Imperial College London

Towards Risk-Aware Decision-Making and Policy Evaluation for Actionable Healthcare

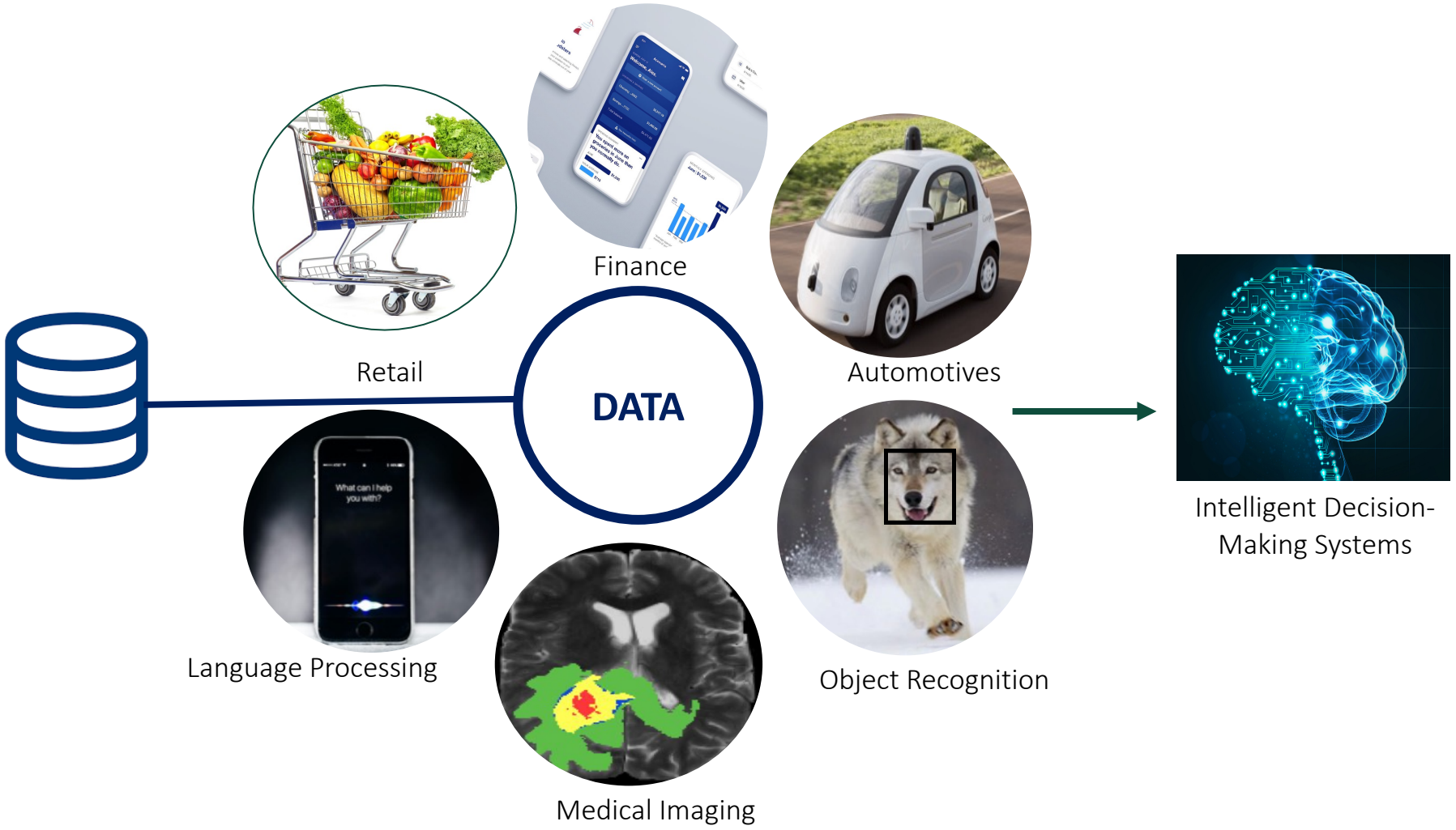
Sonali Parbhoo

Imperial College London

OxML Summer School

7 Aug 2022

Machine Learning is changing the way we make decisions



But machine learning has had limited success in high-risk settings like healthcare

The "inconvenient truth" about AI in healthcare

Trishan Panch, Heather Mattie & Leo Anthony Celi

npj Digital Medicine **2**, Article number: 77 (2019) | Cite this article

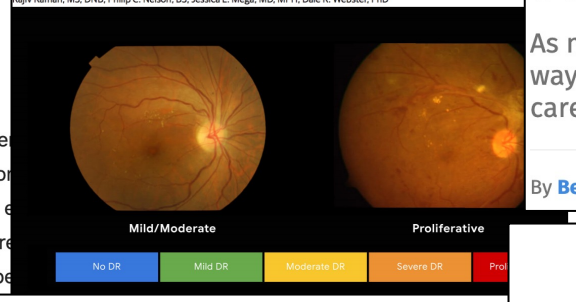
35k Accesses | **24** Citations | **452** Altmetric | Metrics

As the UK sits in painful deadlock over Brexit, it is important to re
that governments are regularly faced with crises, and their respon
create enduring benefit for future generations. Back in 1858, for e
the UK parliament was dealing with another messy crisis: "the gre
In a world before sanitation, the river Thames had become an ope
and as summer blossomed parliament
stench. £2.5 million (about £300 millio
approved to build a network of sewers
particular model of sanitation, develop
other cities around the world and the r
unthinkable that a developed nation w

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

Development and Validation of a Deep Learning Alg for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswa
Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB;
Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD



AI in healthcare - not so fast? Study outlines challenges, dangers for machine learning

As machine learning rapidly expands into healthcare, the ways it "learns" may be at odds with clinical outcomes unless carefully controlled for, a new study shows.

By **Benjamin Harris** | January 16, 2019 | 10:21 AM

Why AI fell short in slowing the spread of COVID-19

Though hopes were high, experts say a lack of high-quality data meant models were both underwhelming and "anti-constructive."

By **Kat Jercich** | July 23, 2020 | 04:35 PM

MEDICAL & BIOTECH

Artificial Intelligence Is Rushing Into Patient Care—And Could Raise Risks

AI systems are not as rigorously tested as other medical devices, and have already made serious mistakes

By Liz Szabo, Kaiser Health News on December 24, 2019

Artificial Intelligence Makes Bad Medicine Even Worse

A new study out from Google seems to show the promise of AI-assisted health care. Actually, it shows the threat.

Why?

High stake decision-making is challenging *even for humans*

What treatment should a patient take?

How should a teacher assist a student?

Why?

High stake decision-making is challenging *even for humans*

What treatment should a patient take?

How should a teacher assist a student?

Only 6% of asthmatics would be eligible for treatments from RCTs[1]

Does a model generalise and to whom will it generalise?

[1]Smith, Saunders et al. Best Care at Lower Cost: The Path to Continuously Learning Healthcare in America, 2013.

Decisions often have *long-term* consequences

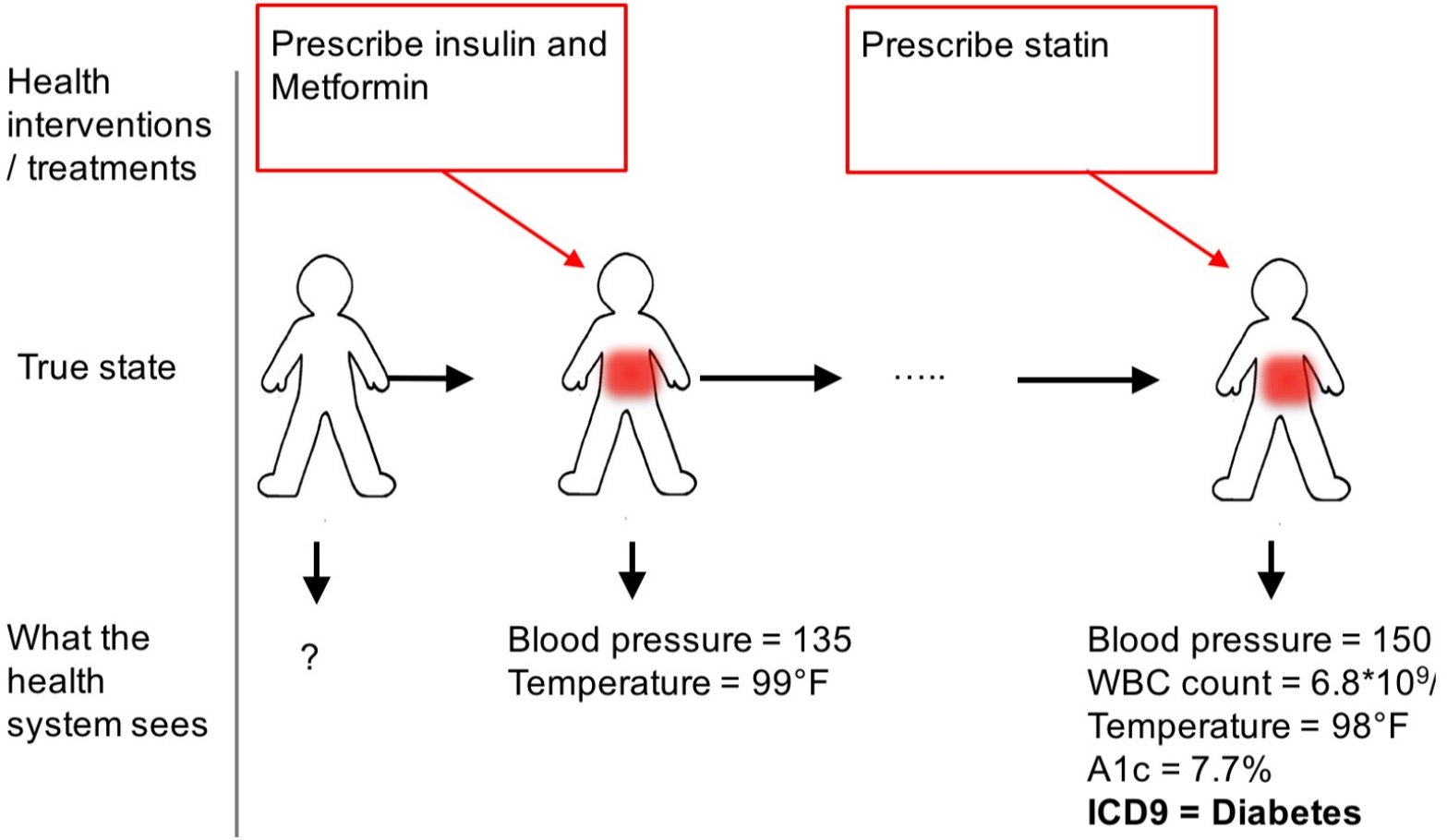


Image credit: David Sontag – ML for Healthcare, 6.871

Decisions often have *long-term* consequences

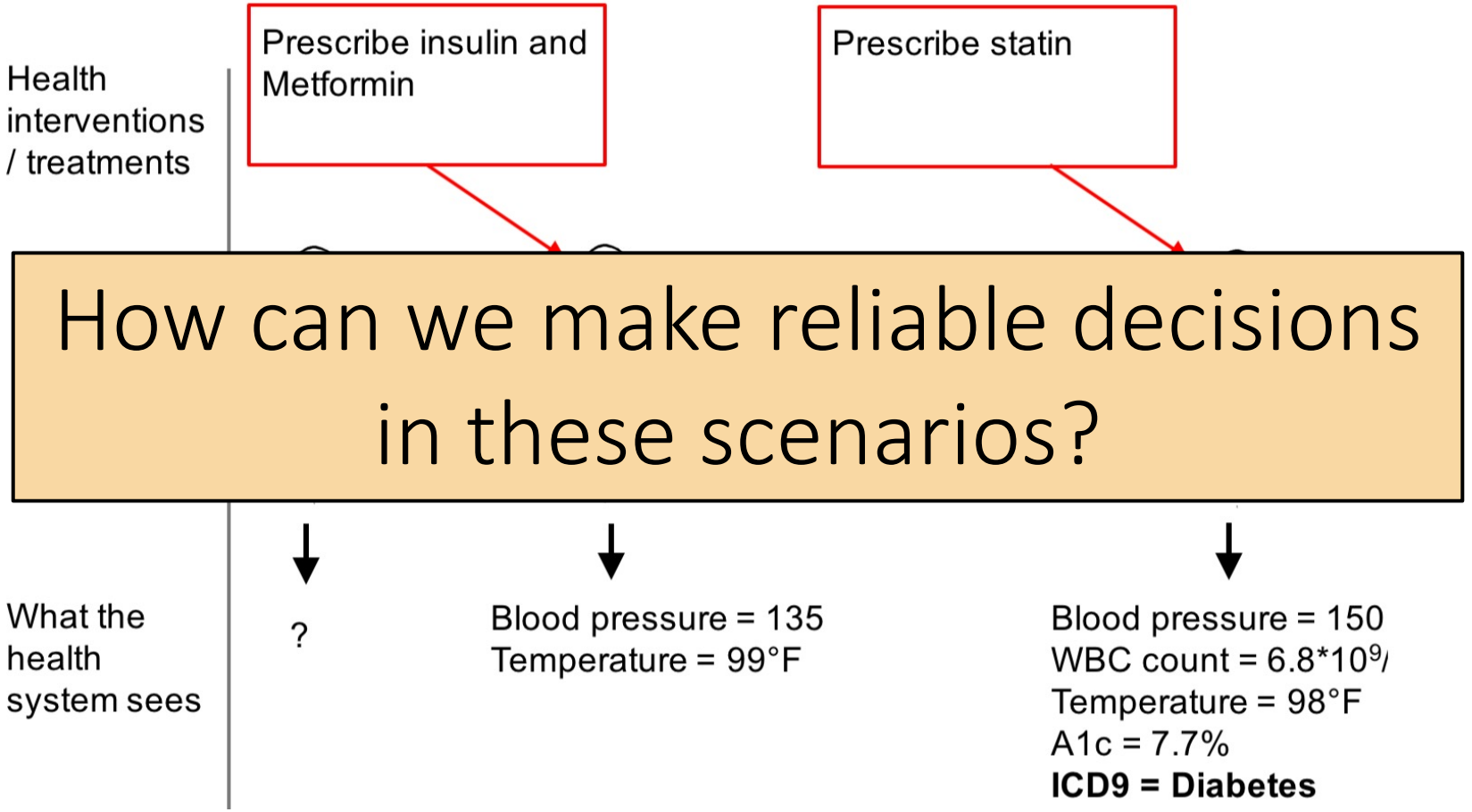
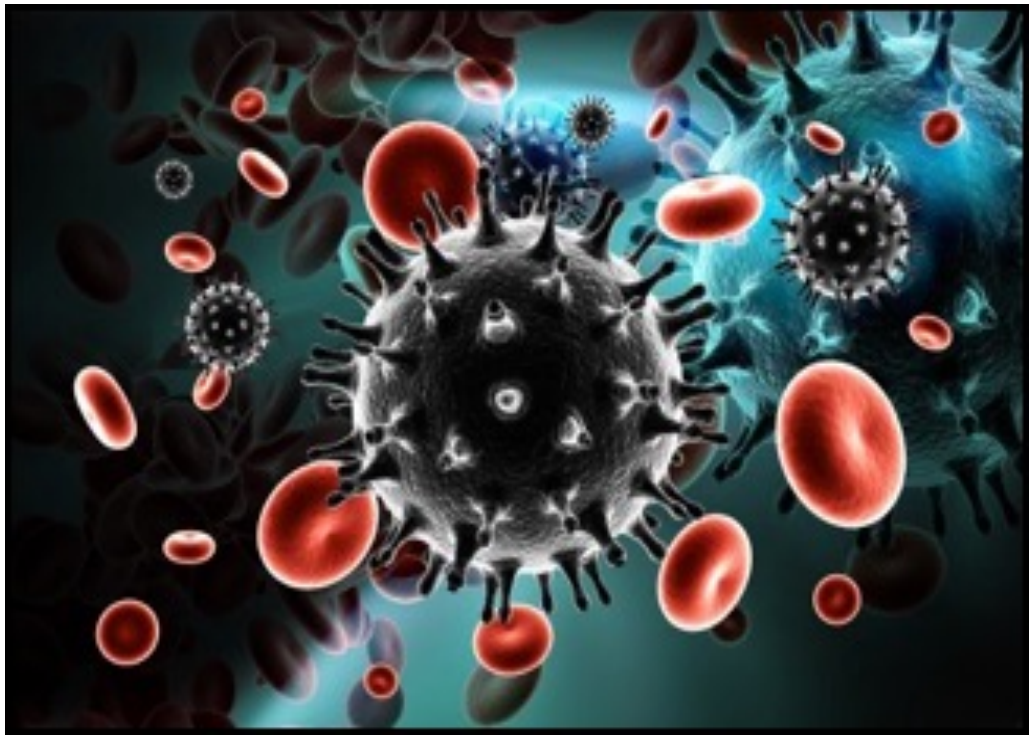


Image credit: David Sontag – ML for Healthcare, 6.871

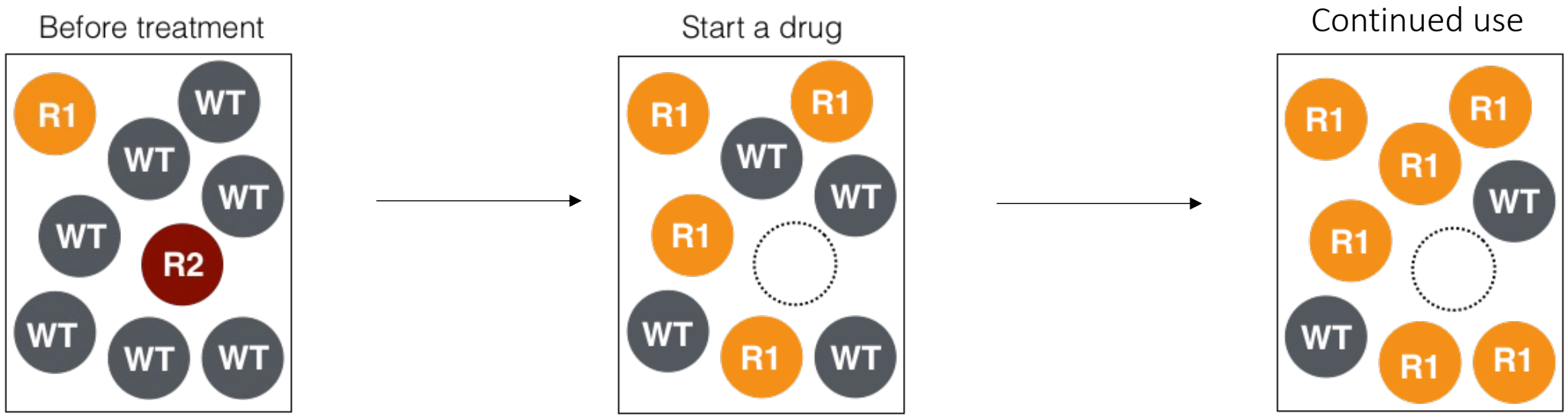
Example: HIV Therapy Management



- Requires life-long treatment with **combinations of antiretrovirals**.
- Virus is highly mutagenic and can become drug resistant.
- Patients exhibit significant **heterogeneity.**

Example: HIV Therapy Management

- Treatments consist of five classes of drugs: NNRTIs, NRTIs, PIs, IIs, FIs
- The virus mutates in response to treatments leading to **drug resistance**.

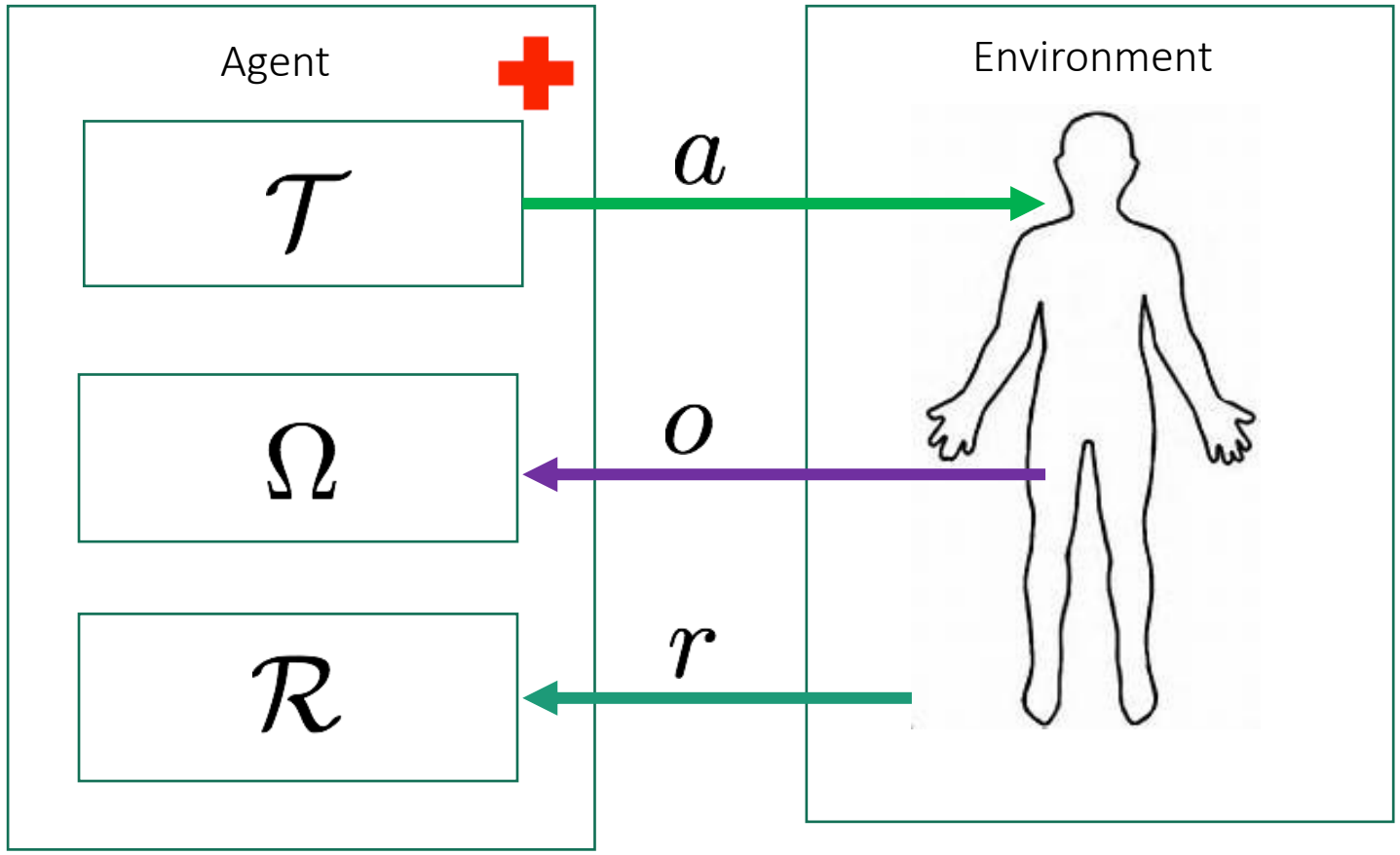


Reinforcement Learning for Healthcare

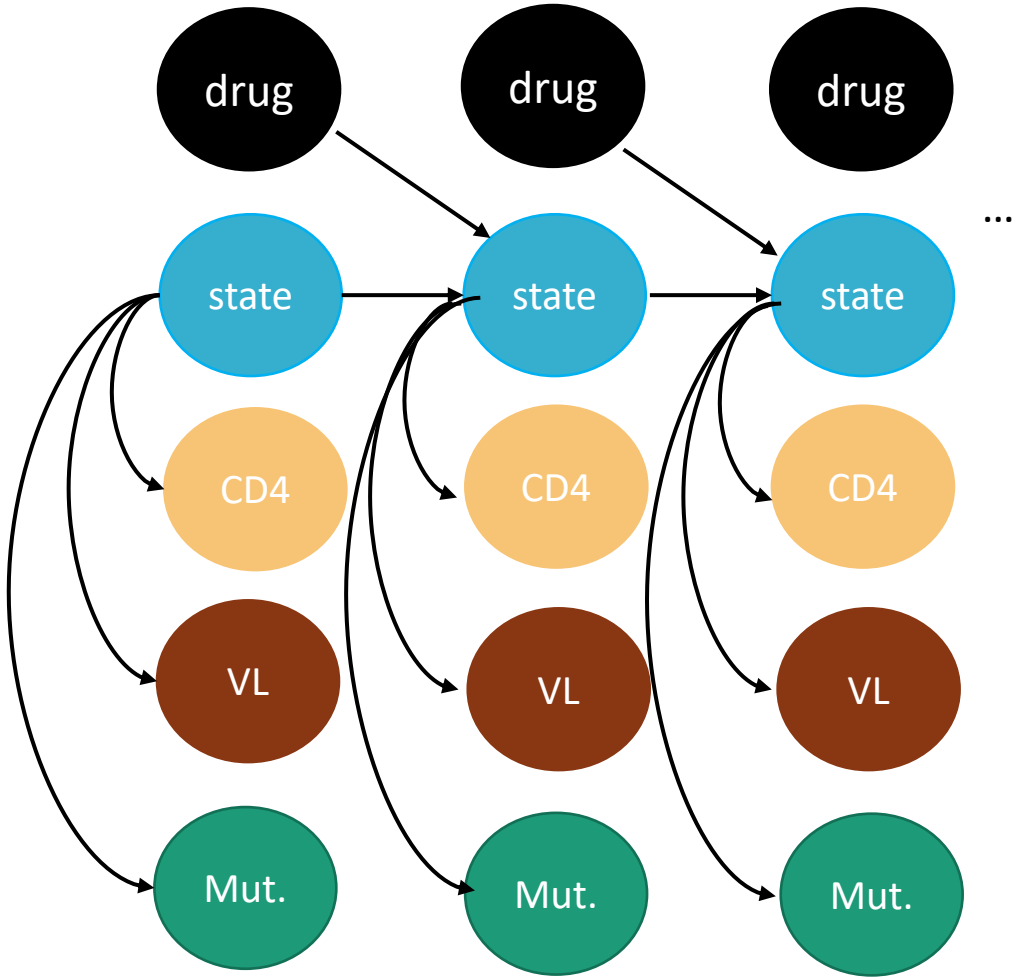
Medicine is inherently a sequential decision making problem:

- Clinicians, with their best understanding of a patient's status, propose a treatment.
- Patient status may change as a result of the treatment.
- Outcomes are a result of prior treatments.

RL makes this process explicit



Solutions: Train a model/learn a value function



Modelling HIV as an RL problem solves the long-term problem (e.g. Ernst 2007; Parbhoo 2014; Marivate 2015), but often in **simplified/simulation settings only.**

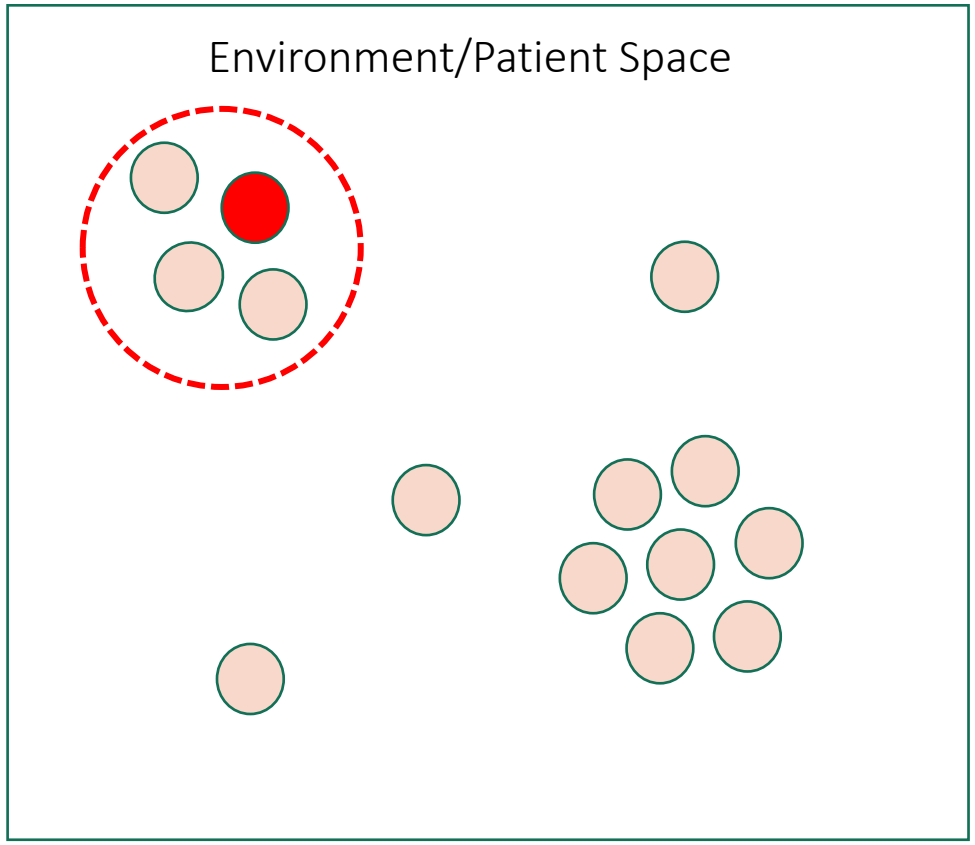
if $V_t > 40$:

$$r_t = -0.7 \log V_t + 0.6 \log T_t - 0.2|M_t|$$

else:

$$r_t = 5 + 0.6 \log T_t - 0.2|M_t|$$

Solutions: Use Non-parametric Methods



- Look at **similar patients** and use their full history to predict immediate outcomes (e.g. Bogojeska et al 2012)

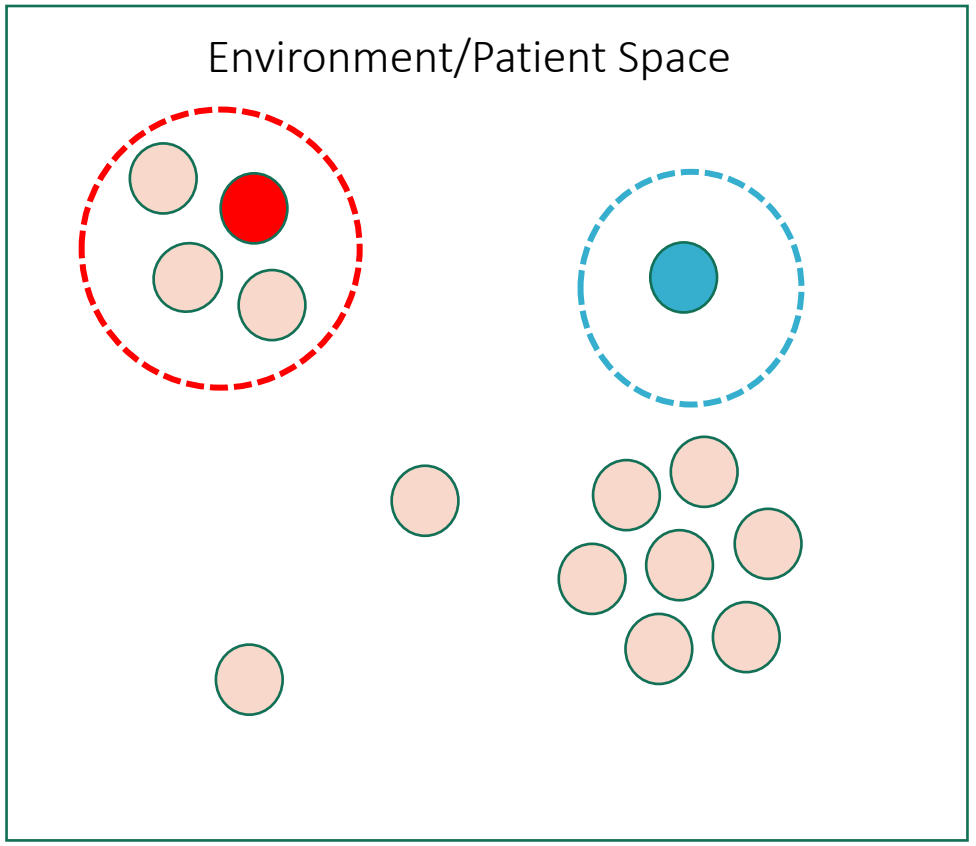
- These methods however, ignore long-term effects:

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

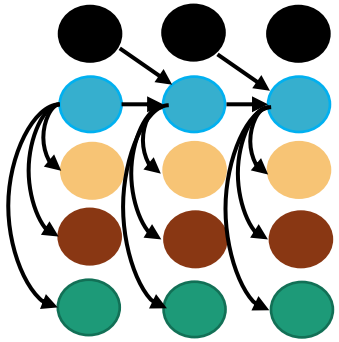
in 21 days

Our insight: These approaches have complementary strengths

Patients that are similar are best modelled by their neighbours

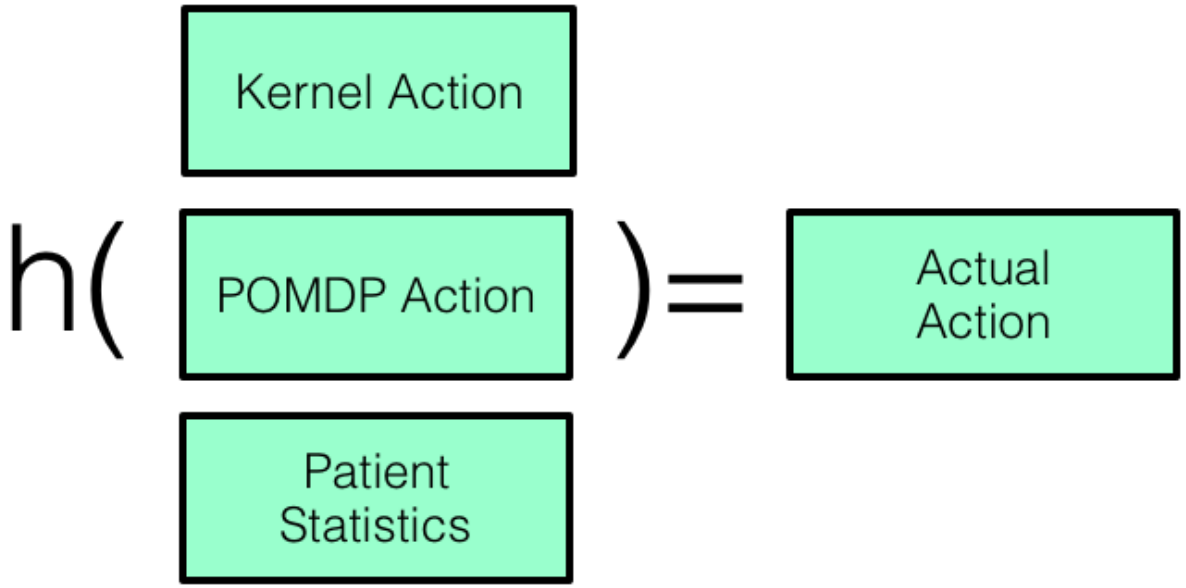


Patients without neighbours are best modelled with a parametric model



Combining Kernel and Model-based RL for HIV Therapy Selection – Parbhoo et al, 2017, AMIA

New Solution: Combine nonparametric and parametric models for better generalisation



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

Combining Kernel and Model-based RL for HIV Therapy Selection – Parbhoo et al, 2017, AMIA

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

Extension: Transfer between EuResist and South African Cohorts

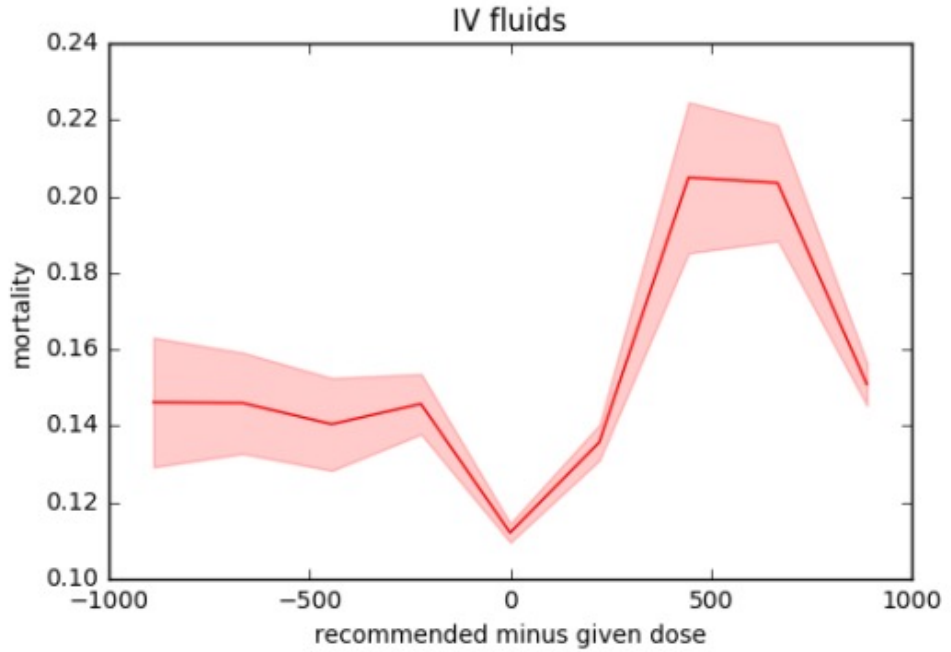
- We also identified when we could transfer policies in well-curated cohorts such as EuResist to South African cohorts that weren't as well curated.

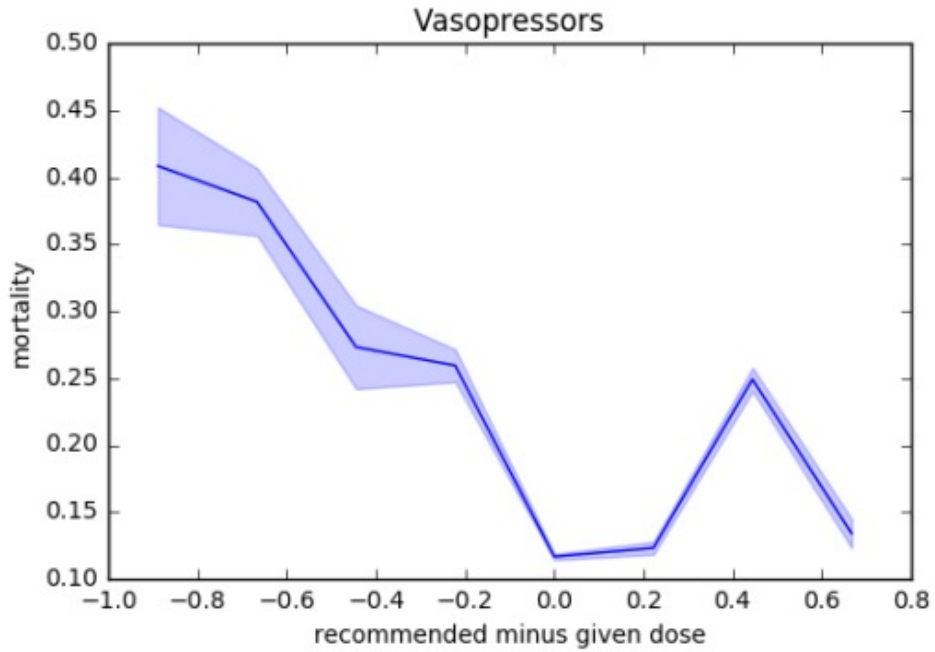
Type	Method	DR	IS	WIS
Behaviour Policy		5.02 ± 1.18		
Local	Kernel	3.56 ± 1.42	1.27 ± 1.14	1.80 ± 1.07
	CEIB	3.29 ± 1.13	3.80 ± 2.41	3.76 ± 2.19
Transfer	Kernel	4.17 ± 1.4	4.18 ± 1.20	4.16 ± 1.71
	CEIB	6.29 ± 0.14	5.17 ± 0.38	5.27 ± 0.29
	Mixture-of-Experts	5.28 ± 0.37	3.42 ± 1.39	4.81 ± 1.25
Local + 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

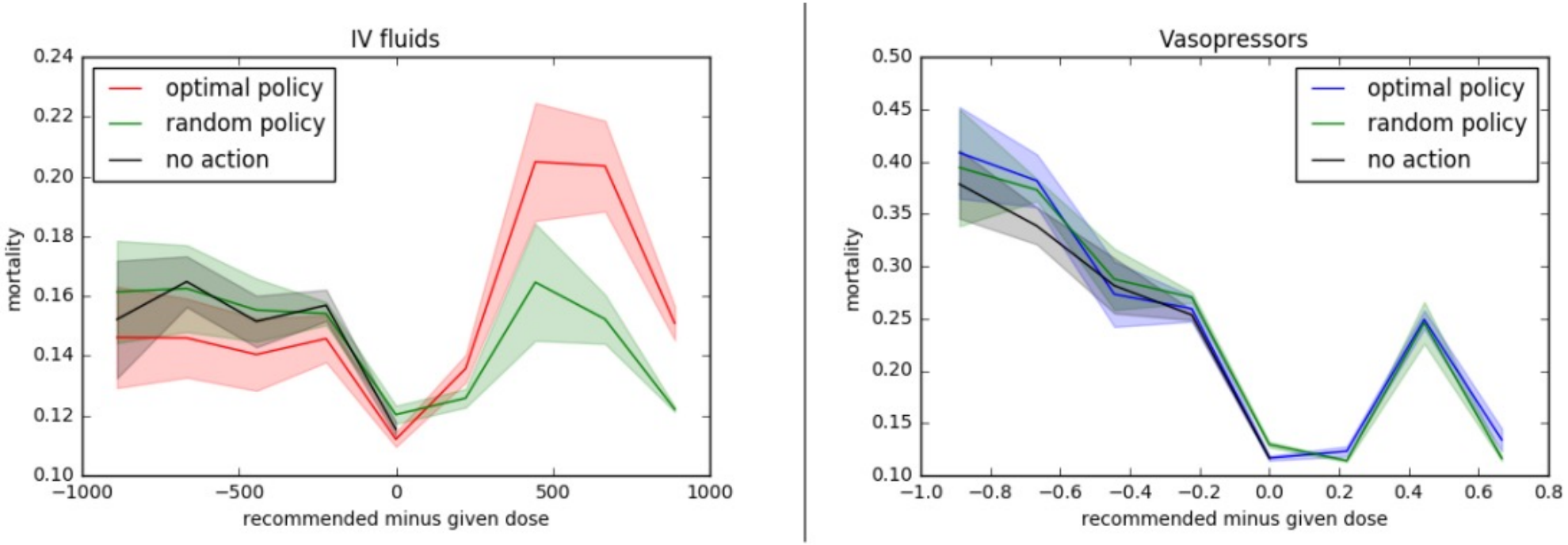




Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

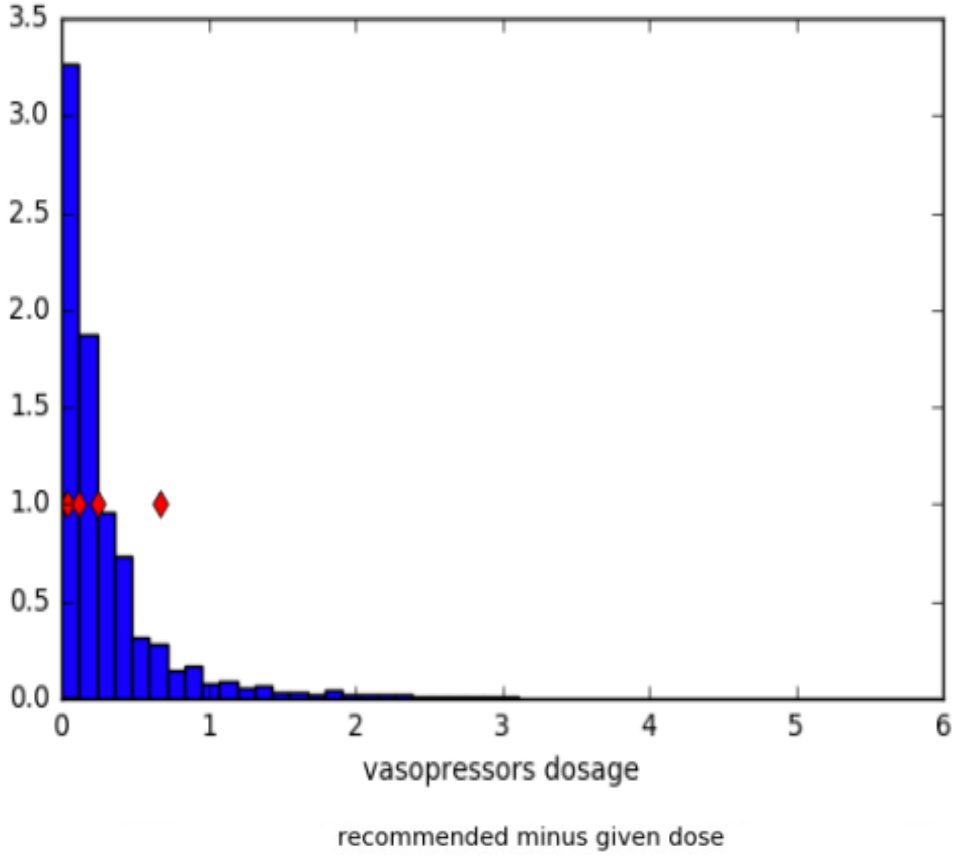
…. but evaluation is brittle

	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



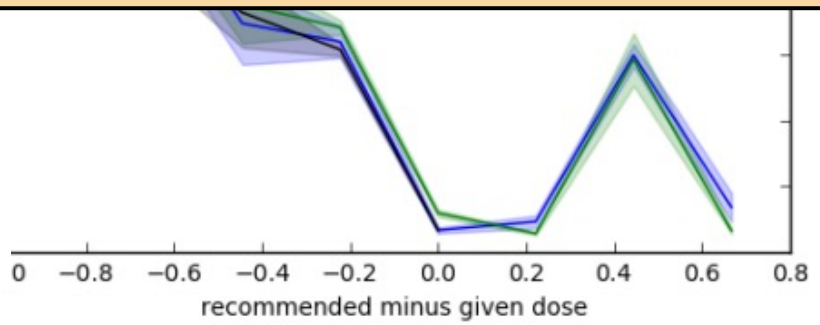
Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

…. but evaluation is brittle



QN	MoE_{V_d,Q_d}	MoE_{V_b,Q_b}
.06	3.93	4.31

This occurs because of the spread of the treatments administered



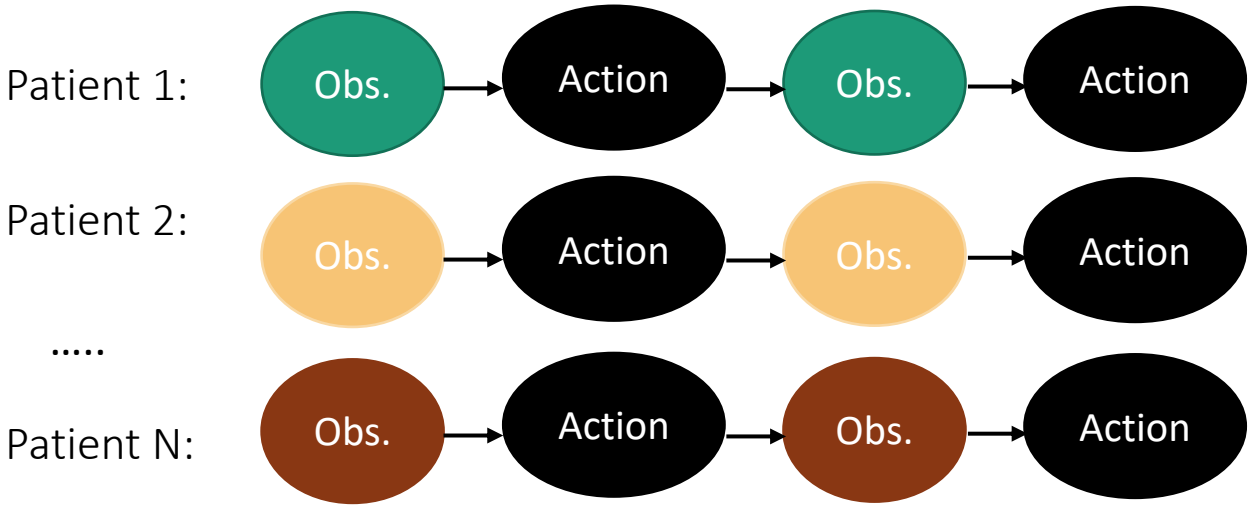
Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

Off-Policy Evaluation

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

Clinician:

π_b



Evaluation Agent:

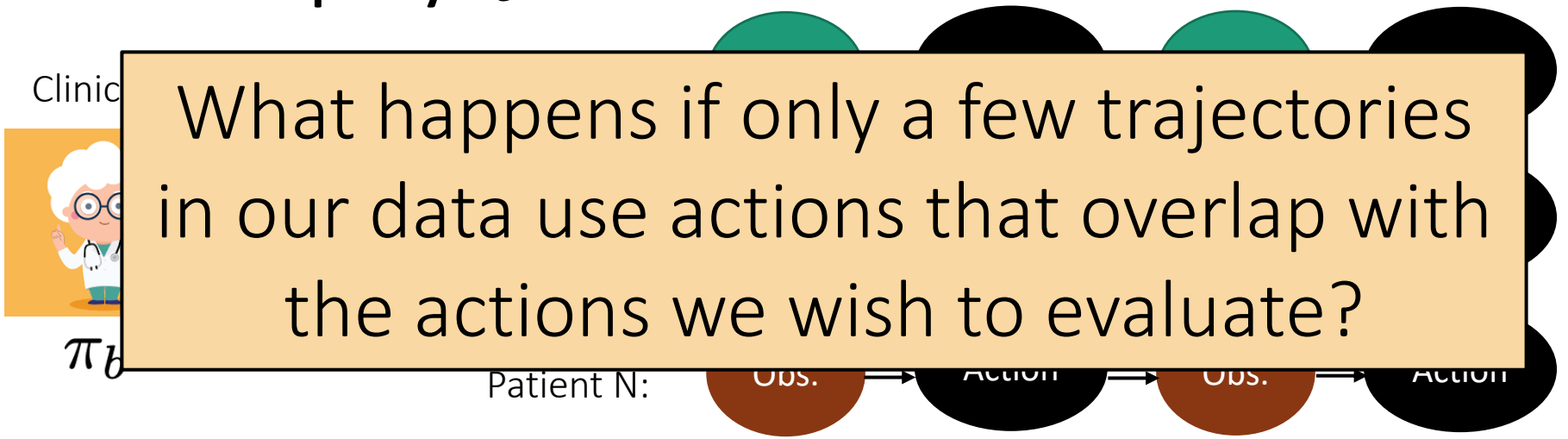
π_e

Reweight trajectories according to how much they overlap using:

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

Off-Policy Evaluation

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



What happens if only a few trajectories in our data use actions that overlap with the actions we wish to evaluate?

Evaluation Agent:

π_e

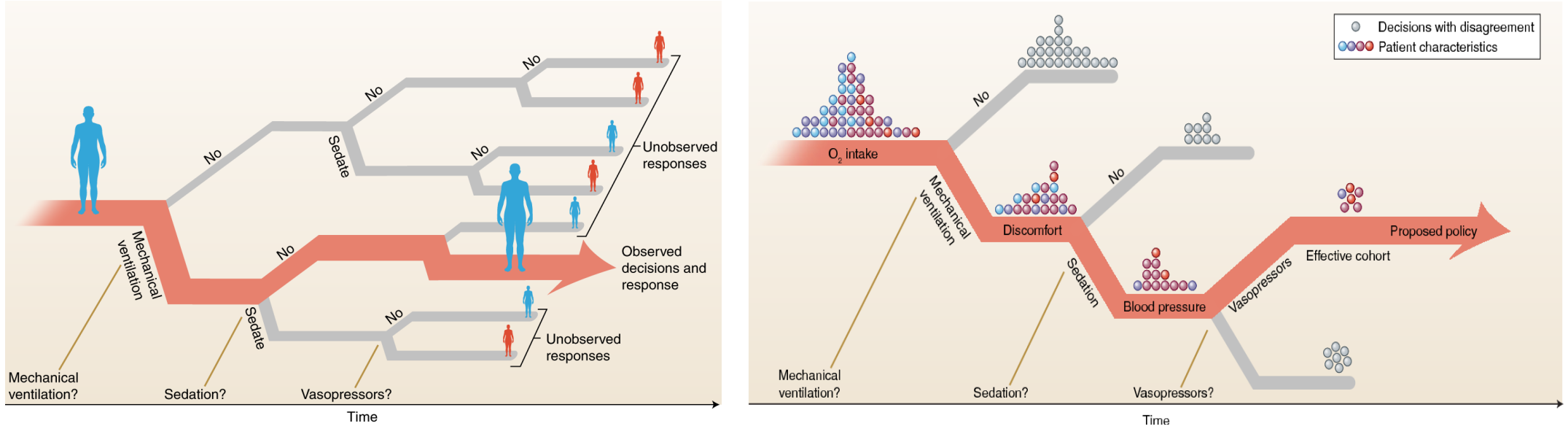
Reweight trajectories according to how much they overlap using:

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

Challenges to using RL for healthcare in practice

Learning suitable treatment policies from observational data is **offline and off-policy**

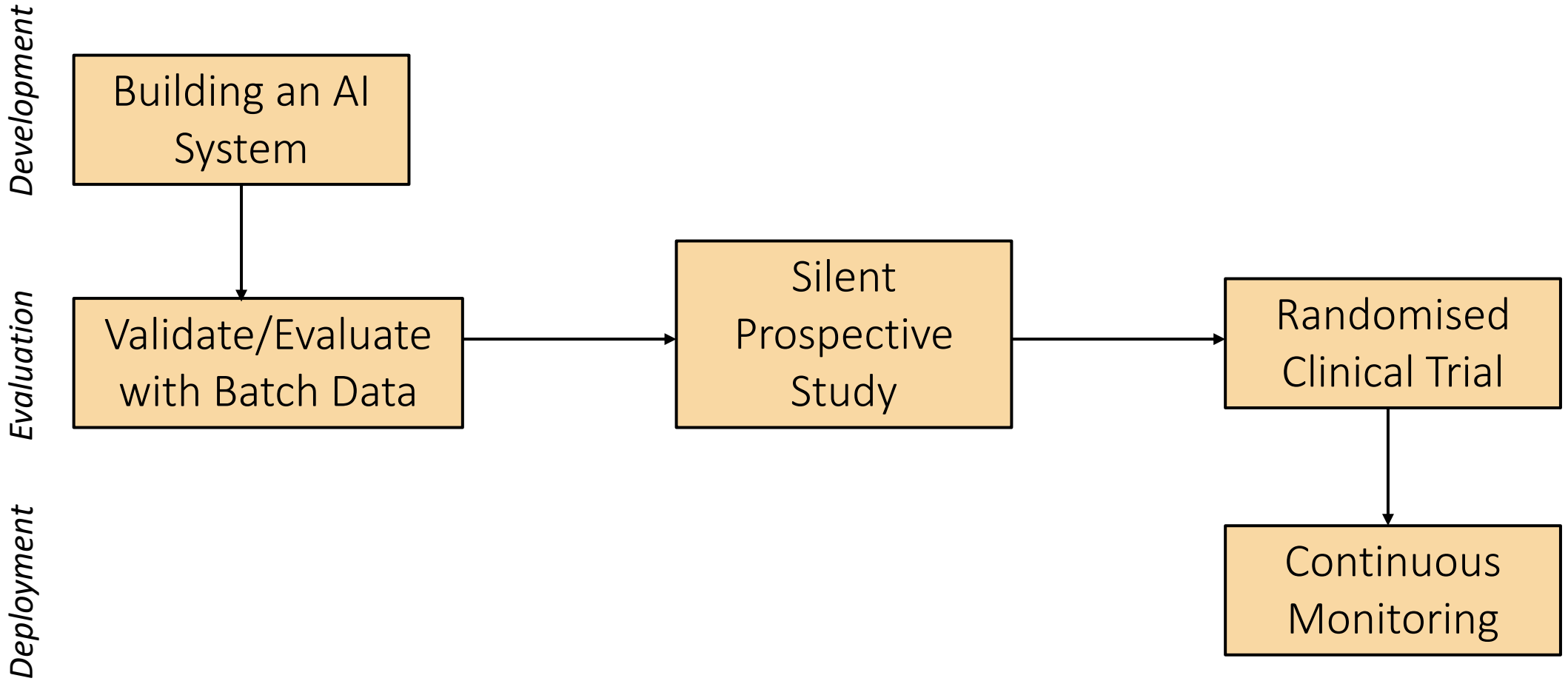
1. Inability to explore
2. Small data (relatively) → shrinking data support as best strategies are discovered



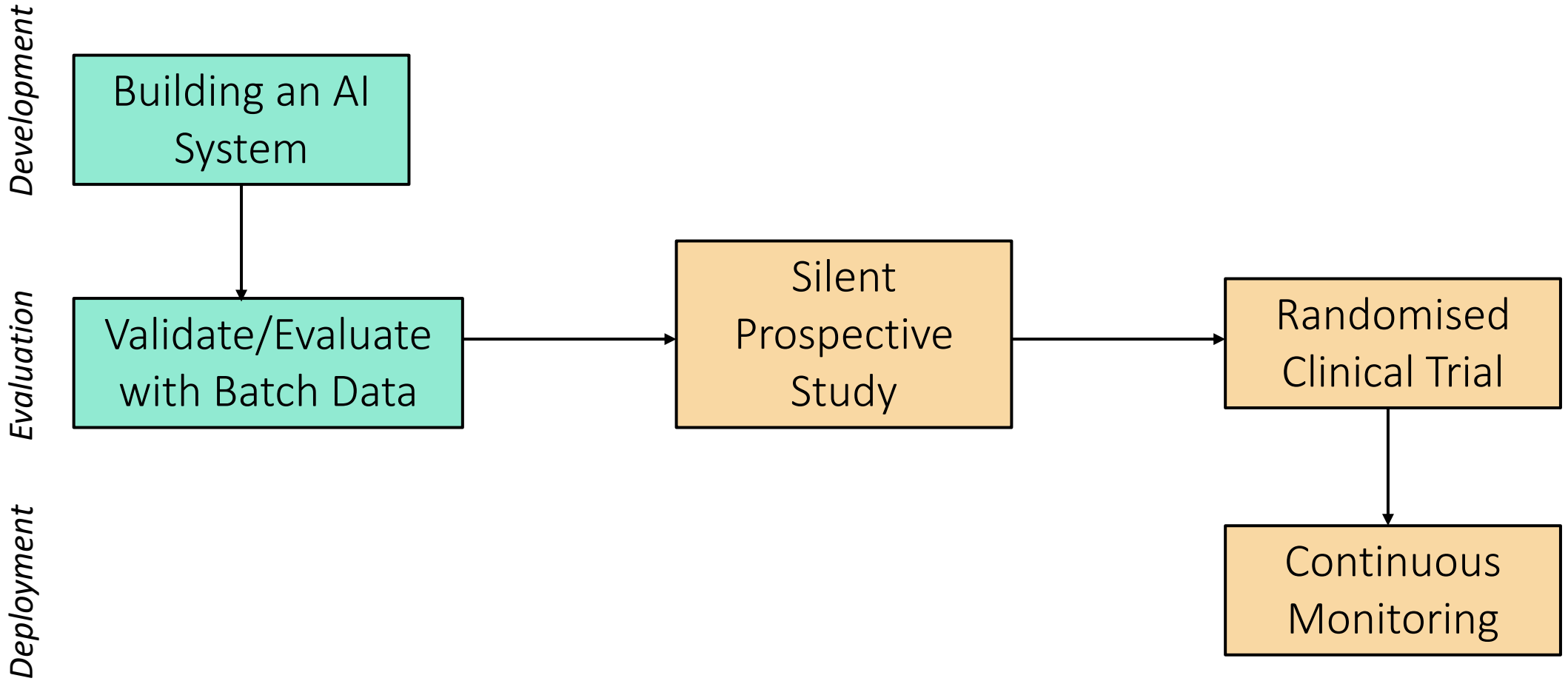
These two limitations severely complicate the development of actionable RL

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

We need methods for responsible model development, evaluation and deployment



We need methods for responsible model development, evaluation and deployment



What should responsible development and evaluation look like?

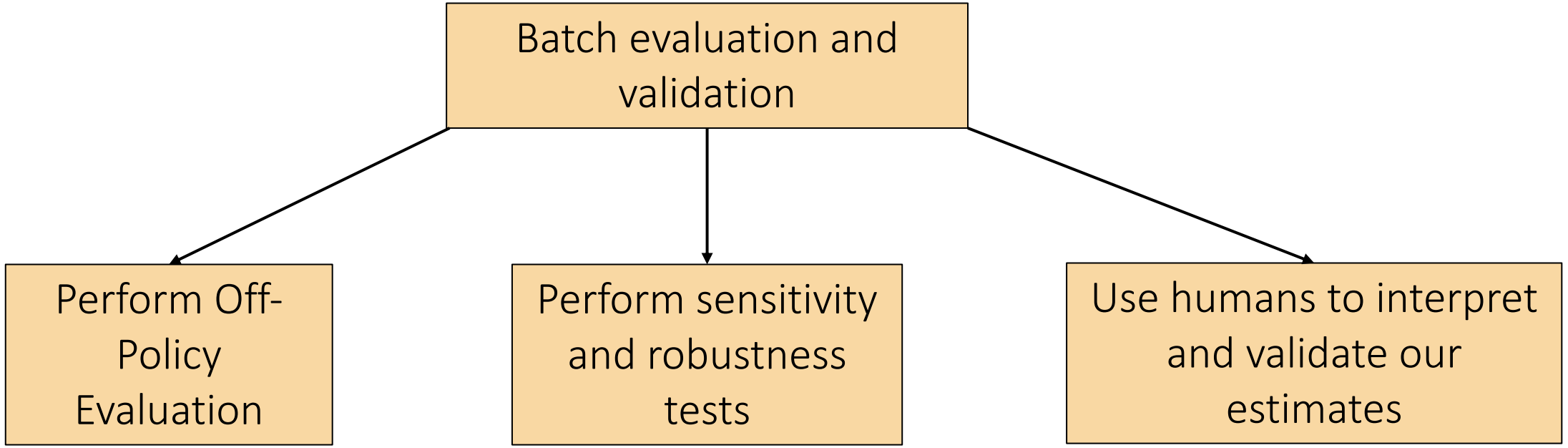


Image Credit: Doshi-Velez, Towards using Batch RL to Identifying Treatment Options, 2022

What should responsible development and evaluation look like?

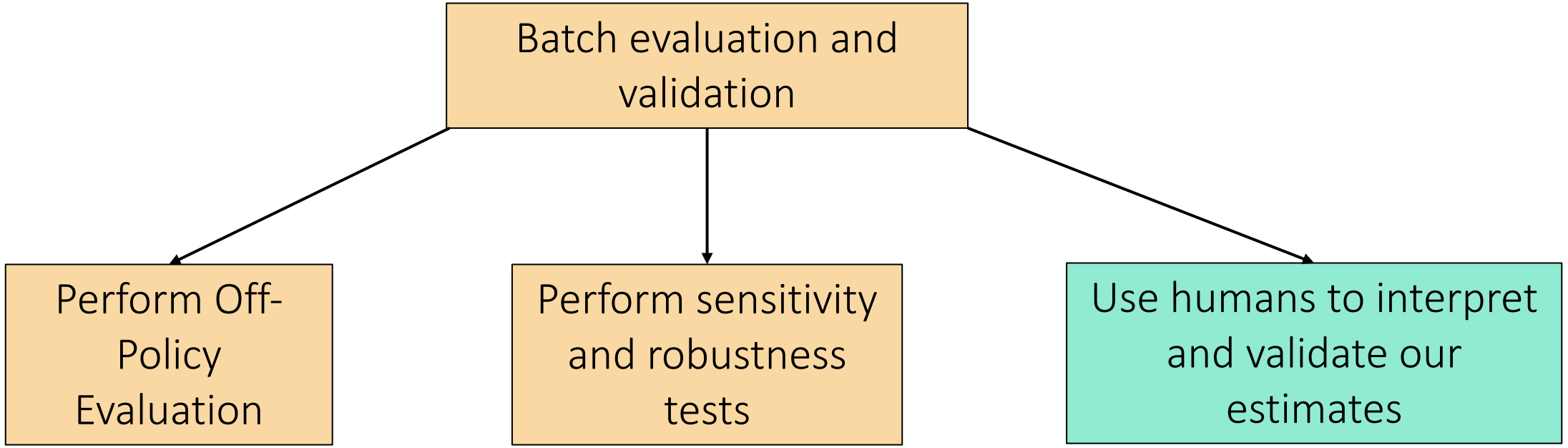
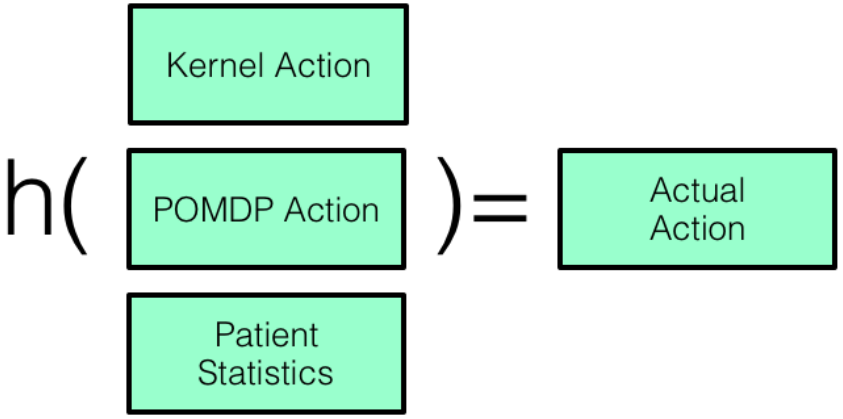


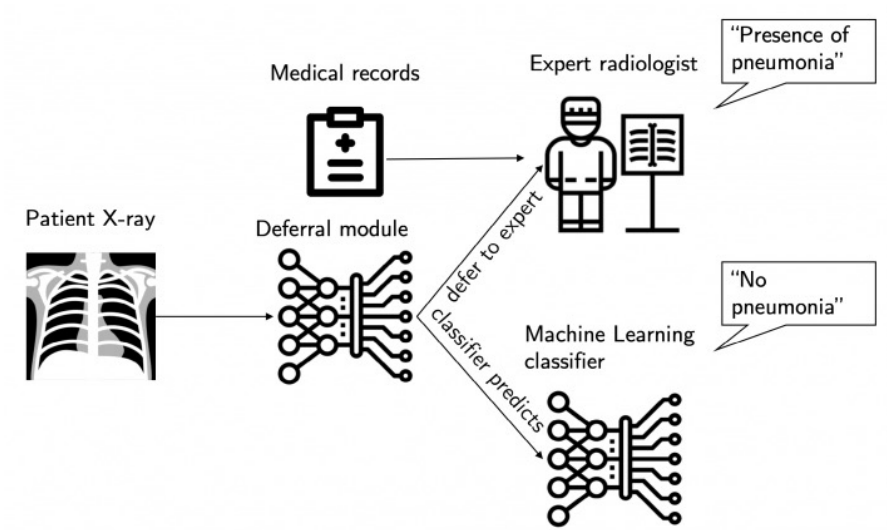
Image Credit: Doshi-Velez, Towards using Batch RL to Identifying Treatment Options, 2022

Learning to Defer to Humans in Uncertainty

- Human expertise is valuable in high-risk settings such as healthcare.
- Domain experts can often **provide additional knowledge** that can help reduce uncertainties for rare cases or outlier patients.
- Though prior works have explored ways to combine humans and models, none of these account for **uncertainty** and **long-term effects**.

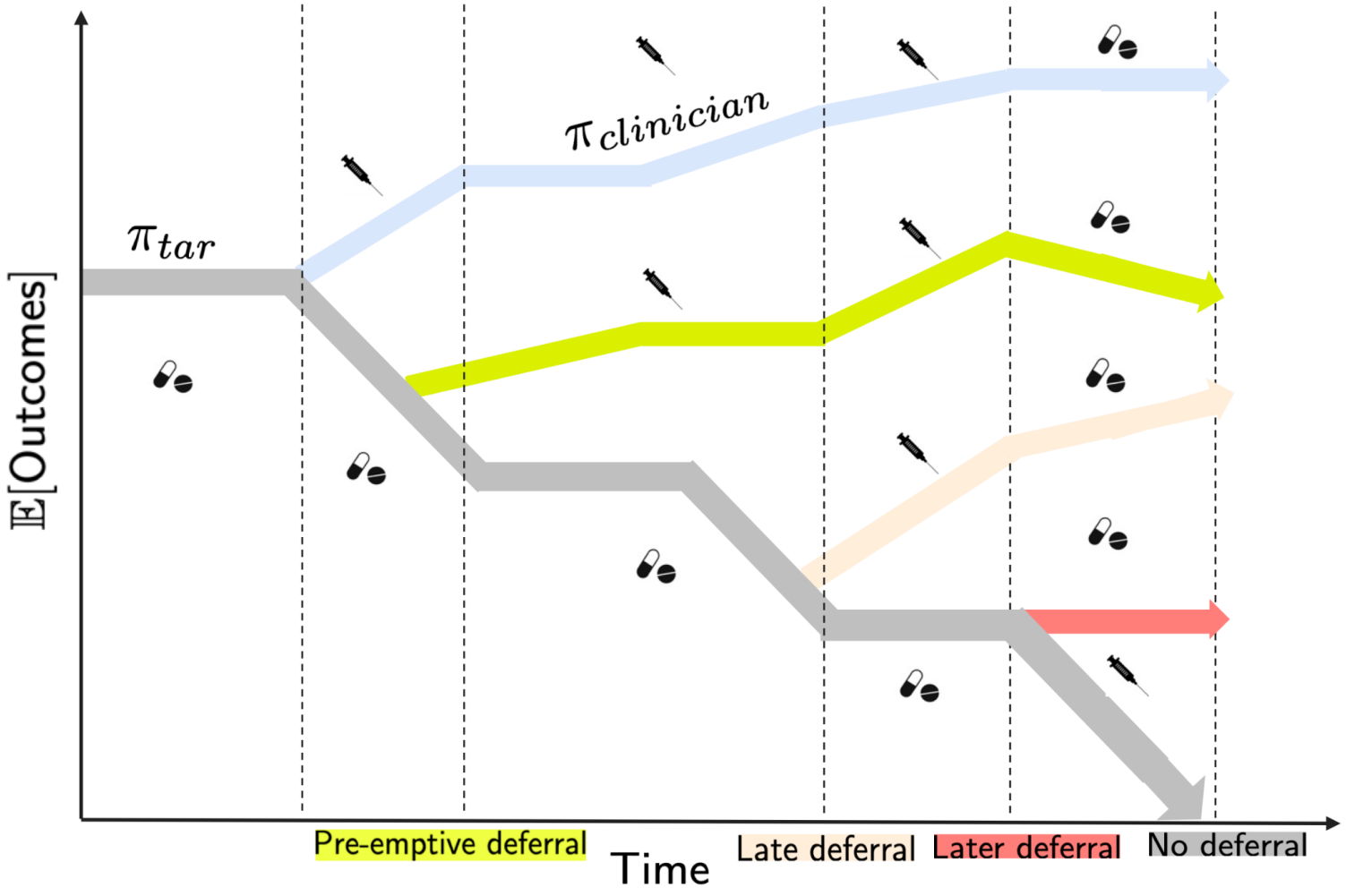


Combining Kernel and Model-based RL for HIV Therapy Selection – Parbhoo et al, 2017, AMIA

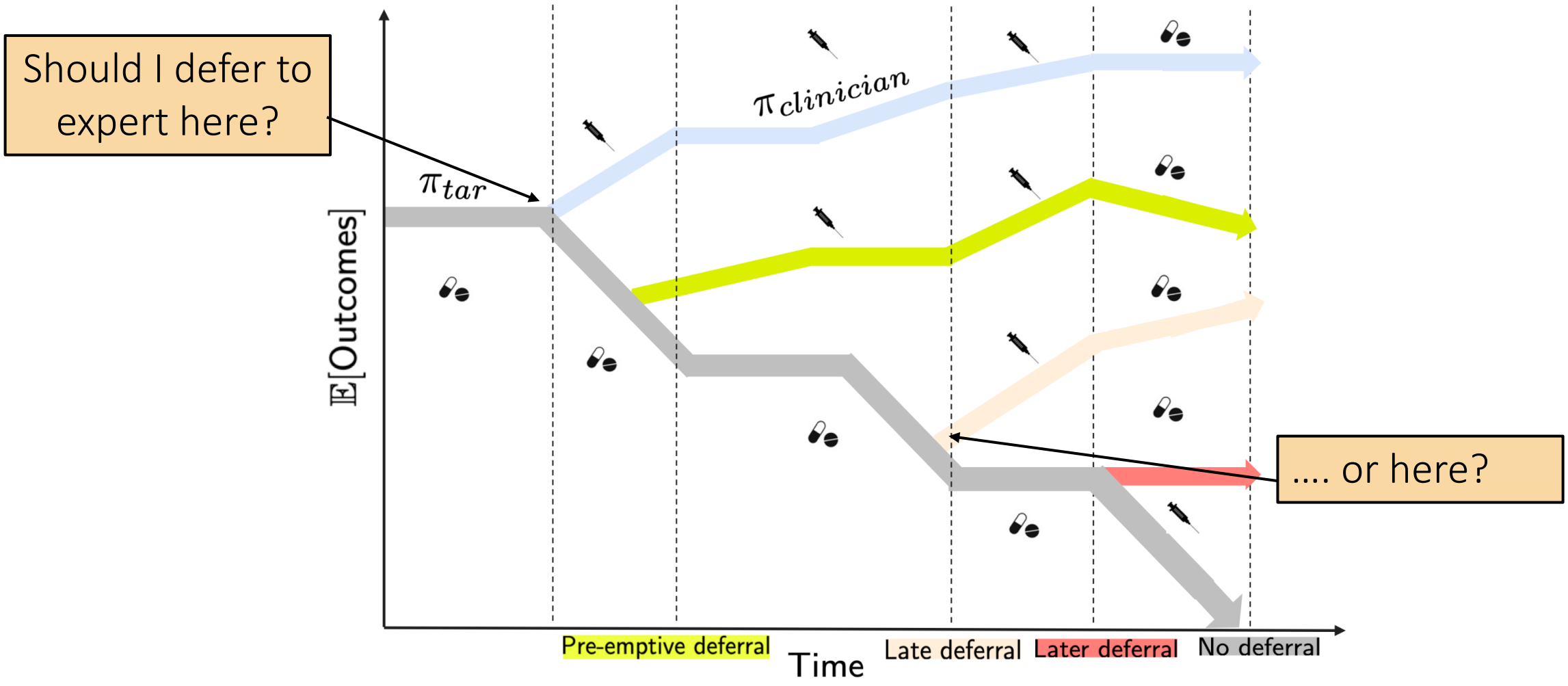


Consistent Estimators for Learning to Defer to an Expert – Mozannar et al, 2020, ICML

Learning to Defer to Humans in Sequential Settings



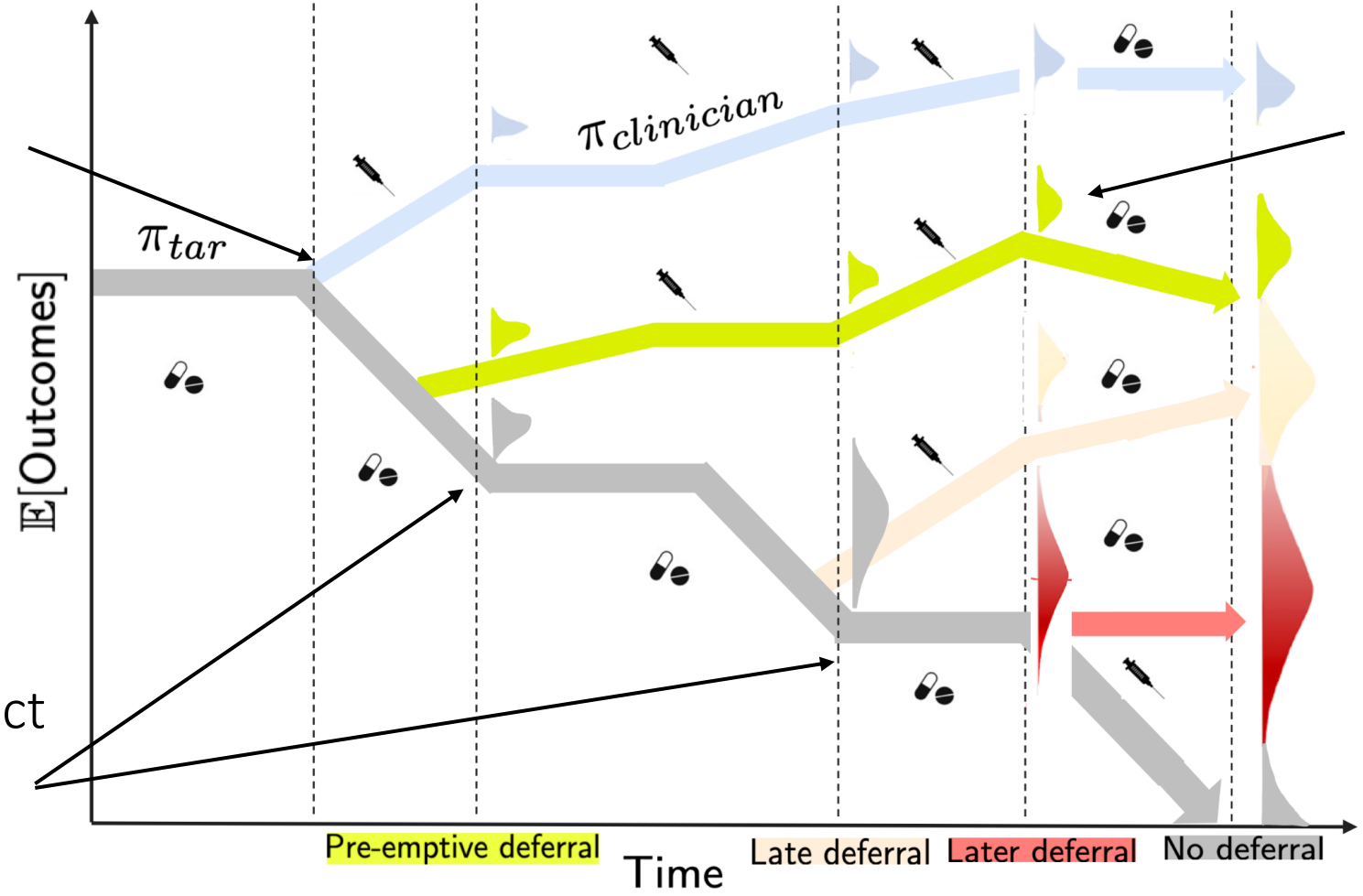
Learning to Defer to Humans in Sequential Settings



Uncertainty-aware Mixtures

Quantify where a **model cannot be trusted**

Quantify impact of **delaying deferral**

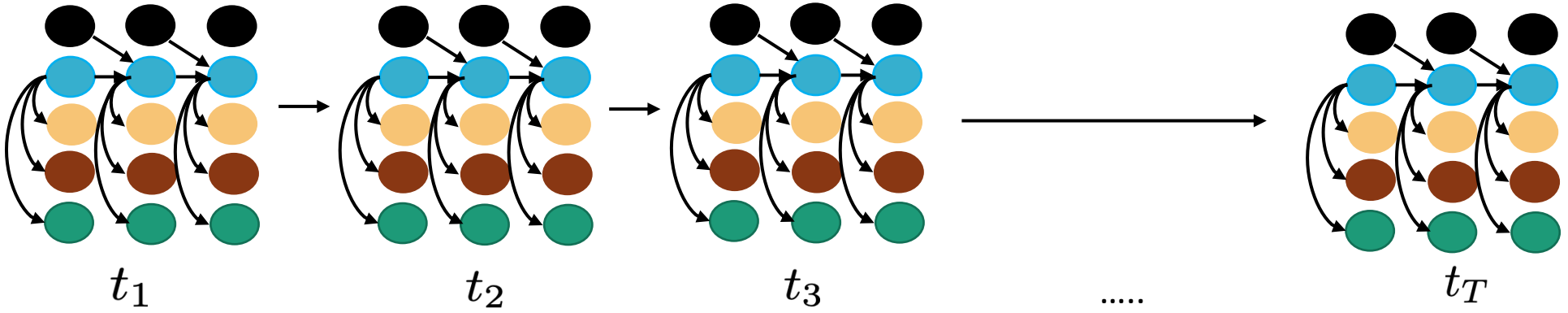


Model **changing dynamics** and **uncertainty**

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* and Doshi-Velez 2022, arXiv preprint.

Uncertainty-aware Mixtures

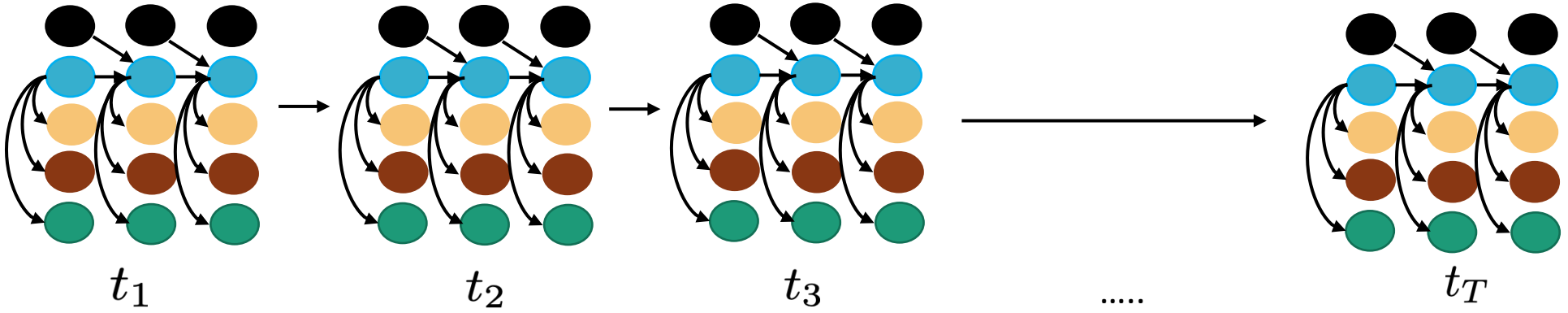
Step 1: Learn a model of **non-stationary** dynamics $\{\mathcal{M}_t\}_t$



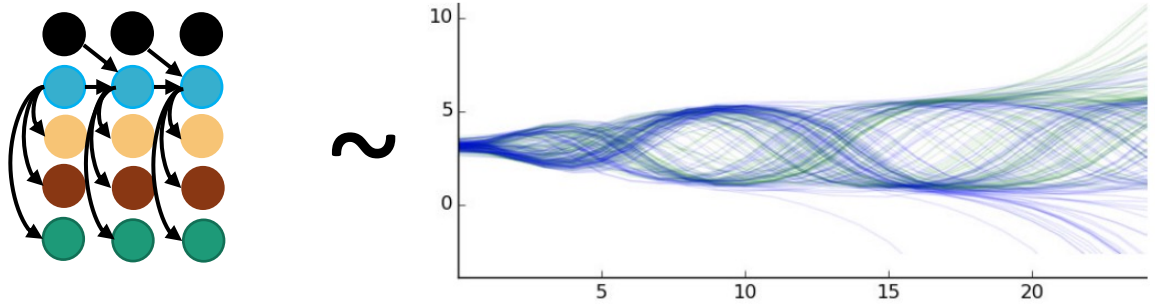
Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Uncertainty-aware Mixtures

Step 1: Learn a model of **non-stationary** dynamics $\{\mathcal{M}_t\}_t$



Step 2: Use posterior sampling to capture **uncertainty over models** $\mathcal{M}_t \sim p(\cdot|\mathcal{D})$



Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Sequential Learning to Defer (SLTD)

Step 3: Quantify the impact of **delaying deferral**

$$\mathbb{E}_{\mathcal{M}\sim p_t(\cdot|\mathcal{D})}[\mathbf{1}(V^{\mathcal{M}}_{\pi_{\text{tar}},t}(s) > V^{\mathcal{M}}_{\perp,t}(s) - c) > \tau]$$

Defer later

Defer now

Repeat this over all states and times

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Sequential Learning to Defer: Algorithm

Algorithm 1 Sequential Learning to Defer

Input: Posterior estimates $\{p_t(\cdot|\mathcal{D})\}_{t=0}^{T}$, target policy π_{tar}, behavior policy π_0.
Initialization: Deferral function $g_{\pi_{\text{tar}}}(s,t) = 0$ for all $s \in \mathcal{S}$ and $t \in \{1, 2, \cdots, T\}$.
for $t \in \{T, T-1, \cdots, 1\}$ **do**
  for $s \in \mathcal{S}$ **do**
    Compute $\{V^{\mathcal{M}}_{\pi_{\text{tar}},t}(s)\}$ and $\{V^{\mathcal{M}}_{\perp,t}(s) - c\}\, \forall \mathcal{M}$
    Update $g_{\pi_{\text{tar}}}(s,t) \leftarrow\approx \mathbb{E}_{\mathcal{M}\sim p_t(\cdot|\mathcal{D})}[\mathbf{1}(V^{\mathcal{M}}_{\pi_{\text{tar}},t}(s) > V^{\mathcal{M}}_{\perp,t}(s) - c) > \tau]$
  end for
end for
return $g_{\pi_{\text{tar}}}(s,t)$

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Decomposing the uncertainty at deferral

- The outcome we are interested in is

$$\mathbb{E}[r_T|s_{t_d},\mu_{t_d}] = \int_{s_{t_d+1}}^{s_T}\int_{a_{t_d}}^{a_T}\int_{\mu_{t_d+1}}^{\mu_T}\int_{\theta_{t_d}}^{T} r(s_T,a_T)\prod_{t'=t_d+1}^{T} p(s_{t'}|\mu_{t'})p(\mu_{t'}|\theta'_t(s_{t'},a_{t'}))\pi_{t'}(a_{t'}|s_{t'})p(\theta_{t'}|\mathcal{D})d\mathbf{s}_{t_d+1}^T d\mathbf{a}_{t_d}^T d\boldsymbol{\mu}_{t_d+1}^T d\boldsymbol{\theta}_{t_d}^T$$

Decompose

$$\underbrace{\mathrm{Var}(r_T|s_{t_d},\mathcal{D})}_{\text{Total Uncertainty}} = \underbrace{\mathbb{E}_{\mu_{t_d}\sim p(\mu_{t_d}|\mathcal{D})}[\mathrm{Var}(r_T|\mu_{t_d},s_{t_d},\mathcal{D})]}_{\text{Irreducible/ Aleatoric Uncertainty}} + \underbrace{\mathrm{Var}_{\mu_{t_d}\sim p(\mu_{t_d}|\mathcal{D})}(\mathbb{E}[r_T|\mu_{t_d},s_{t_d},\mathcal{D}])}_{\text{Epistemic/Modeling Uncertainty}}$$

- First term averages over the variance due to μ_{t_d} and captures **aleatoric uncertainty** at t_d
- Second term captures variance conditioned on knowledge of the model or **epistemic uncertainty** at deferral time.

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

What can these uncertainties tell us?

- **High Epistemic Uncertainty:**
  - Uncertainty of model's prediction is high
  - Possible improvement if additional data in this region could be collected to improve the reducible sources of uncertainty.

- **High Aleatoric Uncertainty:**
  - There is high variability in the patient's dynamics that may need to be managed with careful interventions or is otherwise not manageable.
  - Aleatoric uncertainty is important for safety

- Based on uncertainty, expert may choose to deviate from practice and/or get second opinions.

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Toy Demonstration

- Given 8 states: s6 has reward -5; while others have reward +1
- Each sample starts at s0 and progresses to s7
- Two actions: a0 decreases chances of landing in s6; a1 increases this
- Target policy π_{tar} increases the chances of landing in s6 from {s2, s3, s4} at times 3-8.
- Non-stationarity increases over time, while rewards are positive in {s2, s3, s4} so that poor outcomes only manifest in the future.
- A pre-emptive deferral policy should defer between times 3 and 8.

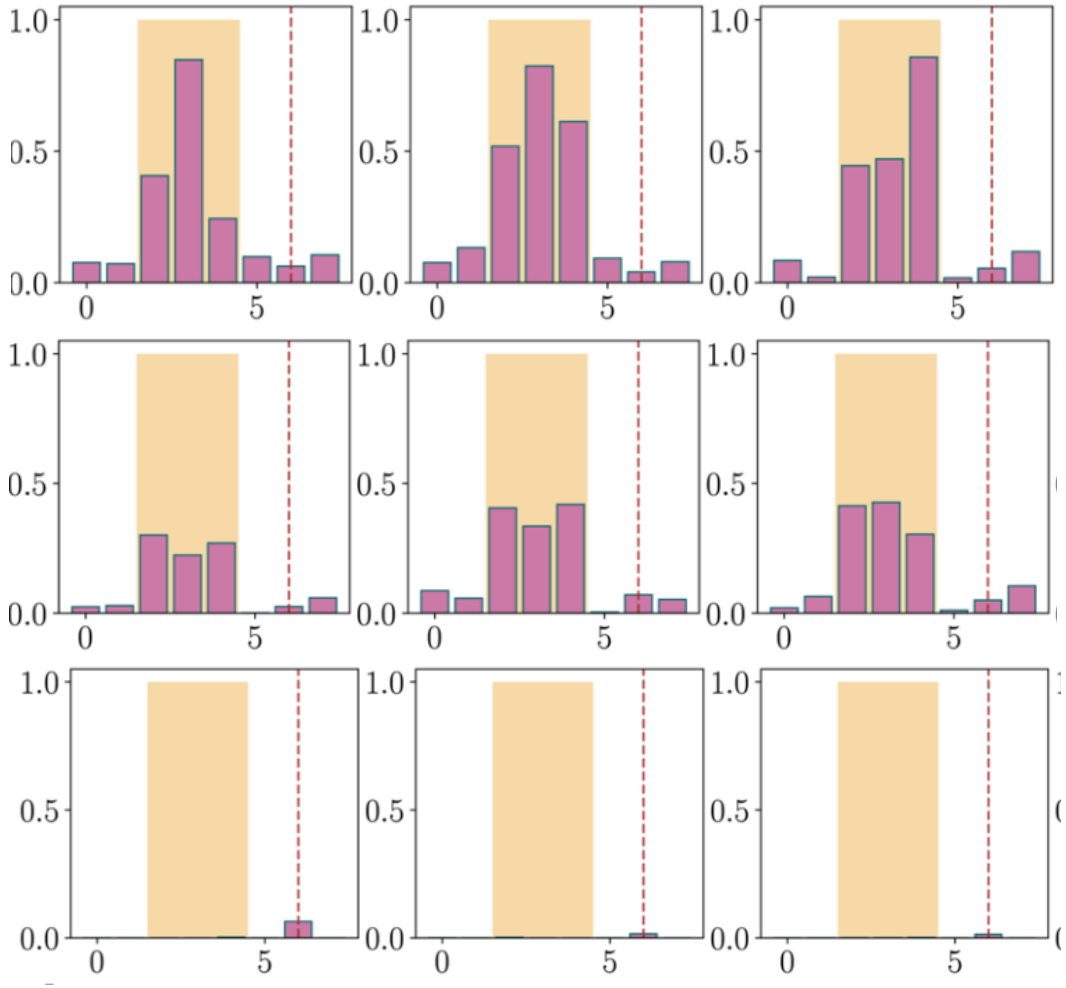
Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

SLTD learns a *pre-emptive* deferral policy

SLTD

Stationary

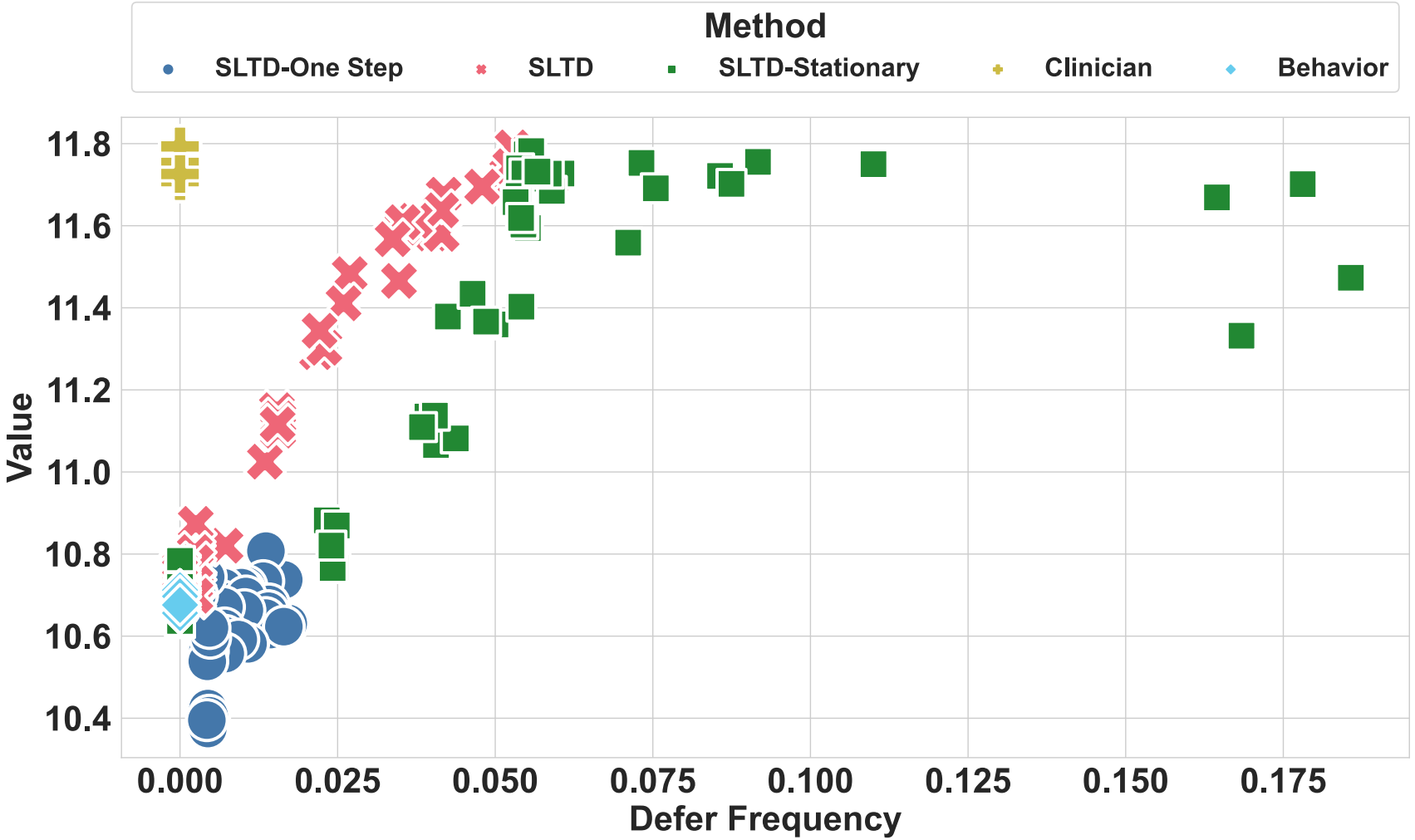
One-step deferral



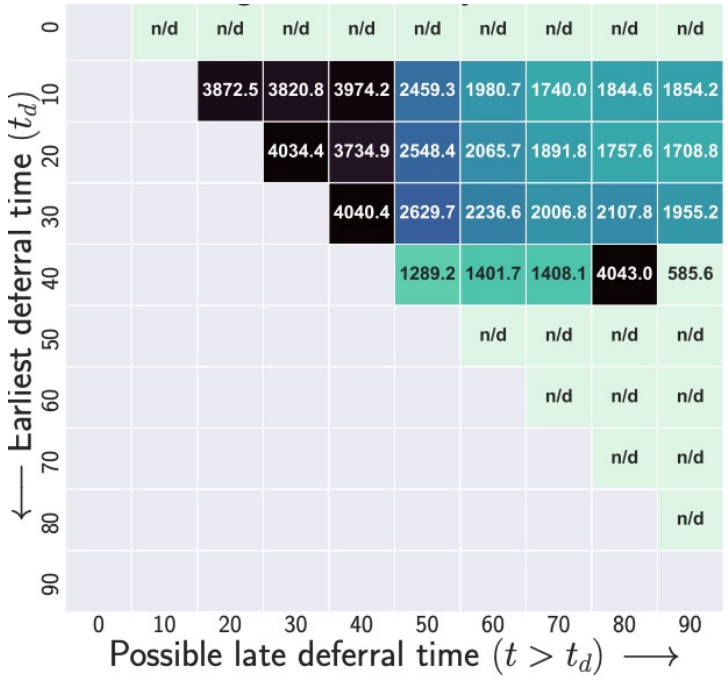
Time t

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

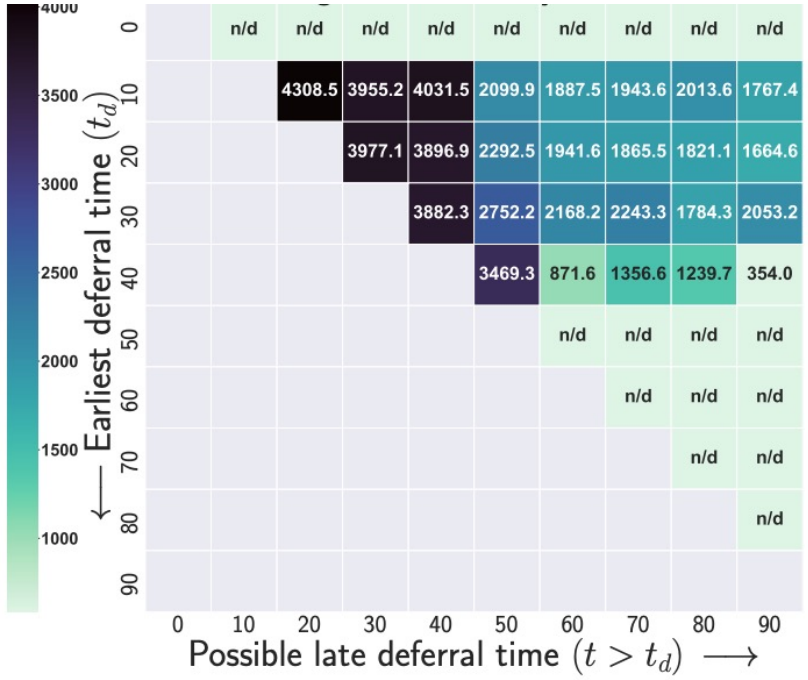
SLTD produces better outcomes with *fewer deferrals*



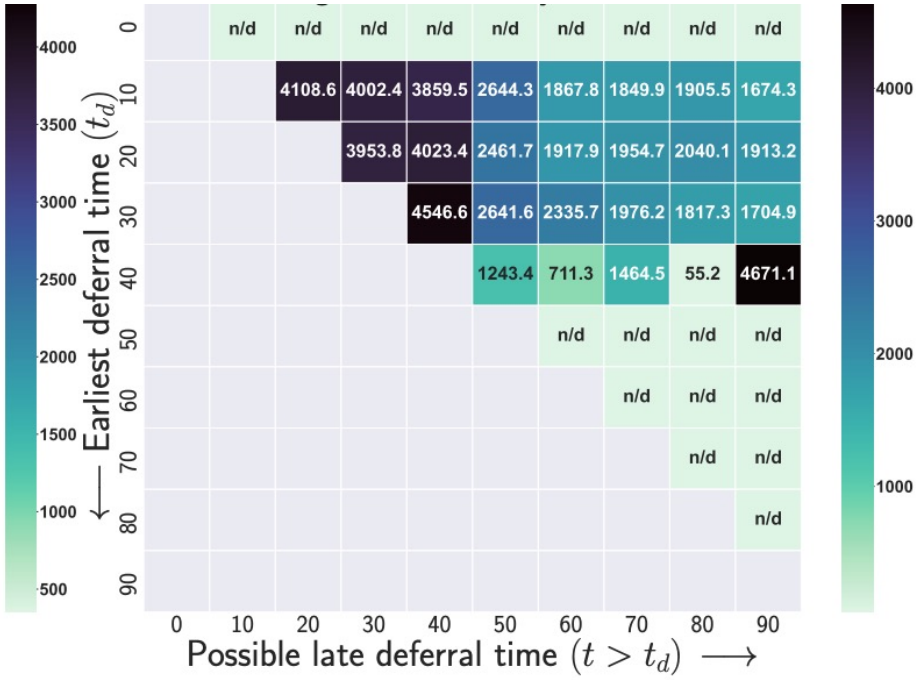
Deferring early potentially *reduces overall uncertainty*



SLTD



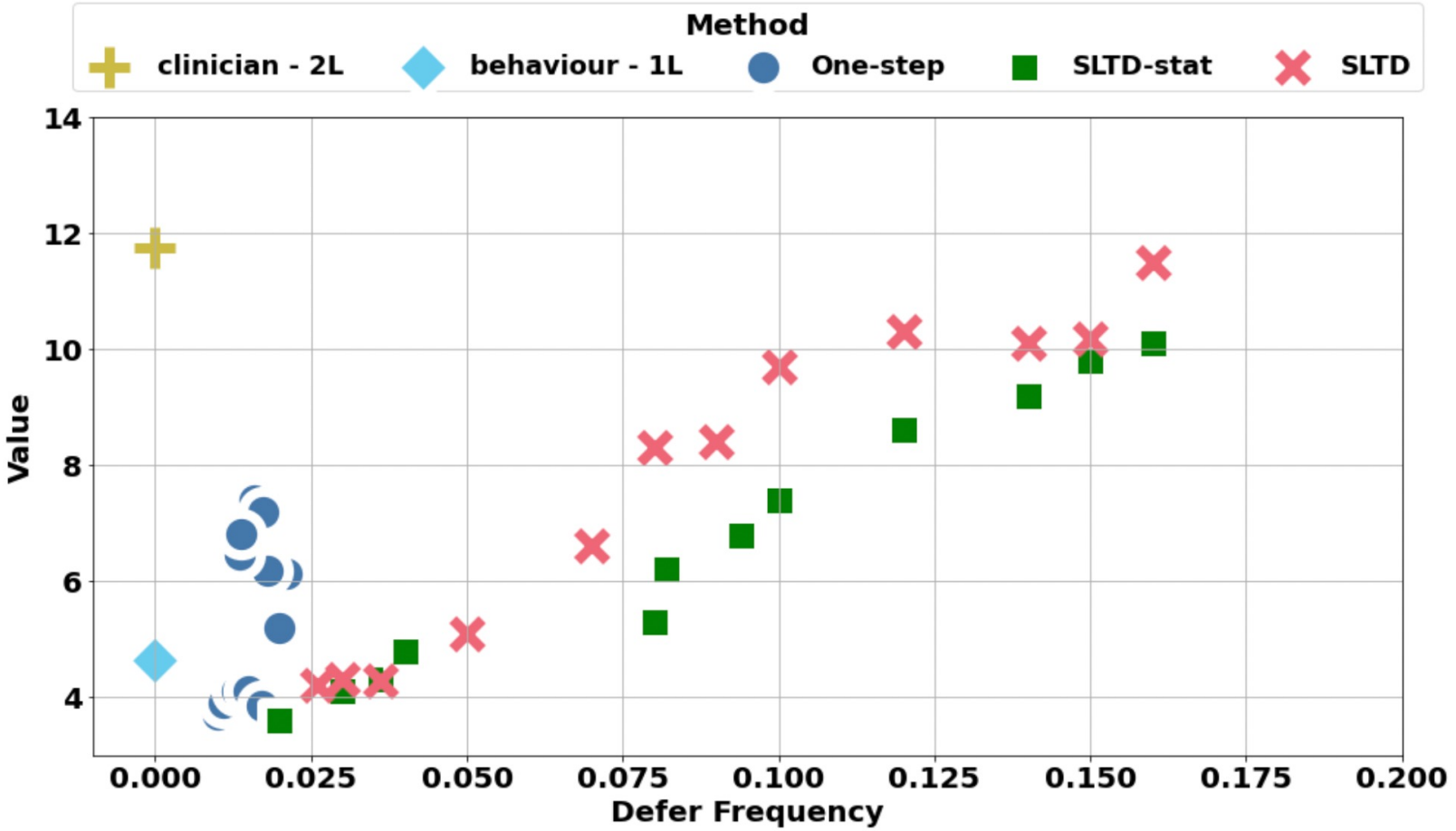
Stationary



One-step

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

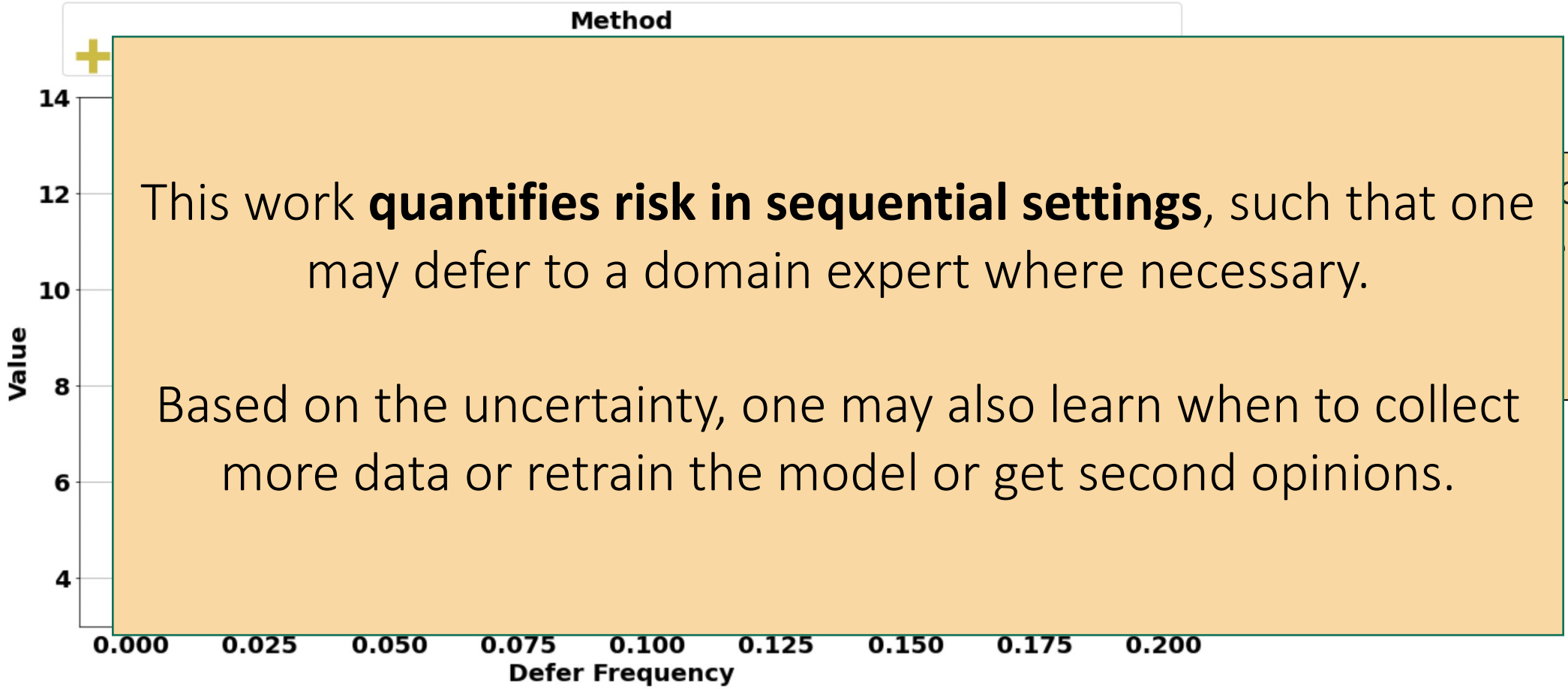
Similar performance for HIV Therapy Management



We get close to clinician policy with +/- 1/8 deferrals i.e. can potentially save 7/8 visits.

Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Similar performance for HIV Therapy Management



This work **quantifies risk in sequential settings**, such that one may defer to a domain expert where necessary.

Based on the uncertainty, one may also learn when to collect more data or retrain the model or get second opinions.

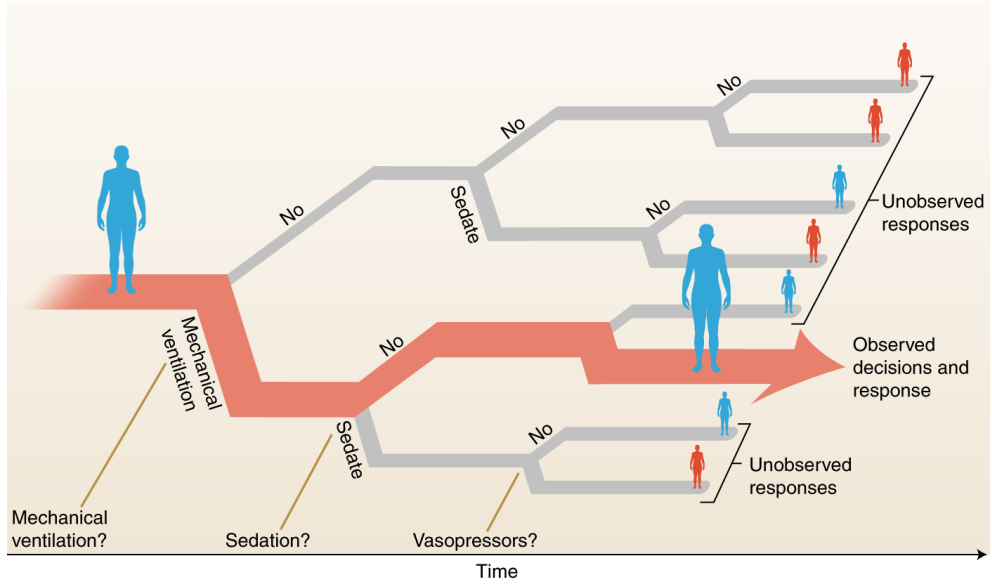
clinician
rrals i.e.
ave 7/8

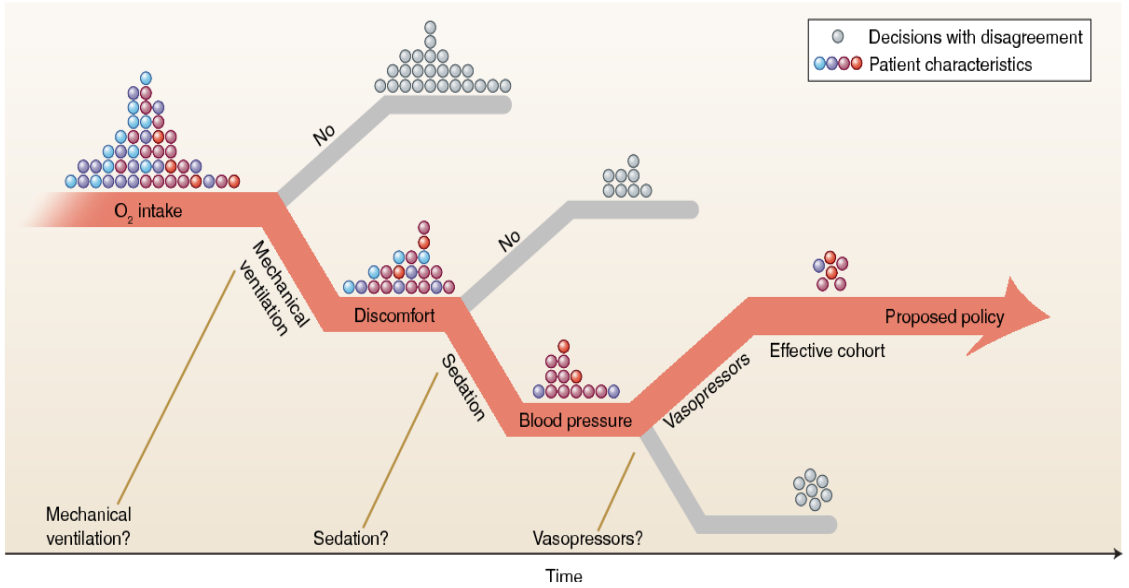
Pre-emptive learning to defer for sequential decision-making in uncertainty – Joshi*, Parbhoo* et al 2021, arXiv preprint.

Challenges to using RL for healthcare in practice

Learning suitable treatment policies from observational data is **offline and off-policy**

1. Inability to explore
2. Small data (relatively) → shrinking data support as best strategies are discovered



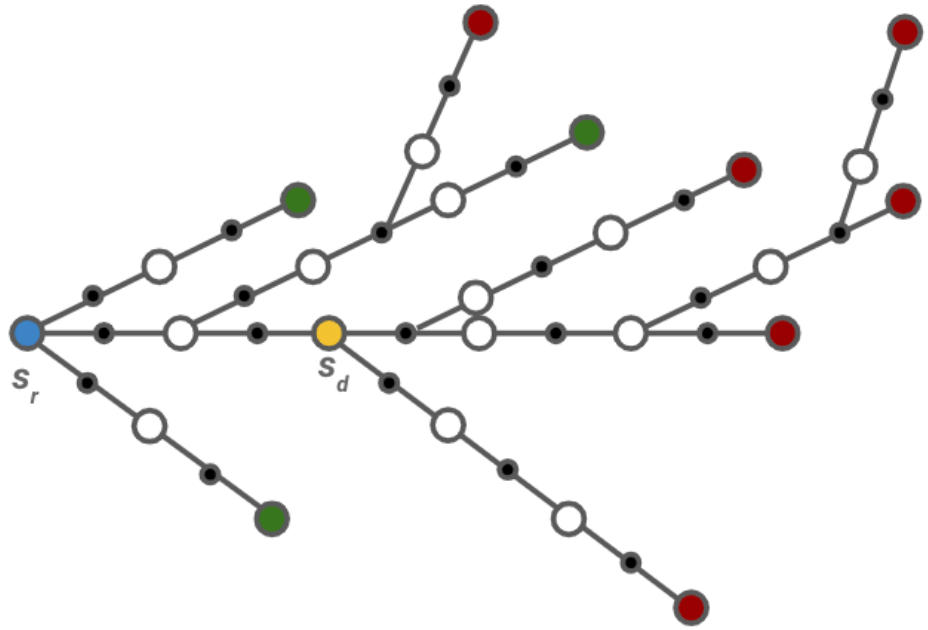


These two limitations severely complicate the development of actionable RL

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

An alternative offline RL paradigm

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



Undesired Terminal State (e.g., mortality)

Desired Terminal State (e.g., recovery)

Dead-end: all trajectories from $\mathbf{s_d}$ reach an an undesired terminal state with probability 1.

Rescue: from $\mathbf{s_r}$ a desired end state is reachable with probability 1

Can we identify all dead-ends and treatments that lead to them?

Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

Treatment Security

To avoid selecting treatments *a* that lead to dead-ends, constrain the policy π

Def: ***Treatment Security***,

$$P_D(s,a) + F_D(s,a) \geq \lambda \implies \pi(s,a) \leq 1 - \lambda$$

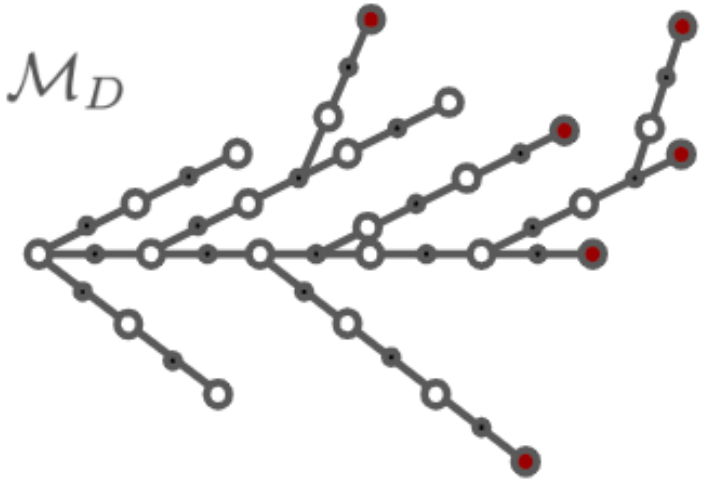
But

- Inferring max λ for all possible (s, a) pairs is intractable
- Accurate estimates of P_D and F_D requires knowledge of all dead-ends and transition probabilities a priori.

Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

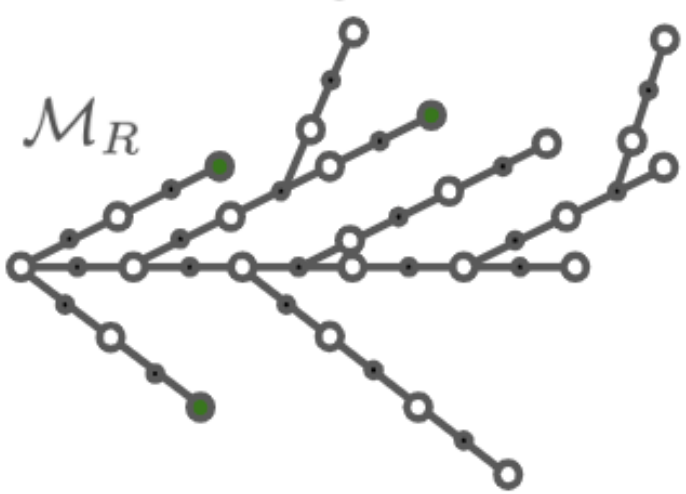
Treatment Security

Separate the MDP into two *decoupled* processes



Zero-out all positive outcomes
No discounting ($\gamma_D = 1$)

$Q_D^* \in [-1, 0]$



Zero-out all negative outcomes
No discounting ($\gamma_R = 1$)

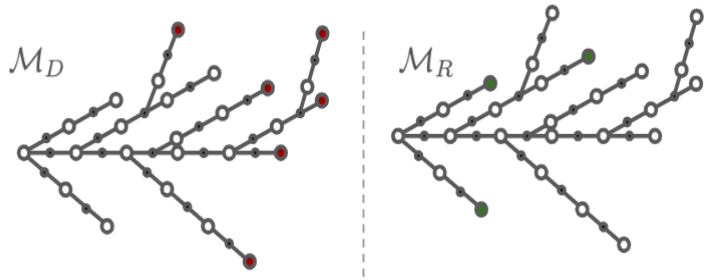
$Q_R^* \in [0, 1]$

Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

A Framework for Treatment Security: Dead-End Discovery (DeD)

As a consequence of $\mathcal{M}_D$

$$-Q_D^*(s,a) = P_D(s,a) + F_D(s,a)$$



This means that $-Q_D^*$ carries special physical meaning

- It is the minimum probability of an *undesired terminal outcome*
- $1 + Q_D^*(s,a)$ is the maximum probability of a *desired terminal outcome*

Using this interpretation of $-Q_D^*$ the treatment security condition is satisfied if

$$P_D(s,a) + F_D(s,a) \geq \lambda \Rightarrow \pi(s,a) \leq 1 - \lambda \quad \Rightarrow \quad \pi(s,a) \leq 1 + Q_D^*$$

Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

A Framework for Treatment Security: Dead-End Discovery (DeD)

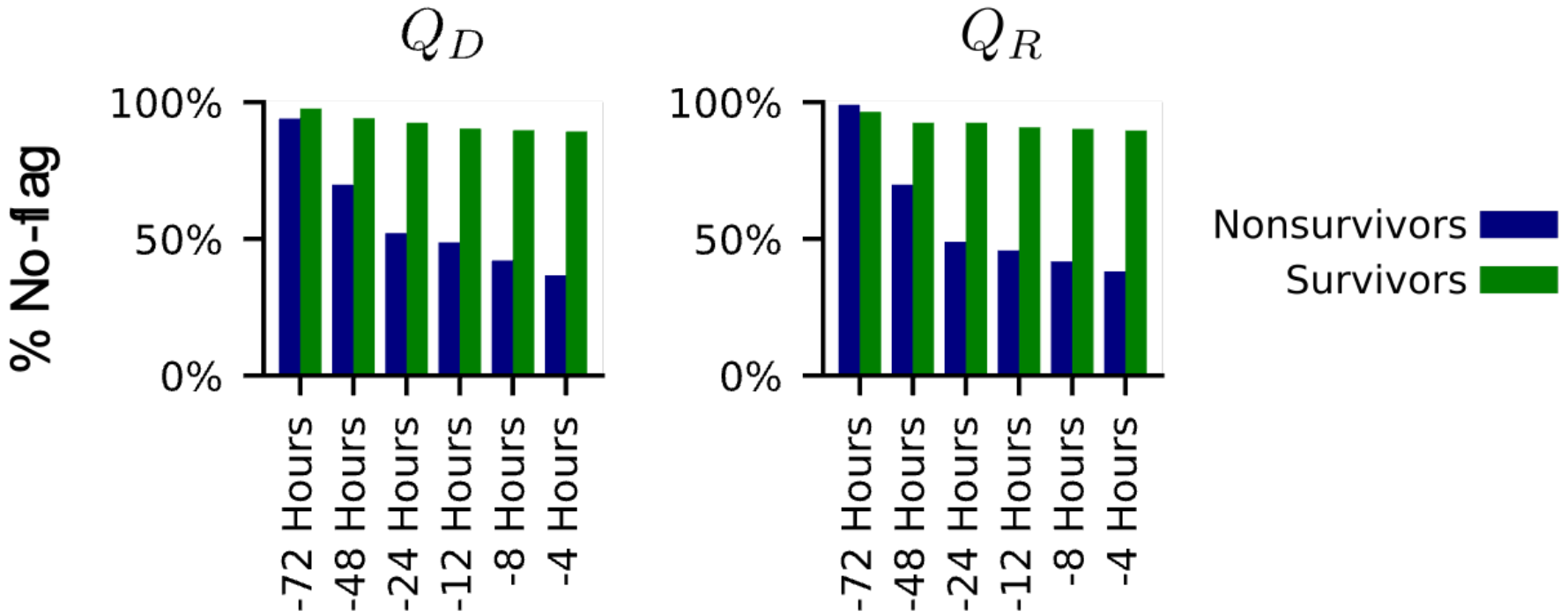
It can be shown:

- $P_D(s,a) + F_D(s,a) = 1$ if and only if $Q^*_D(s,a) = -1$
- $P_R(s,a) + F_R(s,a) = 1$ if and only if $Q^*_R(s,a) = 1$
- If $\pi(s,a) \leq 1 + Q^*_D$ and $P_D(s,a) + F_D(s,a) \geq \lambda$ then $\pi(s,a) \leq 1 - \lambda$ for all λ
- If $\pi(s,a) \geq Q^*_R(s,a)$ and $P_R(s,a) + F_R(s,a) \geq \lambda$ then $\pi(s,a) \geq \lambda$ for all λ
- There exists a threshold $\delta_D \in (-1, 0)$ that separates dead-ends through $Q^*_D(s,a) \geq \delta_D$
- There exists a threshold $\delta_R \in (0, 1)$ that separates rescue states through $Q^*_R(s,a) \geq \delta_R$

Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

DeD framework can help identify dead-ends in septic patients

Clear distinction between patients who survived and those who did not in terms of DeD providing specified risk several hours prior to terminal outcome

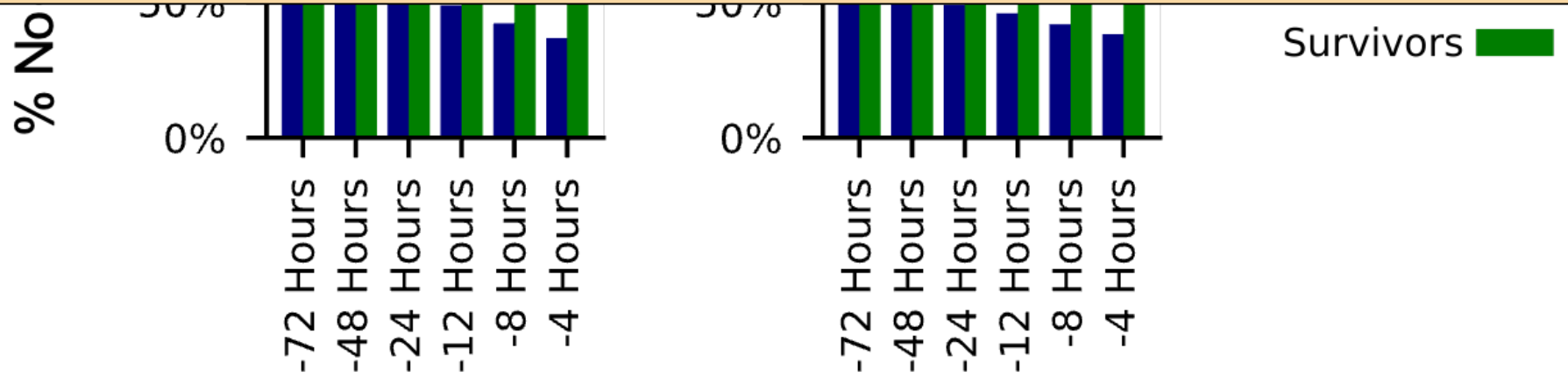


Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

DeD framework can help identify dead-ends in septic patients

Clear distinction between patients who survived and those who did not in terms of DeD

… but there are major limitations. Crucially, DeD uses a naïve definition of risk. We need to account for the **full distribution of value functions**.



Fatemi, M., Killian, T.W., Subramanian, J. and Ghassemi, M., Medical Dead-ends and Learning to Identify High-risk States and Treatments. *Advances in Neural Information Processing Systems*, *34*. (2021)

Potential Sources of Uncertainty in RL

Dynamics

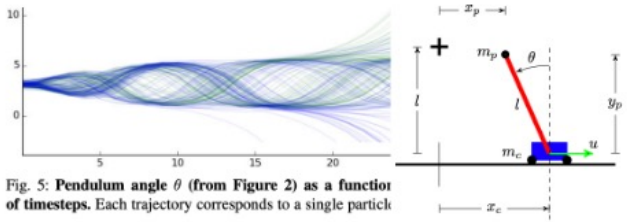


Fig. 5: **Pendulum angle θ (from Figure 2) as a functio of timesteps.** Each trajectory corresponds to a single particl

From Gal, et al; [ICML 2016]

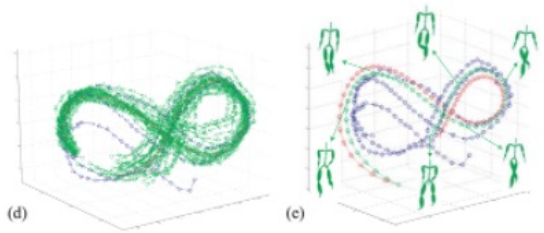
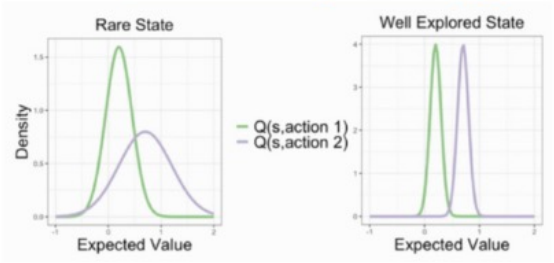


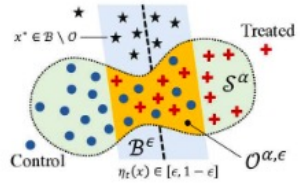
Figure 2: Models learned from a walking sequence of 2.5 gait cycles. The latent positions

From Wang, et al; [NeurIPS 2005]

Actions (Policy)

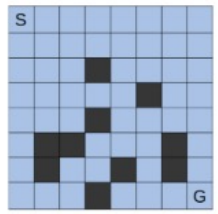


From Sonabend, et al; [NeurIPS 2020]

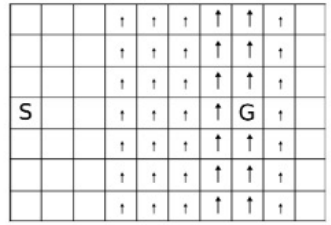


From Oberst, et al; [AISTATS 2020]

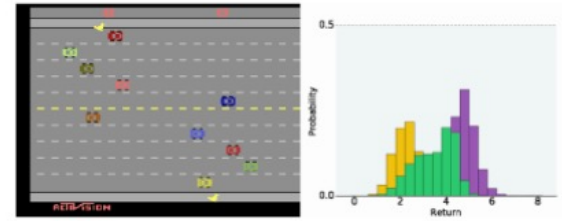
Outcomes (Reward)



Frozen Lake



Windy Grid World



From Bellemare, et al; [ICML 2017]

Wang, Jack, Aaron Hertzmann, and David J. Fleet. "Gaussian process dynamical models." *Advances in neural information processing systems* 18 (2005).

Gal, Yarin, Rowan McAllister, and Carl Edward Rasmussen. "Improving PILCO with Bayesian neural network dynamics models." *Data-Efficient Machine Learning workshop, ICML*. Vol. 4. No. 34. 2016.

Killian, T. W., Daulton, S., Konidaris, G., & Doshi-Velez, F. Robust and efficient transfer learning with hidden parameter markov decision processes. *Advances in neural information processing systems*, *30* (2017).

Sonabend, A., Lu, J., Celi, L. A., Cai, T., & Szolovits, P. Expert-supervised reinforcement learning for offline policy learning and evaluation. *Advances in Neural Information Processing Systems*, *33*, (2020)

Oberst, M., Johansson, F., Wei, D., Gao, T., Brat, G., Sontag, D., & Varshney, K. "Characterization of overlap in observational studies". In *International Conference on Artificial Intelligence and Statistics* (2020)

Matsushima, T., Kondo, N., Iwasawa, Y., Nasuno, K., & Matsuo, Y. Modeling Task Uncertainty for Safe Meta-Imitation Learning. *Frontiers in Robotics and AI*, (2020)

Dabney, W., Ostrovski, G., Silver, D., & Munos, R. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning* (2018)

Stanko, S., & Macek, K. Risk-averse Distributional Reinforcement Learning: A CVaR Optimization Approach. In *IJCCI* (2019)

Bellemare, M. G., Dabney, W., & Munos, R. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning* (2017)

Image Credit: T.W Killian, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

A Short History of Distributional RL

Traditionally, RL utilizes a point estimate for the expected return from any state-action pair.
Bellemare, et al (2017)* introduced an alternative, representing the full return distribution

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

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

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

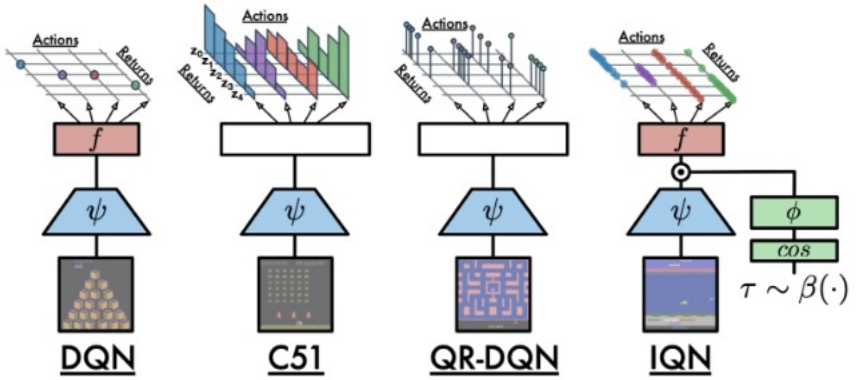
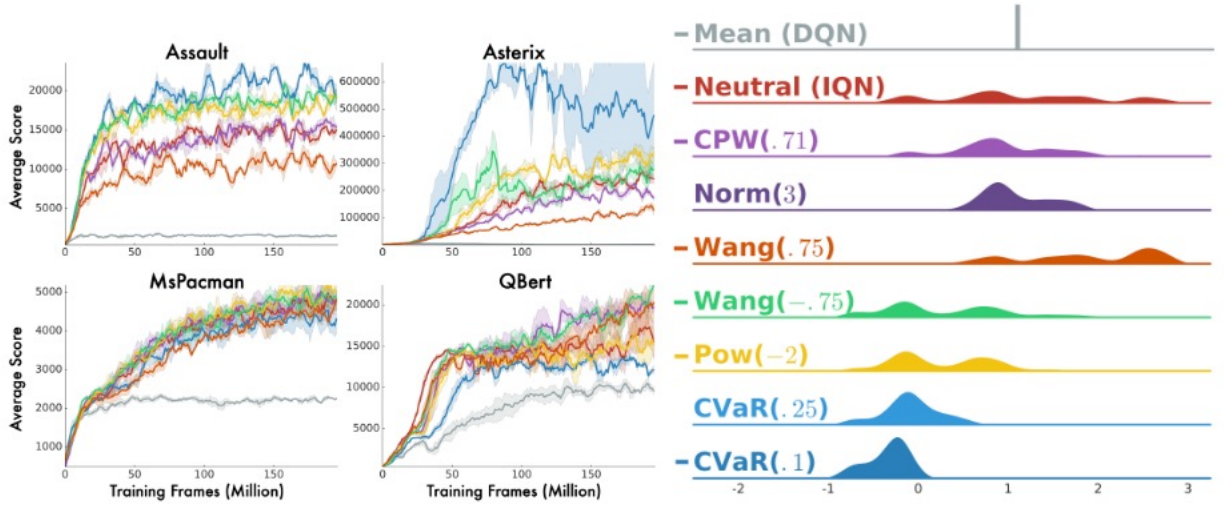


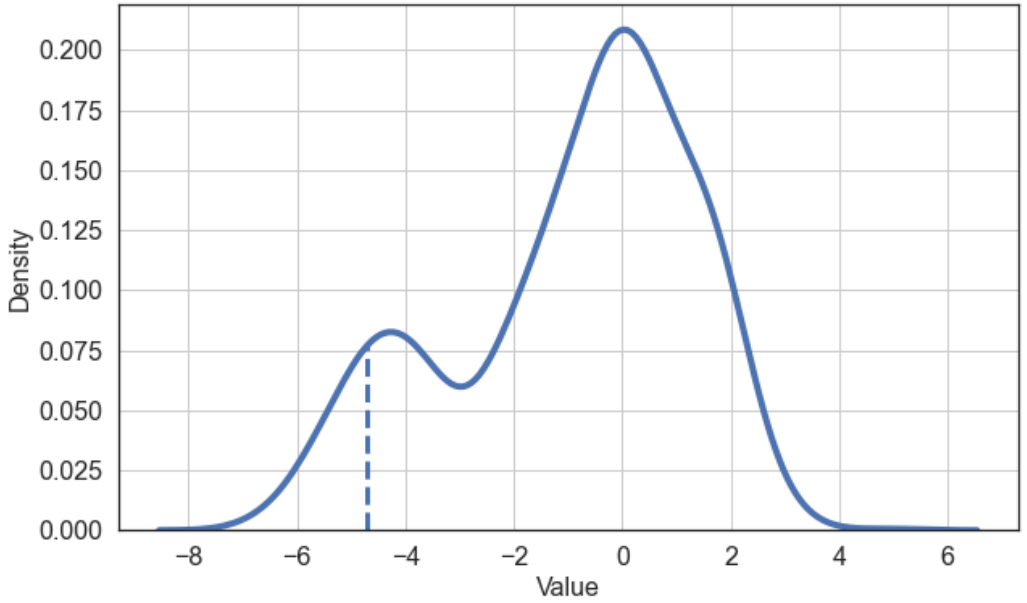
Figure 1. Network architectures for DQN and recent distributional RL algorithms.

From Dabney, et al; [ICML 2018]



From Dabney, et al; [ICML 2018]

Risk estimation with the full distribution



Value-at-risk (VaR):

$$\mathrm{VaR}_\alpha(Z) = \min \{z \mid \alpha \leq F(z)\}$$

Conditional value-at-risk (CVaR):

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

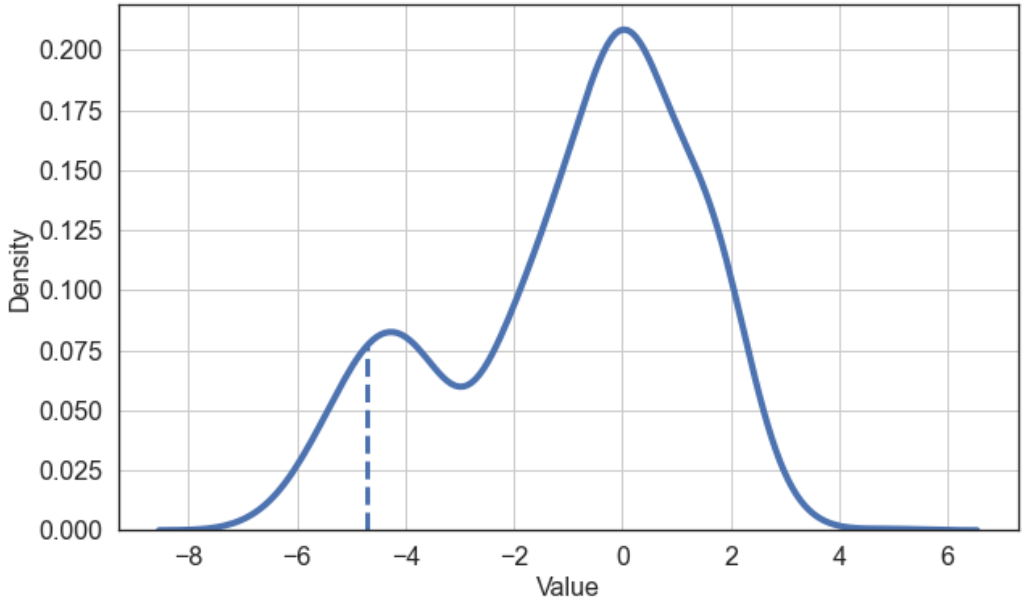
Dual Formulation of CVaR

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

For some state *s* and action *a,* we can assess the estimated value distribution $Z^\pi(s,a)$ and infer the worst-case outcome that occurs with some pre-defined confidence level η (e.g. value-at-risk).

T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

Risk estimation with the full distribution



Value-at-risk (VaR):

$$\mathrm{VaR}_\alpha(Z) = \min\{z \mid \alpha \leq F(z)\}$$

Conditional value-at-risk (CVaR):

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

Dual Formulation of CVaR

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

For some state *s* and action *a,* we can assess the estimated value distribution $Z^\pi(s,a)$ and infer the worst-case outcome that occurs with some pre-defined confidence level η (e.g. value-at-risk).

Then using the distribution that falls below η, we calculate the expected worst case outcome (e.g.CVaR). **This value assesses the risk of possible actions *a***

T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

Risk-Estimation for Medical Dead-Ends
Distributional Treatment Security Condition (DistDeD)

With risk aversion parameter **α** , we see that CVaR lower bounds the expected return:

$$\mathrm{CVaR}_\alpha(Z) \leq \mathbb{E}[Z]$$

Extending this for DeD, we see:

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

We can provide a more robust and risk-averse treatment security condition:

$$\pi(s,a) \leq 1 + \mathrm{CVaR}_\alpha(Z^*(s,a))$$

CVaR helps **lower bound** our choice of which actions to avoid and provides a mathematical means to assess risk.

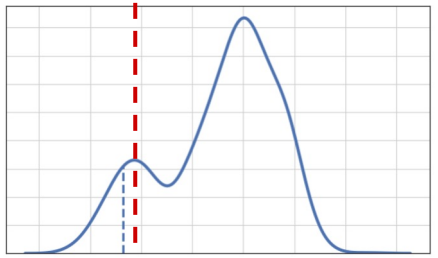
T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

Application of DistDeD to AHE and Septic Patients

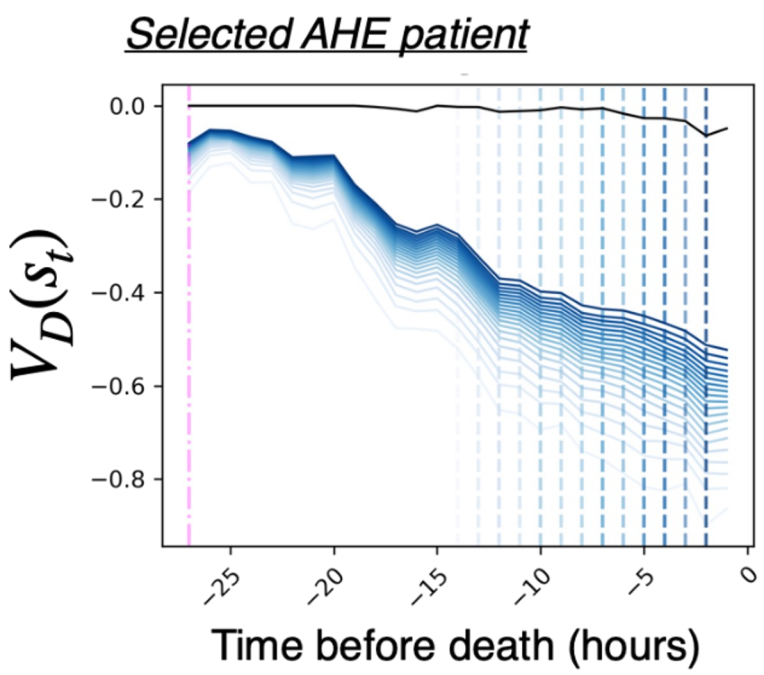
Sepsis is a life-threatening organ dysfunction caused by infection

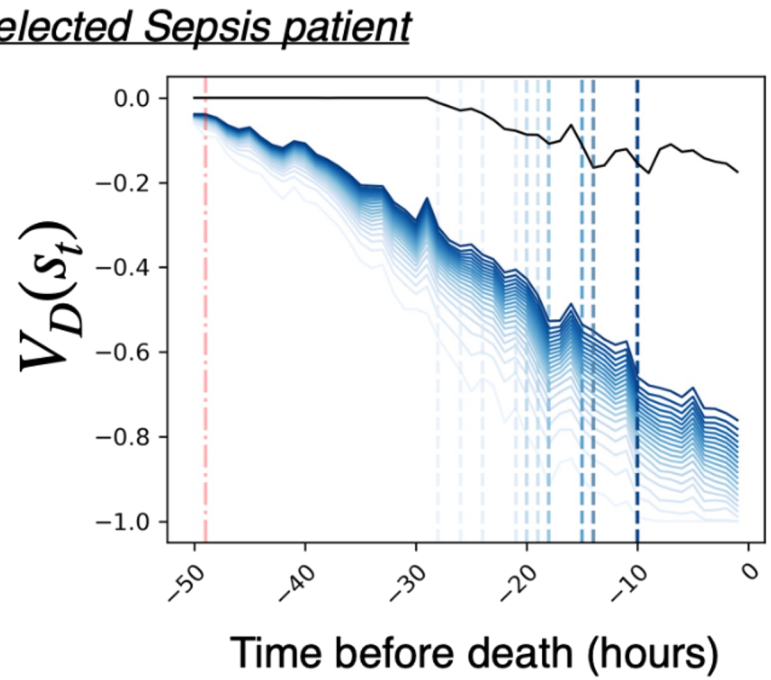
- Using MIMIC-III we extract a cohort of 19,611 patient trajectories
  (72 hours, centered around presumed sepsis onset)
- Patient observations are made up of 47 features, aggregated in 4h windows, comprised of vital signs, lab measurements and demographics
- 25 available treatments
  (combination of 5 volumetric bins for each of vasopressors and IV fluid)
- Terminal outcomes: patient survival (+1) or death (-1)

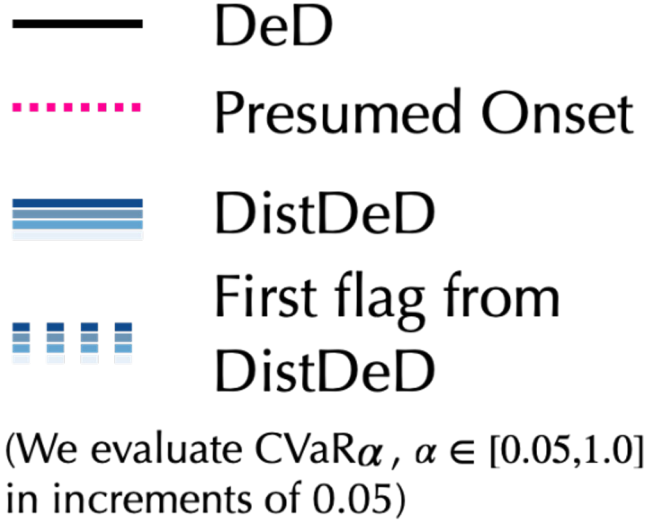
DistDeD provides a *tunable measure of risk*



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



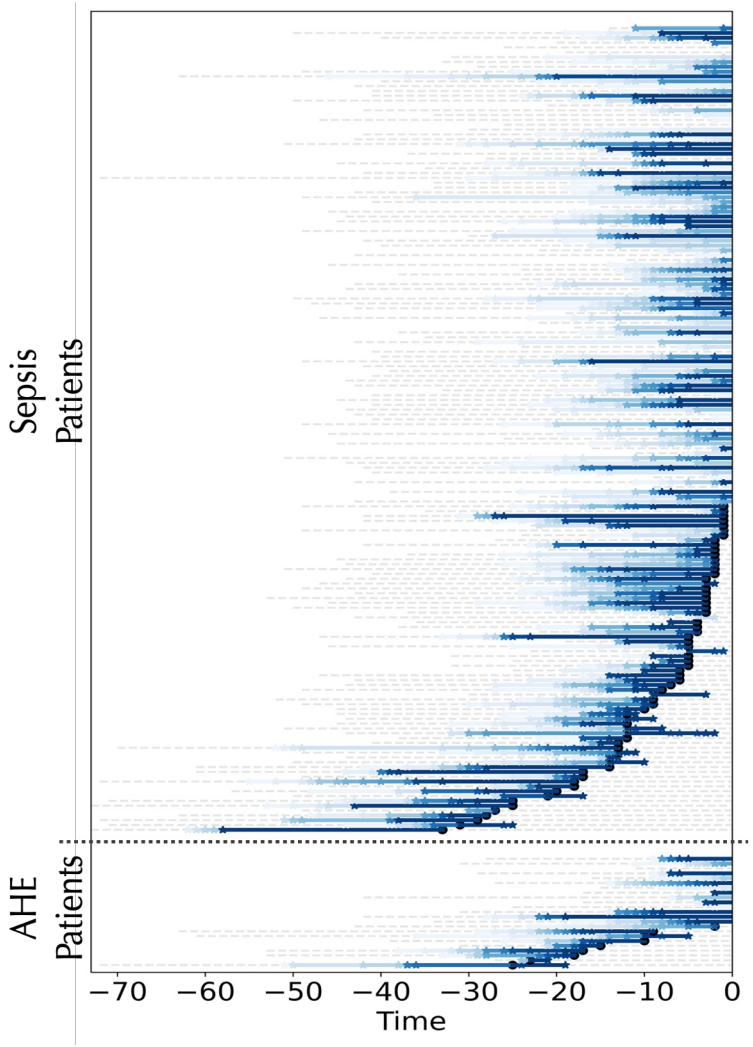


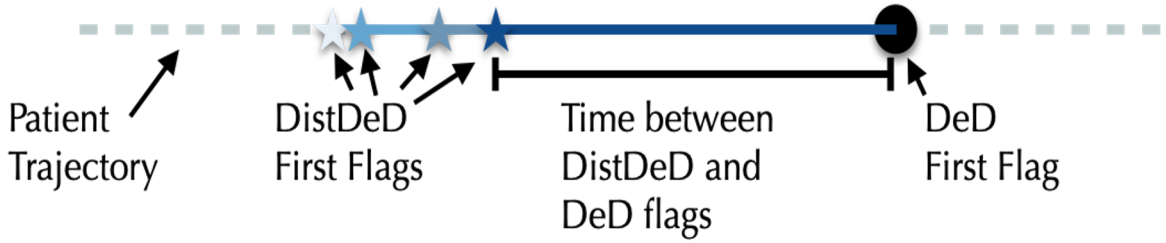


(We evaluate CVaRα, $\alpha \in [0.05,1.0]$ in increments of 0.05)

T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

DistDeD provides an *earlier indication of risk*



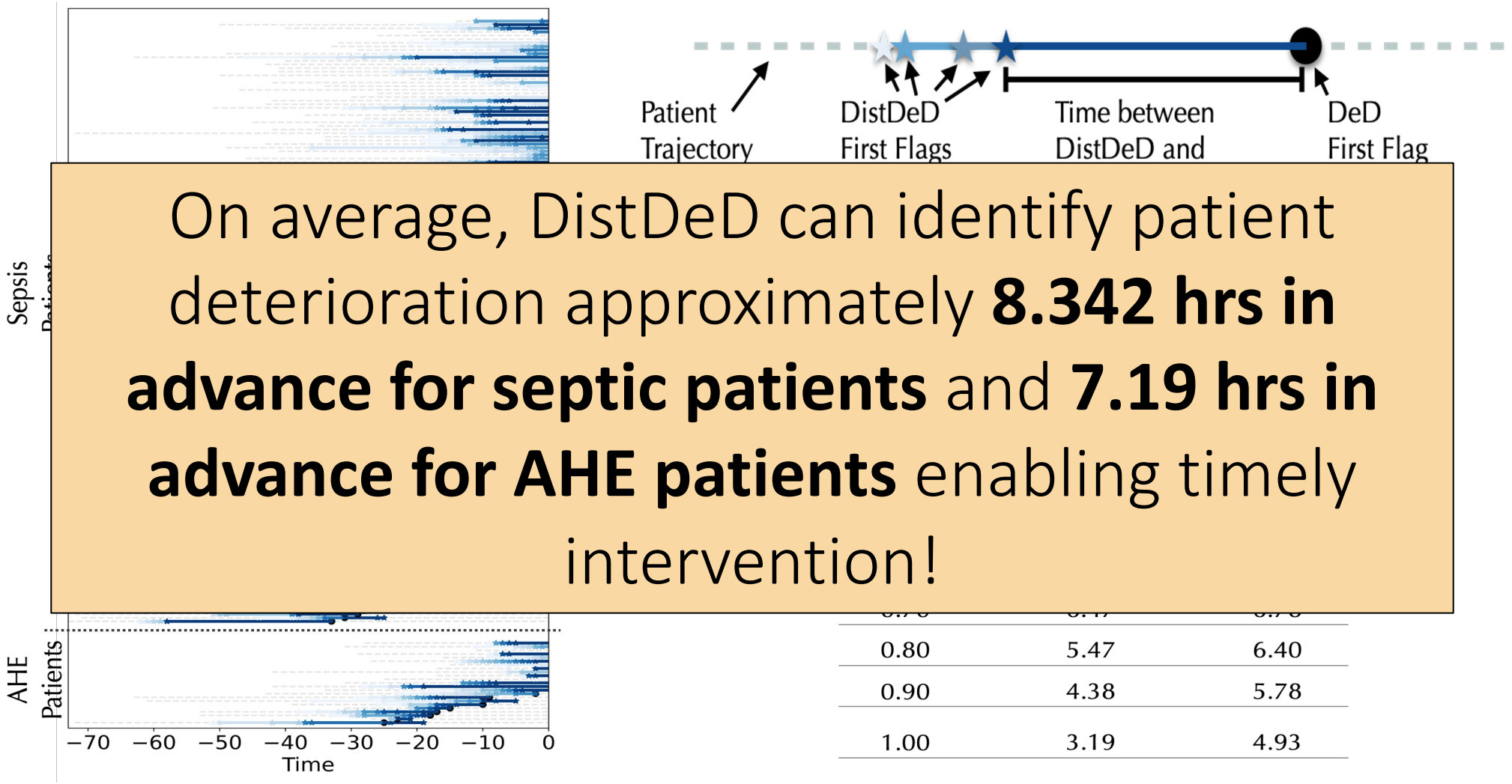


DistDeD Avg. Hours Ahead (by CVaRα setting)

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, RLDM 2022**

DistDeD provides an *earlier indication of risk*



On average, DistDeD can identify patient deterioration approximately **8.342 hrs in advance for septic patients** and **7.19 hrs in advance for AHE patients** enabling timely intervention!

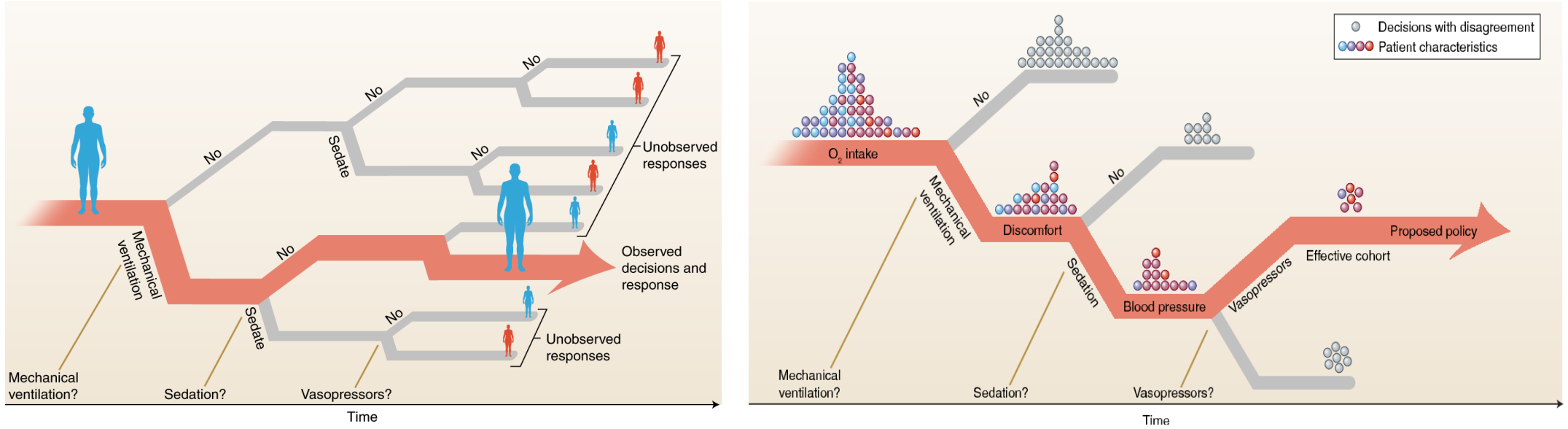
0.80	5.47	6.40
0.90	4.38	5.78
1.00	3.19	4.93

T.W Killian, S Parbhoo and M. Ghassemi,, **Risk-sensitive and robust Dead-end Identification in Safety-Critical Offline RL, RLDM 2022**

Challenges to using RL for healthcare in practice

Learning suitable treatment policies from observational data is **offline and off-policy**

1. Inability to explore
2. Small data (relatively) → shrinking data support as best strategies are discovered



These two limitations severely complicate the development of actionable RL

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

SLTD and DeD cannot always be used in place of evaluation

- **DeD Framework doesn't fully support intermediate outcomes**
  - requires careful definition of what to avoid
- **Both methods do not account for sources of confounding and bias**
  - Value functions are prone to *overestimation*; not assessed for sensitivity to patient demographics

SLTD and DeD cannot always be used in place of evaluation

- **DeD Framework doesn't fully support intermediate outcomes**
  - requires careful definition of what to avoid
- **Both methods do not account for sources of confounding and bias**
  - Value functions are prone to *overestimation*; not assessed for sensitivity to patient demographics

What are the alternatives in these settings?

Robustness Checks for OPE

Are the results reproducible across different sites or rewards?

E.g. Use two cohorts and 3 different OPE methods

EuResist

	DR	IS	WIS
Random	-2.31 ± 1.42	-3.48 ± 1.36	-2.80 ± 1.27
Short-term kernel	2.17 ± 1.4	2.18 ± 1.20	2.16 ± 1.71
Long-term kernel	9.47 ± 1.70	5.72 ± 1.81	6.97 ± 1.29
POMDP	6.04 ± 2.18	4.15 ± 2.28	6.67 ± 1.74
Mixture-of-experts	$\mathbf{11.83 \pm 1.26}$	$\mathbf{12.50 \pm 1.19}$	$\mathbf{11.07 \pm 1.21}$

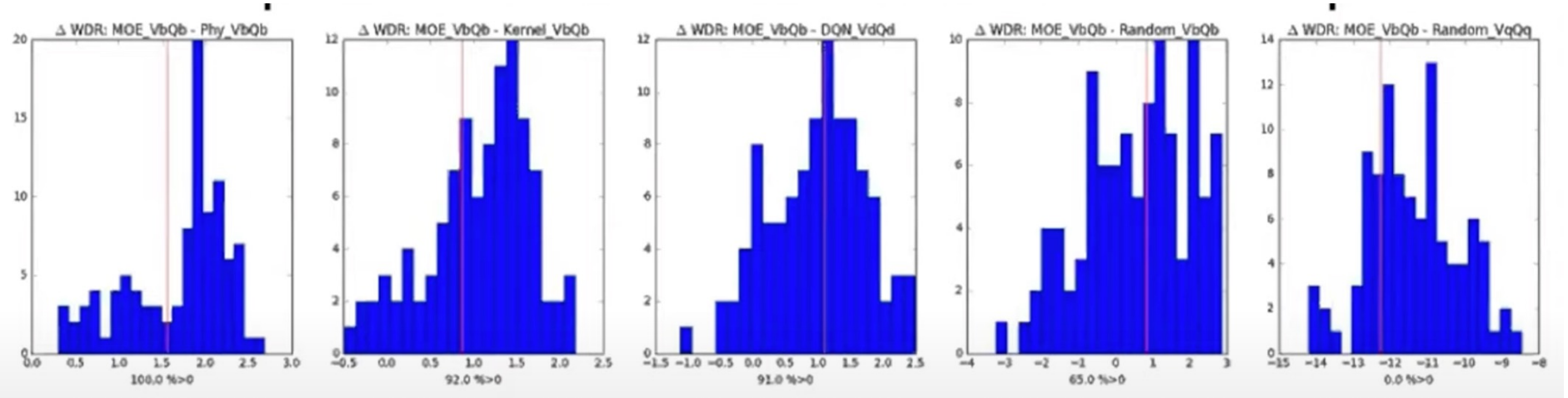
Swiss HIV Cohort Study

	DR	IS	WIS
Random	-6.33 ± 3.47	-5.57 ± 2.17	-6.18 ± 3.24
ST Kernel	1.64 ± 1.86	2.03 ± 1.81	2.17 ± 1.74
LT Kernel	9.67 ± 1.49	7.38 ± 1.72	7.64 ± 1.92
POMDP	5.46 ± 2.05	6.72 ± 2.88	7.76 ± 2.10
Mixture-of-experts	$\mathbf{10.73 \pm 1.02}$	$\mathbf{13.59 \pm 1.57}$	$\mathbf{11.83 \pm 1.31}$

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

Sensitivity Analysis for Parameter Choices

How does evaluation vary with different choices of control variates?

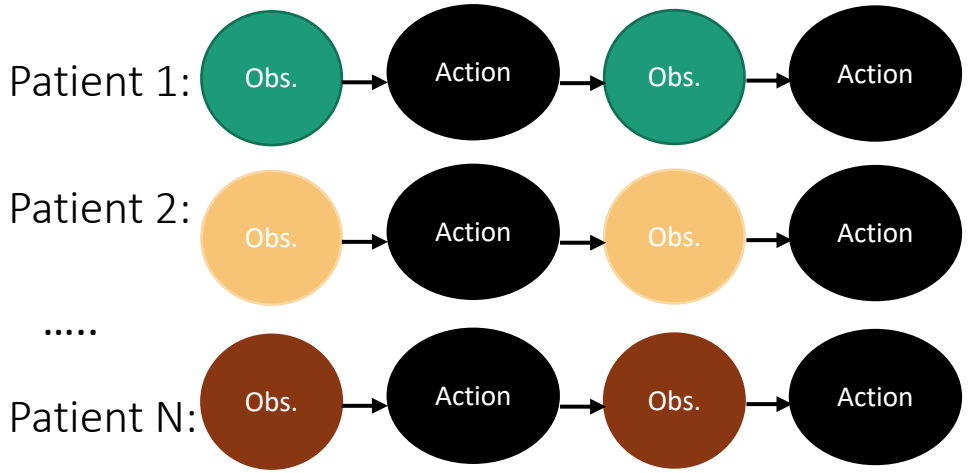


Evaluating Reinforcement Learning Algorithms in Observational Health Settings– Gottesman et al 2017, Nature Medicine

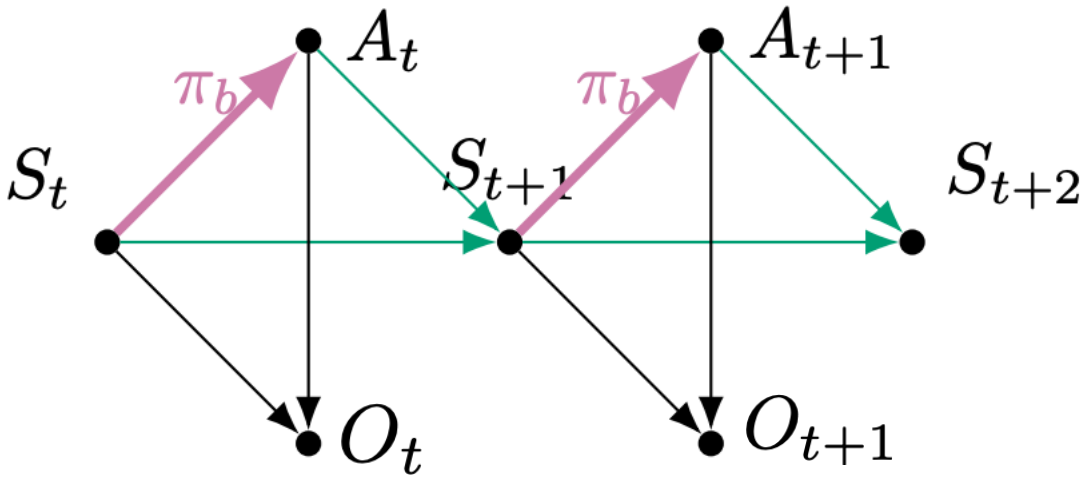
Moving Forward: From OPE to *Causal* OPE Estimates

Reweight trajectories according to how much they overlap using:

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



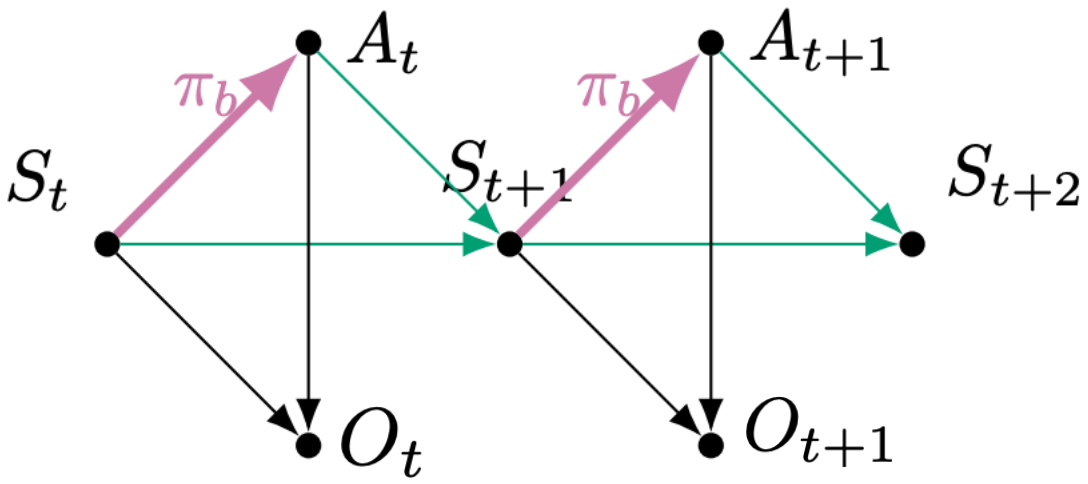
Standard OPE **ignores** the data-generating mechanism resulting in states, actions and observations.

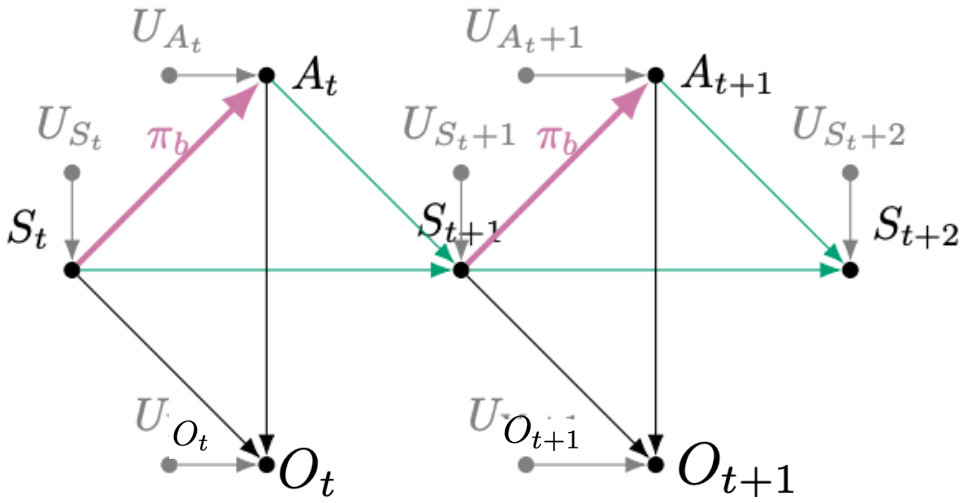


But the data-generating mechanism is important to understand when an OPE estimate makes sense and who it is applicable to

Moving Forward: From OPE to *Causal* OPE Estimates

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

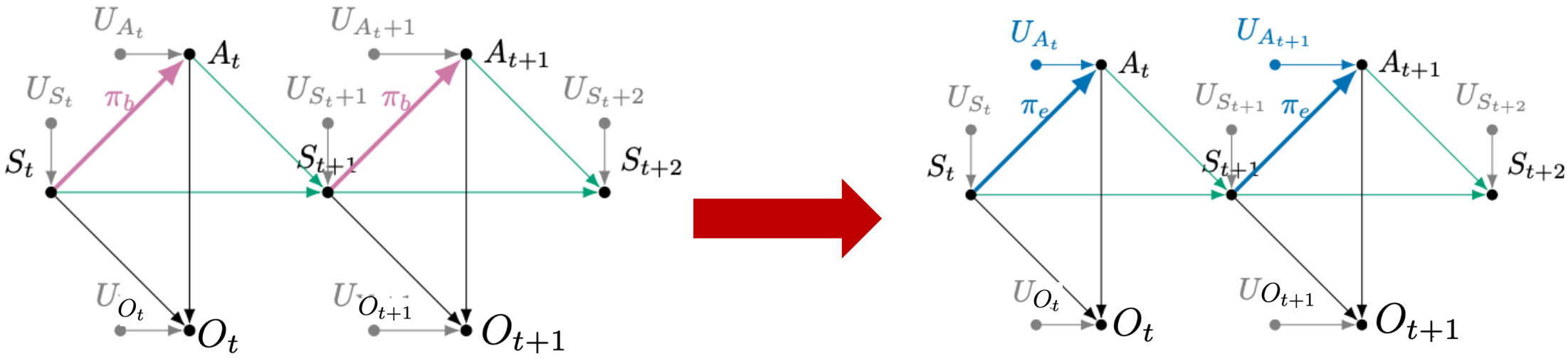




Parbhoo et al, Generalizing OPE from a Causal Perspective, arXiv pre-print, 2022.

Moving Forward: From OPE to *Causal* OPE Estimates

If we treat OPE as a type of soft intervention then:

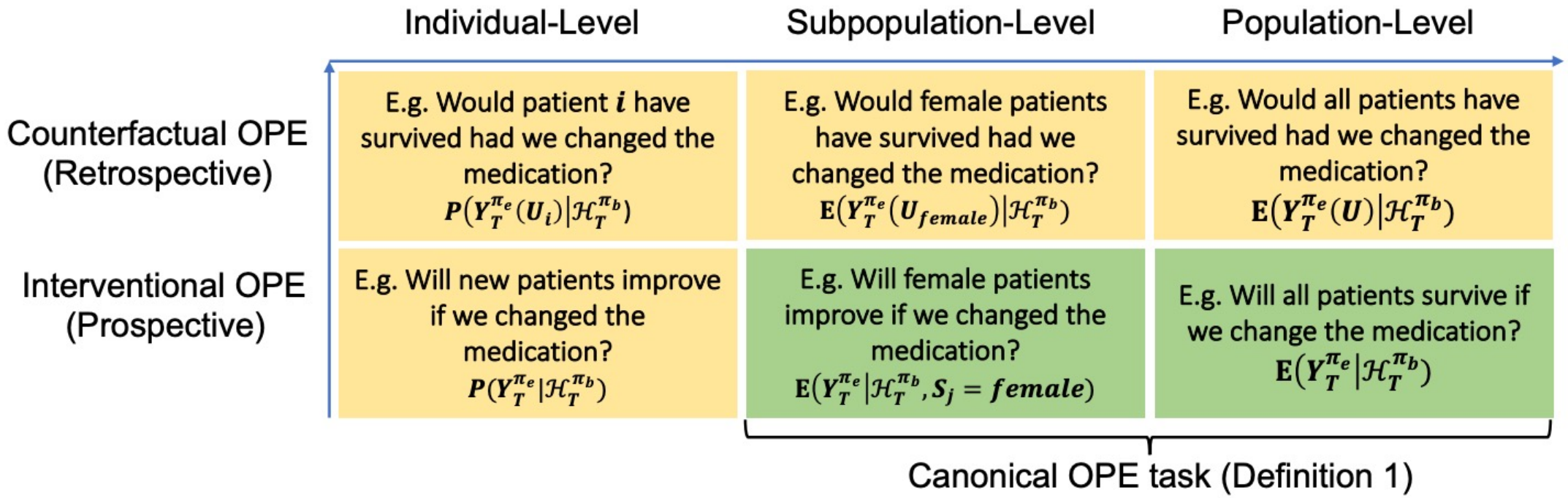


Depending on whether we want to compute **retrospective** or **prospective** estimates, the latent factors **U** may play a significant role the quality of the performance.

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

Moving Forward: From OPE to *Causal* OPE Estimates

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



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

Risk estimation *matters* for Causal OPE too!

Generalizing OPE as a causal estimand highlights two challenges:

1. Identifiability from observational data
2. Estimation challenges

Difficulty in performing OPE manifests as **induced uncertainty** in the OPE process.

Depending on the nature of uncertainty, we may be able to outline where human expertise may be helpful to mitigate uncertainty or where more assumptions or data are required.

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

What should responsible development and evaluation look like?

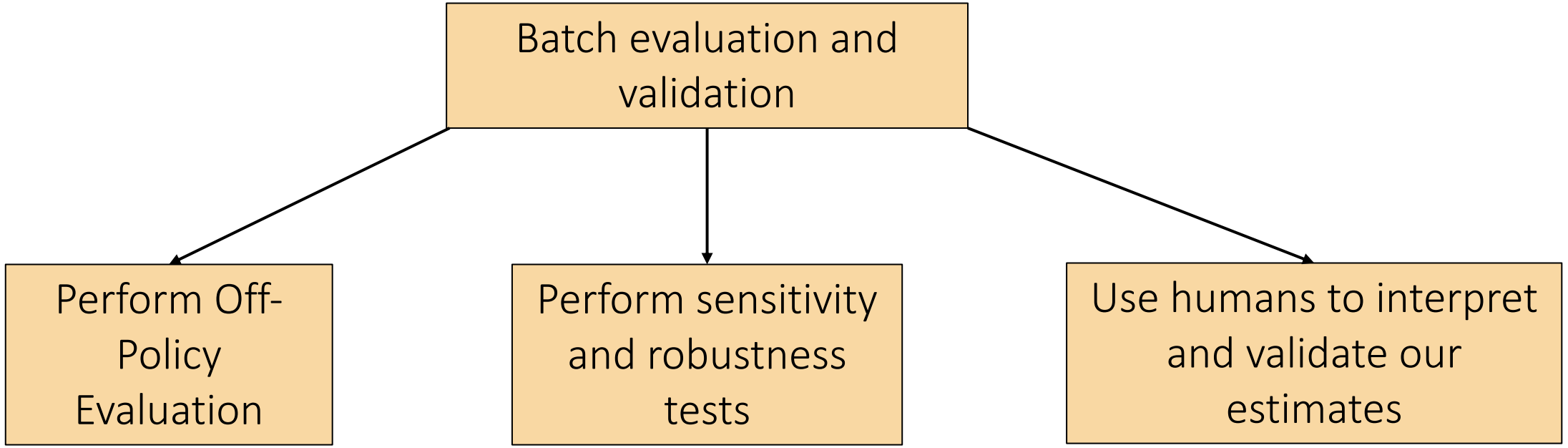
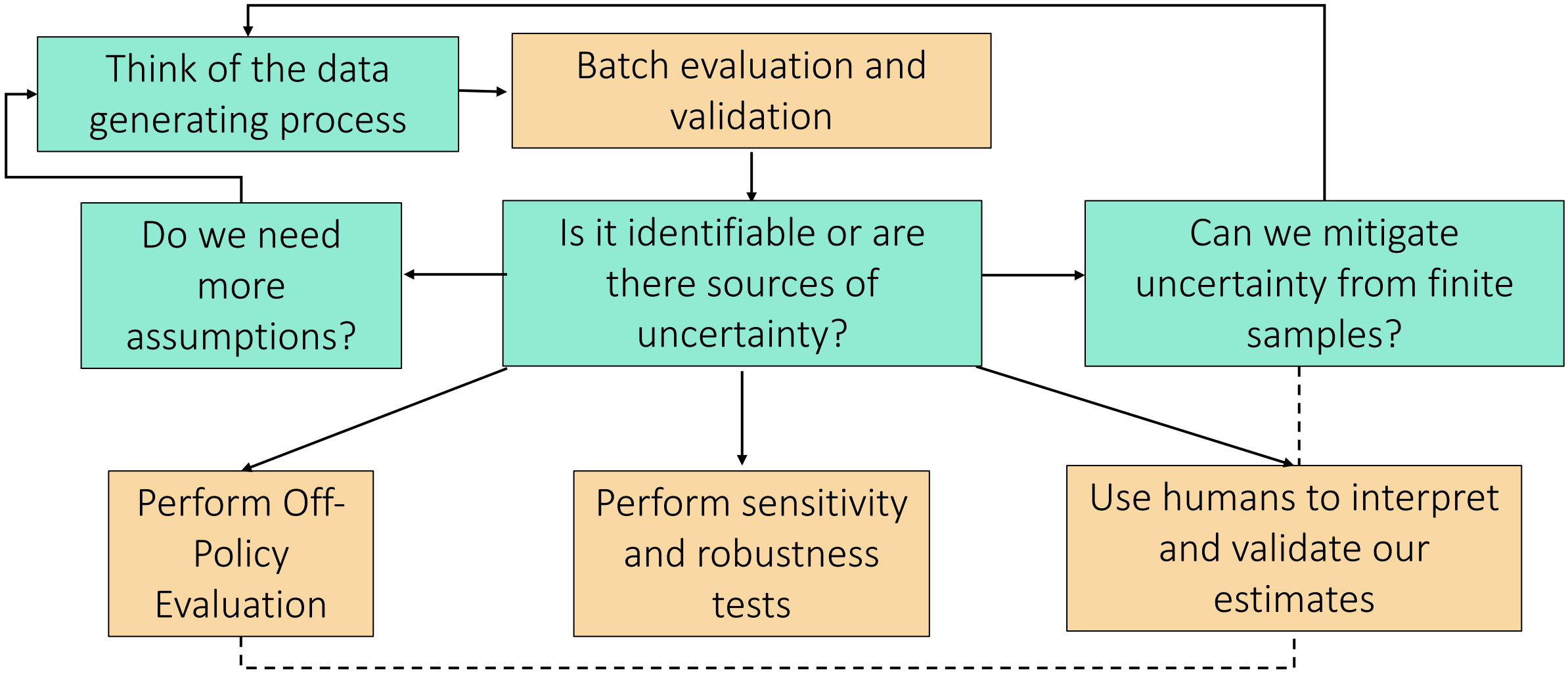


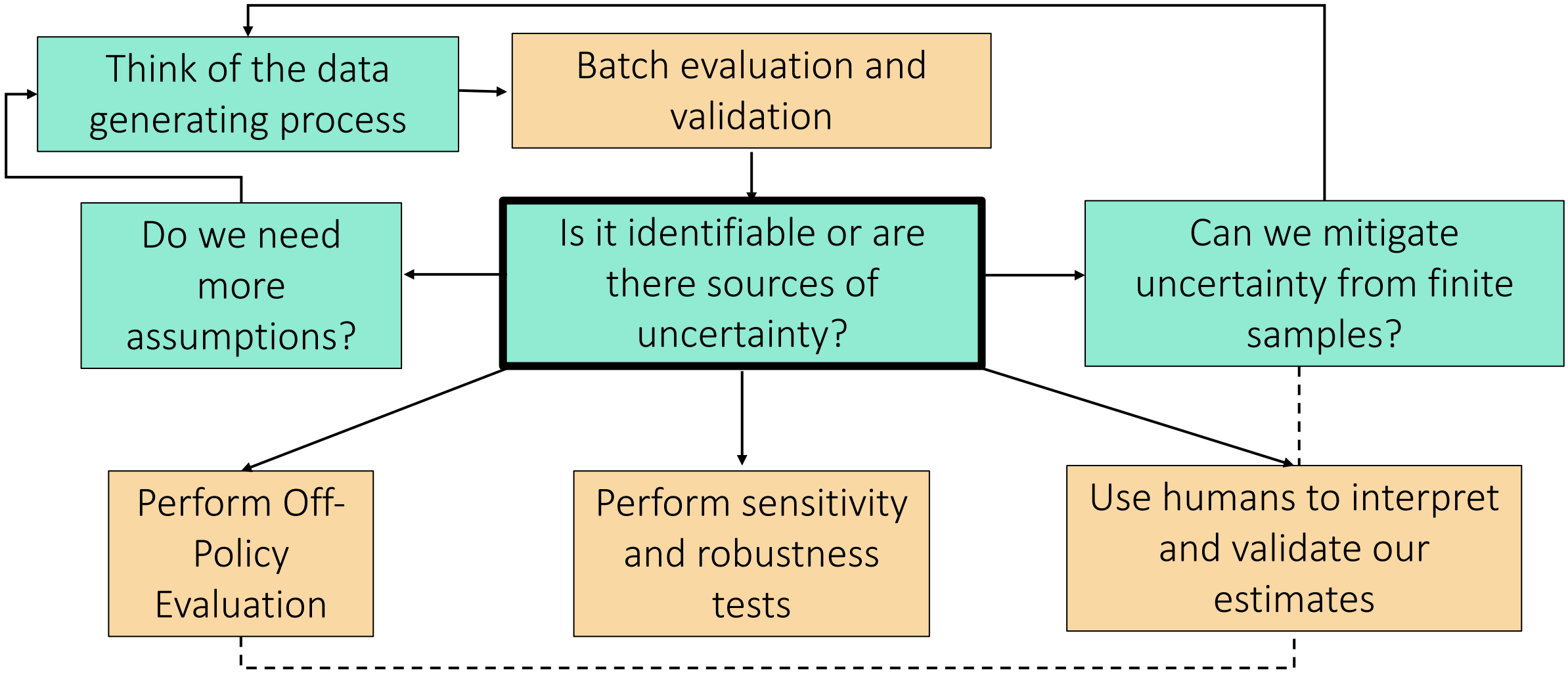
Image Credit: Doshi-Velez, Towards using Batch RL to Identifying Treatment Options, 2022

Causal OPE for responsible development and evaluation



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

Causal OPE for responsible development and evaluation



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

Summary

- High stake decisions have **long-term consequences** and are often applied in exceptional situations where generalisation is hard.
- RL has a lot of potential in healthcare.
- Evaluation should **tailored to the application** and the **nature of the query** for careful validation.
- Uncertainty and risk-estimation **are key** for careful evaluation.

Collaborators: Finale Doshi-Velez, Volker Roth, Maurizio Zazzi, Mario Wieser, Jasmina Bogojeska, Huldrych Gunthard, Omer Gottesman, Matthieu Komorowski, Aldo Faisal, David Sontag, Leo Celi, Shalmali Joshi, Taylor Killian, Marzyeh Ghassemi.

We’re hiring……

If you are interested in learning more about our research and engaging with us or working with our lab, please email us on

s.parbhoo@imperial.ac.uk