Health Care Breakout Session 3:
Treating Sepsis with Deep RL

Matt Engelhard

Today

- Guided discussion of “Deep Reinforcement Learning for Sepsis Treatment” (Raghu et al., 2017)
  - Overview
  - Formulation as RL
  - Evaluation Strategy
  - Results and their Significance

- Highlight applications of reinforcement learning to medical applications of sequential decision-making and adaptive trial design

Deep Reinforcement Learning for Sepsis Treatment

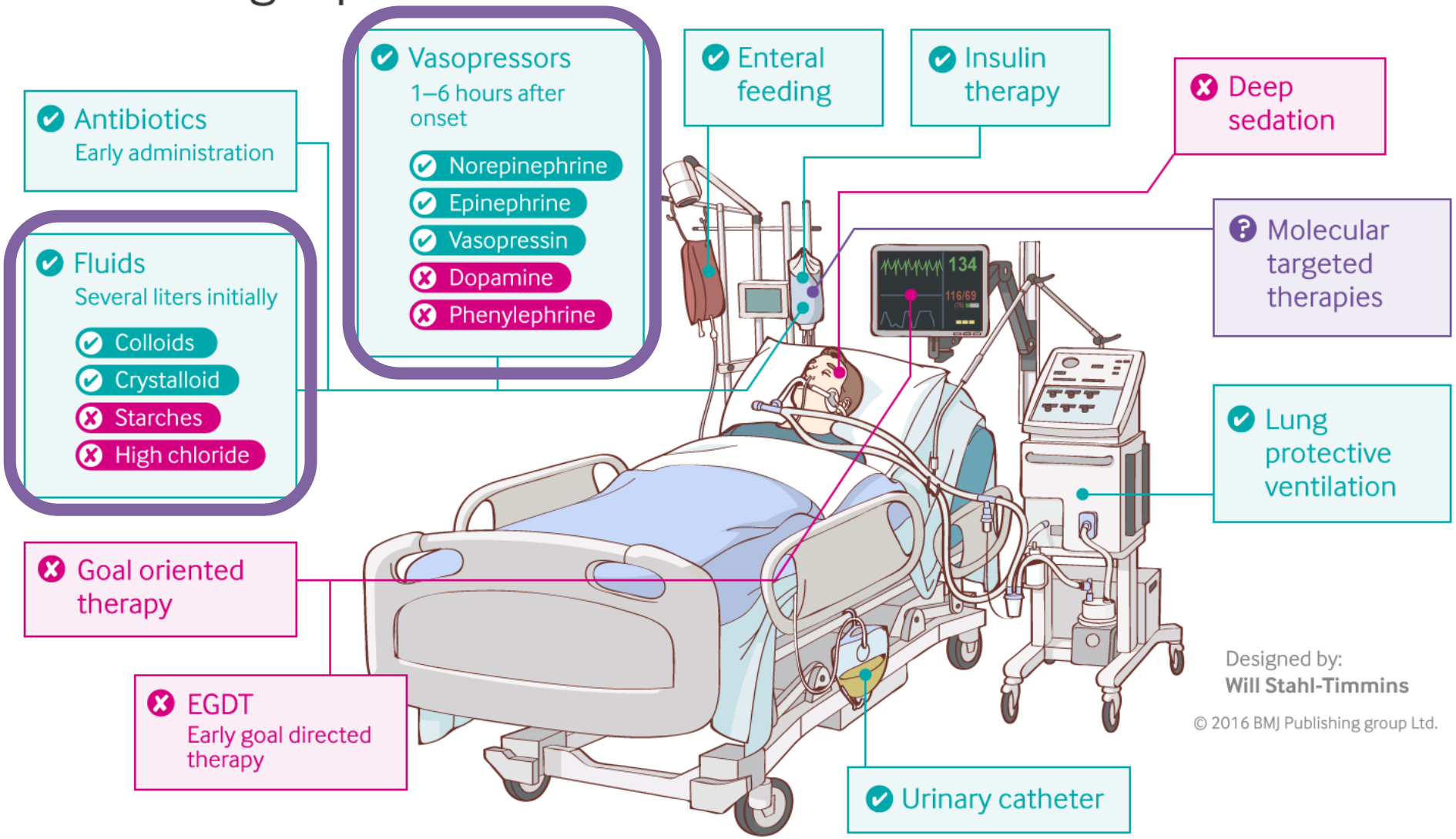
Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M.

arXiv:1711.09602. 2017 Nov 27

OVERVIEW:

- Deep Q-Learning (described in morning lecture 2) is applied to learn a policy for sepsis treatment
- 17,898 patients from the MIMIC-III Database
- The learned policy appears to lead to greater survival in low SOFA score group (more data), but not in high SOFA score group (less data)

Treating sepsis: the latest evidence



"Uncertainties still exist regarding the optimal type of fluid, the optimal volume, and the best way to monitor the response to therapy."

Gotts JE, Matthay MA. Sepsis: pathophysiology and clinical management. bmj. 2016 May 23;353(i1585).

Familiar Methods…

- Approximate $Q^*(s, a)$ with a type of Deep Q Network
- Train in Tensorflow via stochastic gradient descent with Adam optimizer
- Code available at https://github.com/aniruddhraghu/sepsisrl

FORMULATION AS RL

Data and Preprocessing

- Multiparameter Intelligent Monitoring in Intensive Care (MIMIC-III) Database
- 17,898 patients fulfilling Sepsis-3 criteria

  "suspected infection (prescription of antibiotics and sampling of bodily fluids for microbiological culture) combined with evidence of organ dysfunction, defined by a Sequential Organ Failure Assessment (SOFA) score greater or equal to 2"
- Data aggregated into windows of 4 hours, with the mean or sum being recorded (as appropriate) when several data points were present in one window

"Sepsis should be defined as life-threatening organ dysfunction caused by a dysregulated host response to infection. For clinical operationalization, organ dysfunction can be represented by an increase in the Sequential [Sepsis-related] Organ Failure Assessment (SOFA) score of 2 points or more, which is associated with an in-hospital mortality greater than 10%."

Singer M, Deutschman CS, Seymour CW, et al. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA*. 2016;315(8):801–810. doi:10.1001/jama.2016.0287

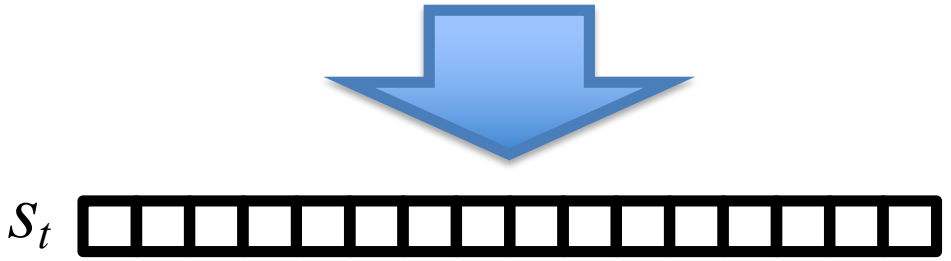
State Vector s_t

Demographics/Static: Shock Index, Elixhauser, SIRS, Gender, Re-admission, GCS - Glasgow Coma Scale, SOFA - Sequential Organ Failure Assessment, Age

Lab Values: Albumin, Arterial pH, Calcium, Glucose, Hemoglobin, Magnesium, PTT - Partial Thromboplastin Time, Potassium, SGPT - Serum Glutamic-Pyruvic Transaminase, Arterial Blood Gas, BUN - Blood Urea Nitrogen, Chloride, Bicarbonate, INR - International Normalized Ratio, Sodium, Arterial Lactate, CO2, Creatinine, Ionised Calcium, PT - Prothrombin Time, Platelets Count, SGOT - Serum Glutamic-Oxaloacetic Transaminase, Total bilirubin, White Blood Cell Count

Vital Signs: Diastolic Blood Pressure, Systolic Blood Pressure, Mean Blood Pressure, PaCO2, PaO2, FiO2, PaO/FiO2 ratio, Respiratory Rate, Temperature (Celsius), Weight (kg), Heart Rate, SpO2 Intake and Output Events: Fluid Output - 4 hourly period, Total Fluid Output, Mechanical Ventilation

Miscellaneous: Timestep



Actions a_t

IV fluid administration:

{0, Quartile1, Q2, Q3, Q4}

$\times$

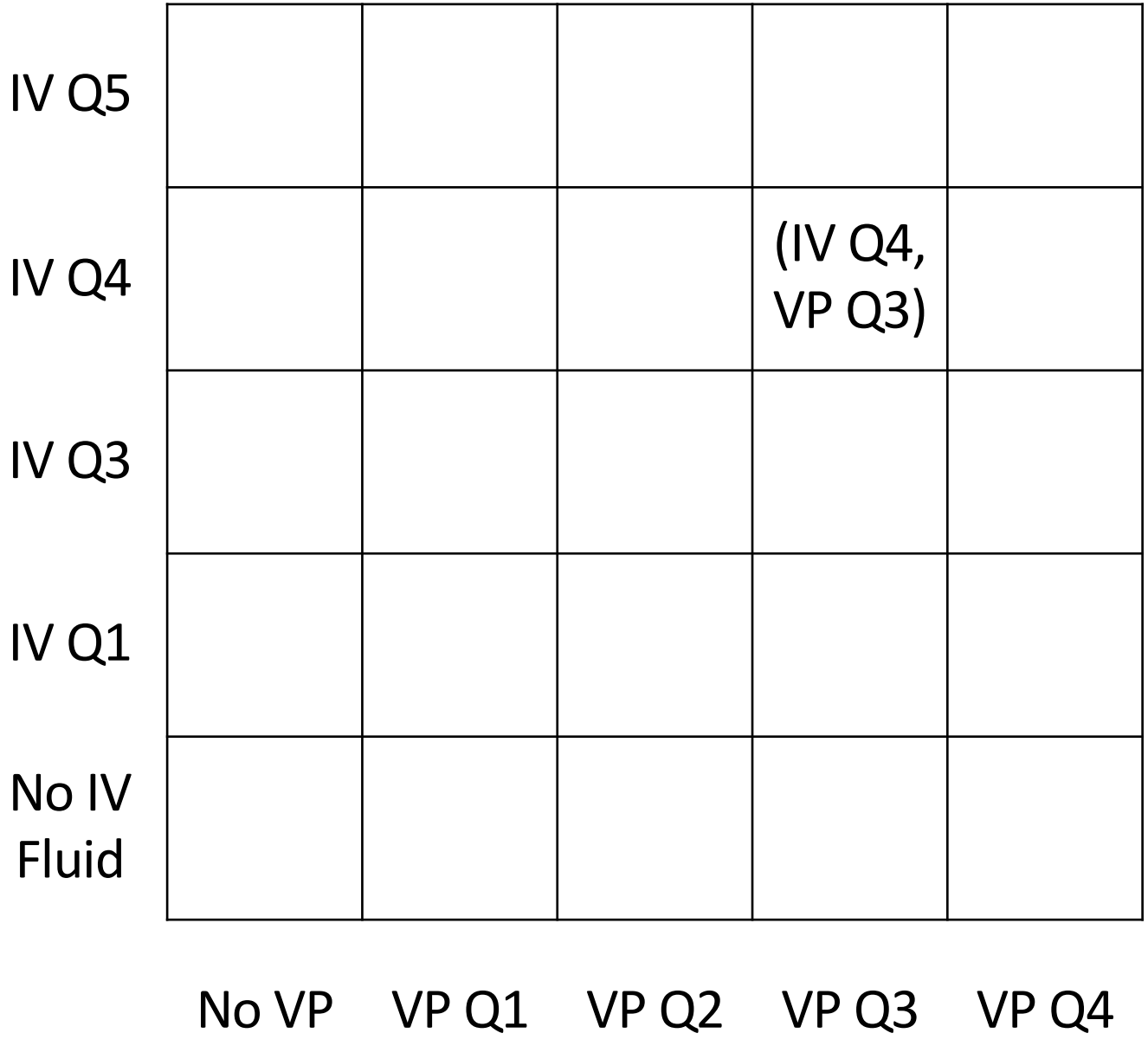
maximum vasopressor dosage

{0, Quartile1, Q2, Q3, Q4}

Actions a_t

(total IV in, max VP in)

at each time *t*



Which vasoactive drugs?

The effect of vasoactive drugs on mortality in patients with severe sepsis and septic shock. A network meta-analysis of randomized trials.

Belletti A[1], Benedetto U[2], Biondi-Zoccai G[3], Leggieri C[4], Silvani P[5], Angelini GD[6], Zangrillo A[7], Landoni G[8].

Author information

Abstract

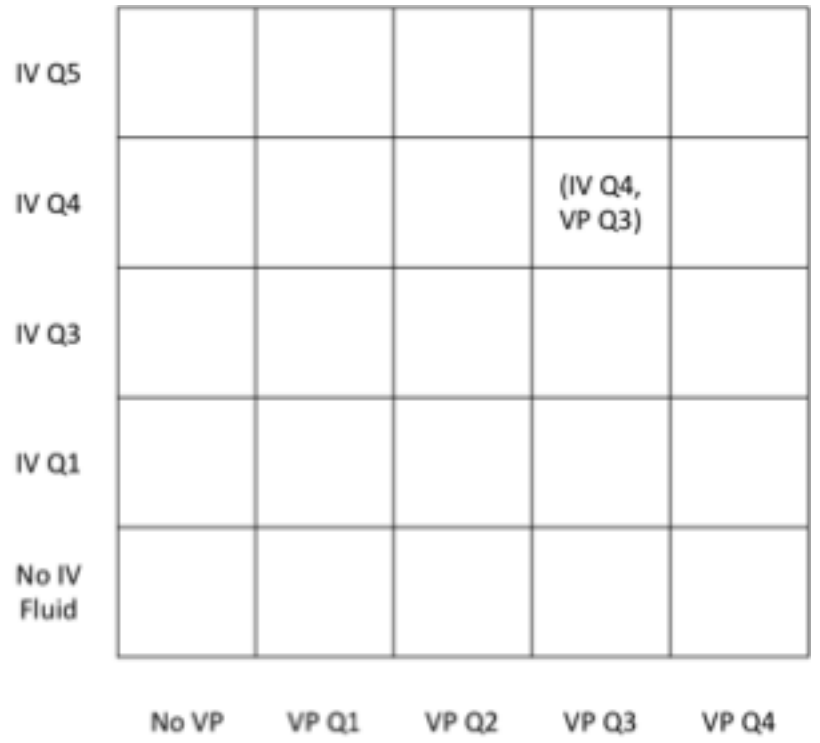
PURPOSE: Inotropes and vasopressors are cornerstone of therapy in septic shock, but search for the best agent is ongoing. We aimed to determine which vasoactive drug is associated with the best survival.

MATERIALS AND METHODS: PubMed, BioMedCentral, Embase, and the Cochrane Central Register were searched. Randomized trials performed in septic patients with at least 1 group allocated to an inotrope/vasopressor were included. Network meta-analysis with a frequentist approach was performed.

RESULTS: The 33 included studies randomized 3470 patients to 16 different comparators. As compared with placebo, levosimendan (odds ratio [OR], 0.17, 95%; confidence interval [CI], 0.05-0.60), dobutamine (OR, 0.30; 95% CI, 0.09-0.99), epinephrine (OR, 0.35; 95% CI, 0.13-0.96), vasopressin (OR, 0.37; 95% CI, 0.16-0.89), and norepinephrine plus dobutamine (OR, 0.4; 95% CI, 0.11-0.96) were significantly associated with survival. Norepinephrine improved survival compared with dopamine (OR, 0.81; 95% CI, 0.66-1.00). Rank analysis showed that levosimendan had the highest probability of being the best treatment.

CONCLUSIONS: Among several regimens for pharmacological cardiovascular support in septic patients, regimens based on inodilators have the highest probability of improve survival.

Q: How well does this action space represent the choices available to clinicians when treating sepsis?



Reward r_t

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}\left(\left(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}\right) \& \left(s_{t+1}^{\text{SOFA}} > 0\right)\right) + C_1\left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right) + C_2 \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right)$$

If SOFA score is unchanged and greater than zero, receive penalty of $|C_0|$

Receive reward/penalty of $|C_1|$ for decrease/increase

Receive reward/penalty up to $|C_2|$ for decreases/increases in lactate

*C_0, C_1, and C_2 are negative, so they can be viewed as penalties

Reward r_t

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}\left(\left(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}\right)\&\left(s_{t+1}^{\text{SOFA}} > 0\right)\right) + C_1\left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right) + C_2 \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right)$$

"We experimented with multiple parameters and opted to use $C_0 = -0.025$, $C_1 = -0.125$, $C_2 = -2$."

"At terminal timesteps, we issue a reward of +15 if a patient survived their ICU stay, and a negative reward of −15 if they did not."

"We opted to use tanh to cap the maximum lactate-related reward/penalty to |C2|."

Reward r_t

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}\left(\left(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}\right)\&\left(s_{t+1}^{\text{SOFA}} > 0\right)\right) + C_1\left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right) + C_2 \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right)$$

C_0 = −0.025, C_1 = −0.125, C_2 = −2

+15 if a patient survived their ICU stay. −15 if not.

Q: How well does the reward function reflect the goals of clinicians when treating sepsis?

$$r(s_t, s_{t+1}) = C_0 \mathbb{1}\left(\left(s_{t+1}^{\text{SOFA}} = s_t^{\text{SOFA}}\right)\&\left(s_{t+1}^{\text{SOFA}} > 0\right)\right) + C_1\left(s_{t+1}^{\text{SOFA}} - s_t^{\text{SOFA}}\right) + C_2 \tanh\left(s_{t+1}^{\text{Lactate}} - s_t^{\text{Lactate}}\right)$$

***C0* = −0.025, *C1* = −0.125, *C2* = −2**

+15 if a patient survived their ICU stay. −15 if not.

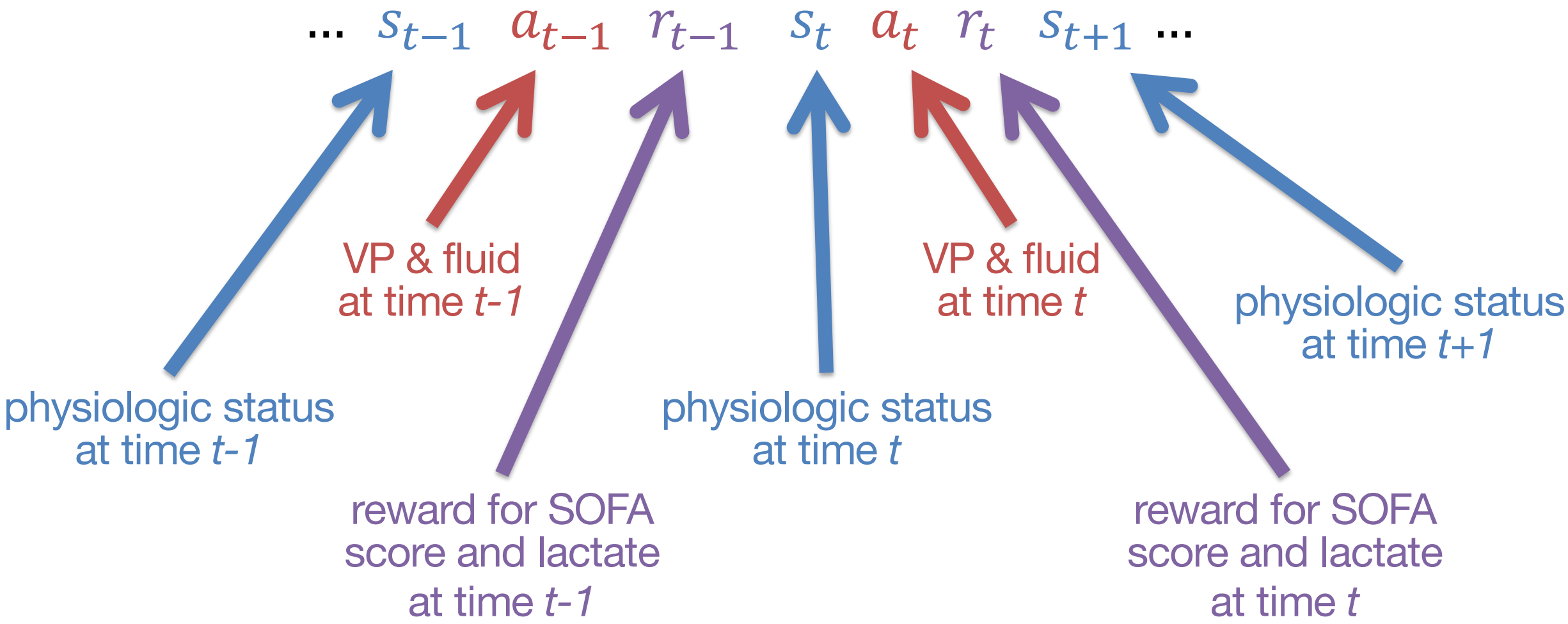
short-term goals:

- central venous pressure (8-12 mm Hg)
- mean arterial pressure (65-90 mm Hg)
- urine output (0.5 mL/kg/h)
- central venous oxygen saturation (70% target in early goal directed therapy)

versus longer-term goals:

- organ dysfunction
- death

State, Action, Reward

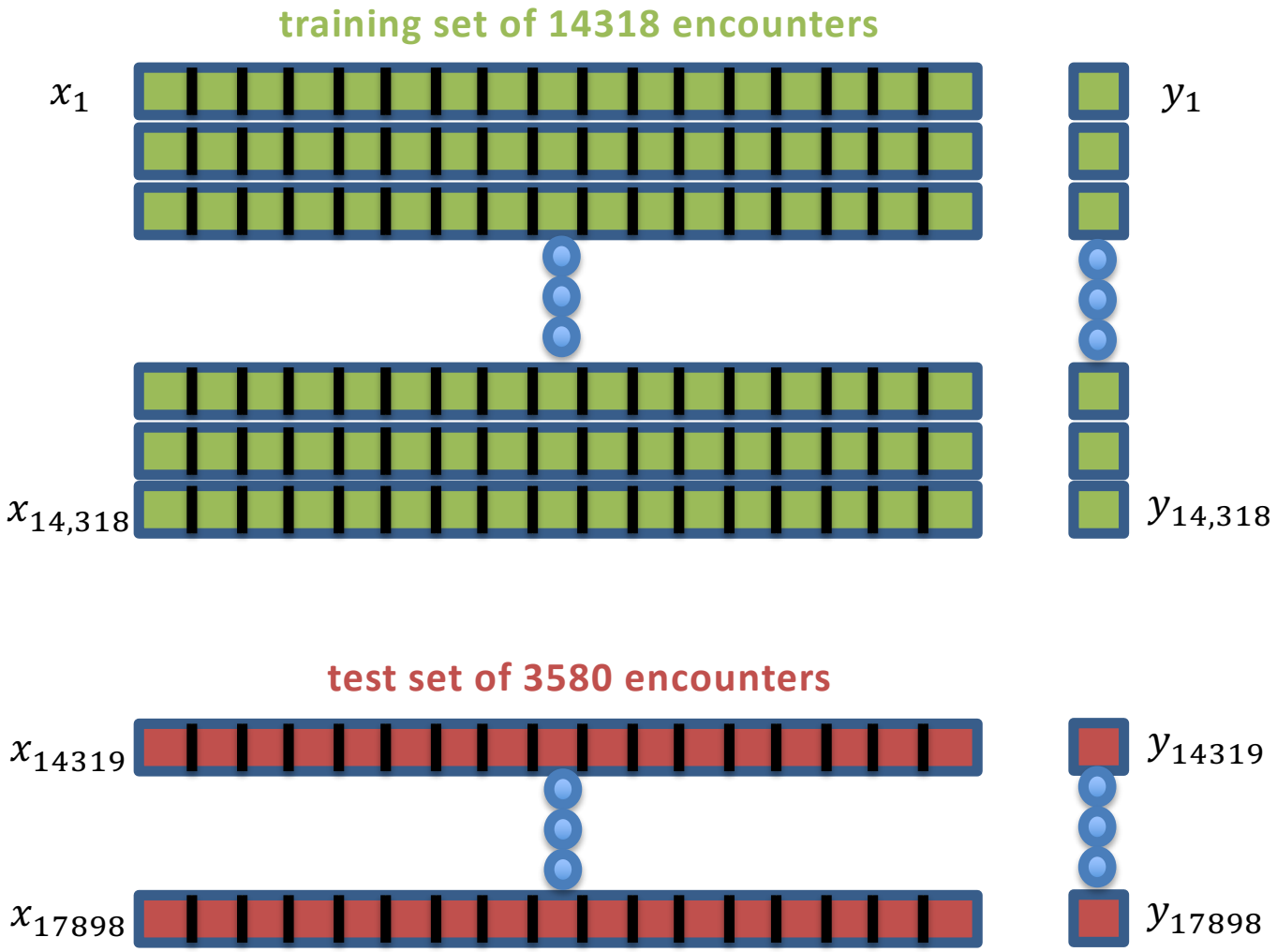


Q: How well does this model reflect reality?
(either in medicine or in general)

$$\ldots \quad s_{t-1} \quad a_{t-1} \quad r_{t-1} \quad s_t \quad a_t \quad r_t \quad s_{t+1} \quad \ldots$$

EVALUATION STRATEGY

Evaluating on a Held-Out Test Set



- 80/20 train/set split (by encounter)
- ensure that a proportionate number of patient outcomes are present in both sets (outcome-adaptive assignment)

Q: Suppose different encounters from the same (very unfortunate) patients were split between test and training splits. Would this compromise results?

Evaluation Measures

- Classification performance?
- Mean Square Error?

- ➢ Compare outcomes between the learned policy and competing policies.
- ➢ In this case, the “competing policy” is real physician decision-making

RESULTS

Results: Learned Policy vs Physician Policy

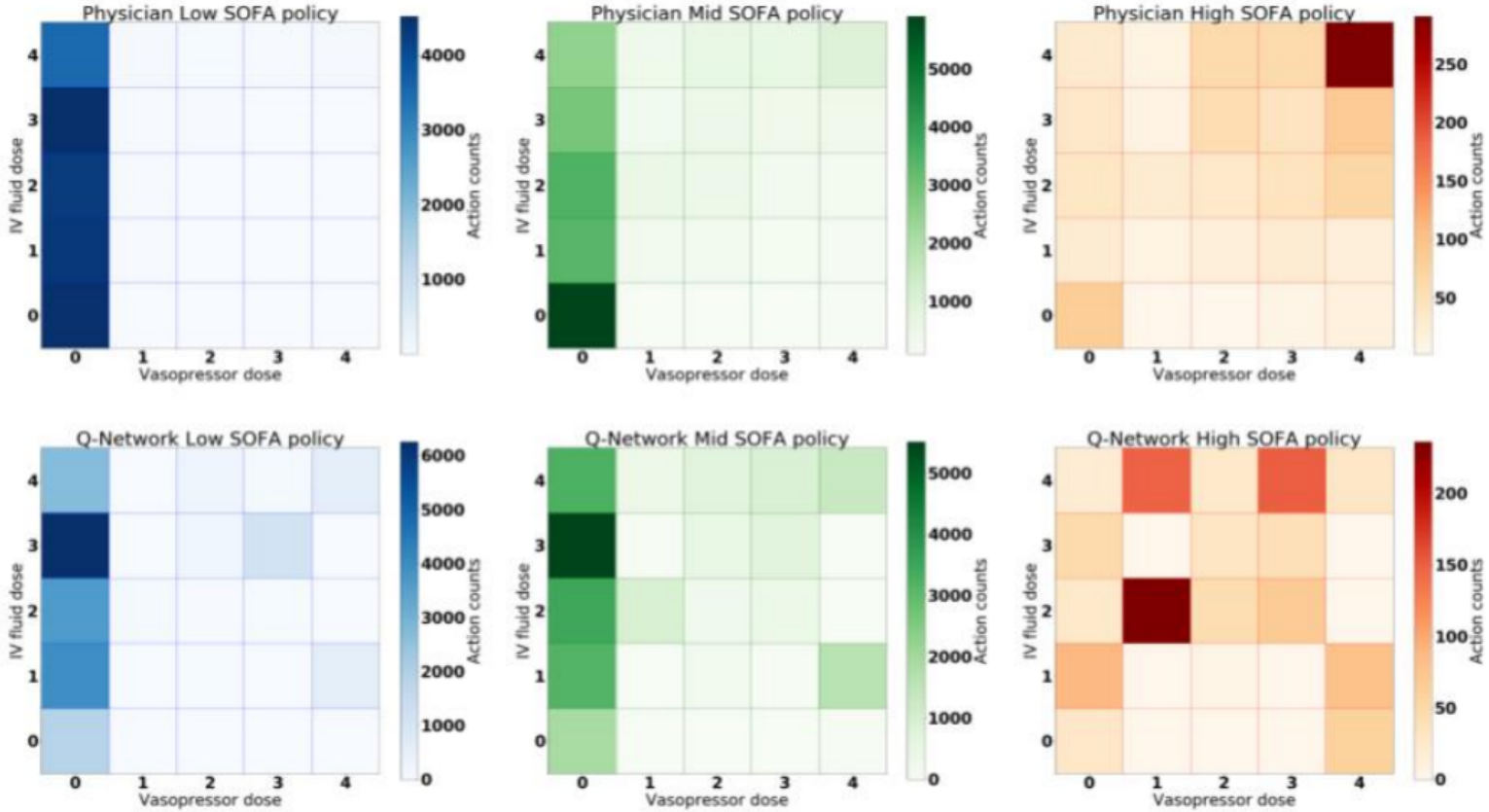


Figure 1: Policies learned by the different models, as a 2D histogram, where we aggregate all actions selected by the physician and model on the test set over all relevant timesteps. The axes labels index the discretized action space, where 0 represents no drug given, and 4 the maximum of that particular drug. The model learn to prescribe vasopressors sparingly, a key feature of the physician's policy.

Q: How could the policy be more effectively visualized (e.g. mapping from physiologic states to actions)?

Could variables important to the choice of action could be highlighted? (much like the saliency map yesterday)

Results: Observed Mortality

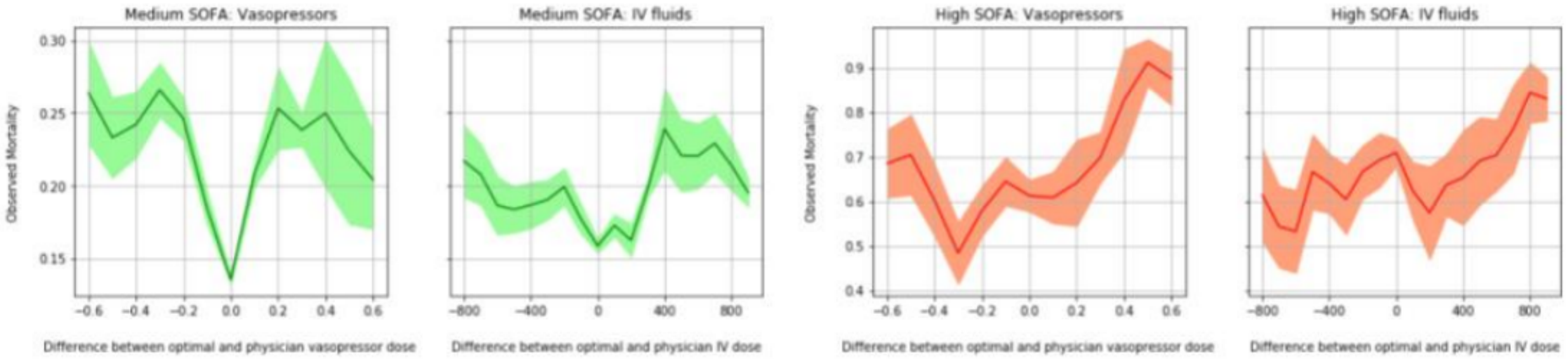
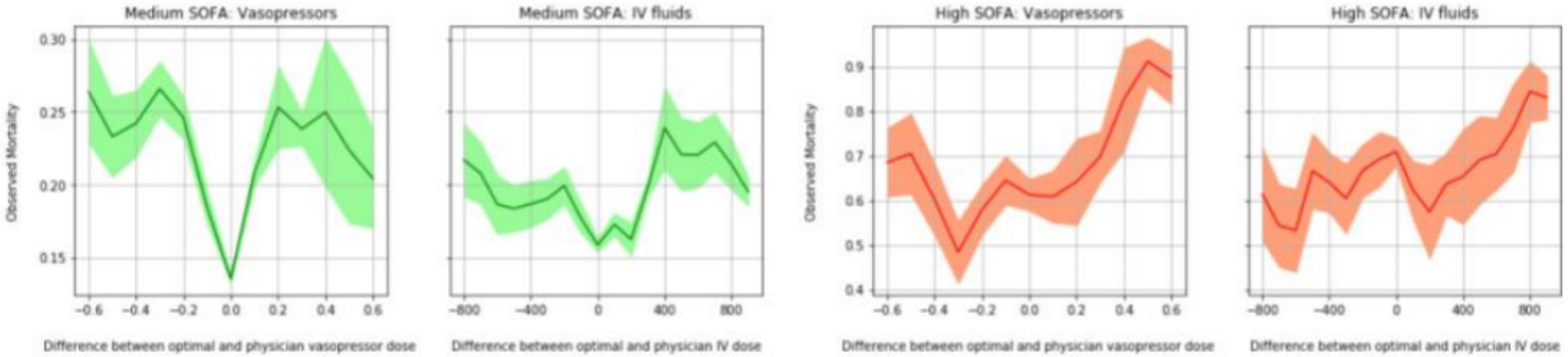


Figure 2: Comparison of how observed mortality (y-axis) varies with the difference between the dosages recommended by the optimal policy and the dosages administered by clinicians (x-axis) on a held-out test set. For every timestep, this difference was calculated and associated with whether the patient survived or died in the hospital, allowing the computation of observed mortality. We see low mortality with medium SOFA scores for when the difference is zero, indicating that when the physician acts according to the learned policy in this regime we observe more patient survival.

Q: Figure 2 shows that for medium SOFA scores, observed mortality is lowest when physician actions match the learned policy. However, this is not true for high SOFA scores. How does this observation relate to model generalizability and the training/test split? How does it relate to the reward function?



Q: In this work, a reinforcement learning agent is trained using data from MIMIC-III. With this setup, can the agent learn to take actions *never taken* by a physician? Why or why not? Can it learn to take actions *not typically taken* by a physician? After attending the Day 3 lectures, relate your reasoning to the Q-Learning equation.

Q-Learning:

$$Q^{new}(s,a) \leftarrow (\alpha - 1) \cdot Q^{old}(s,a) + \alpha \cdot [\, r(s,a,s') + \gamma \cdot \max_{a'} Q^{old}(s',a') \,]$$

Q: Suppose the feature vector representing the patient's state s_t did not contain a specific physiological variable used by clinicians when treating sepsis. Would this compromise the policy learned by the agent? Why or why not, and what other factors might be relevant?

Q-Learning:

$$Q^{new}(s,a) \leftarrow (\alpha - 1) \cdot Q^{old}(s,a) + \alpha \cdot [\, r(s,a,s') + \gamma \cdot \max_{a'} Q^{old}(s',a') \,]$$

Limitations

- Observational data
  - decisions and rewards may have been affected by characteristics not recorded / quantified
  - fundamental but difficult to correct; prospective testing requires physician supervision
- Assumptions encoded in model, e.g. state space, action space, reward, deep Q-network
- Limited data in the high SOFA regime

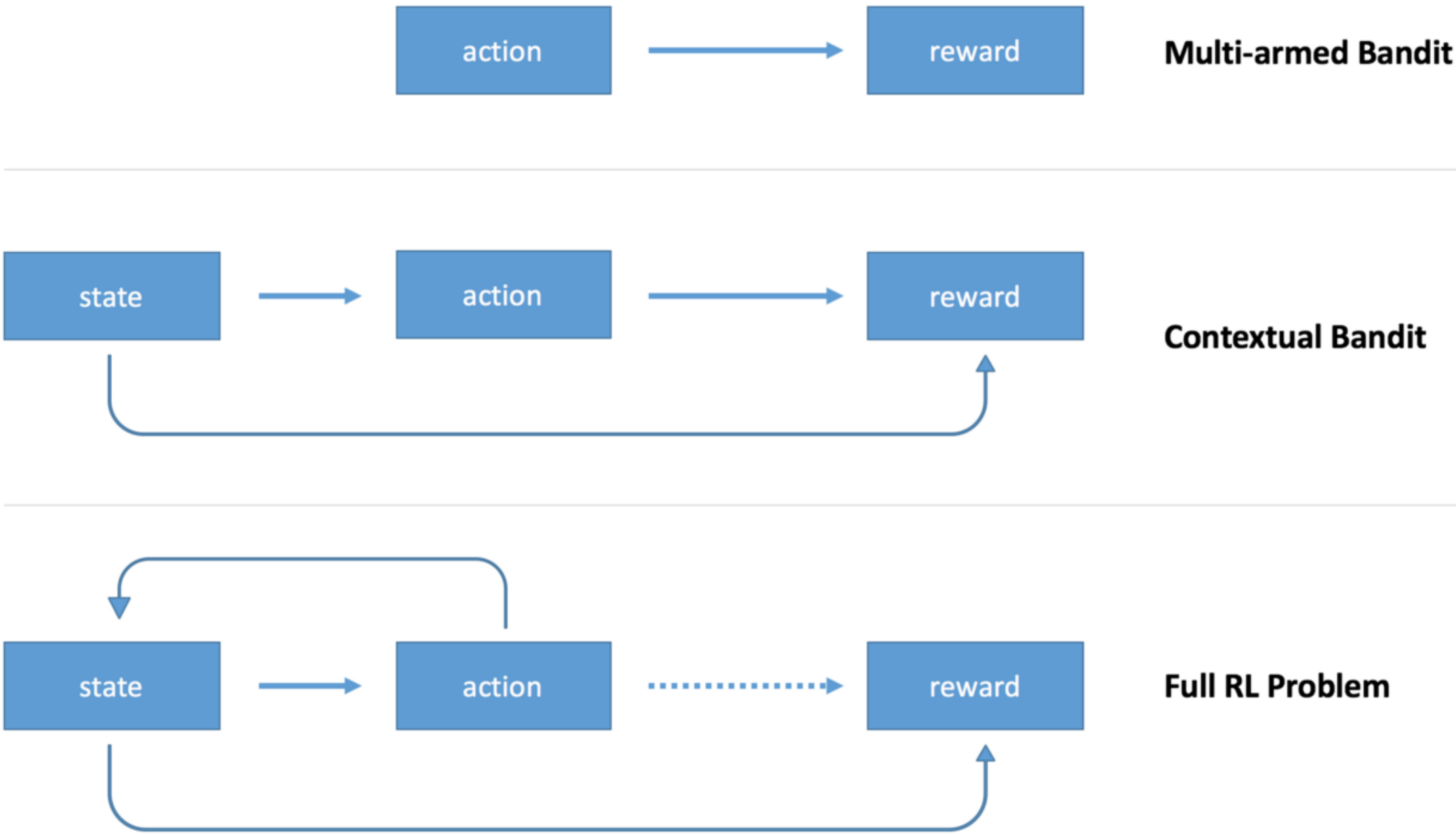
Q1: Is the comparison to physicians a fair and/or reasonable one?

Q2: Are there other important methodological limitations?

Q3: What are the barriers to adoption in clinical practice?

Reinforcement Learning and Related Algorithms

SURVEY OF CLINICAL APPLICATIONS



From https://medium.com/emergent-future/

Multi-armed Bandit

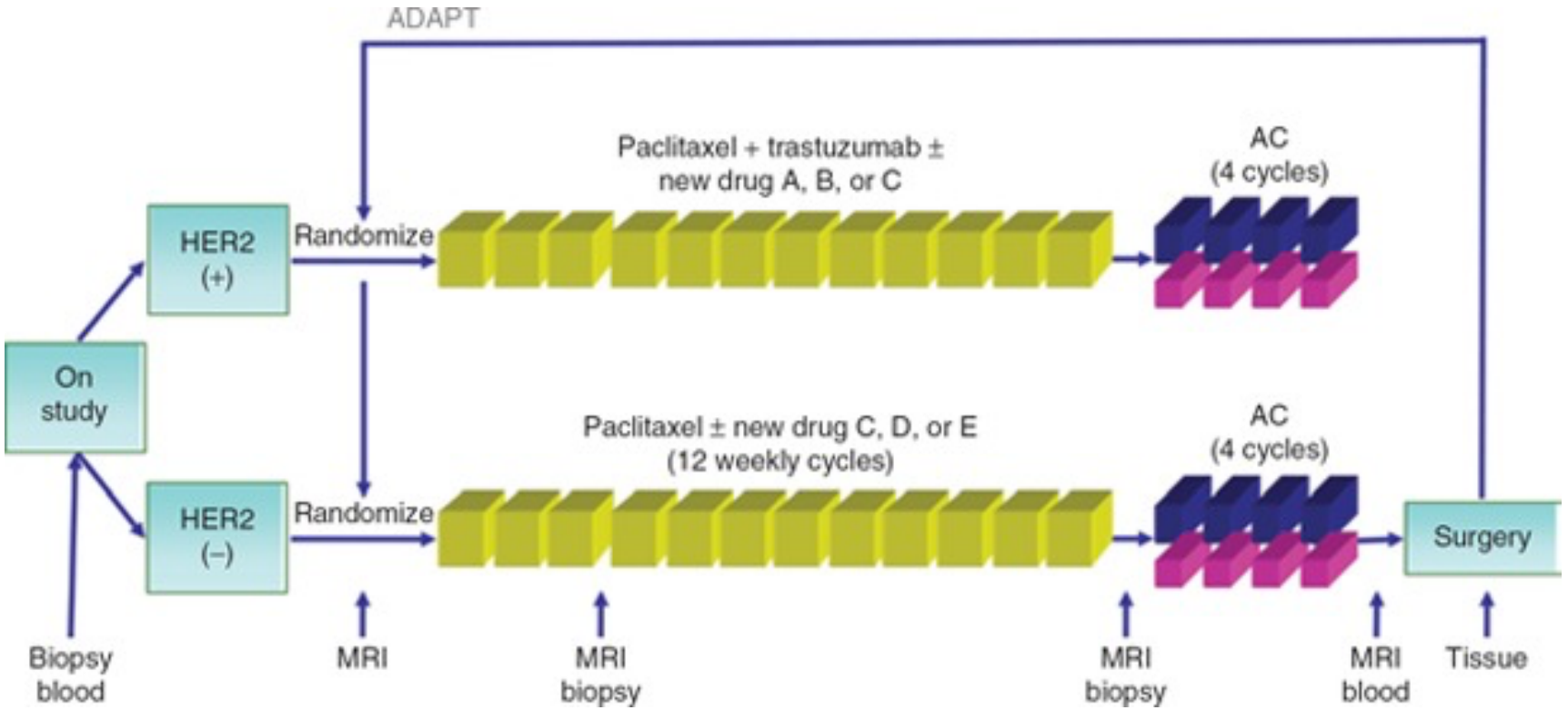
Application: Optimal Allocation of Clinical Trial Participants

"An explicit assumption is the goal to treat patients effectively, in the trial as well as out. That is controversial (…)"

(Stangl, Inoue and Irony, 2012)

Multi-armed Bandit

Barker, A. D., et al. "I-SPY 2: an adaptive breast cancer trial design in the setting of neoadjuvant chemotherapy." *Clinical Pharmacology & Therapeutics* 86.1 (2009): 97-100.



Multi-armed Bandit

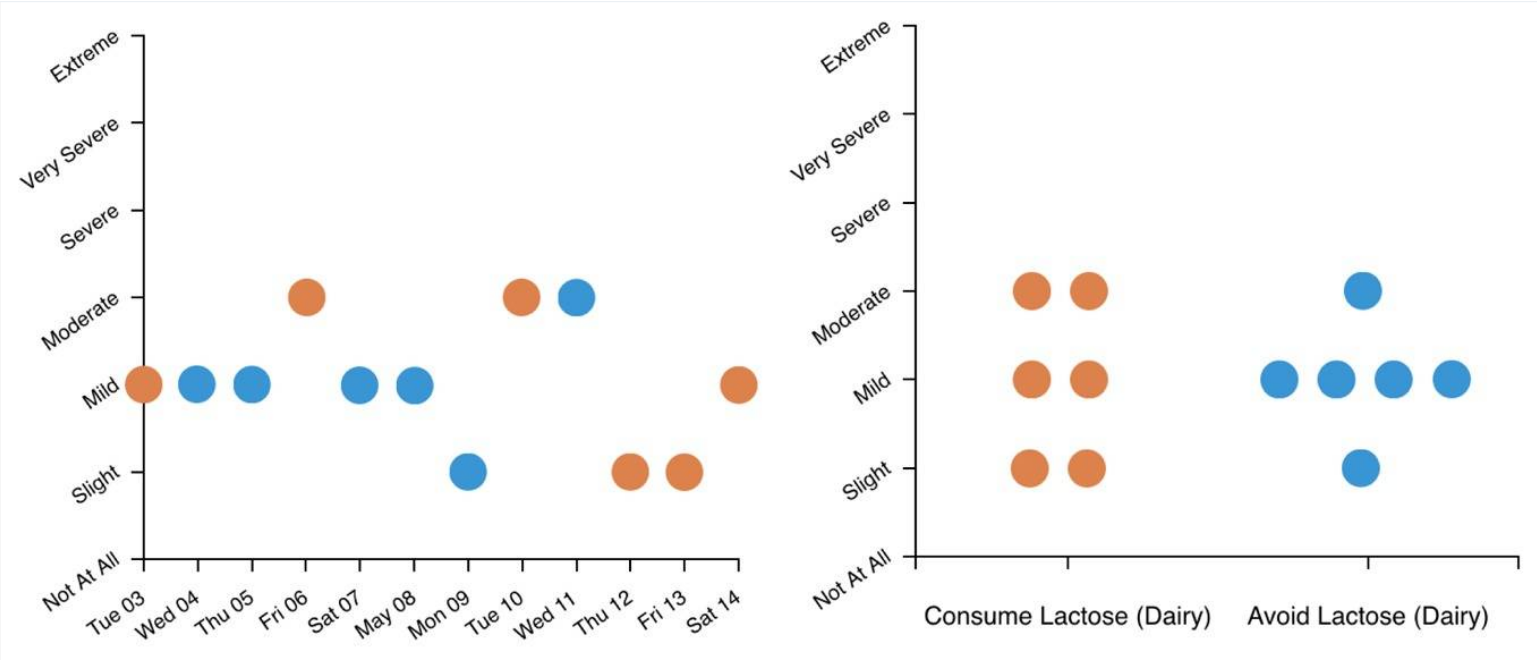
Application: Optimal Allocation of Clinical Trial Participants

Challenges:

- Modifications to meet frequentist trial design requirements (e.g. maintain higher degree of *exploration* to control type-I error rate)
- Time delay between allocation and result
- Deterministic allocation can make the trial more vulnerable to bias (e.g. participant drift, manipulation by a sponsor)

Multi-armed Bandit

Application: Self-Experimentation and other Quasi-Experimental Study Designs

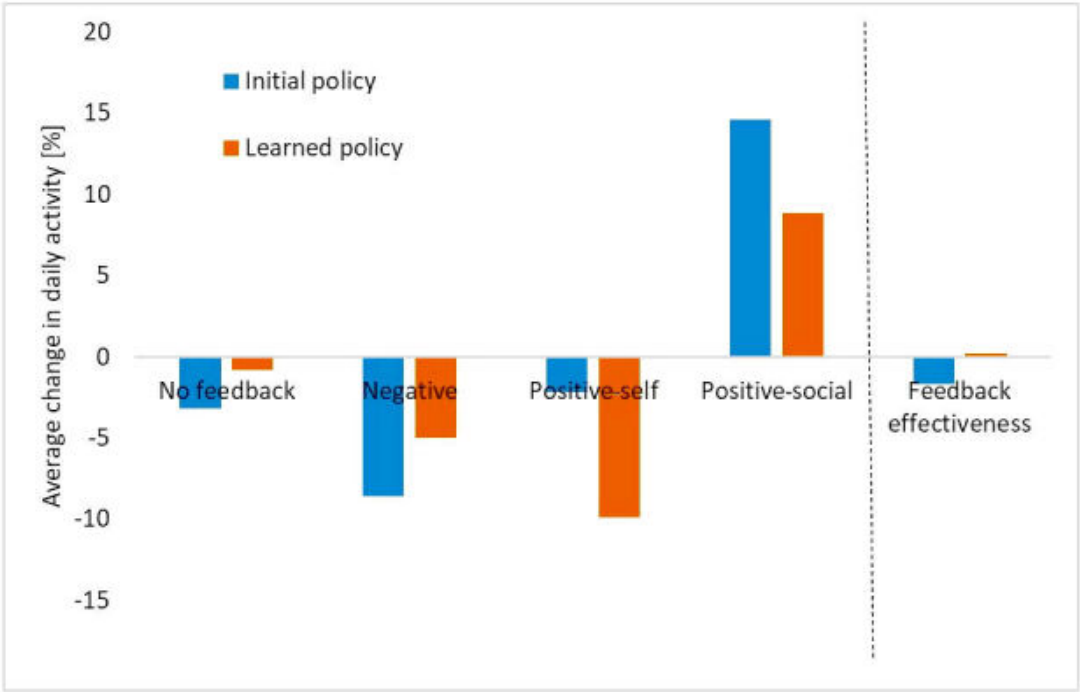


TummyTrials visualizes self-experimentation both as a timeline (left) and by trend in experimental condition (right).

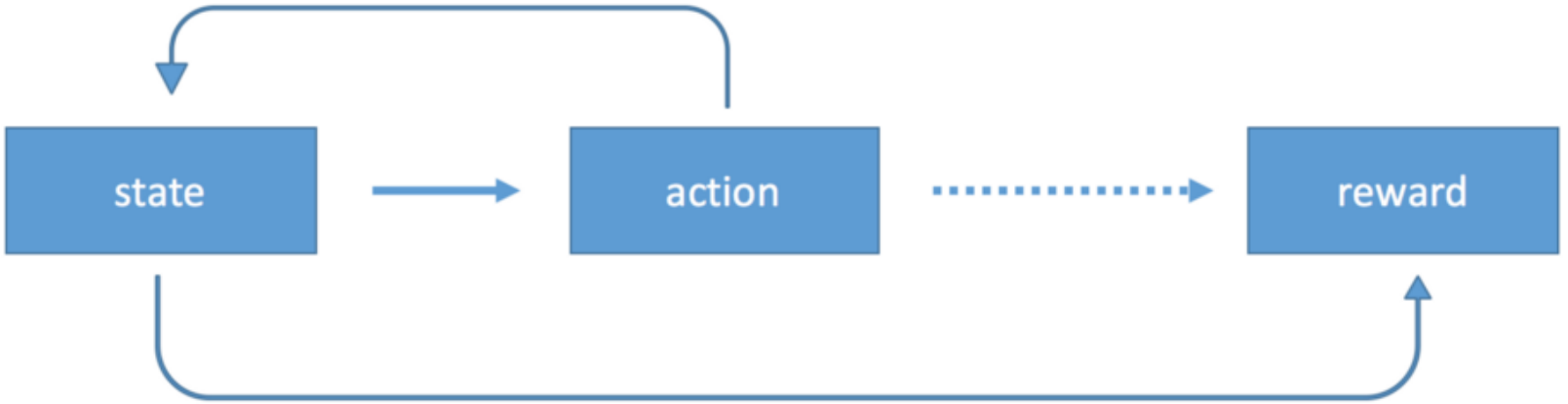
TummyTrials: A Feasibility Study of Using Self-Experimentation to Detect Individualized Food Triggers.
Karkar R, Schroeder J, Epstein DA, et al.
SIGCHI Conference 2017;2017:6850-6863.

Contextual Bandit

- **personalized adaptive trial design**
- **personalized or context-sensitive feedback**



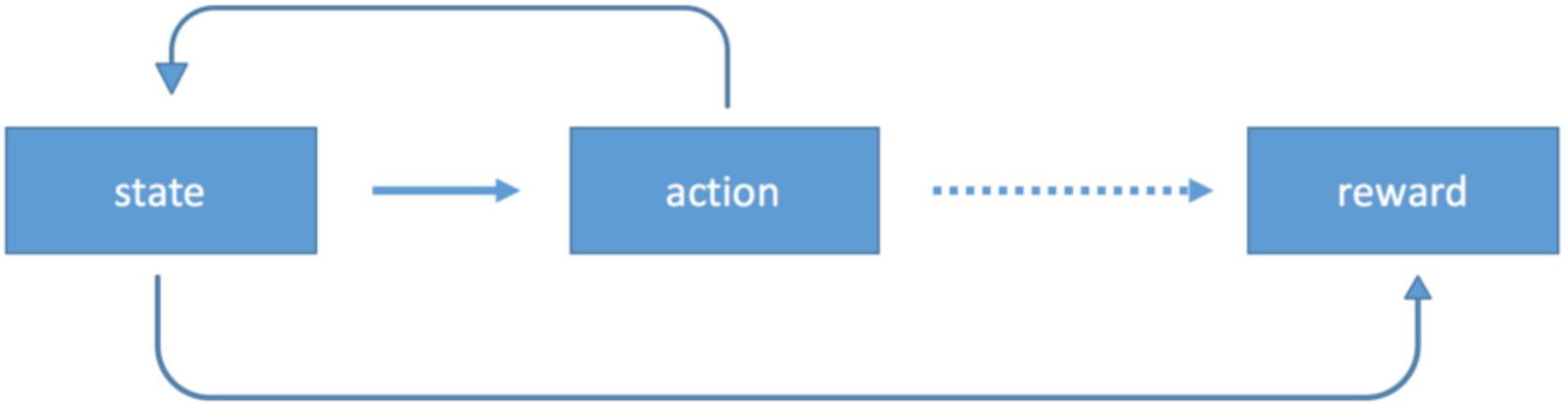
Yom-Tov, Elad et al. "**Encouraging Physical Activity in Patients With Diabetes: Intervention Using a Reinforcement Learning System**." Ed. Gunther Eysenbach. *Journal of Medical Internet Research* 19.10 (2017): e338. *PMC*. Web. 27 June 2018.



Full RL Problem

Sequential Clinical Decision-making:

A reinforcement learning approach to weaning of mechanical ventilation in intensive care units.
Prasad, Niranjani, et al. arXiv:1704.06300 (2017).

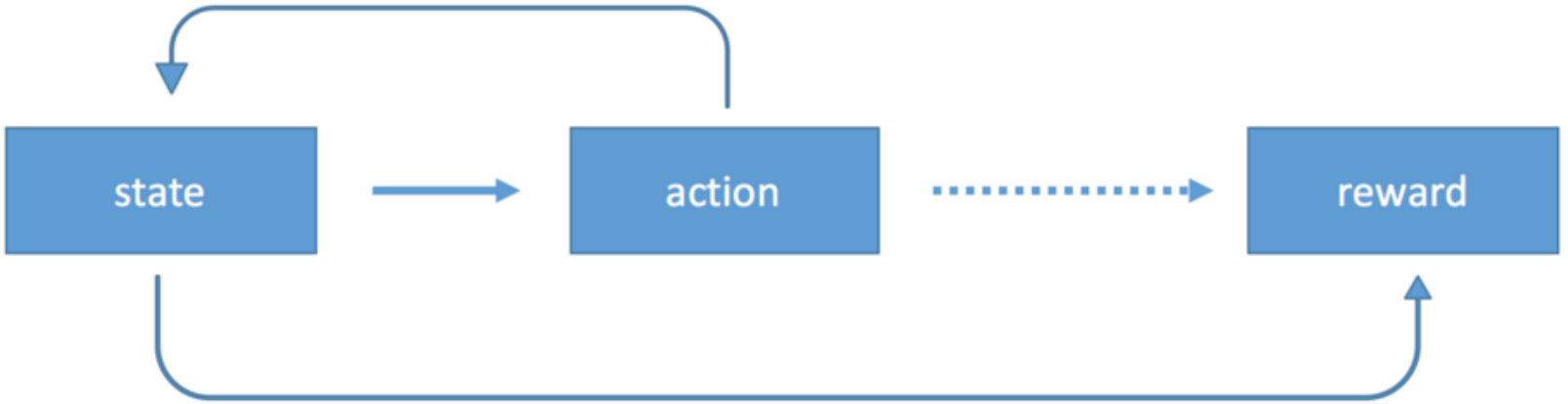


Full RL Problem

Sequential Clinical Decision-making:

Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach.

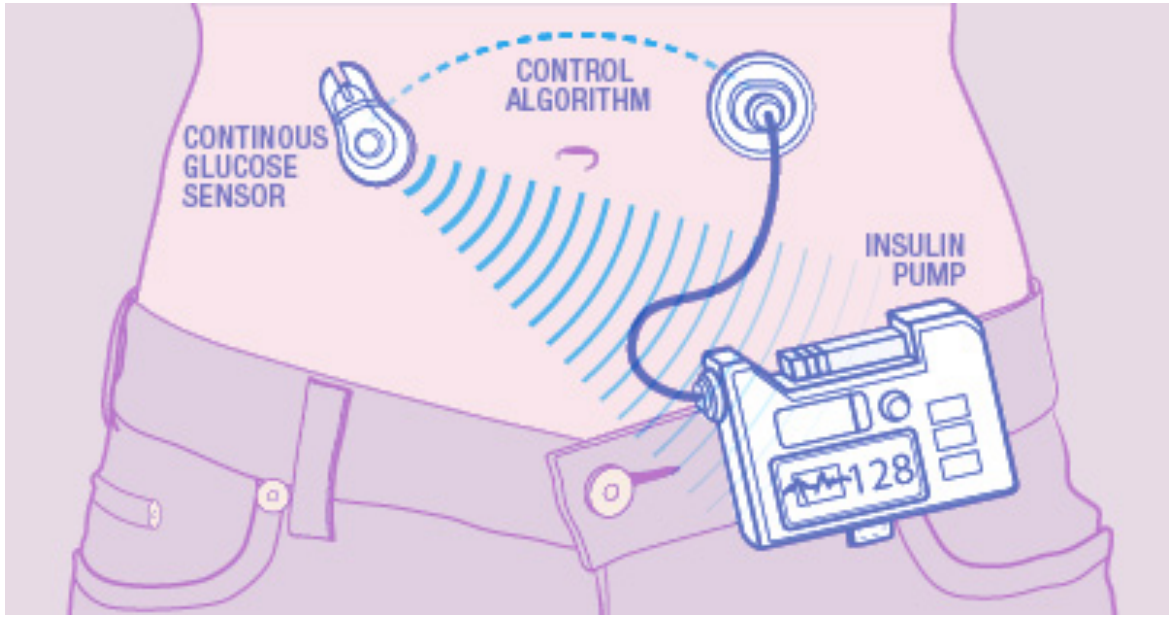
Nemati S, Ghassemi M, and Clifford G. Engineering in Medicine and Biology Society (EMBC), 2016



Full RL Problem

Sequential Clinical Decision-making:

Closed-loop blood glucose control
("artificial pancreas")



https://www.mayo.edu/research/labs/artificial-pancreas/overview

THANK YOU!

Questions or ideas? Please contact me at m.engelhard@duke.edu