-function

Define
$Q^{\pi}_{i,k}(o,a) = \mathbb{E}^{\pi}(R_{i,t} \mid O_{i,k} = o, A_{i,k} = a, u_i, v_k)$.

Theorem 1 Under the additive assumption,

$$Q^{\pi}_{i,k}(o,a) = \theta_{i,k} + \lambda_{t,k} + r_k(o,a),$$

where $\theta_{i,k}$ and $\lambda_{t,k}$ are non-stochastic.

Estimand

We focus on the individual- and time-specific value:

$$\eta_{i,t}^{\pi} \equiv \mathbb{E}^{\pi}(R_{i,t} \mid O_{i,1}, U_i, V_1).$$

Sepsis data:

- Individualization enables tailored interventions.
- Timing is related to disease progression: early intervention for sepsis within the first 6–12 hours is crucial.

Other Estimands

Individual- and time-specific value:

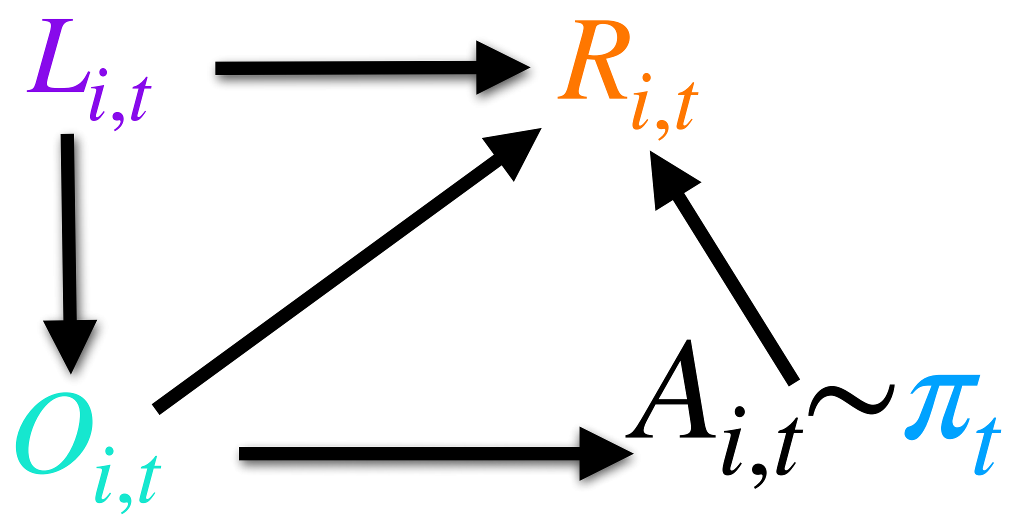
$$\eta_{i,t}^{\pi} \equiv \mathbb{E}^{\pi}(R_{i,t} \mid O_{i,1}, U_i, V_1).$$

Other interests:

- Individual-specific value: $\eta_i^{\pi} \equiv \frac{1}{T}\sum_{t=1}^{T} \eta_{i,t}^{\pi}.$
- Time-specific value: $\eta_t^{\pi} \equiv \frac{1}{N}\sum_{i=1}^{N} \eta_{i,t}^{\pi}.$
- Average reward: $\eta^{\pi} \equiv \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} \eta_{i,t}^{\pi}.$

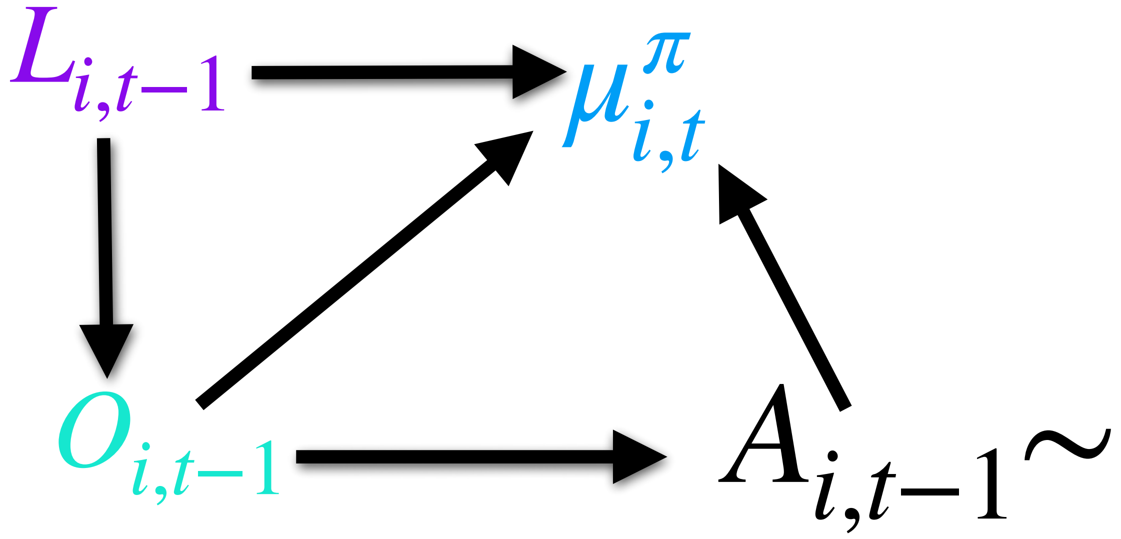
Backward Induction with Two-way Fixed Effects

Step 1



$$\mu^\pi_{i,t} = \mathbb{E}^\pi(R_{i,t} \mid O_{i,t}, u_i, v_t) = \sum_a Q^\pi_{i,t}(O_{i,t}, a)\pi_t(a \mid O_{i,t})$$

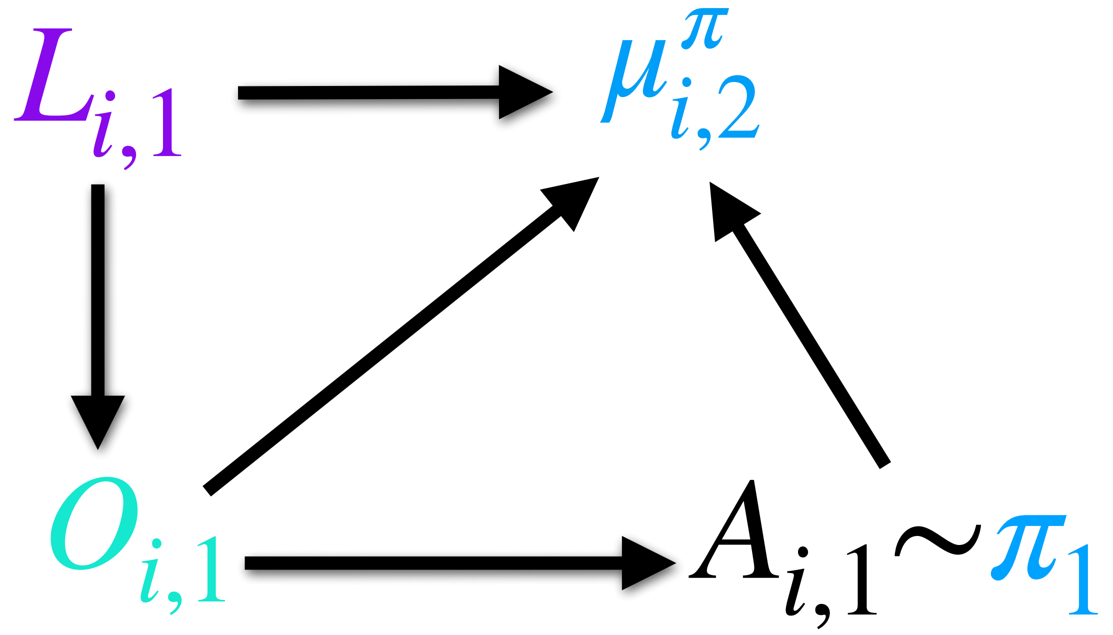
Step



$$\mu^\pi_{i,t-1} = \sum_a \underbrace{\mathbb{E}(\mu^\pi_{i,t} \mid O_{i,t-1}, a, u_i, v_{t-1})}_{Q^\pi_{i,t}(O_{i,t},a)} \pi_{t-1}(a \mid O_{i,t-1})$$

⋮ ⋮ ⋮ ⋮

Step $t+1$ (Target)



$$\eta^\pi_{i,t} = \mu^\pi_{i,1} = \sum_a \underbrace{\mathbb{E}(\mu^\pi_{i,2} \mid O_{i,1}, a, u_i, v_1)}_{Q^\pi_{i,1}(O_{i,1},a)} \pi_1(a \mid O_{i,1})$$

Algorithm

Pseudocode for Estimating $\eta_{i,t}^{\pi}$

1. Set $\widehat{\mu}_{i,t+1}^{\pi} = R_{i,t}$.
2. **for** $k = t, t-1, \cdots, 1$ **do**
3. Solve

$$(\widehat{\theta}_{i,k}, \widehat{\lambda}_{t,k}, \widehat{r}_k) = \text{argmin}_{\theta_{i,k},\lambda_{t,k},r_k} \sum_{i,j} \left[\widehat{\mu}_{i,k+1}^{\pi} - \theta_{i,k} - \lambda_{t,k} - r_k(O_{i,j}, A_{i,j}) \right]^2$$

4. $\widehat{Q}_{i,k}^{\pi}(o,a) = \widehat{\theta}_{i,k} + \widehat{\lambda}_{t,k} + \widehat{r}_k(o,a)$
5. Compute $\widehat{\mu}_{i,k}^{\pi} = \sum_a \widehat{Q}_{i,k}^{\pi}(O_{i,k}, a)\pi(a \,|\, O_{i,k})$
6. **end for**
7. Output: $\widehat{\eta}_{i,t}^{\pi} = \widehat{\mu}_{i,1}^{\pi}$

Uniform Convergence Rate

Theorem 2 Under some regularity conditions,

$$\max_{i,t} \left| \widehat{\eta}^{\pi}_{i,t} - \eta^{\pi}_{i,t} \right| = O_p\left(\sqrt{\log(NT)/\min(N,T)} \right).$$

Recap: Literature Review

	Heterogeneous	Non-stationary	Non-Markov	$t \to \infty$
Fitted-Q		✓	✓	$t \ll N$
Importance Sampling		✓	✓	fixed t
Double RL	✓			✓
Our Method	✓	✓	✓	✓

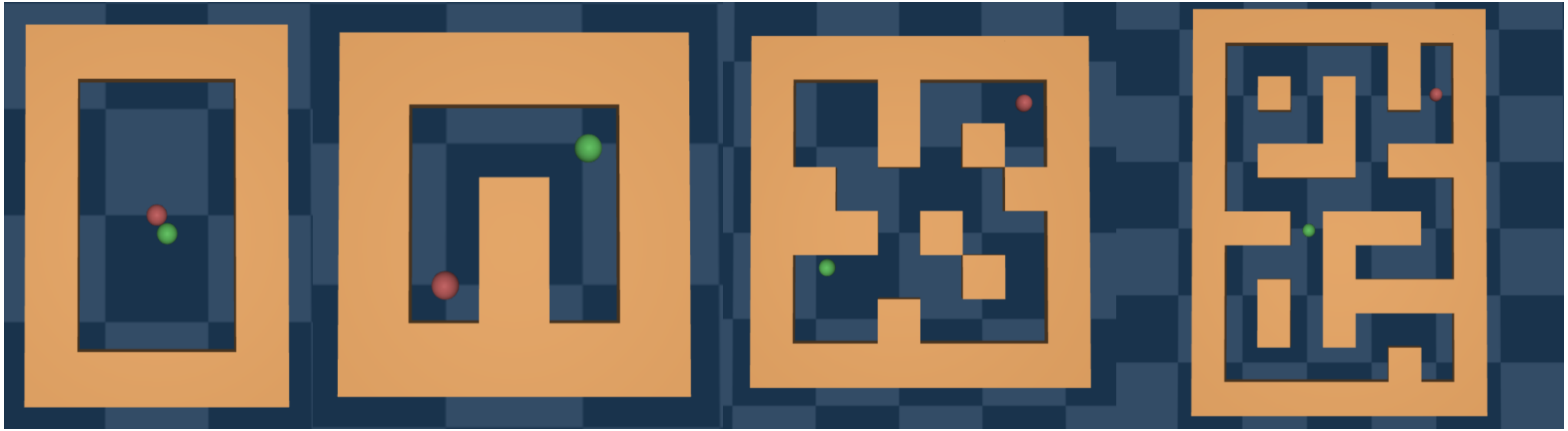
Asymptotically Normality

Theorem 3 Under some regularity conditions, we have

$$\sqrt{\min(N,T)}\sigma^{-1}\left(\widehat{\eta}^{\pi}_{i,t}-\eta^{\pi}_{i,t}\right)\xrightarrow{D}\mathcal{N}(0,1).$$

Numerical Study I: D4RL

D4RL dataset is specifically designed for evaluating RL algorithms.



Maze2D task, the 4 settings differ in maze layouts and the level of difficulty.

D4RL Results

Table 1: MSEs of the estimated value (four targets) using our proposed methods and other competing methods for Maze2D with $N = T = 20$ over 20 replications. The best method with smallest MSE in each column were highlighted with blue.

	Maze2D-open				Maze2D-umaze				Maze2D-medium				Maze2D-large			
	η^π	η^π_i	η^π_t	$\eta^\pi_{i,t}$	η^π	η^π_i	η^π_t	$\eta^\pi_{i,t}$	η^π	η^π_i	η^π_t	$\eta^\pi_{i,t}$	η^π	η^π_i	η^π_t	$\eta^\pi_{i,t}$
2OPE	0.01	0.45	0.30	0.75	0.02	0.47	0.28	0.73	0.00	0.45	0.31	0.76	0.00	0.45	0.32	0.77
DM1	0.01	0.49	0.34	0.83	0.03	0.52	0.33	0.82	0.01	0.51	0.35	0.85	0.00	0.50	0.36	0.85
DM2	3.75	4.25	3.72	4.23	2.98	3.49	3.93	4.43	0.75	1.26	1.12	1.64	0.55	1.07	0.96	1.48
IS1	0.66	1.17	1.26	3.63	0.42	0.93	0.39	2.06	0.35	0.87	0.62	2.56	0.62	1.13	1.12	3.34
IS2	1.52	2.03	6.10	10.12	1.81	2.32	4.65	8.06	0.93	1.44	3.43	6.67	1.28	1.80	5.22	8.94
IS3	0.01	0.52	0.35	0.85	0.03	0.54	0.33	0.84	0.01	0.52	0.35	0.87	0.00	0.52	0.36	0.87
DR1	0.25	2.99	0.44	7.03	0.99	12.81	3.11	60.60	0.15	1.80	0.38	7.45	0.21	1.41	0.28	4.31
DR2	0.25	3.09	1.16	13.04	0.13	2.68	0.65	9.86	0.18	2.35	0.64	8.82	0.21	1.95	0.64	8.06
DR3	0.01	0.51	0.36	0.86	0.03	0.54	0.33	0.84	0.01	0.52	0.36	0.87	0.00	0.52	0.36	0.88

Numerical Study II: Sensitivity Analysis

Table 2: A summary of environments in the sensitivity analysis.

Environment	I	II	III	IV
Reward	additive	additive	interactive	interactive
Transition	clustering	interactive	additive	interactive

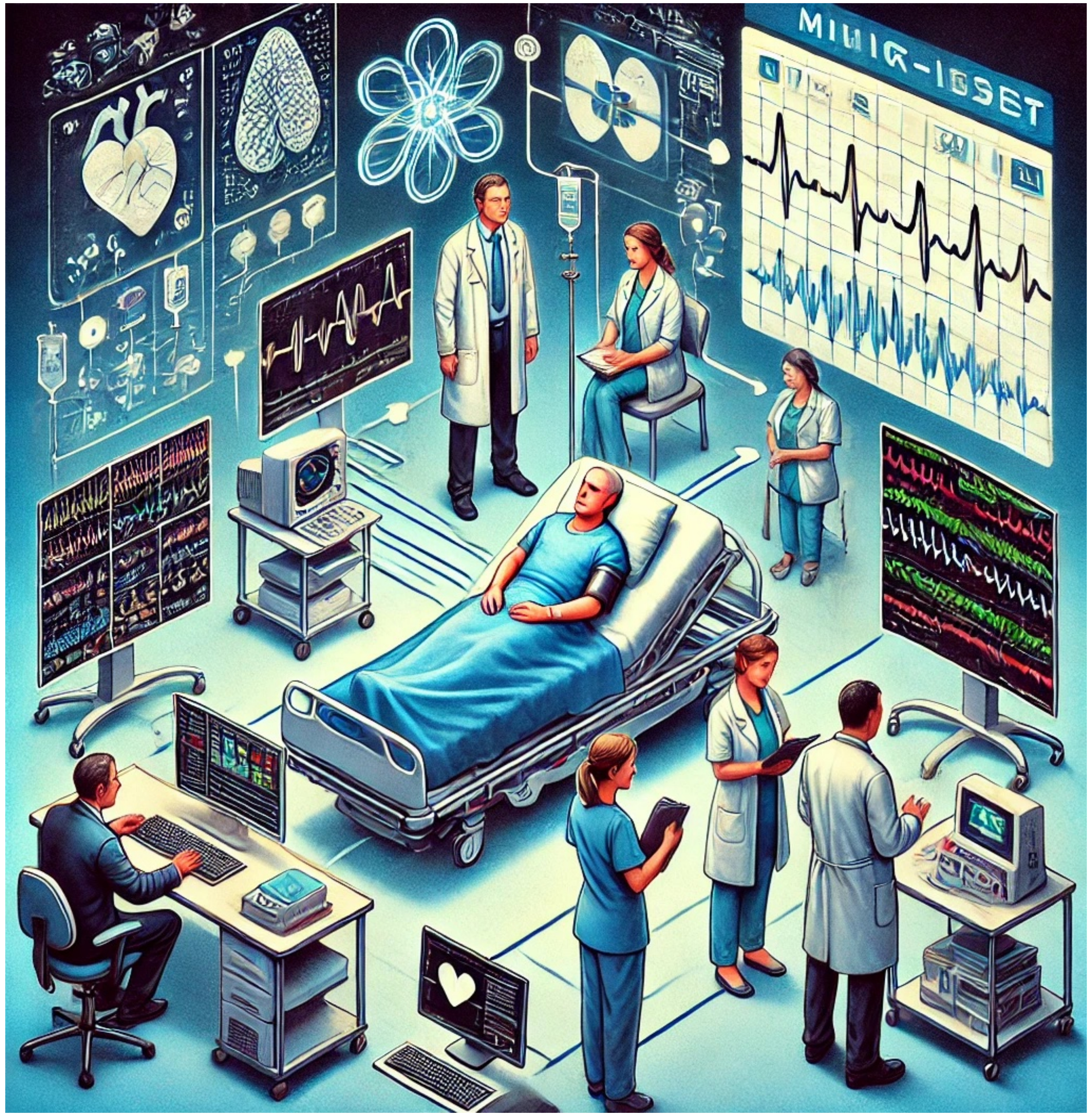
The additive assumption is violated in each scenario.

Sensitivity Analysis Results

Table 3: MSEs of the estimated values using our proposed methods with other competing methods. The best method with the smallest MSE in each column is highlighted in blue.

	Scenario 1				Scenario 2				Scenario 3				Scenario 4			
	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$	η^π	η_i^π	η_t^π	$\eta_{i,t}^\pi$
2OPE	0.01	0.48	0.17	3.54	0.66	0.73	0.10	1.78	0.41	0.51	0.07	4.65	0.03	0.17	0.05	9.00
DM1	0.01	1.40	0.85	4.02	0.01	1.26	1.26	3.06	0.04	0.51	0.37	4.36	0.02	0.02	0.08	8.37
DM2	0.37	1.77	0.98	4.16	0.39	1.64	0.85	2.31	0.81	1.27	0.51	4.08	0.77	0.77	0.73	8.25
IS1	0.20	1.60	0.58	5.25	0.84	2.08	0.55	3.41	0.63	1.09	0.63	4.74	0.30	0.30	0.78	8.83
IS2	0.05	1.45	0.13	4.82	1.26	2.50	0.31	3.43	0.93	1.40	0.33	4.48	0.41	0.41	0.36	8.23
IS3	2.89	4.28	3.14	6.32	7.04	8.29	4.63	6.09	7.47	7.94	5.17	8.74	6.48	6.49	5.72	13.24
DR1	0.16	1.56	0.35	5.14	0.53	1.78	0.29	3.00	0.67	1.14	0.33	4.54	0.24	0.24	0.37	8.36
DR2	0.23	1.63	0.85	6.32	0.58	1.82	0.25	3.67	0.95	1.41	0.60	5.07	0.22	0.22	0.37	8.38
DR3	2.21	3.61	2.54	5.72	7.02	8.26	4.61	6.08	6.84	7.31	4.66	8.24	5.54	5.54	4.86	12.38

Review: Sepsis Data



Sepsis

Longitudinal data of sepsis patients,

$$N = 500,\ T = 50.$$

- **Treatment**: in  fl ids vs. 

s.

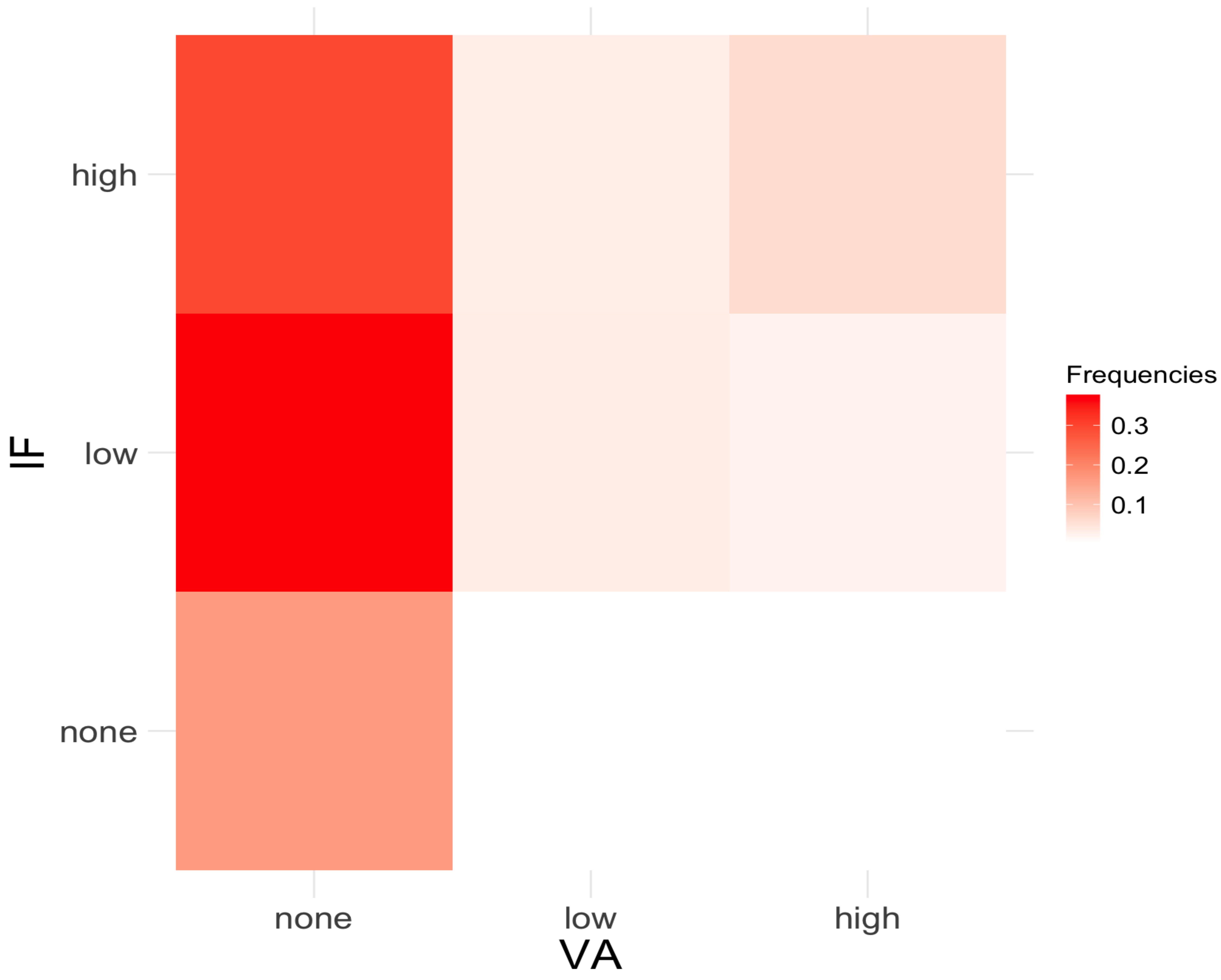
- **Reward**: SOFA score: measures organ failure.
- **Observations**: gender, age, weight, etc.

Target Policies

Compare two policies:

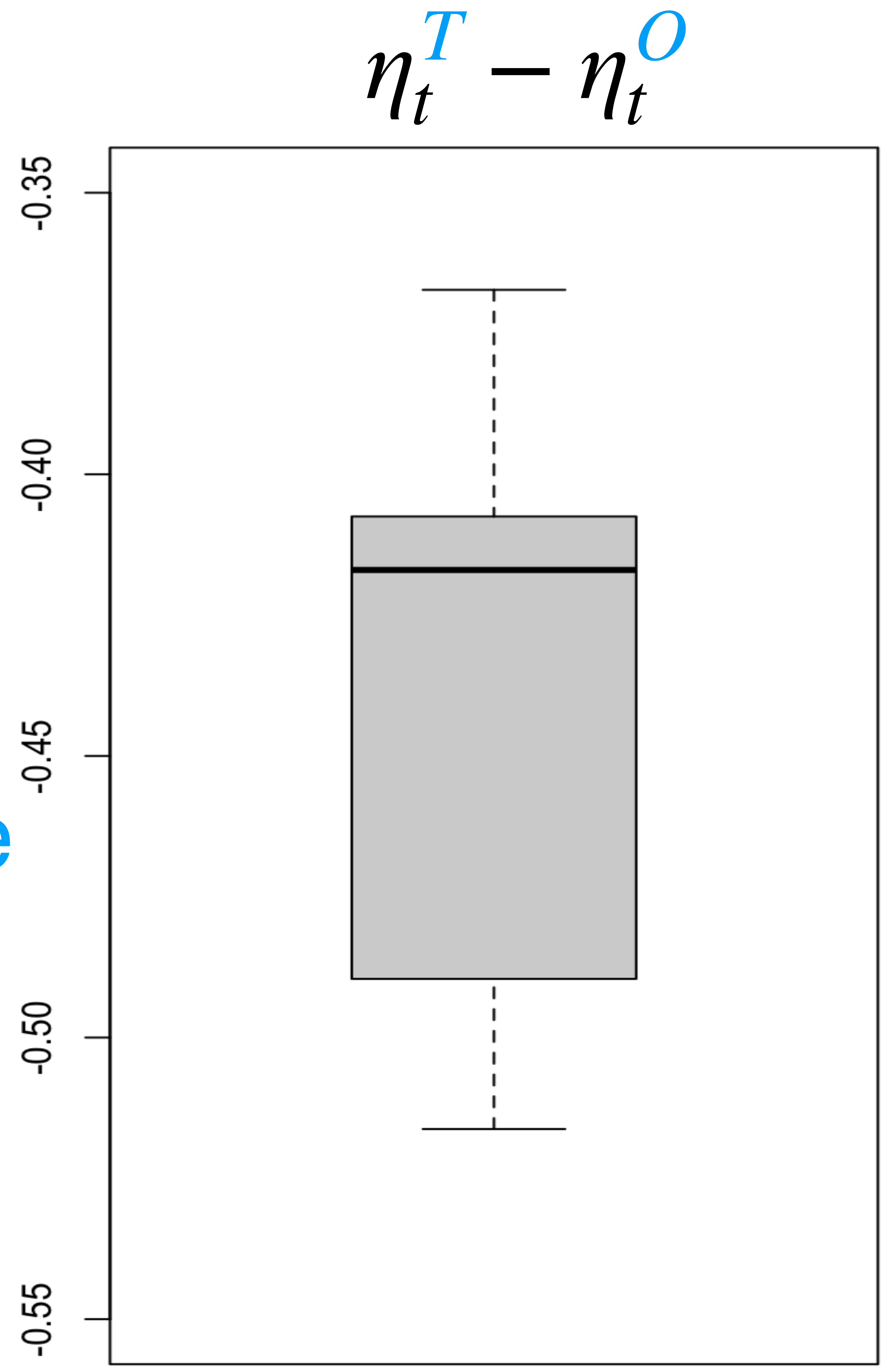
- One size fits all policy π^{O} : always low IF.
- Tailored policy π^{T}: a low IF if SOFA < 11; a high IF dose otherwise.

Frequency of Three Dose Levels in Physician Strategies

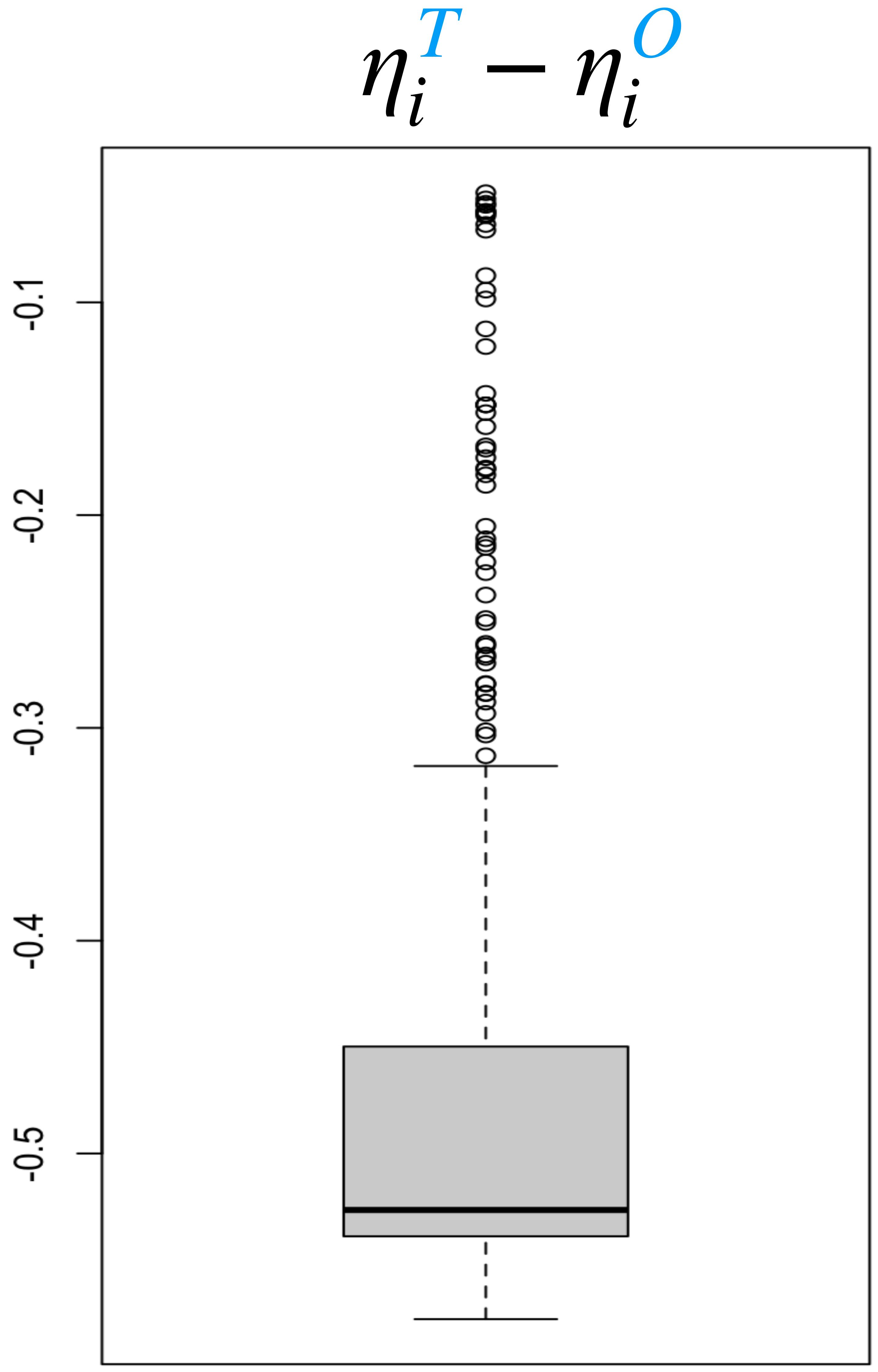


Results

Value difference



Value difference over t



Value difference over i

Summary

This work addresses violations of the Markov, stationary, and homogeneity assumptions.

Future Works

- RL with interactive effects.
- RL under a confounded environment.

Research Overview

My research focuses on addressing challenges arising from real-world applications.

- PhD Research: decision making with high-dimensional data.
- Current Research:
  - RL under partial identification.
  - Decision making with fairness constraint.

Thank you!!

Bian, Z., Shi, C., Qi, Z., and Wang, L. (2024). “Off-policy evaluation in doubly inhomogeneous environments”. Journal of the American Statistical Association, in press.

Policy Evaluation VS. Policy Learning

- Two tasks, each with its own distinct importance.
  - Policy learning: obtain the optimal policy.
  - Policy evaluation is fundamental to RL:
    i. Policy learning usually involves OPE;
    ii. Policy/algorithm comparison: statistical inference.

Error Propagation

$$\min_{\theta_{i,k},\lambda_{t,k},r_k} \sum_{i,j} \left[\widehat{\mu}^{\pi}_{i,k+1} - \theta_{i,k} - \lambda_{t,k} - r_k(O_{i,j}, A_{i,j}) \right]^2$$

Issue: outcome $\widehat{\mu}^{\pi}_{i,k+1}$ is estimated.

As the number of iterations ↑, $\widehat{\mu}^{\pi}_{i,k+1}$ becomes unstable.

Under our setting, the Bellman error decays exponentially, preventing error accumulation.

$\Longrightarrow$ We can learn when $t \to \infty$.

Early Stopping

The Bellman error decays exponentially.

- Early stopping can be applied.
  - No need to run t iterations when t is large.
  - Theoretically, $\log(Nt)$ iterations is sufficient.

Assumptions about the Behavior Policy

Depends on the algorithms:

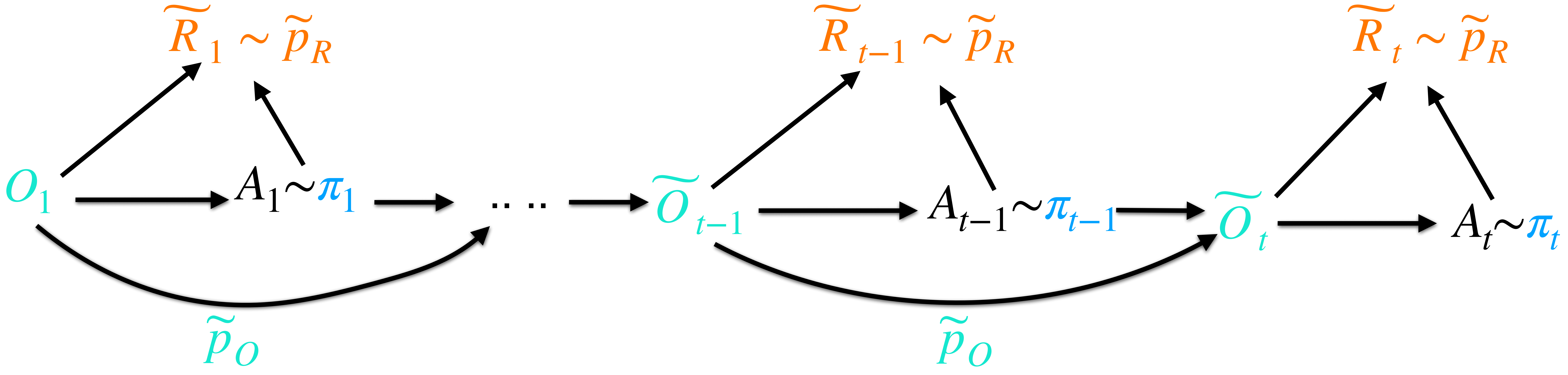
- Importance sampling: $\frac{\pi}{\pi^b}$ is bounded.
- Q-learning: depends on the parametrization.
    - Linear approximation: invertible matrix.

π and π^b cannot "differ" significantly.

Standard Causal Assumptions

- No unmeasured confounders.
- Positivity.
- No interference.

Extension: Model-based Approach



1. Using working model $\widetilde{p}_O(O_{i,t+1} \mid O_{i,t}, U_i, V_t)$ and $\widetilde{p}_R$ $(R_{i,t} \mid O_{i,t}, U_i, V_t)$, e.g., VAE, EM, etc, to generate $(\widetilde{O}_{i,t}, \widetilde{A}_{i,t}, \widetilde{R}_{i,t})$ using $\widetilde{p}_O$, π, and $\widetilde{p}_R$.
2. Evaluate value using Monte Carlo.

Uniform Convergence Rate

- Assume Q-function is Hölder smooth.
- Use linear sieve (e.g., B-splines, wavelet) to approximate the Q-function.

Theorem 2 Under some regularity conditions,

$$\max_{i,t} \left| \widehat{\eta}^{\pi}_{i,t} - \eta^{\pi}_{i,t} \right| = O(L^{-s/d}) + O_p\left(\sqrt{\log(NT)/\min(N,T)}\right).$$

- L: number of basis functions.
- s: smoothness parameter.
- d: dimension.

Relaxing the Additive Assumption

The transition and the reward is additive in u_i, v_t and (o, a):

- $$p(O_{i,t+1} \,|\, u_i, v_t, o, a)$$
  $$= \omega_u p_{u_i}(O_{i,t+1} \,|\, u_i(a)) + \omega_v p_{v_t}(O_{i,t+1} \,|\, v_t(a)) + \omega_0 p_0(O_{i,t+1} \,|\, o, a),$$
  with $\omega_u + \omega_v + \omega_0 = 1$.
- $R_{i,t} = \theta_i(A_{i,t}) + \lambda_t(A_{i,t}) + r(A_{i,t}, O_{i,t}) + \varepsilon_{i,t}$.