IMPERIAL

Mind Your Evaluation: On the Challenges with Off-Policy Evaluation for Real World Decision-Making

Sonali Parbhoo

s.parbhoo@imperial.ac.uk

Imperial College London

RLC I Cant Believe It's Not Better Workshop

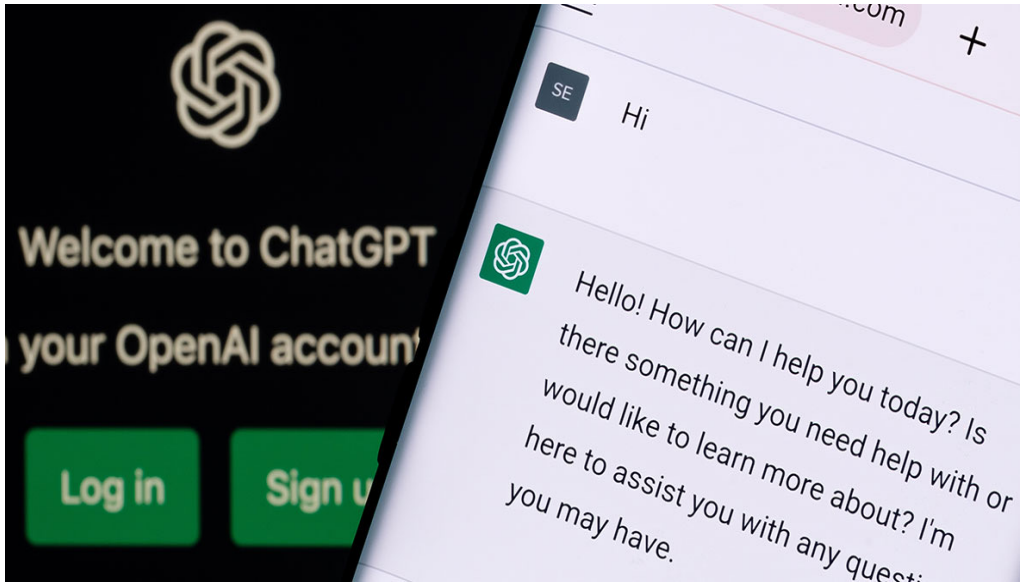
Aug 9, 2024

Collaborators: Aaman Peter Rebello, Ritam Majumdar, Shengpu Tang, Jack Teversham, Omer Gottesman, Taylor Killian, Shalmali Joshi, Abhishek Sharma, Leo Benac, Mario Wieser, Jasmina Bogojeska, Maurizio Zazzi, Huldrych Günthard, Matthieu Komorowski, Jenna Wiens, Marzyeh Ghassemi, Leo Celi, Volker Roth, Finale Doshi-Velez.

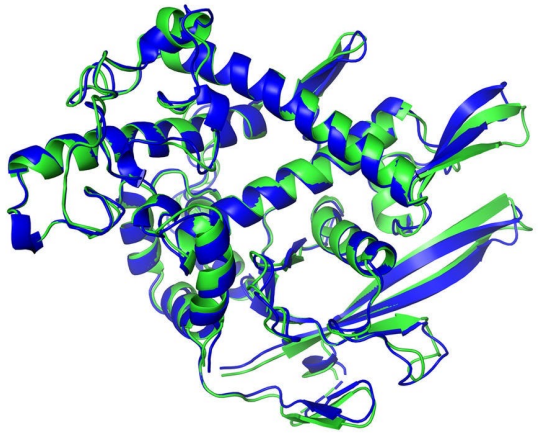
Success in Reinforcement Learning

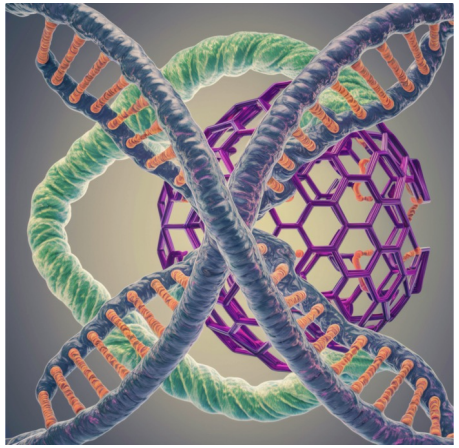




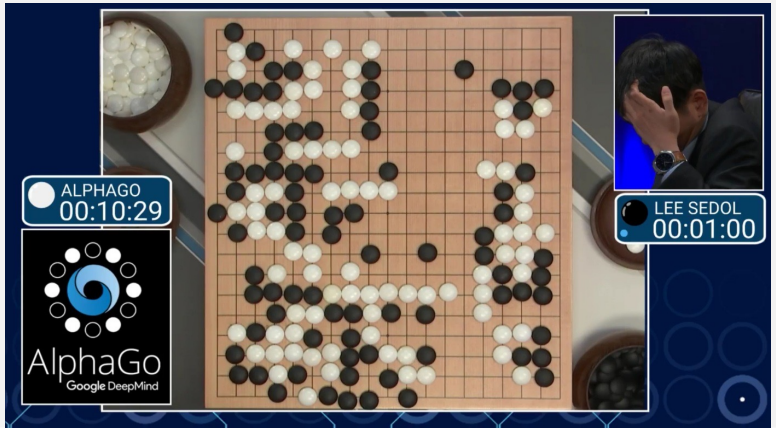
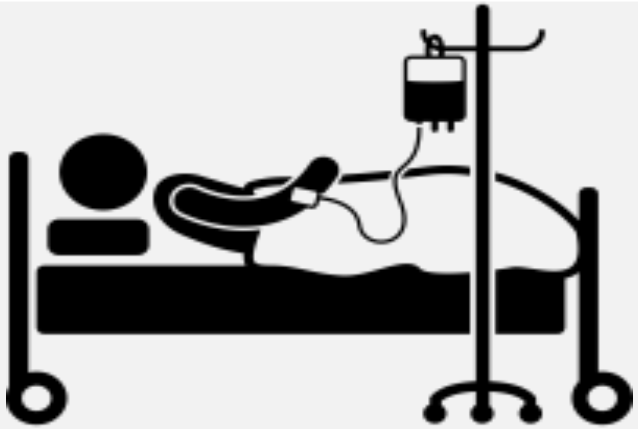








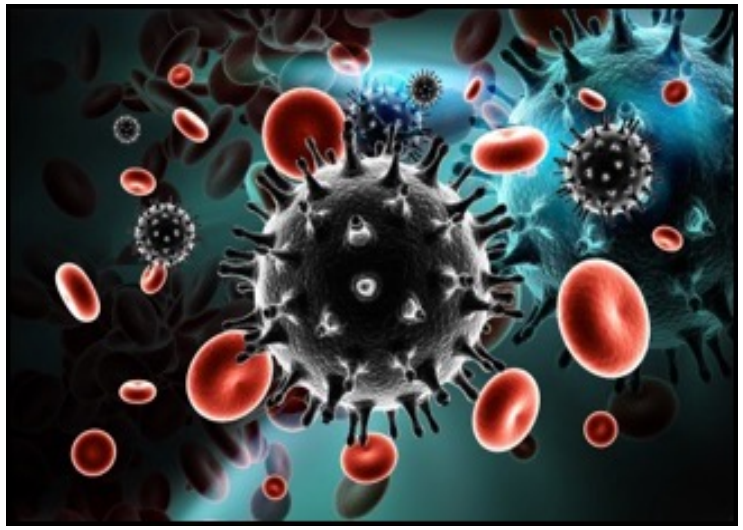
Making decisions in real life settings like healthcare is not so easy

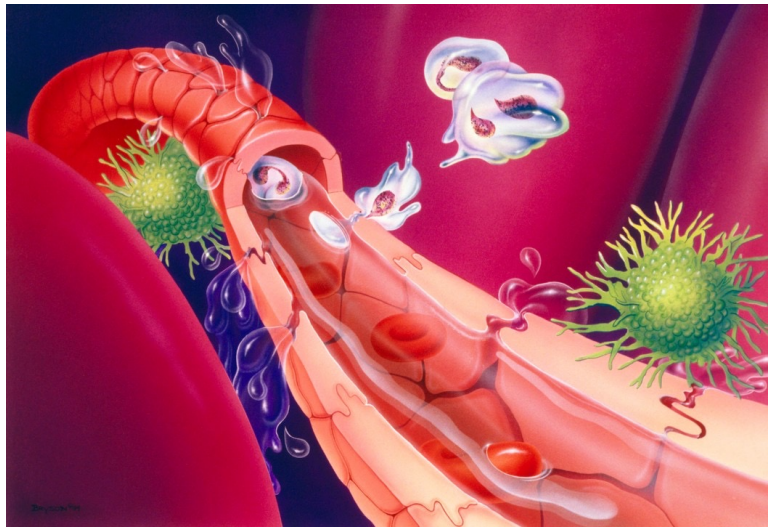
	AlphaGo	Healthcare
Problem Setting		
Interaction with Environment	Yes	No
Large, Relevant Training Samples	Yes	No
Simple objectives	Yes – Win/Lose	No – Healthy? Safe? Cost Effective?
Nonstationary	No	Yes

Why RL for Real Life?

Real-life decisions have **long-term consequences** that manifest downstream. RL is a natural framework that, unlike other forms of machine learning, makes this explicit.

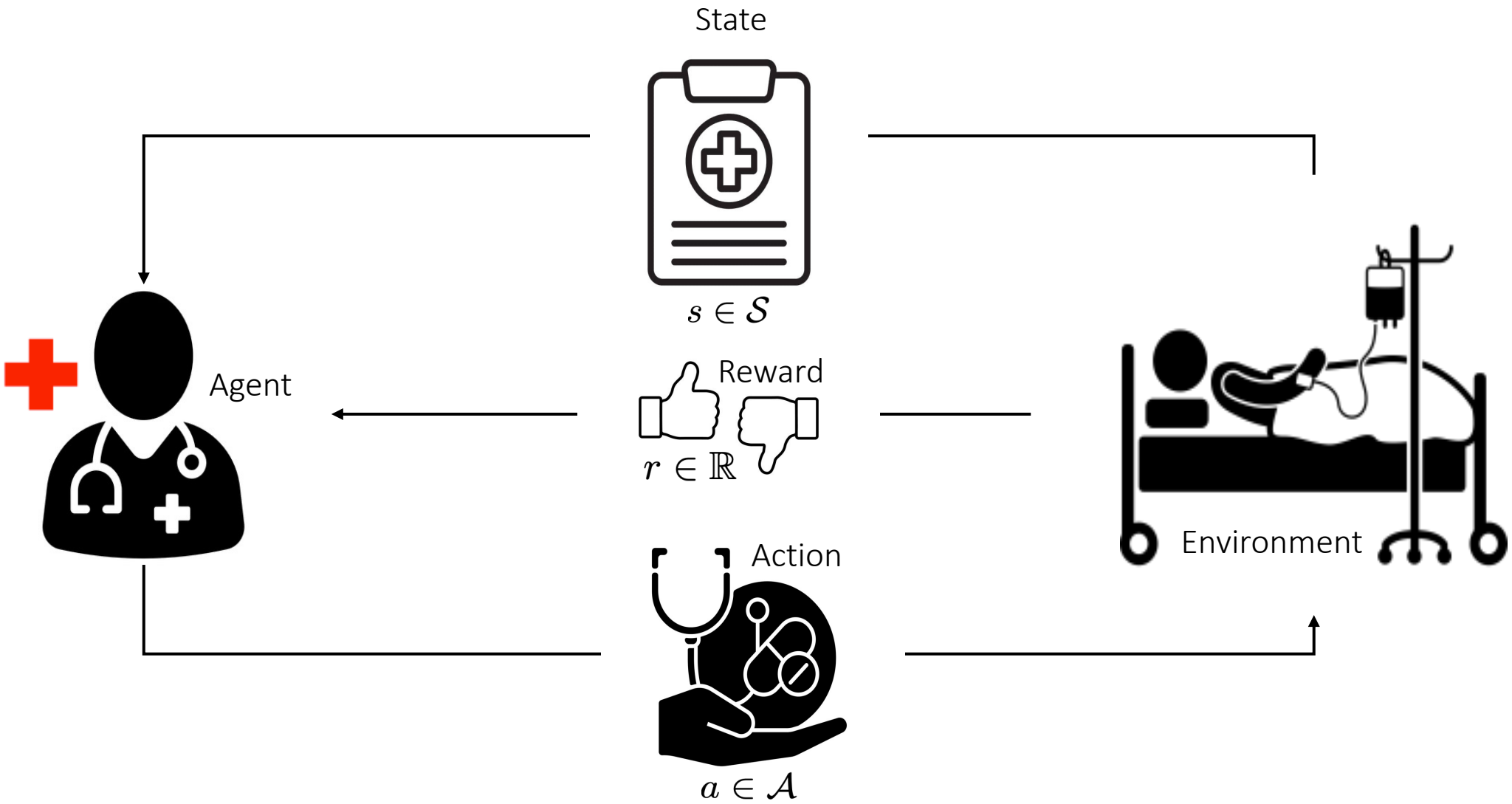
- RL enables directly **learning which actions to take** to improve outcomes.



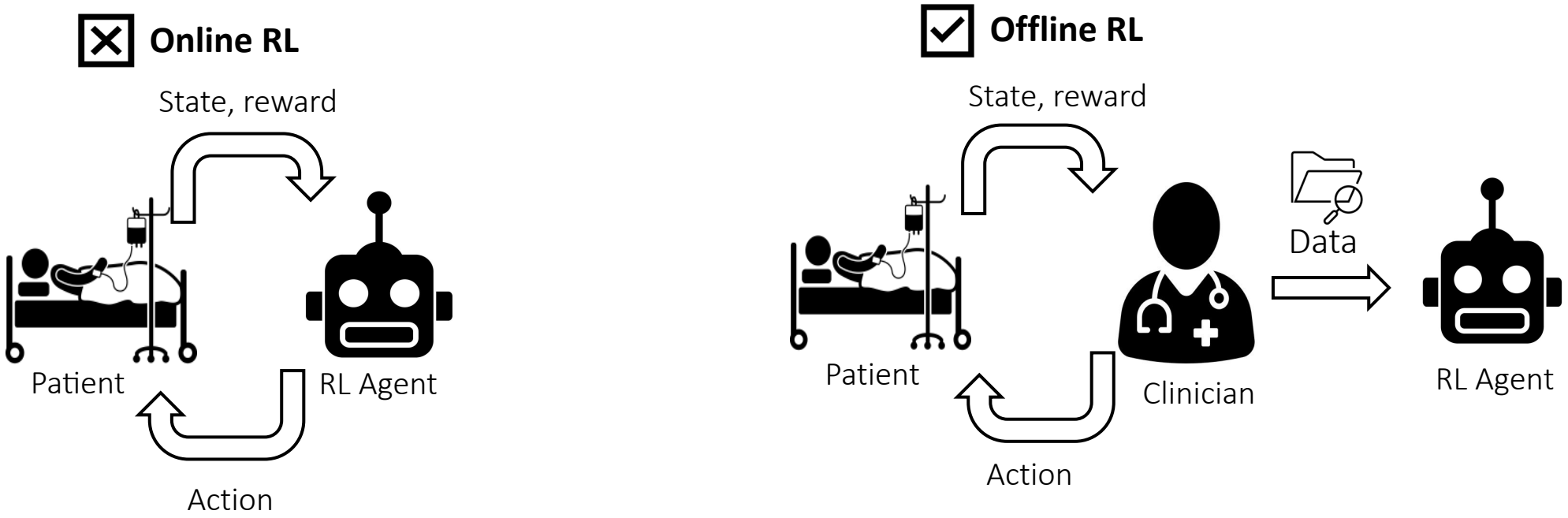


E.g. What antiretrovirals should we prescribe a patient?
How can we improve a septic patient's chance of organ dysfunction?

Example: Treatment Selection as an RL problem



How can we evaluate real life decisions?



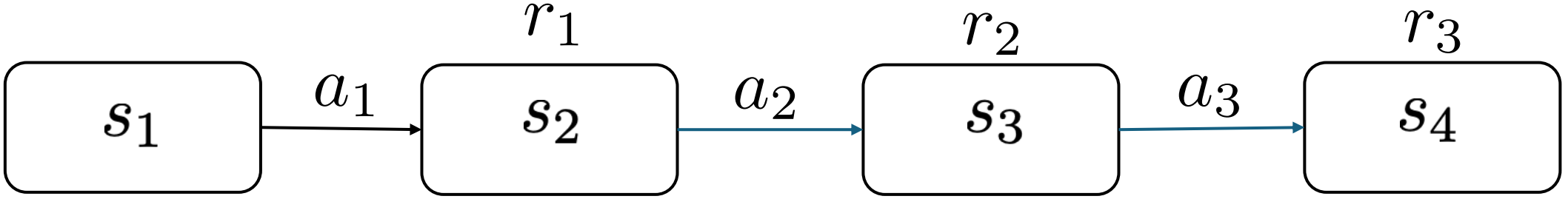
1.Inability to explore

2.Small data (relatively) → shrinking data support as best strategies are discovered

Off-Policy Evaluation

Given data collected by some **behaviour policy** π_b , can we estimate the value of **another evaluation policy** π_e ?

- Markov Decision Process (MDP): $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \mu)$
- Data Trajectories:



- $s_1 \sim \mu$
- $s_{t+1} \sim P(\cdot|s_t, a_t)$
- $r_t = R(s_t, a_t) + \text{noise}$
- $a_t \sim \pi_b(\cdot|s_t)$

(Unknown) parameters of the environment

Goal: Estimate $V^{\pi_e} = \mathbb{E}[\sum_{t=1}^{T} r_t | a_t \sim \pi_e(\cdot|s_t)]$

Solutions in the Sequential Setting

- Recall data trajectory looks like $(s_1, a_1, r_1, s_2, \ldots, s_T, a_T, r_T)$

1. Turn into a bag of (s, a, r, s') tuples
   Regress $(s, a) \rightarrow r$ for $\hat{R}$ and $(s, a) \rightarrow \mathbf{1}_{s'}$ for $\hat{P}$ and use the estimated model parameters to compute $\hat{V}^{\pi_e}$

2. Perform (Per-Decision) Importance Sampling: Reweight trajectories according to

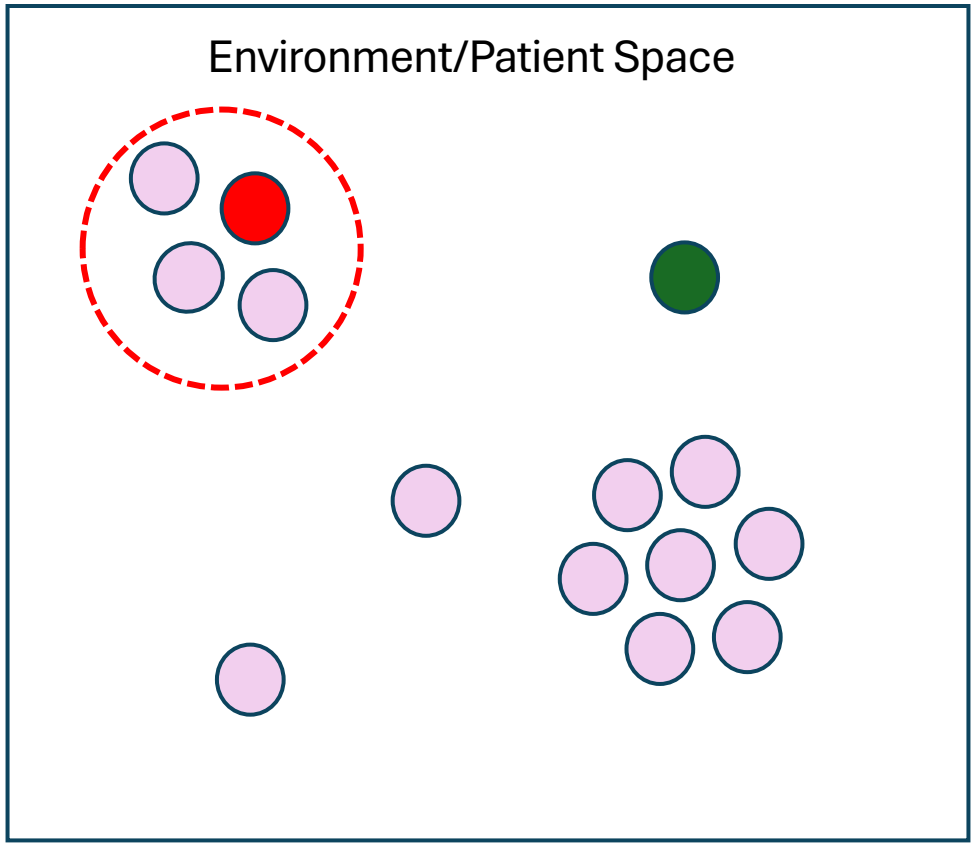
$$\rho_t = \frac{\pi_e(a_t|s_t)}{\pi_b(a_t|s_t)}, \; \rho_{1:t} = \prod_{t'=1}^{t} \rho_{t'}, \quad \hat{V}^{\pi_e} = \sum_{t=1}^{T} \rho_{1:t} r_t$$

Evaluating the performance of a mixture model in RL using Off-Policy Evaluation

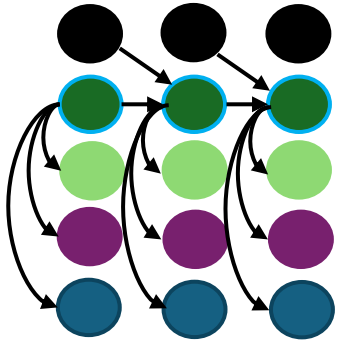
Patients that are similar are best modelled by their neighbours

$$f(history, drugs) = l(VL < 40)$$

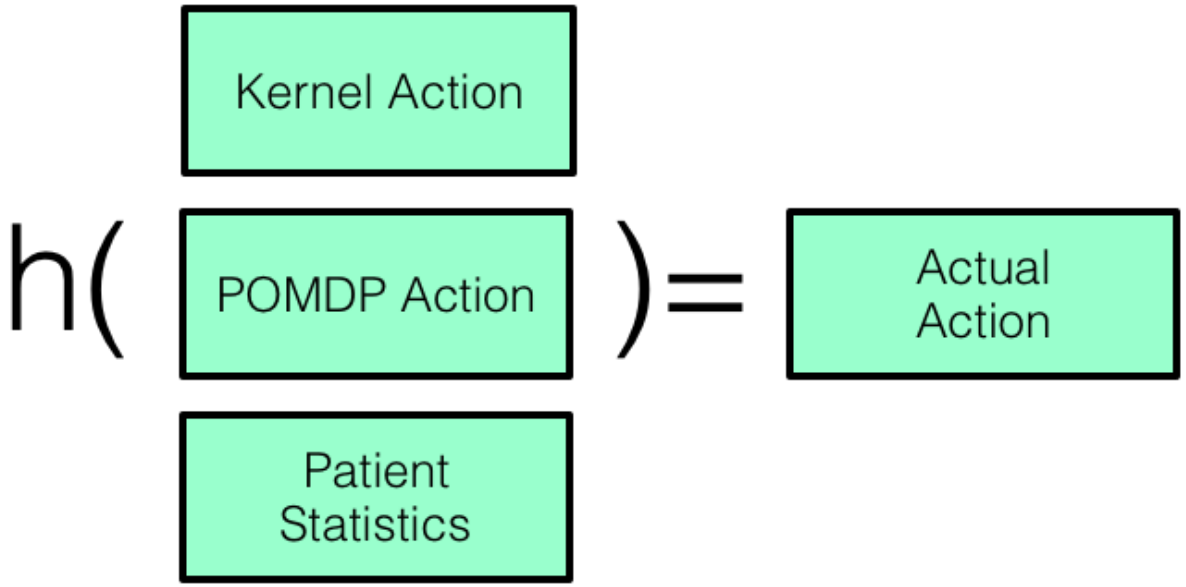
in 21 days



Patients without neighbours are best modelled with a parametric model



Evaluating the performance of a mixture model in RL using Off-Policy Evaluation



We can ensemble the solutions as a Mixture-Of-Experts to learn the appropriate action choice for each patient.

Application to HIV Therapy Management

- Cohort: 32960 patients from EuResist Database
- Observations: CD4s, VLs, mutations, clinical variables
- Actions: 312 most frequently occurring drug combinations (from 25 drugs)

Baseline	Value
Random Policy	-7.31 +/- 3.72
Neighbour Policy	9.35 +/- 2.61
Model Policy	3.37 +/- 2.15
Policy-Mixture Policy	**11.52 +/- 1.31**
Model-Mixture Policy	**12.47+/- 1.38**

Combining Kernel and Model-based RL for HIV Therapy Selection – Parbhoo et al, 2017, AMIA; Kernelised Dynamic Mixing Models – Parbhoo et al, 2018, PLOS One

Similar performance observed across another cohort

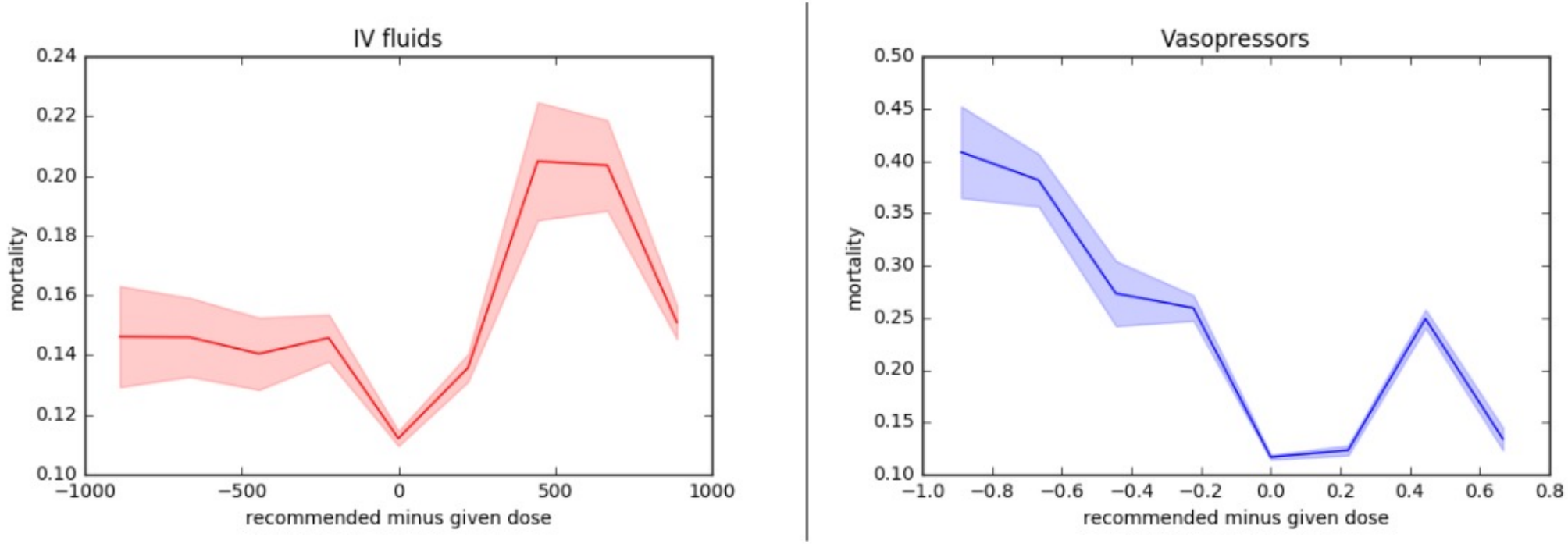
- We also identified when we could transfer policies in well-curated cohorts such as EuResist to South African cohorts that weren't as well curated.

Type	Method	DR	IS	WIS
Behaviour Policy		5.02 ± 1.18		
Local	Kernel	3.56 ± 1.42	1.27 ± 1.14	1.80 ± 1.07
	CEIB	3.29 ± 1.13	3.80 ± 2.41	3.76 ± 2.19
Transfer	Kernel	4.17 ± 1.4	4.18 ± 1.20	4.16 ± 1.71
	CEIB	6.29 ± 0.14	5.17 ± 0.38	5.27 ± 0.29
	Mixture-of-Experts	5.28 ± 0.37	3.42 ± 1.39	4.81 ± 1.25
Local + Transfer	**Ours**	$\mathbf{8.96 \pm 0.39}$	$\mathbf{10.64 \pm 1.2}$	$\mathbf{10.62 \pm 1.67}$

Transfer Learning from Well-Curated to Less Resourced Populations with HIV – Parbhoo et al 2020, MLHC

Similar mixtures have been used for sepsis management

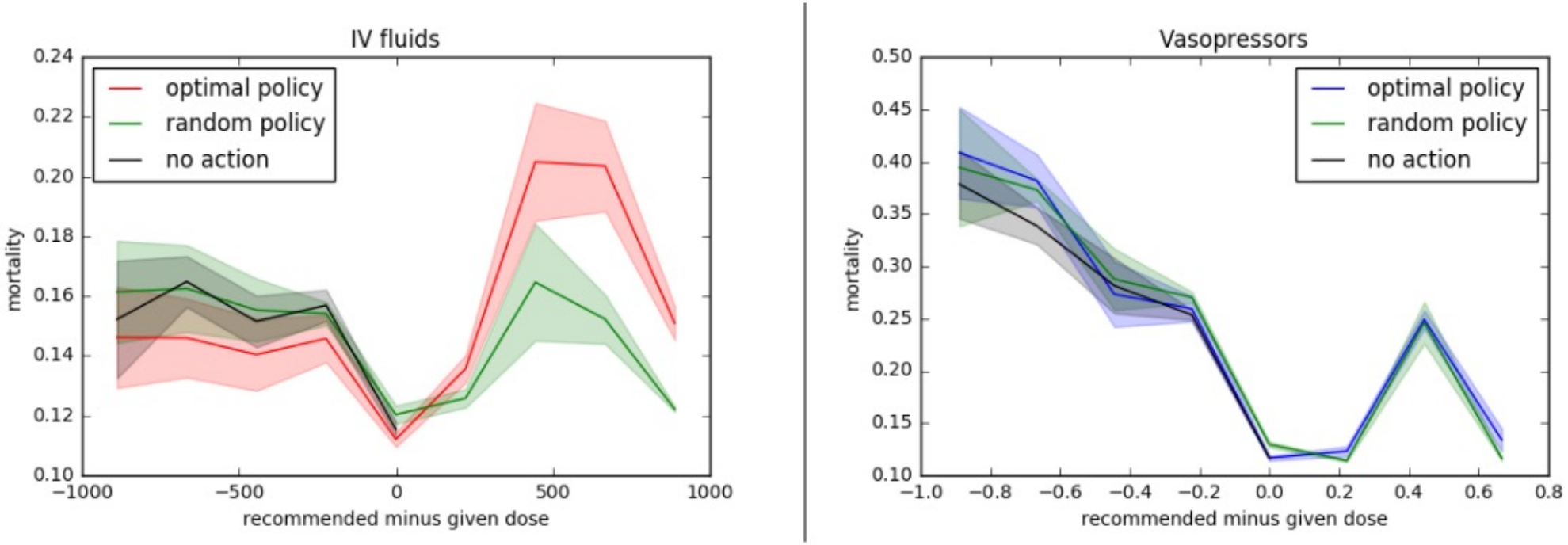
	Physician	Kernel	DQN	MoE_{V_d,Q_d}	MoE_{V_b,Q_b}
non-recurrent encoded	3.76	3.73	4.06	3.93	4.31
recurrent encoded	3.76	4.46	4.23	5.03	5.72



Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

…. but evaluation is unreliable

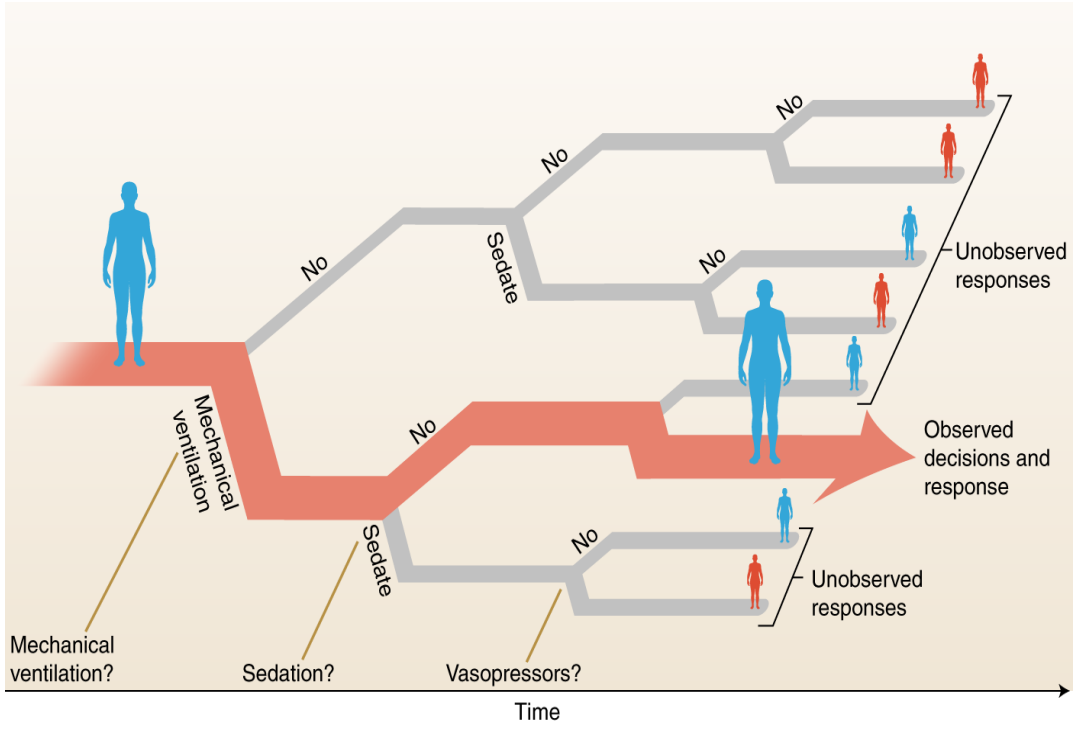
	Physician	Kernel	DQN	MoE_{V_d,Q_d}	MoE_{V_b,Q_b}
non-recurrent encoded	3.76	3.73	4.06	3.93	4.31
recurrent encoded	3.76	4.46	4.23	5.03	5.72



Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

Challenges for OPE in practice

- Inability to explore
- Small data → shrinking data support as best strategies are discovered



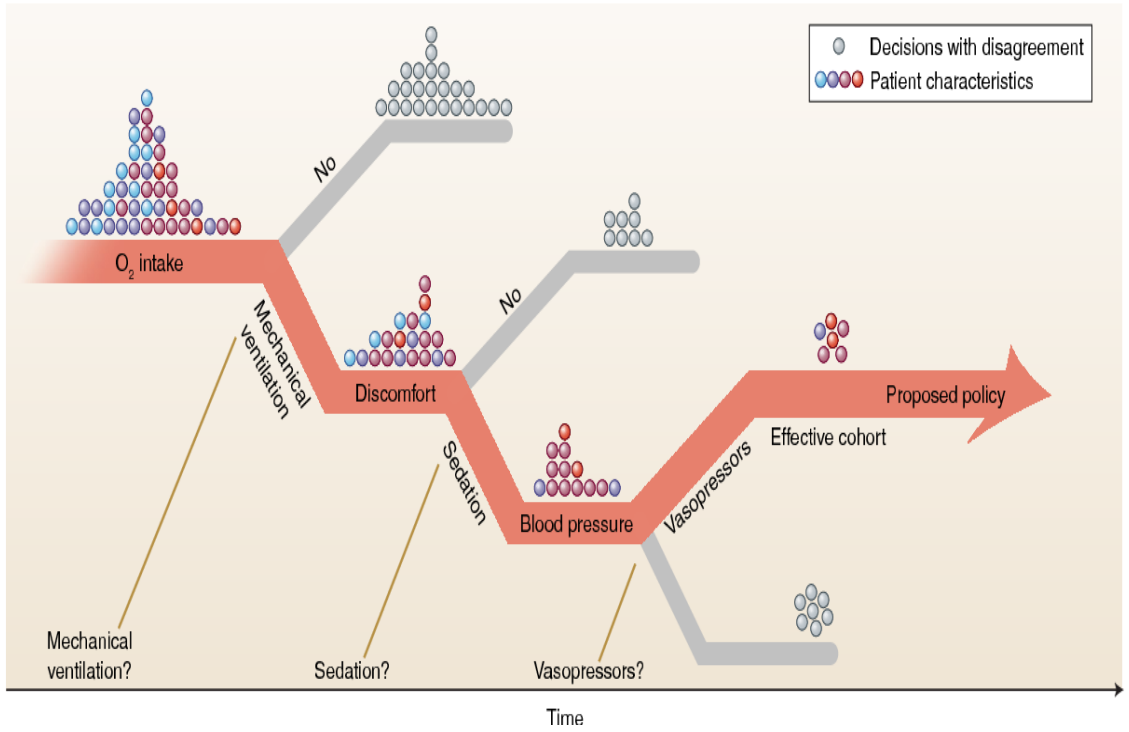
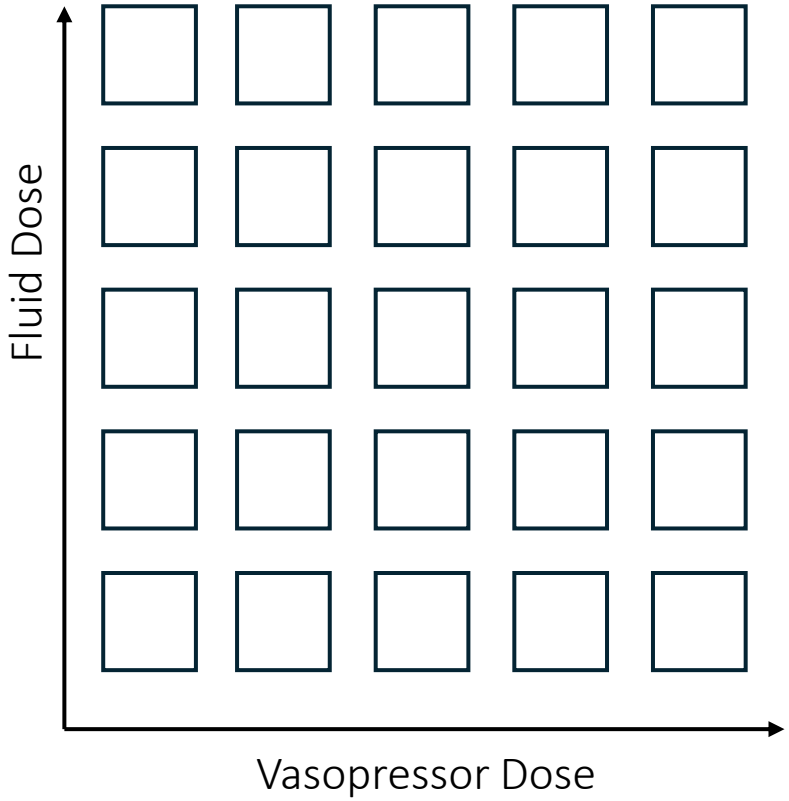


Image Credit: *Gottesman, Omer, et al. "Guidelines for reinforcement learning in healthcare." Nat Med 25.1 (2019): 16-18.*

Factored Action Spaces

Sepsis Management

Komorowski et al 2018, Nature Medicine



Ventilator Weaning

Prasad et al 2017, UAI

$$\mathcal{A} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix} \right\}$$ ventilator on/off, sedation level

Respiratory Distress

Jabbour et al 2023, JAMA

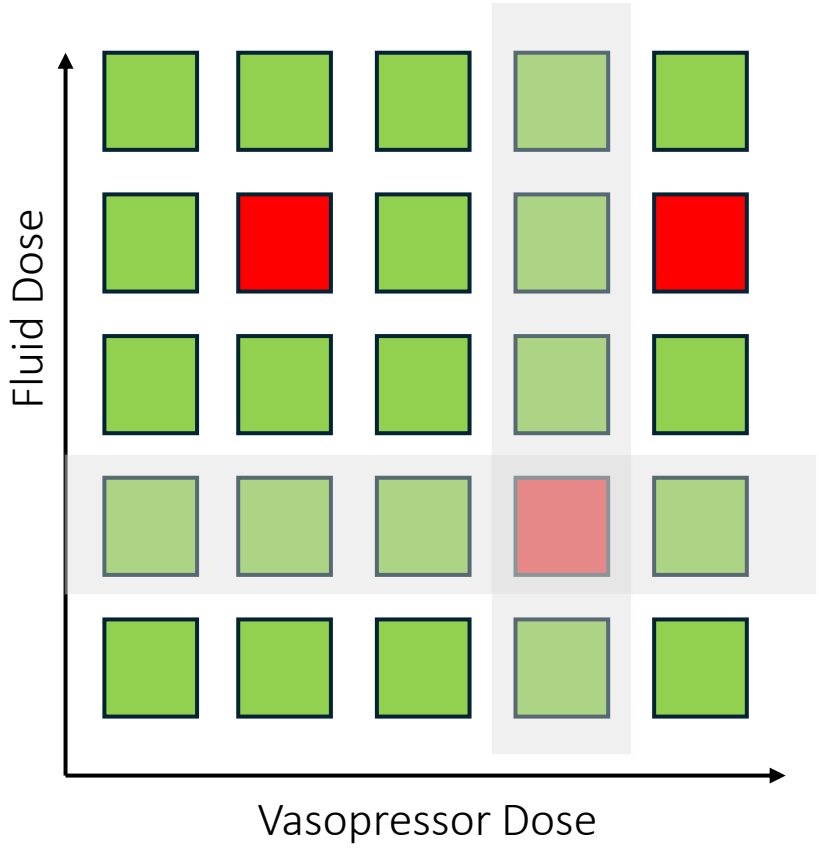
$$\mathcal{A} = \underbrace{\{0,1\}}_{\text{steroids}} \times \underbrace{\{0,1\}}_{\text{fluids}} \times \underbrace{\{0,1\}}_{\text{antibiotics}} \times \underbrace{\{0,1\}}_{\text{diuretics}} \times \underbrace{\{0,1\}}_{\text{anticoagulants}}$$

Komorowski et al, 'The AI Clinician Learns Optimal Treatment Strategies for sepsis in intensive Care", Nature Medicine, 2018

Prasad et al, 'An RL approach to weaning mechanical ventilation in ICUs", UAI, 2017

Komorowski et al, 'Measuring Impact of AI in diagnosis of hospitalized patients", JAMA 2023

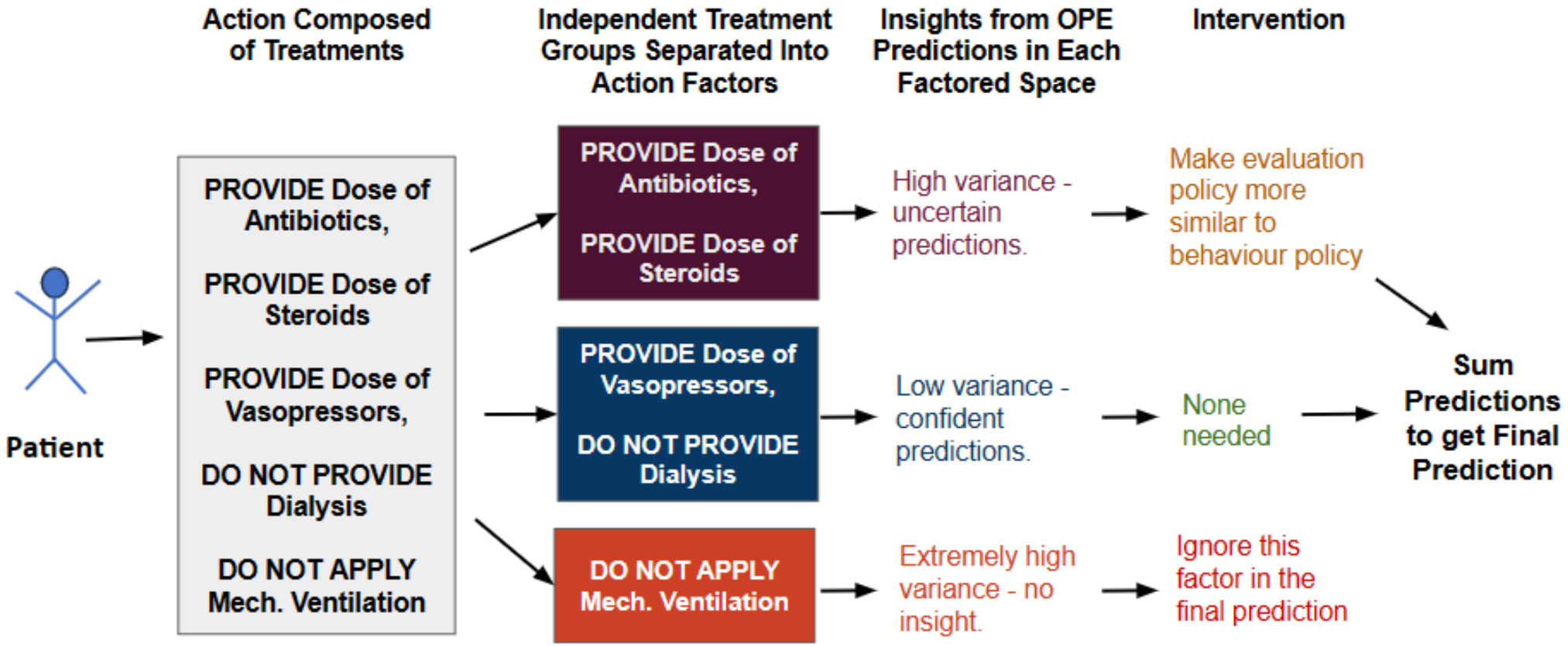
Our Insight: Aggregating Similar Action Factors Can Help OPE



Can we leverage the structure of the action space to improve OPE?

Yes!

Our Solution: Decompose evaluation based on action factors that account for different subspaces



Rebello, A, Tang, S., Wiens, J and Parbhoo, S, "Leveraging Factored Action Spaces for Off-Policy Evaluation", in submission 2024.

Our Solution: Decompose evaluation based on action factors that account for different subspaces

- **Decompose each action** into constituent factors. These may either be determined from using human expertise or learnt.

Intuition: It is easier to induce overlap across some factored dimensions than across an entire trajectory

IS reweighting:

$$\hat{V}^{PDIS}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n)}_{0:t} \cdot r^{(n)}_t$$

$$\rho^{(n)}_{0:t} = \prod_{t'=0}^{t}\frac{\pi_e(a^{(n)}_{t'+1},\ s^{(n)}_{t'})}{\pi_b(a^{(n)}_{t'+1},\ s^{(n)}_{t'})}$$

Decomposed IS reweighting:

$$\hat{Q}^{DecIS}_{\pi_e} = \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:T} \cdot r^{(n),d}_t$$

$$\rho^{(n),d}_{0:T} = \prod_{t=0}^{T}\frac{\pi^d_e(a^{(n),d}_t|z^{(n),d}_t)}{\pi^d_b(a^{(n),d}_t|z^{(n),d}_t)}$$

Rebello, A, Tang, S., Wiens, J and Parbhoo, S, “Leveraging Factored Action Spaces for Off-Policy Evaluation”, in submission 2024.

A Family of Factored Estimators

IS: $$\hat{V}^{IS}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\rho^{(n)}_{0:T}\sum_{t=0}^{T}\gamma^t \cdot r^{(n)}_t$$

DecIS: $$\hat{Q}^{DecIS}_{\pi_e} = \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:T} \cdot r^{(n),d}_t$$

$$\rho^{(n),d}_{0:T} = \prod_{t=0}^{T}\frac{\pi^d_e(a^{(n),d}_t|z^{(n),d}_t)}{\pi^d_b(a^{(n),d}_t|z^{(n),d}_t)}$$

PDIS: $$\hat{V}^{PDIS}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n)}_{0:t} \cdot r^{(n)}_t$$

DecPDIS: $$\hat{Q}^{DecPDIS}_{\pi_e} = \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:t} \cdot r^{(n),d}_t$$

$$\rho^{(n),d}_{0:t} = \prod_{t'=0}^{t}\frac{\pi^d_e(a^{(n),d}_{t'}|z^{(n),d}_{t'})}{\pi^d_b(a^{(n),d}_{t'}|z^{(n),d}_{t'})}$$

PDWIS: $$\hat{V}^{PDWIS}_{\pi_e} = \frac{\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n)}_{0:t} \cdot r^{(n)}_t}{\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n)}_{0:t}}$$

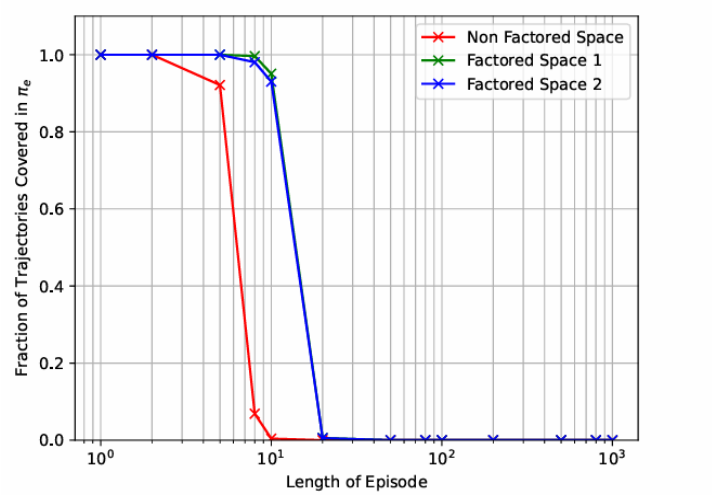
DecPDWIS: $$\hat{Q}^{DecPDWIS}_{\pi_e} = \sum_{d=1}^{D}\frac{\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:t} \cdot r^{(n),d}_t}{\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:t}}$$

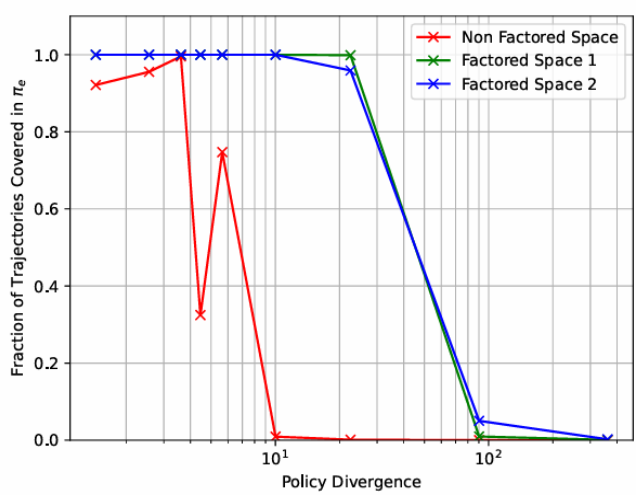
DR: $$\hat{V}^{DR}_{\pi_e} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n)}_{0:t} \cdot r^{(n)}_t - \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \left(\rho^{(n)}_{0:t} \cdot \hat{Q}^{DM}_{\pi_b}(s^{(n)}_t, a^{(n)}_{t+1}) - \rho^{(n)}_{0:t-1} \cdot \hat{V}^{DM}_{\pi_b}(s^{(n)}_t)\right)$$

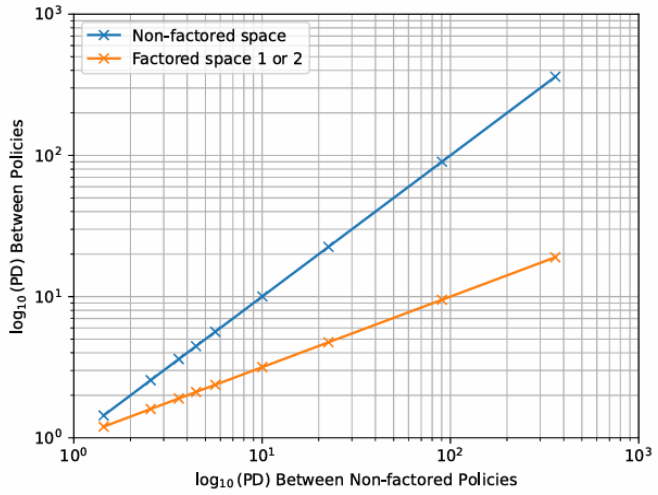
DecDR: $$\hat{V}^{DecDR}_{\pi_b} = \frac{1}{N}\sum_{d=1}^{D}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho^{(n),d}_{0:t} \cdot r^{(n),d}_t - \sum_{d=1}^{D}\frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \left(\rho^{(n),d}_{0:t} \cdot \hat{Q}^{DM,d}_{\pi^d_b}(z^{(n),d}_t, a^{(n),d}_{t+1}) - \rho^{(n)}_{0:t-1} \cdot \hat{V}^{DM,d}_{\pi^d_b}(z^{(n),d}_t)\right)$$

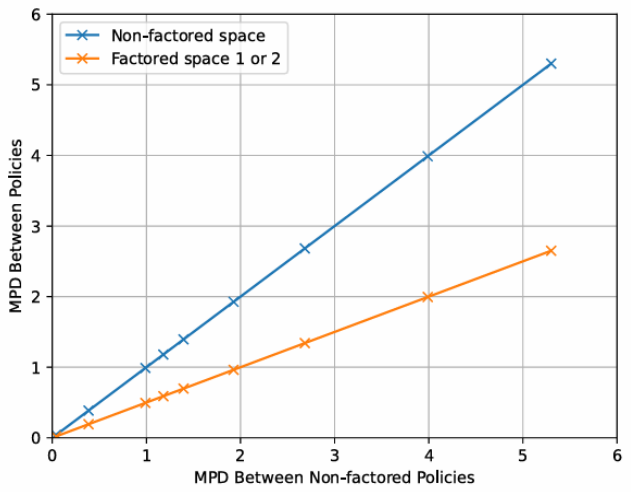
Rebello, A,, Tang, S., Wiens, J and Parbhoo, S, “Leveraging Factored Action Spaces for Off-Policy Evaluation”, arXiv pre-print, 2024.

Factored OPE estimators have better overlap





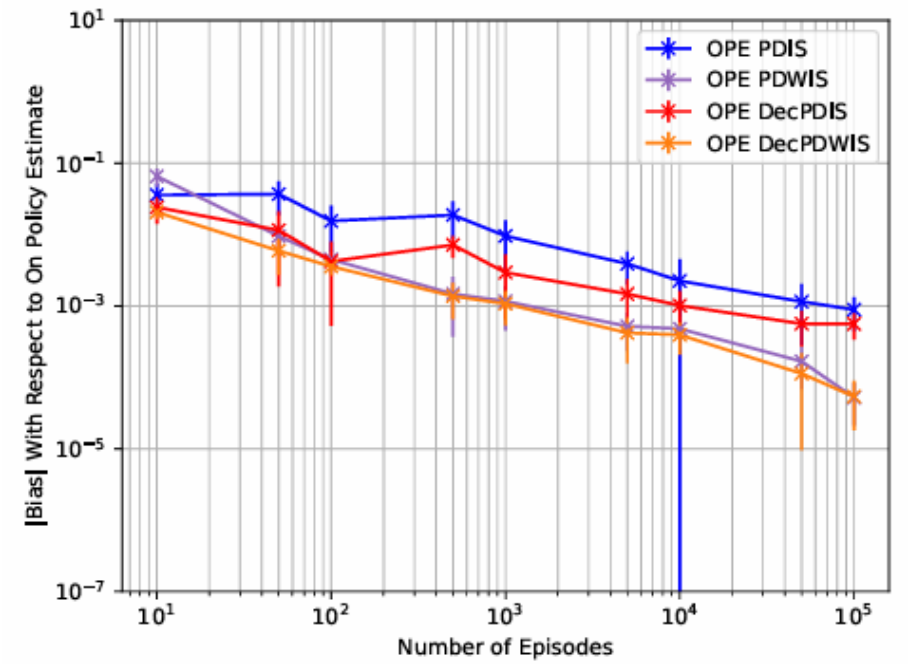


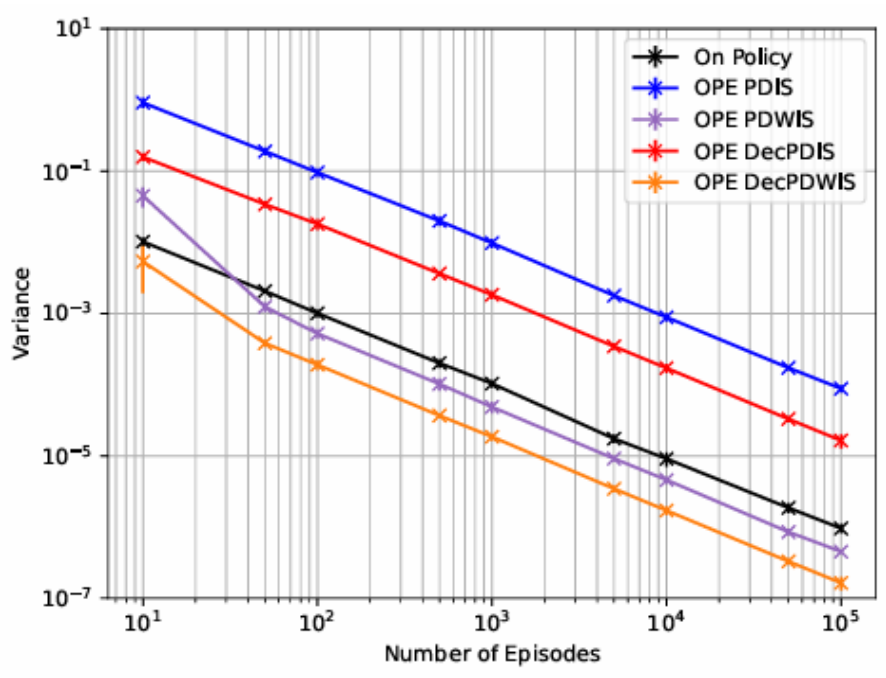


Rebello, A, Tang, S., Wiens, J and Parbhoo, S, “Leveraging Factored Action Spaces for Off-Policy Evaluation”, in submission 2024.

Factored estimators have lower variance and bias

By grouping patients by similar action factors, we obtain better statistical guarantees and performance with less data.

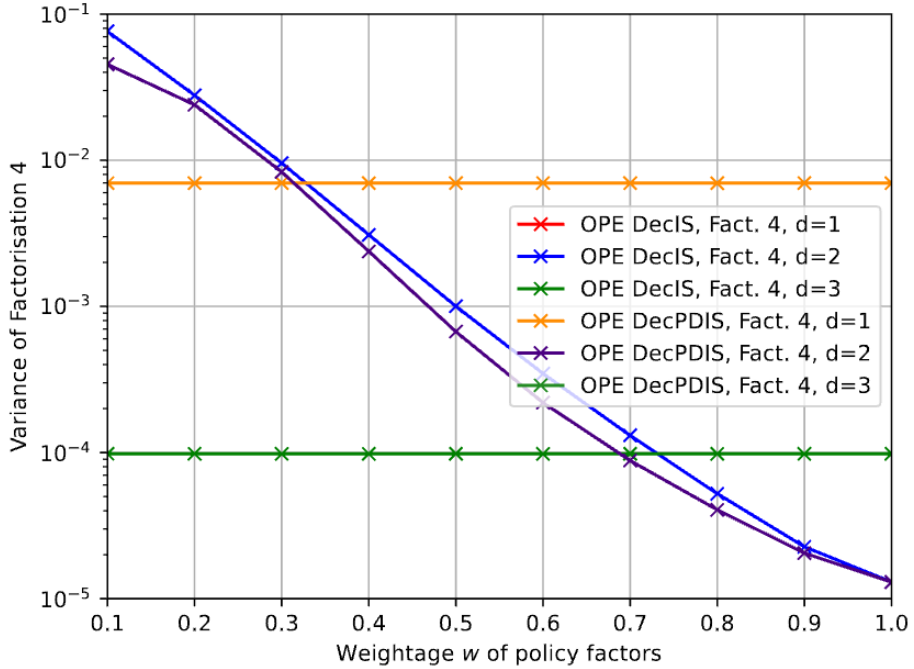


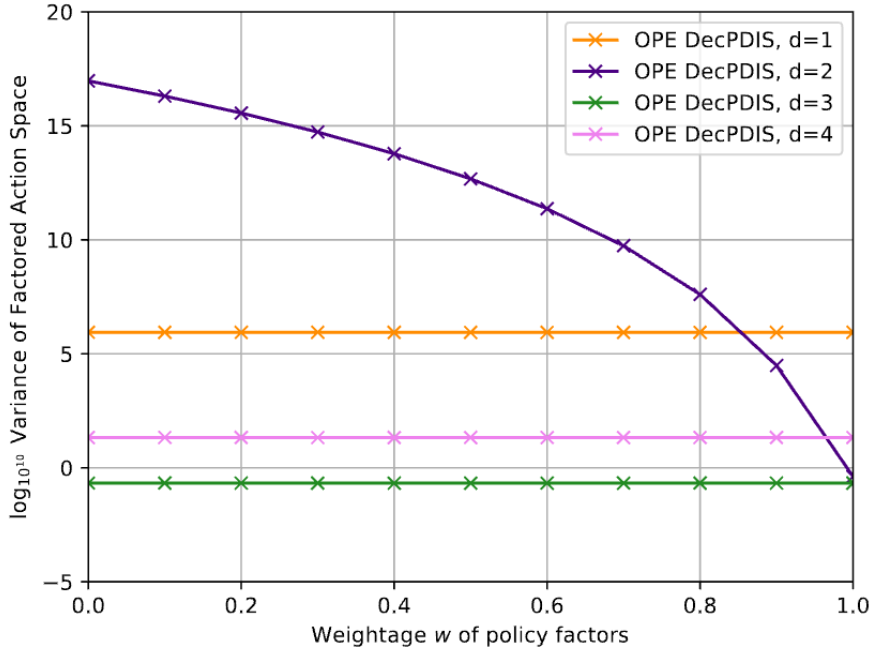


Rebello, A, Tang, S., Wiens, J and Parbhoo, S, “Leveraging Factored Action Spaces for Off-Policy Evaluation”, in submission 2024.

Factors can be intervened on to remedy high variance OPE

We can isolate those subproblems that contribute to the highest variance in OPE and intervene on those factors to remedy the issues. We might choose to ignore those factors or specifically elicit human expertise for those factors accordingly.

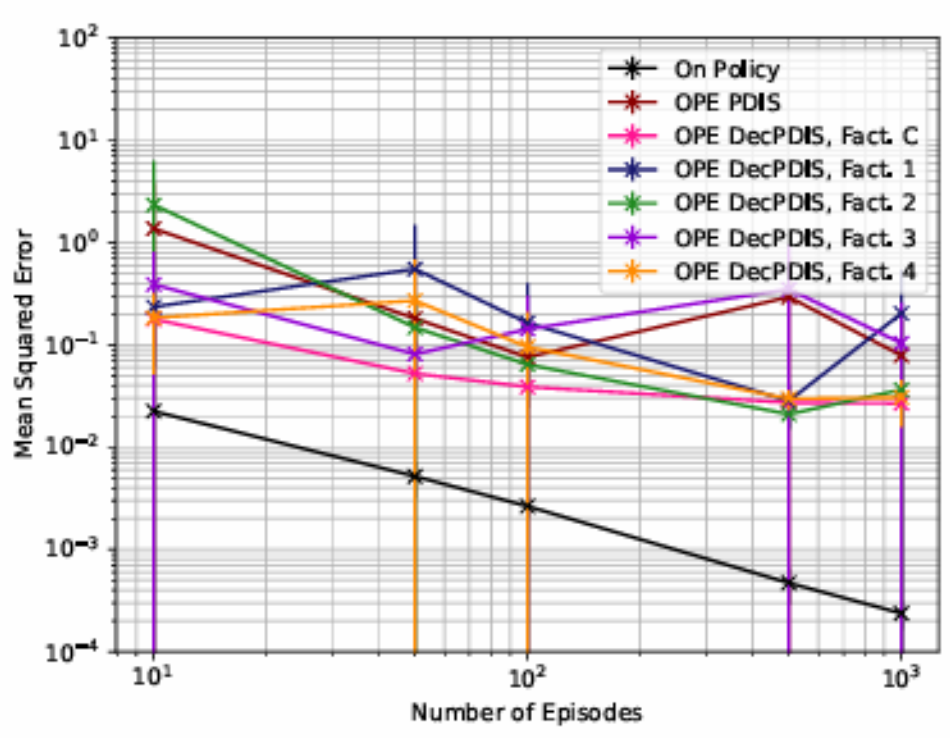


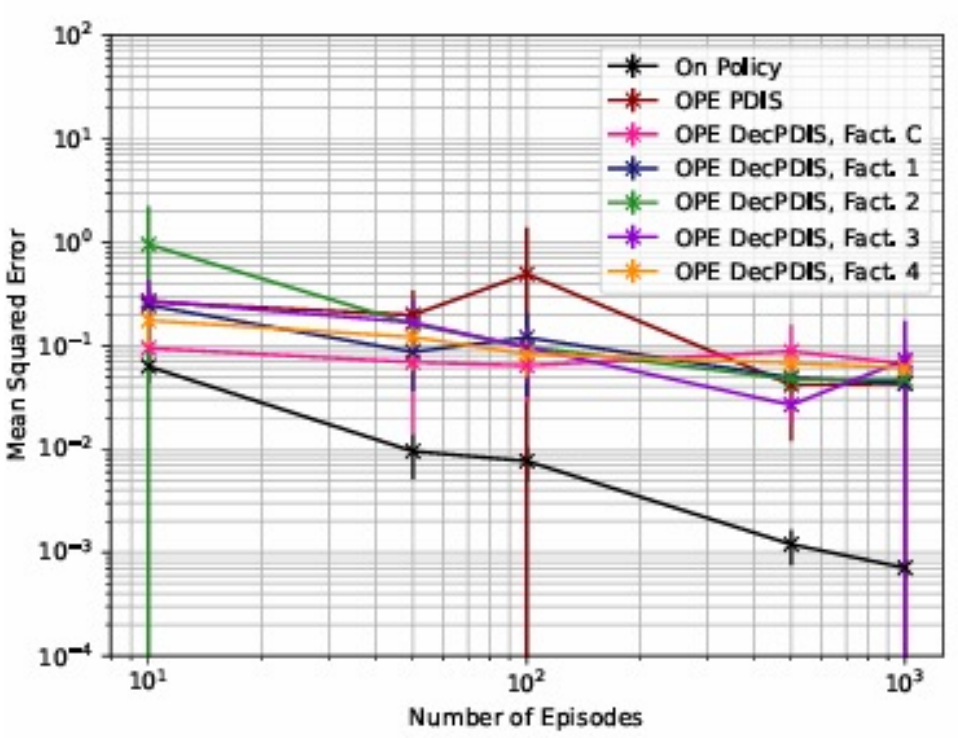


Rebello, A, Tang, S., Wiens, J and Parbhoo, S, "Leveraging Factored Action Spaces for Off-Policy Evaluation", in submission 2024.

Performance of Factored OPE varies with how we learn factors

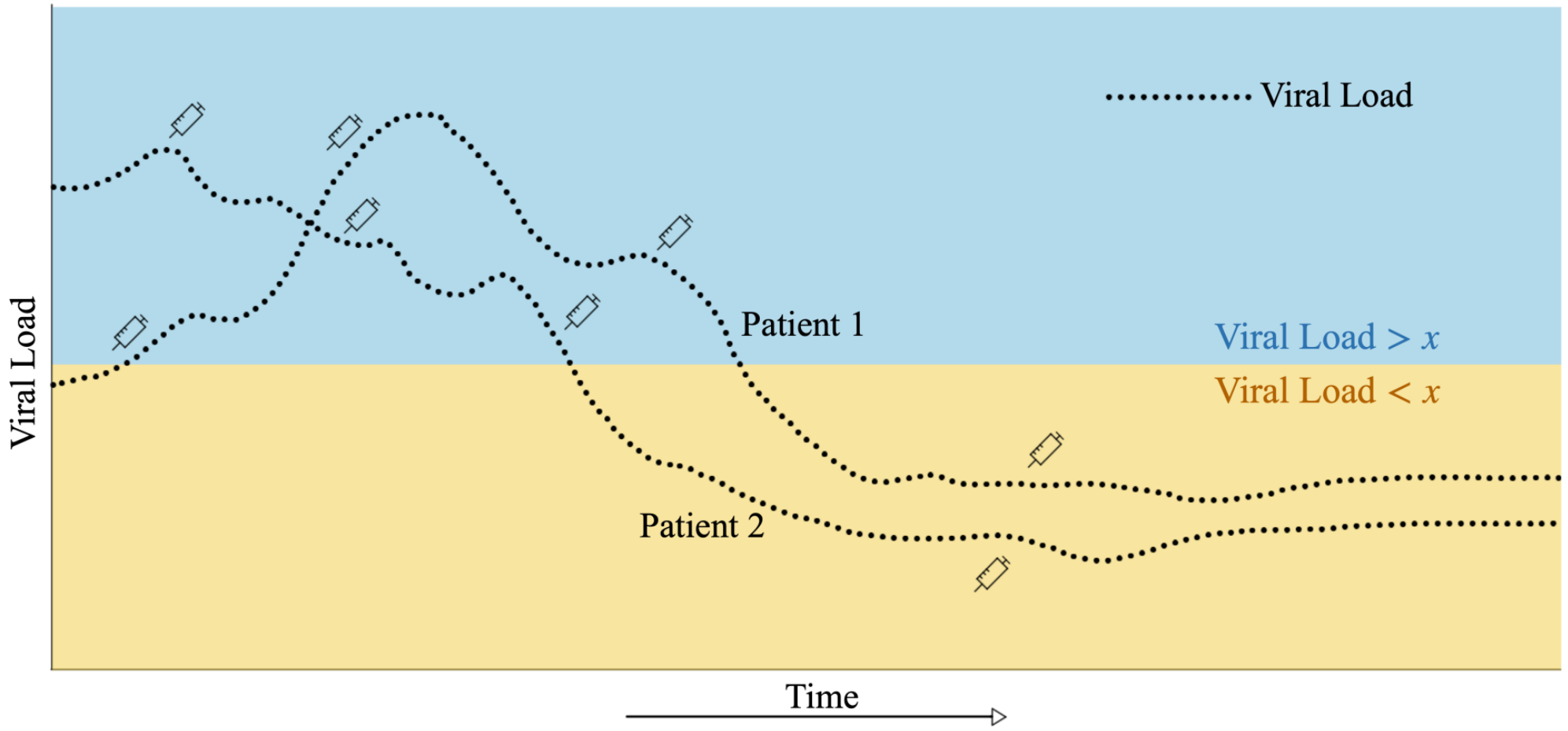
Factorisation 1: (ABX, VA), VE, **Factorisation 2**: (ABX, VE), VA, **Factorisation 3**: (VE, VA), ABX, **Factorisation 4**: ABX, VA, VE. **Factorisation C** learned through clustering





Rebello, A, Tang, S., Wiens, J and Parbhoo, S, “Leveraging Factored Action Spaces for Off-Policy Evaluation”, arXiv pre-print, 2024.

Our insight: Similar ideas can be used to aggregate states into interpretable concepts for evaluation



Our Solution: Decompose evaluation based on interpretable concepts that account for similar states

- Summarise policies in terms of **interpretable concepts**

Intuition: It is easier to induce overlap between policies across concepts rather than states and actions

IS reweighting:

$$\hat{V}_{\pi_e}^{PDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \cdot \rho_{0:t}^{(n)} \cdot r_t^{(n)}$$

$$\rho_{0:t}^{(n)} = \prod_{t'=0}^{t}\frac{\pi_e(a_{t'+1}^{(n)},\ s_{t'}^{(n)})}{\pi_b(a_{t'+1}^{(n)},\ s_{t'}^{(n)})}$$

Concept Based IS reweighting:

$$\hat{V}_{\pi_e}^{CPDIS} = \frac{1}{N}\sum_{n=1}^{N}\sum_{t=0}^{T}\gamma^t \rho_{0:t}^{(n)} r_t^{(n)}$$

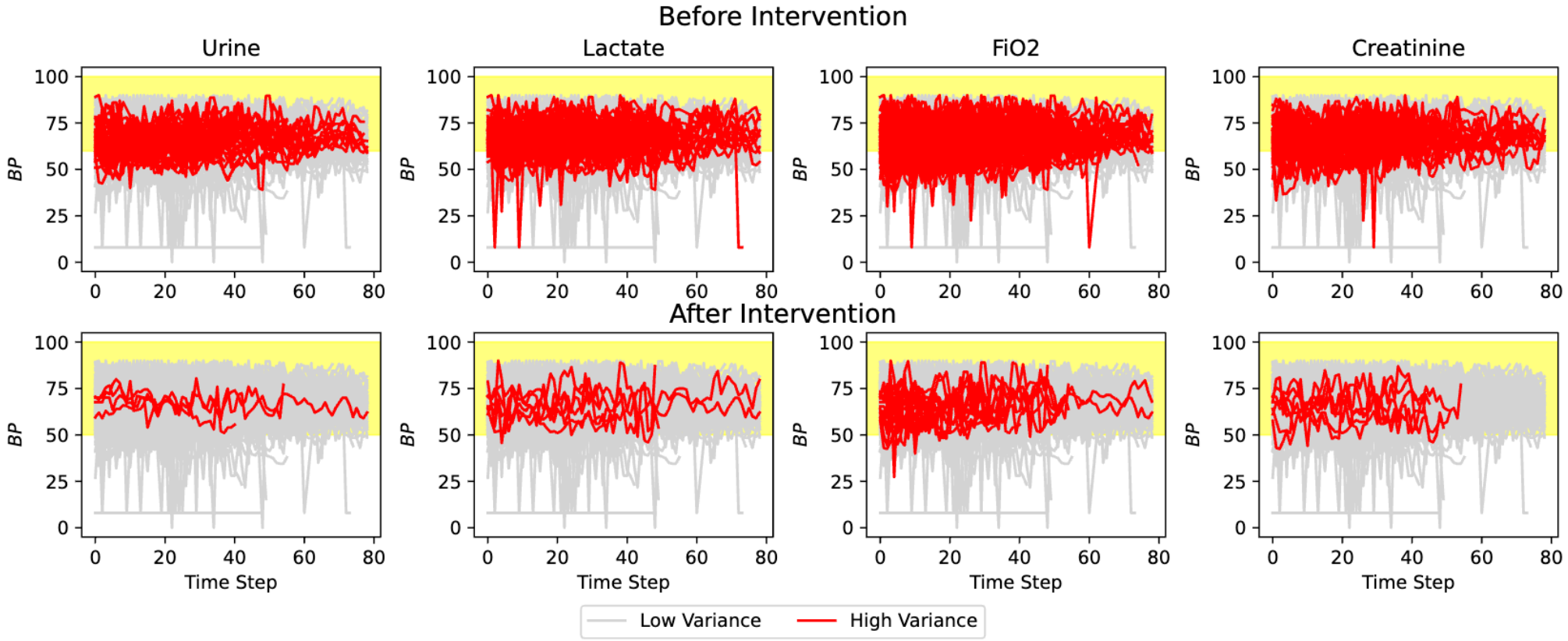
$$\rho_{0:t}^{(n)} = \prod_{t'=0}^{t}\frac{\pi_e(a_{t'}^{(n)}|c_{t'}^{(n)})}{\pi_b(a_{t'}^{(n)}|c_{t'}^{(n)})}$$





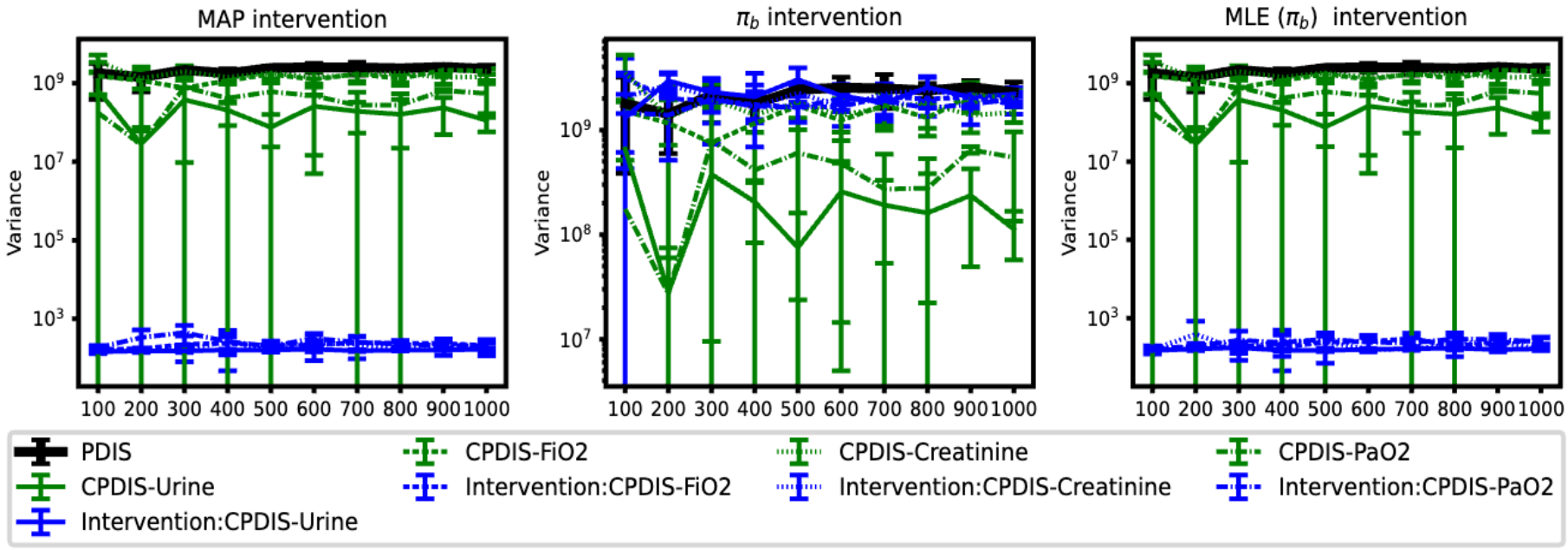
Concept Based Estimators Have Similar Benefits

We can isolate those concepts that contribute to the highest variance in OPE and intervene on those concepts to remedy the issues.



Concept Based Estimators Have Similar Benefits

We can isolate those concepts that contribute to the highest variance in OPE and intervene on those concepts to remedy the issues.



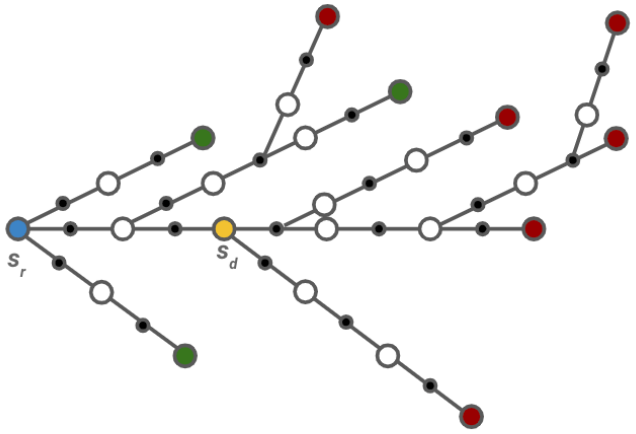
Learn which actions to avoid via risk sensitive RL

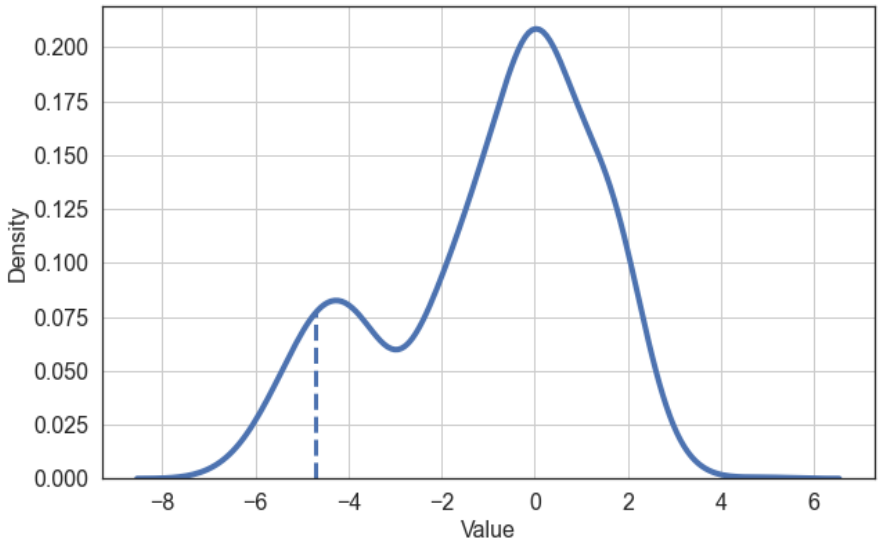
- Rather than the difficult task of learning optimal policies that suggest what to do, we instead learn **what treatments to avoid**

$$Z^\pi = \sum_{t=0}^{\infty} \gamma^t r_t$$

$$\mathcal{T}^\pi Z(s,a) \overset{D}{=} r + \gamma Z(s',a')$$

$$\text{where } r \sim R(\cdot|s,a), s' \sim P(\cdot|s,a), a' \sim \pi(\cdot|s')$$





Conditional value-at-risk

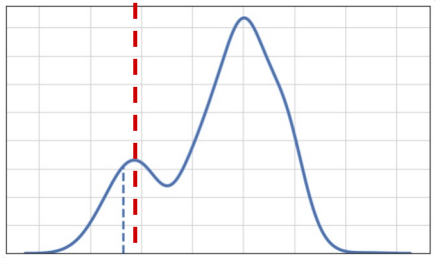
$$\text{CVaR}_\alpha(Z) = \frac{1}{\alpha}\mathbb{E}[(Z - \text{VaR}_\alpha(Z))^-] + \text{VaR}_\alpha(Z)$$

$$\text{CVaR}_\alpha(Z) = \min_{\xi \in \mathcal{U}_{\text{CVaR}}(\alpha,\mathbb{P})} \mathbb{E}_\xi[Z]$$

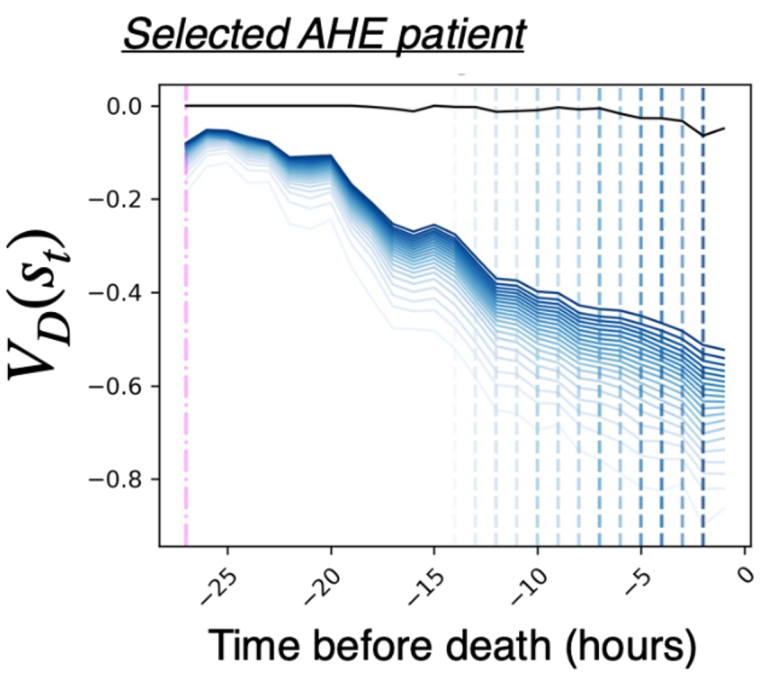
$$-\text{CVaR}_\alpha(Z_D^*(s,a)) \geq -\mathbb{E}[Z^*(s,a)] = -Q_D^*(s,a) = P_D(s,a) + F_D(s,a)$$

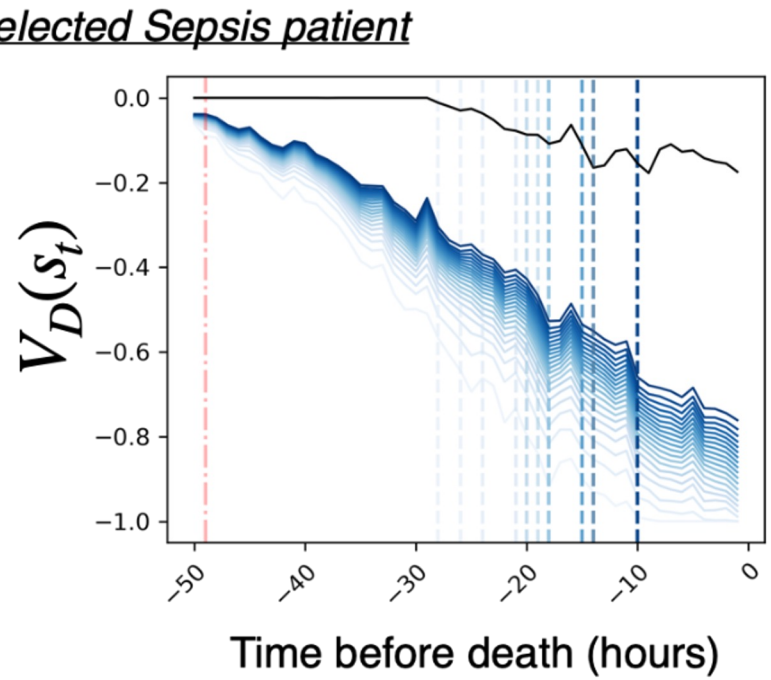
T.W Killian, S Parbhoo* and M. Ghassemi*, **Risk-sensitive and Robust Dead-end Identification in Safety-Critical Offline RL, TMLR 2023.**

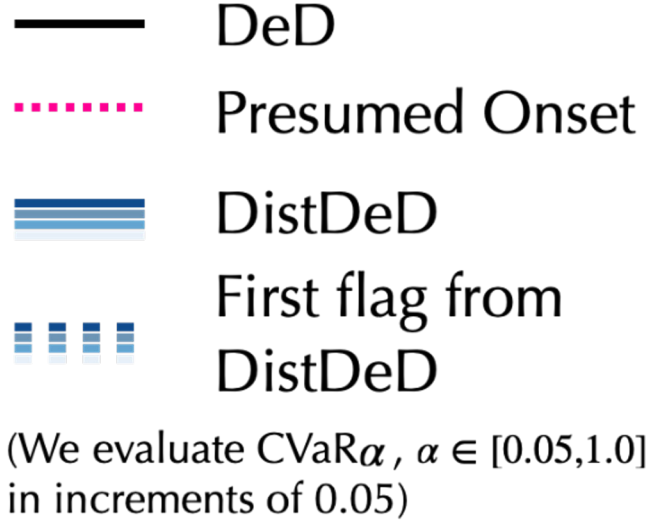
DistDeD provides a tunable measure of risk



By modeling the full distribution of outcomes, DistDeD enables expert to modulate conservatism on a per-case basis depending on their own risk characterisation



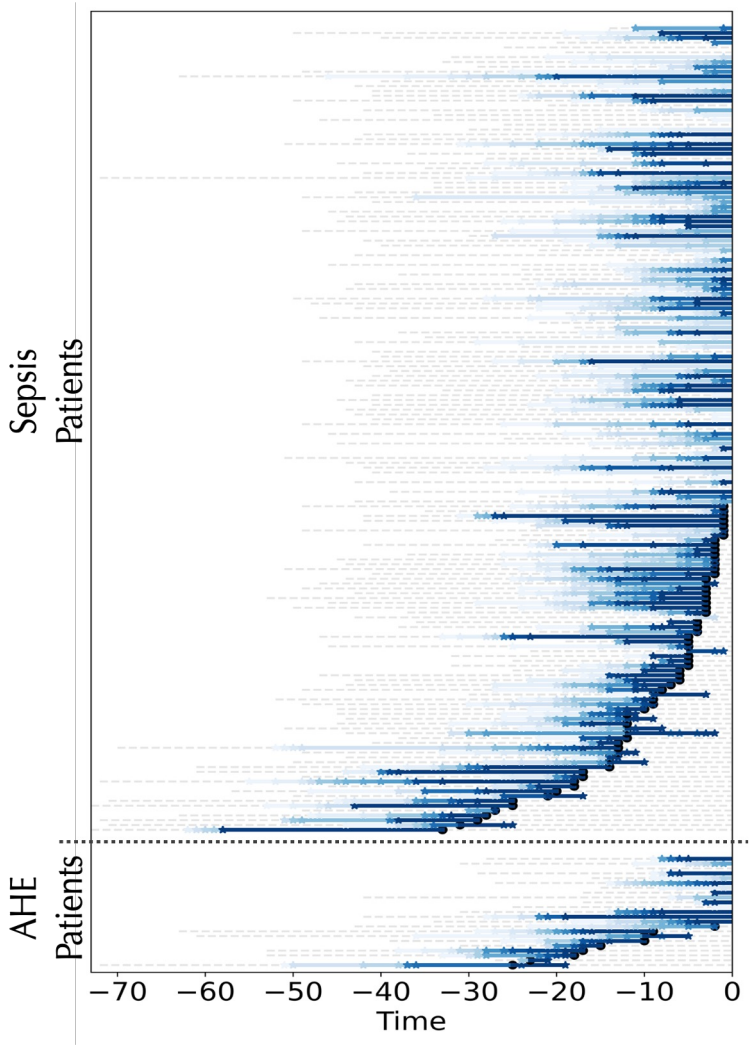


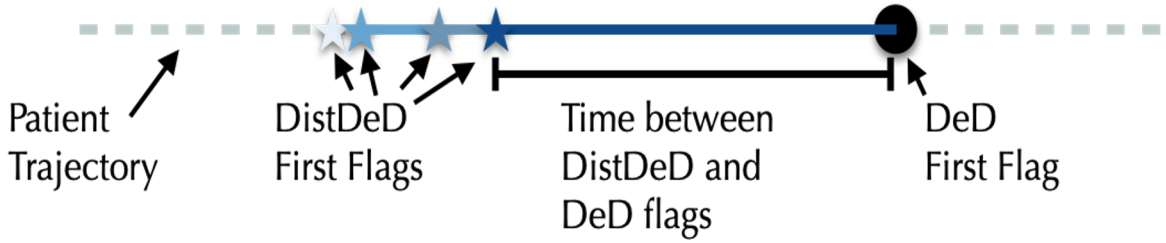


(We evaluate CVaRα, $\alpha \in [0.05,1.0]$ in increments of 0.05)

T.W Killian, S Parbhoo* and M. Ghassem*i,, **Risk-sensitive and Robust Dead-end Identification in Safety-Critical Offline RL, TMLR 2023.**

DistDeD provides an earlier indication of risk





DistDeD Avg. Hours Ahead (by CVaRα setting)

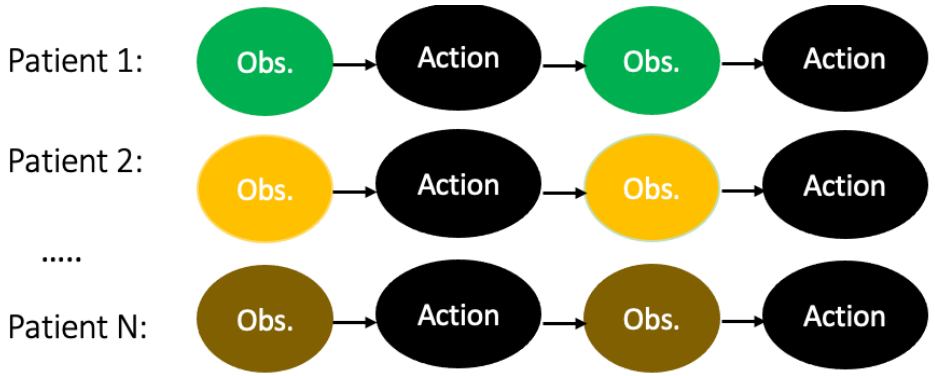
CVaRα risk quantile (α)	AHE Patients	Sepsis Patients
0.10	11.29	14.47
0.20	9.75	11.31
0.30	8.50	9.65
0.40	8.00	8.74
0.50	7.67	7.91
0.60	7.18	7.45
0.70	6.47	6.78
0.80	5.47	6.40
0.90	4.38	5.78
1.00	3.19	4.93

T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and Robust Dead-end Identification in Safety-Critical Offline RL, TMLR 2023.**

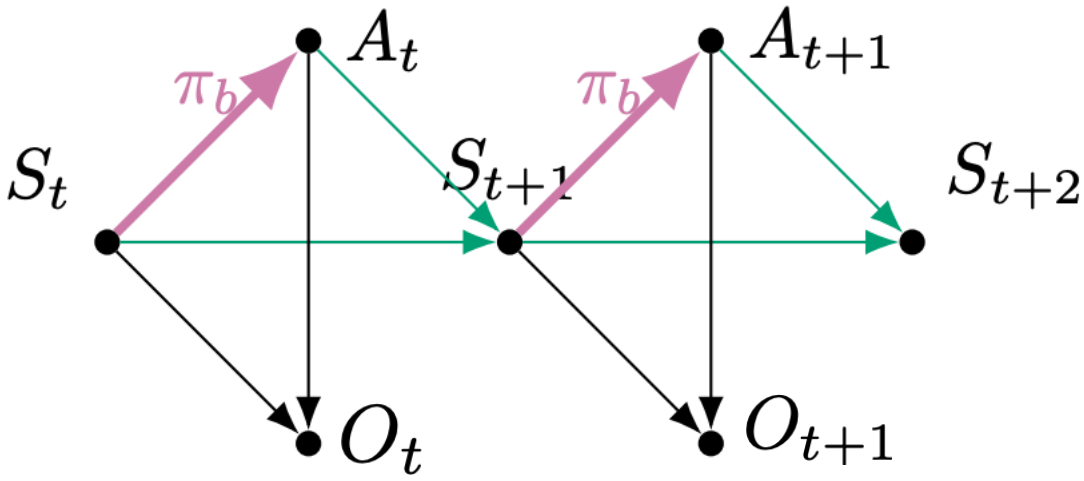
From Evaluation to Causally-Aware Evaluation

Reweight trajectories according to how much they overlap using:

$$\rho_n = \prod_t \frac{\pi_e(a_{tn}|s_{tn})}{\pi_b(a_{tn}|s_{tn})}$$



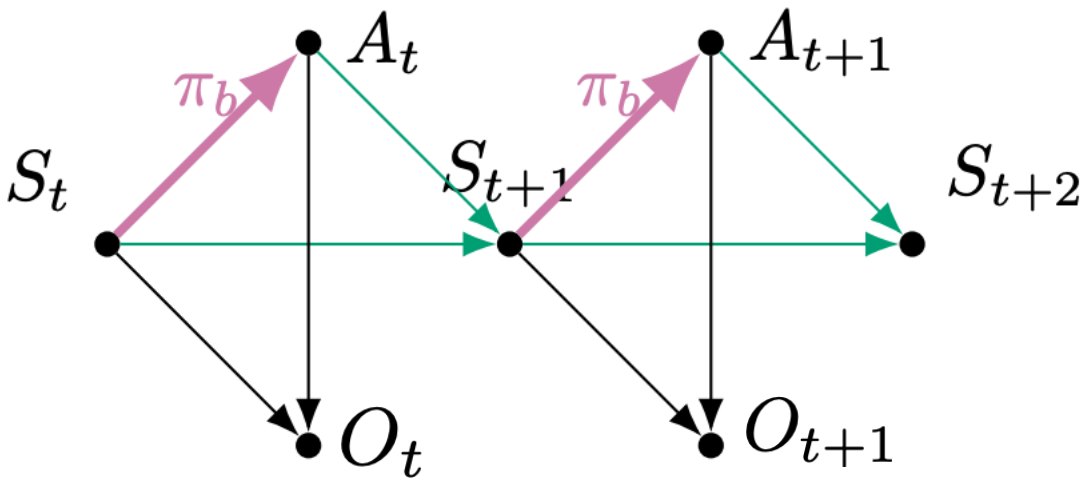
Standard OPE **ignores** the data-generating mechanism resulting in states, actions and observations.

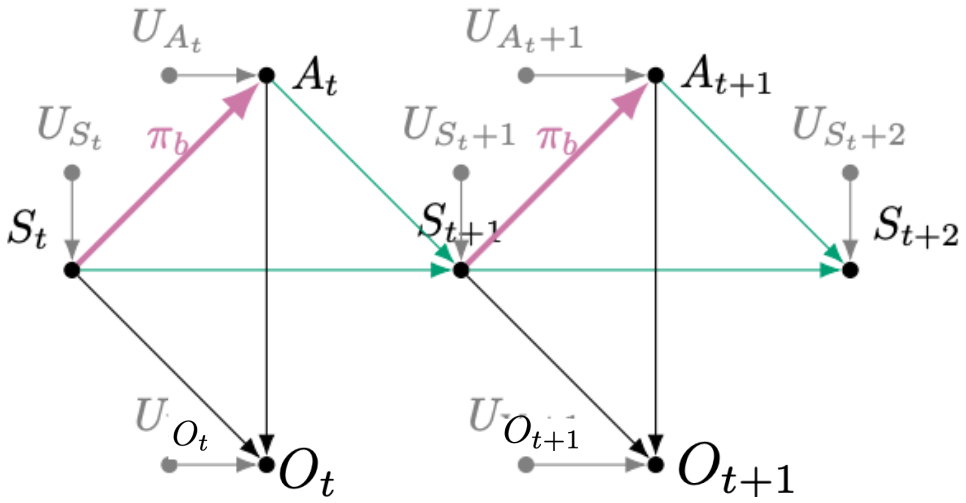


But the data-generating mechanism is important to understand when an OPE estimate makes sense and who it is applicable to

From Evaluation to Causally Aware Evaluation

The data-generating mechanism is important to understand when an OPE estimate makes sense and who it is applicable to

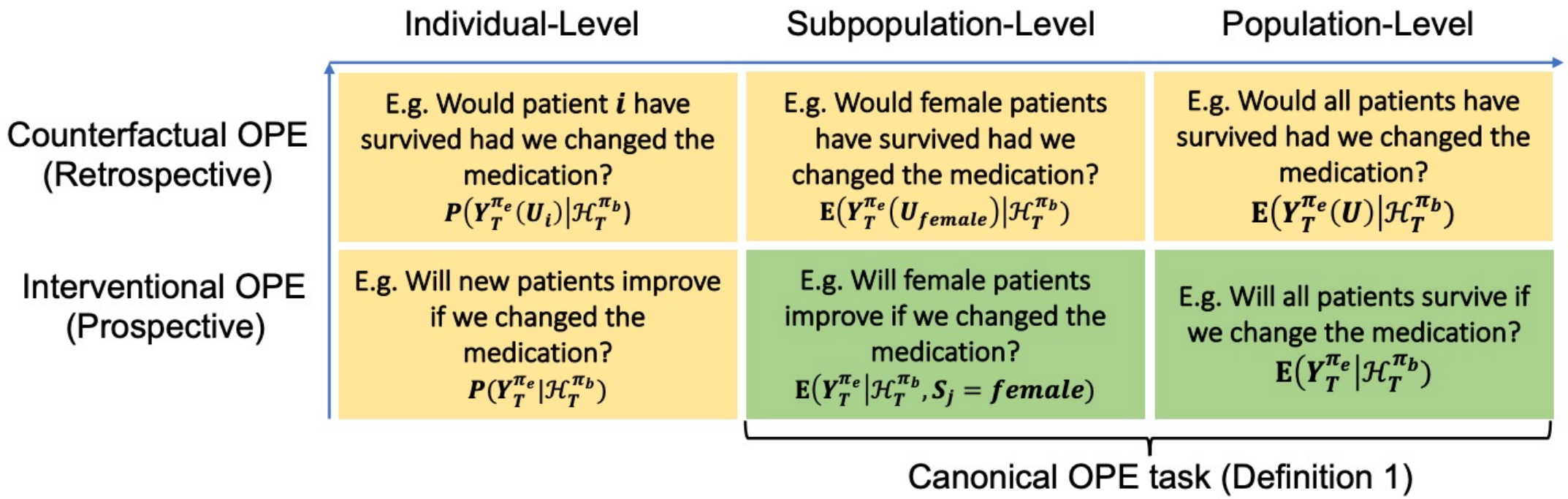




Parbhoo et al, Generalizing OPE from a Causal Perspective, arXiv pre-print, 2022.

From Evaluation to Causally Aware Evaluation

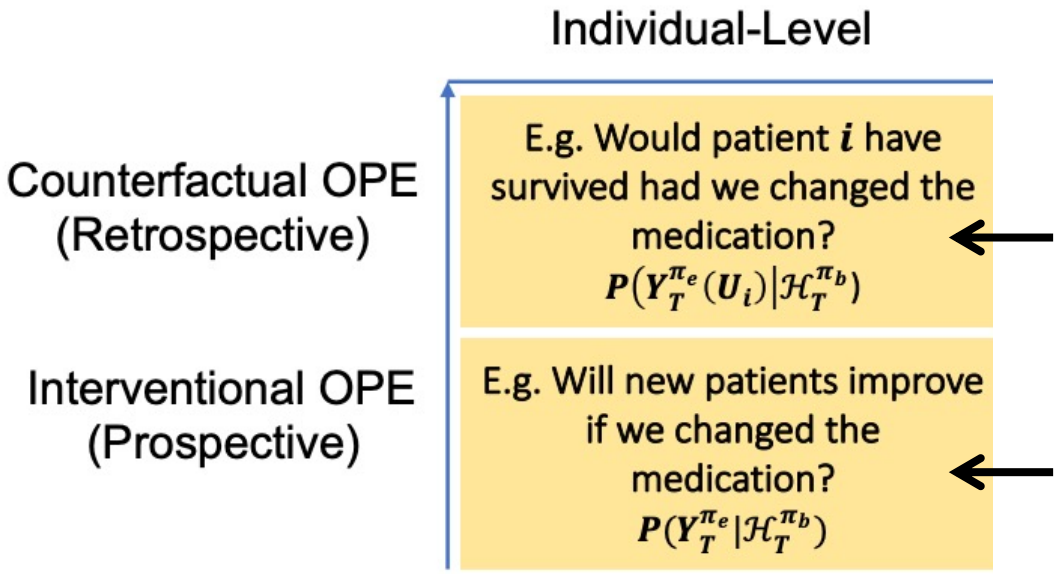
Specifically, retrospective and prospective queries **require different estimands**:



Parbhoo et al, Generalizing OPE from a Causal Perspective, arXiv pre-print, 2022.

From Evaluation to Causally-Aware Evaluation

Specifically, retrospective and prospective queries **require different estimands**:



Need to account for underlying **hidden factors**

Learn task-relevant representations that capture these factors (Sharma, Benac, Parbhoo, Doshi-Velez 2024)

Need to account for history by capturing the **transition dynamics**

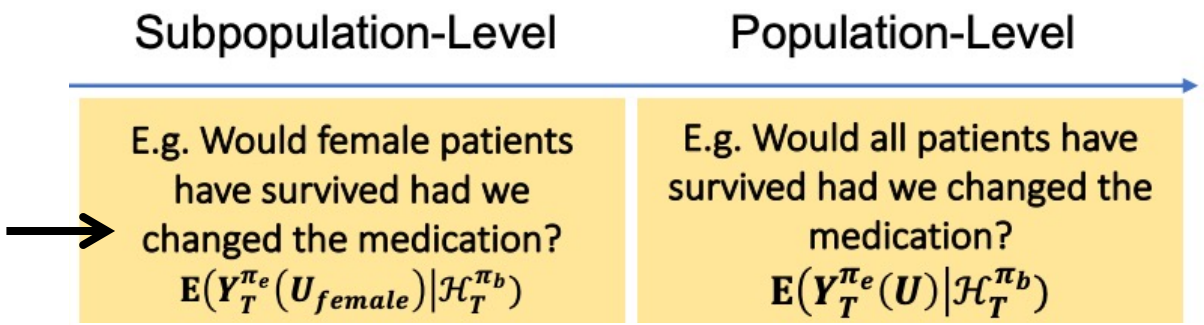
Parbhoo et al, Generalizing OPE from a Causal Perspective, arXiv pre-print, 2022.

From Evaluation to Causally-Aware Evaluation

Specifically, retrospective and prospective queries **require different estimands**:

Need to identify those **human concepts** or factors that capture underlying process **for specific subgroups**

Concept Driven OPE (Teversham, Majumdar, Parbhoo 2024)
Reward is (almost) enough (Jagota, Parbhoo 2024)
Shaping Control Variates for OPE (Jagota, Parbhoo, 2024)



Otherwise, need to identify where we might need to collect more information and **bring humans in the loop**

Learning to Defer to Humans for Sequential Decision-Making in Uncertainty (Parbhoo, Joshi, Doshi-Velez 2023)

Parbhoo et al, Generalizing OPE from a Causal Perspective, arXiv pre-print, 2022.

Summary

- **Evaluation should tailored to the application and the nature of the query for careful validation.**
- Factored action spaces are recommended for use in OPE when:
  - Data is limited
  - When the problem is **clearly decomposable into independent sub-problems**
  - Isolating certain action features can greatly improve the variance and MSE
- In other scenarios, it may be more useful to determine which **actions to avoid**
- In choosing between approaches, one should carefully consider what their objectives are as well as account for potential hidden factors and human concepts/invariances.
- Please feel free to reach out at: s.parbhoo@imperial.ac.uk