

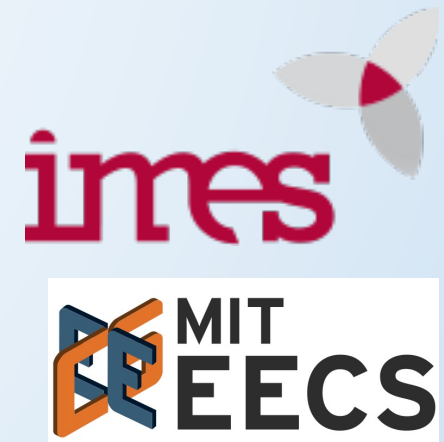
Clinical Reinforcement Learning

Irene Y. Chen

 @irenetrampoline



6.S897 / HST.956 Machine Learning for Healthcare: Recitation 7



Housekeeping

1. Final projects

- Start early
- Reach out to professors and TAs if you need help

2. HW5 and HW6

- Final stretch!
- HW6 going out Thurs

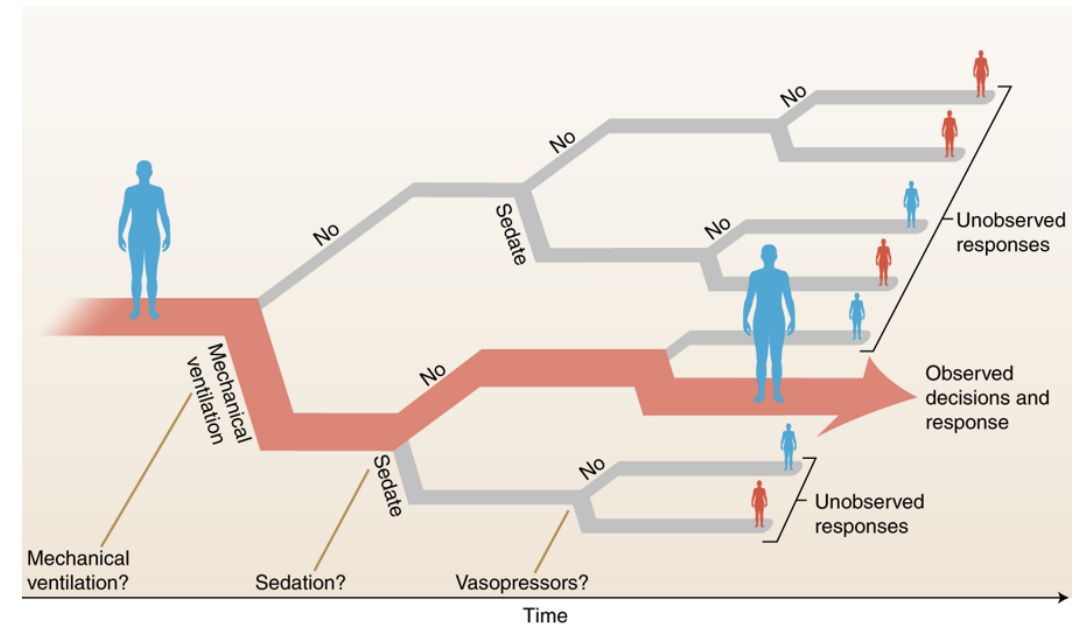
Agenda for today

- ~~1. Housekeeping~~
2. Review lecture material [15 mins]
3. Smoking cessation two-stage example [15 mins]
4. Broader discussion [15 mins]

Goal: 1) contextualize RL lectures this week, 2) balance optimism and skepticism about RL in healthcare

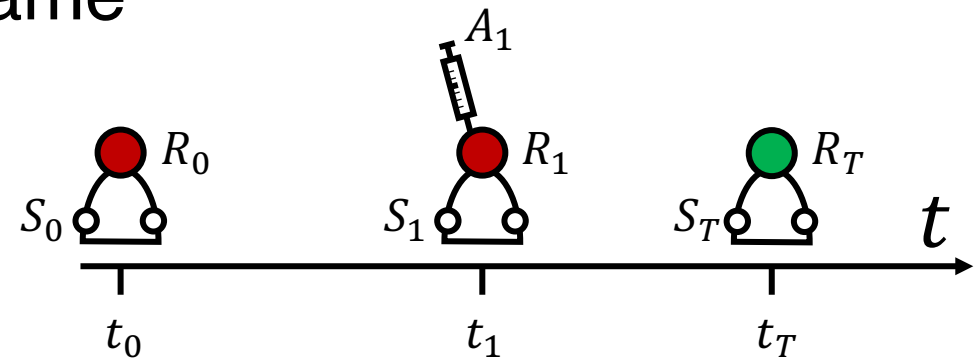
Why clinical reinforcement learning?

- Recent wins from AlphaGo, AlphaStar, and other video games
- Computational gains and methodological advances mean we can model more complex state and action spaces.
- With tools learned so far, we can only make static decisions.
- How can we make dynamic treatment policies?



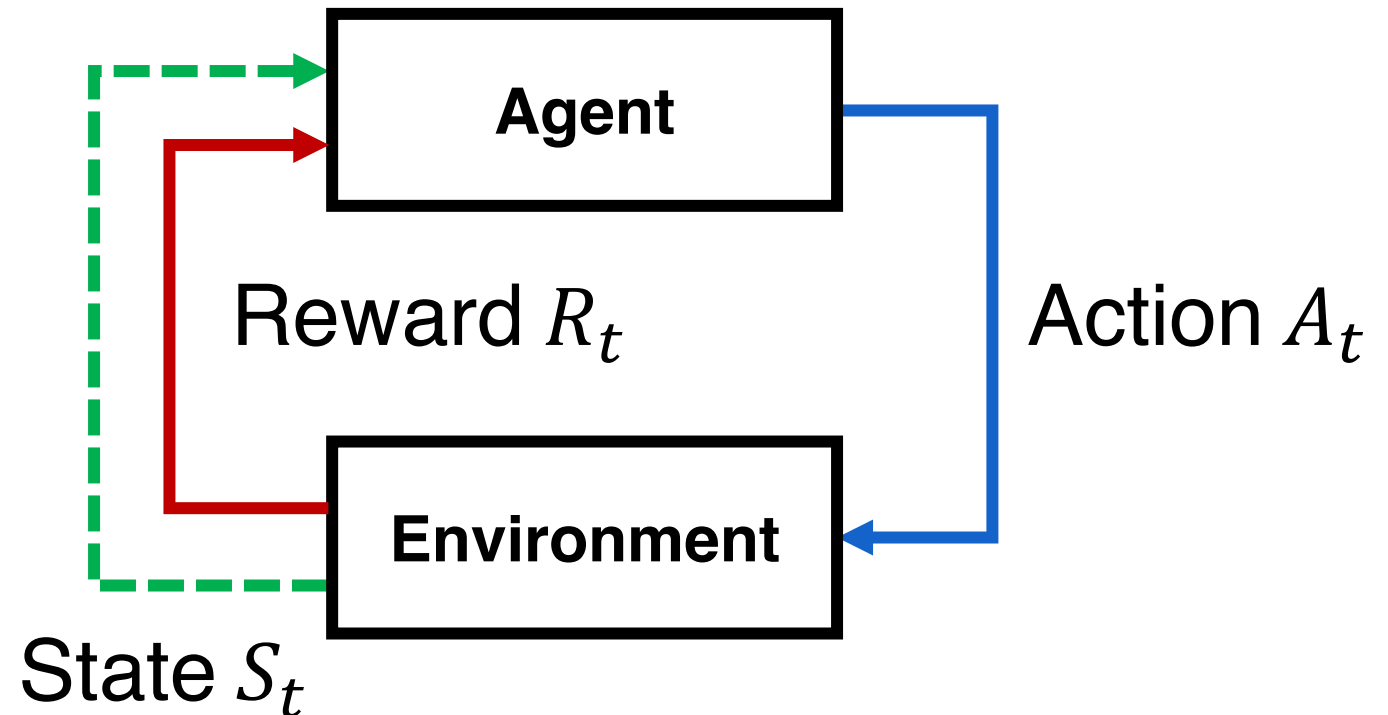
Great! Now let's treat patients

- Patient **state** at time S_t is like the game board
- Medical **treatments** A_t are like the actions
- **Outcomes** R_t are the rewards in the game
- What could **possibly** go wrong?



Decision processes

- An **agent** repeatedly, at times t takes **actions** A_t to receive **rewards** R_t from an **environment**, the **state** S_t of which is (partially) observed



Model-based RL

Value-based RL

Policy-based RL

Want to learn

Tools

Useful for
observational
data?

Relevant
Lectures

Model-based RL

Value-based RL

Policy-based RL

Want to learn

Transitions

$$p(S_t | S_{t-1}, A_{t-1})$$

Value/return

$$p(G_t | S_t, A_t)$$

Policy

$$p(A_t | S_t)$$

Tools

Useful for
observational
data?

Relevant
Lectures

Model-based RL

Value-based RL

Policy-based RL

Want to learn

Transitions

$$p(S_t | S_{t-1}, A_{t-1})$$

Value/return

$$p(G_t | S_t, A_t)$$

Policy

$$p(A_t | S_t)$$

Tools

Q-learning, G-
estimation

Useful for
observational
data?

Yes

Relevant
Lectures

Fredrik
Johansson (L16)

Model-based RL

Value-based RL

Policy-based RL

Want to learn	Transitions $p(S_t \mid S_{t-1}, A_{t-1})$	Value/return $p(G_t \mid S_t, A_t)$	Policy $p(A_t \mid S_t)$
Tools	G-computation, MDP estimation	Q-learning, G- estimation	
Useful for observational data?	Yes	Yes	
Relevant Lectures	Barbara Dickerman (L17)	Fredrik Johansson (L16)	

Model-based RL

Value-based RL

Policy-based RL

Want to learn	Transitions $p(S_t S_{t-1}, A_{t-1})$	Value/return $p(G_t S_t, A_t)$	Policy $p(A_t S_t)$
Tools	G-computation, MDP estimation	Q-learning, G- estimation	REINFORCE, marginal structural models
Useful for observational data?	Yes	Yes	No
Relevant Lectures	Barbara Dickerman (L17)	Fredrik Johansson (L16)	AlphaGo on Netflix

Recap: Fredrik Johansson (Lecture 16)

- “Assign value to a **state-action pair** and maximize over time”
- Similar to **covariate adjustment** (from causal inference) with regression as a moving target
- Solve **Bellman equations** with dynamic programming

Recap: Barbara Dickerman (Lecture 17)

- “Simulate a **weighted average of risks** and then analyze”
- **G-formula** assesses a given policy based on observational data
- **MC sampling** to estimate risk over 10k population, **bootstrap** to get confidence intervals
- Sensitivity analysis for the **confounder of serious medical condition**

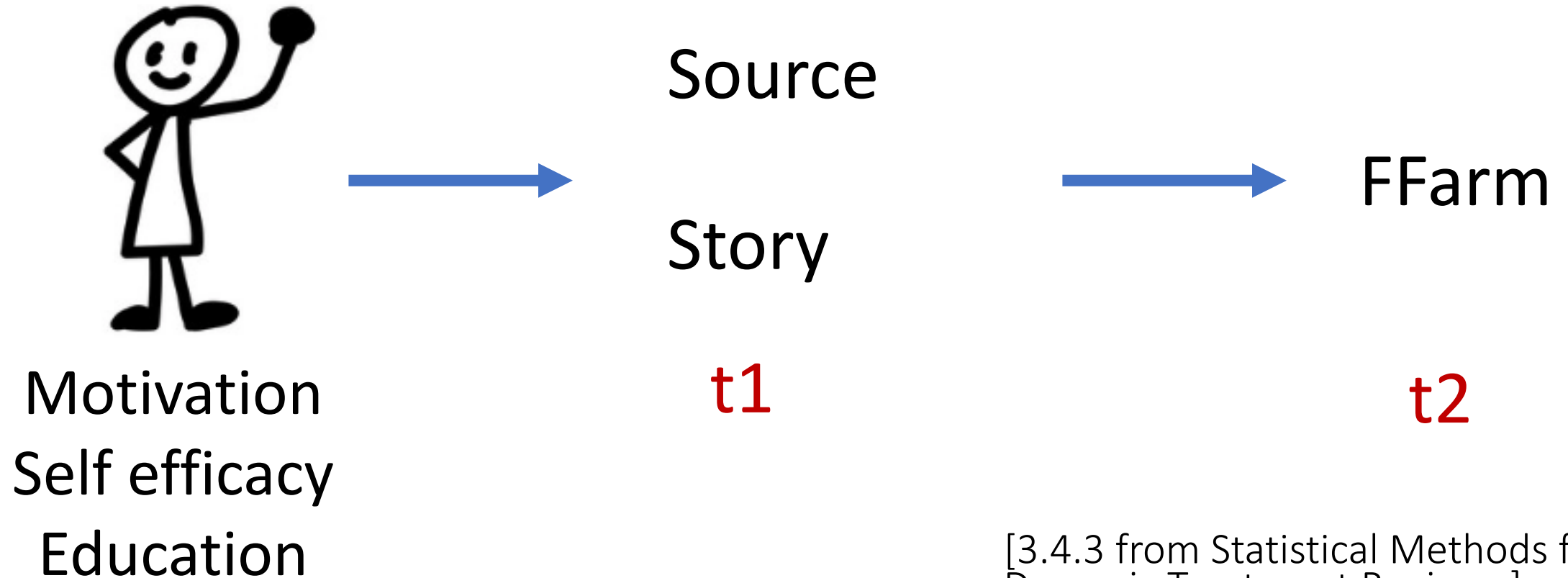
FAQs: Clinical Reinforcement Learning

- **Where does the deep learning come in?**
 - Anywhere we have a probability function, we can estimate the probability density
- **When should we use model-based vs value-based learning?**
 - Model-based when you can build simulator; value-based otherwise
- **Why can we beat the world's best player in Go but not solve a problem of when to use vassopressors for sepsis?**
 - Review lectures 16 and 17.
- **All of these papers and approaches have big limitations.**
 - Yes, yes they do. We make assumptions and then sensitivity analyses.

Agenda for today

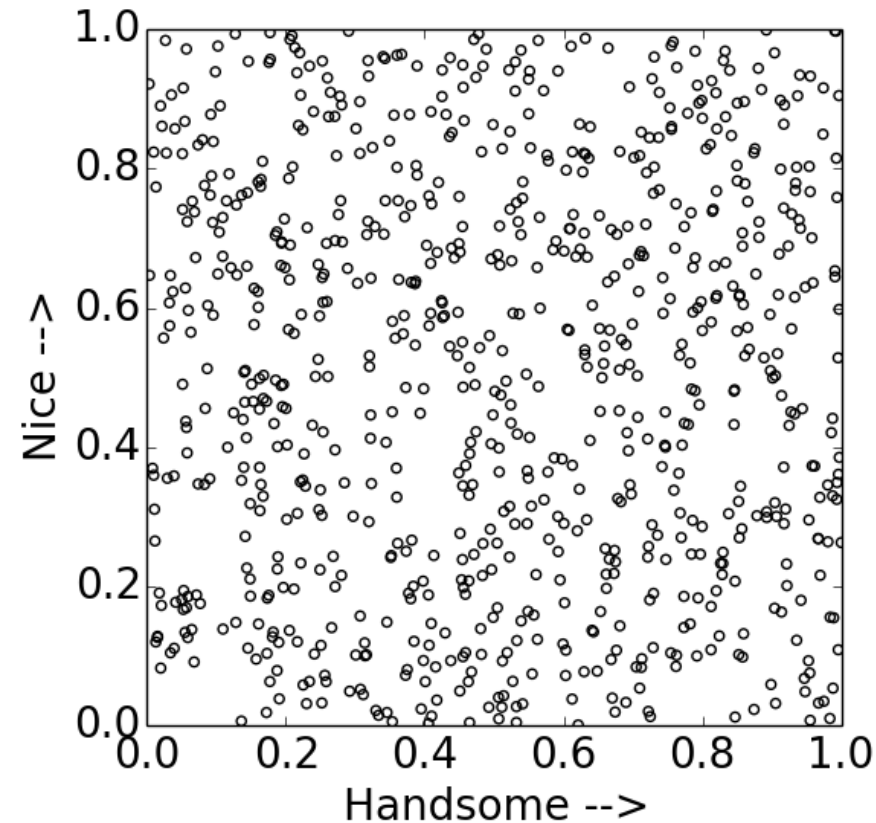
- ~~1. Housekeeping~~
- ~~2. Review lecture material [15 mins]~~
3. Smoking cessation two-stage example [15 mins]
4. Broader discussion [15 mins]

Smoking Cessation: Two steps



[3.4.3 from Statistical Methods for Dynamic Treatment Regimes]

Interlude: Berkman's paradox

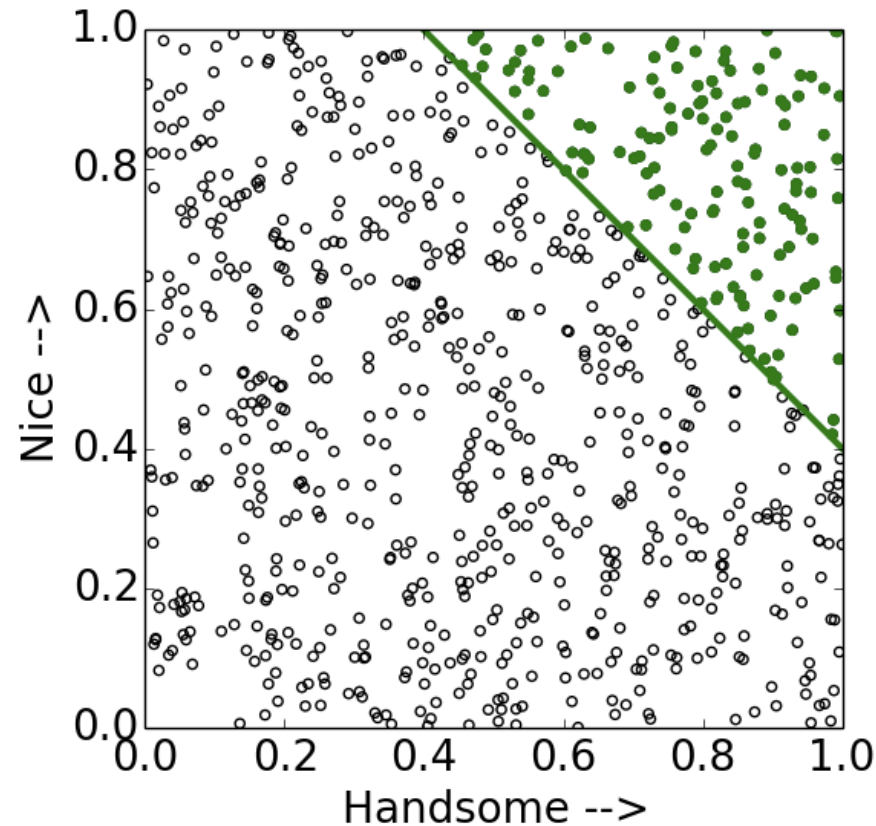


$$P(A | B) = A$$
$$P(B | A) = B$$

Q: Are handsome guys really jerks?

[corysimon.github.io]

Interlude: Berkman's paradox

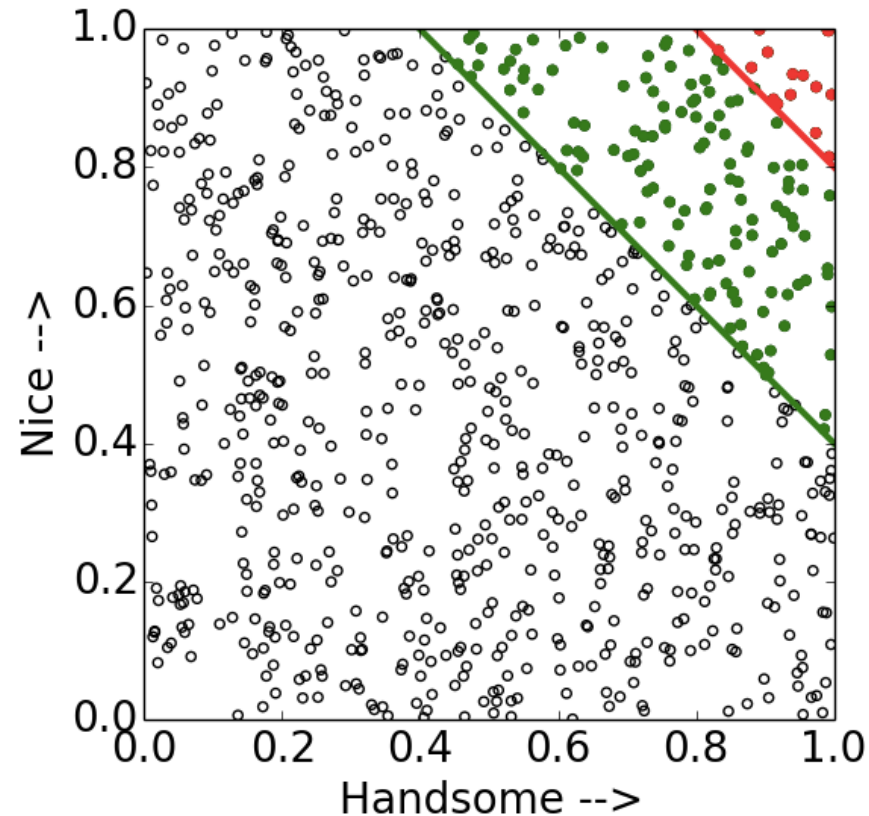


$$P(A | B, A \text{ or } B) < P(A | A \text{ or } B)$$

Dating criterion is subset of field.

[corysimon.github.io]

Interlude: Berkman's paradox



$$P(A | B, A \text{ or } B) < P(A | A \text{ or } B)$$

Therefore, negative correlation in variables in subset despite uncorrelated overall.

[corysimon.github.io]

Back to Smoking Cessation

- Clearly naïve estimation at the end is wrong.
- How can we estimate the impact of the first stage of treatment?
- Assuming linear (with some interaction terms) models, how can set up a two-stage analysis?



Smoking Cessation: Two steps

- Actions:
 - A1: Source x Story
 - A2: FFarm
- Observed:
 - O1: motivation, self efficacy, education at t0
 - O2: quit status, reduction in avg cigarettes smoked, num months not smoked at t1
 - O3: same variables as O2 but at t2
- Outcome:
 - Y: quit status at the end of the study
 - PQ6Quitstatus: quit at stage 1
 - PQ6Quitstatus: quit at stage 2

[3.4.3 from Statistical Methods for Dynamic Treatment Regimes]

1. Fit stage 2 regression ($n = 281$) of FF6Quitstatus using the model:

$$\begin{aligned} \text{FF6Quitstatus} = & \beta_{20} + \beta_{21} \times \text{motivation} + \beta_{22} \times \text{source} \\ & + \beta_{23} \times \text{selfefficacy} + \beta_{24} \times \text{story} \\ & + \beta_{25} \times \text{education} + \beta_{26} \times \text{PQ6Quitstatus} \\ & + \beta_{27} \times \text{source} \times \text{selfefficacy} \\ & + \beta_{28} \times \text{story} \times \text{education} \\ & + \left(\psi_{20} + \psi_{21} \times \text{PQ6Quitstatus} \right) \times \text{FFarm} + \text{error}. \end{aligned}$$

Actions

Outcomes

[3.4.3 from Statistical Methods for
Dynamic Treatment Regimes]

2. Construct the pseudo-outcome (\hat{Y}_1) for the stage 1 regression by plugging in the stage 2 estimates:

$$\begin{aligned}\hat{Y}_1 = & \text{PQ6Quitstatus} + \hat{\beta}_{20} + \hat{\beta}_{21} \times \text{motivation} + \hat{\beta}_{22} \times \text{source} \\ & + \hat{\beta}_{23} \times \text{selfefficacy} + \hat{\beta}_{24} \times \text{story} \\ & + \hat{\beta}_{25} \times \text{education} + \hat{\beta}_{26} \times \text{PQ6Quitstatus} \\ & + \hat{\beta}_{27} \times \text{source} \times \text{selfefficacy} + \hat{\beta}_{28} \times \text{story} \times \text{education} \\ & + \left[\hat{\psi}_{20} + \hat{\psi}_{21} \times \text{PQ6Quitstatus} \right]\end{aligned}$$

3. Fit stage 1 regression ($n = 1,401$) of the pseudo-outcome using a model of the form:

$$\begin{aligned}\hat{Y}_1 = & \beta_{10} + \beta_{11} \times \text{motivation} + \beta_{12} \times \text{selfefficacy} + \beta_{13} \times \text{education} \\ & + \left(\psi_{10}^{(1)} + \psi_{11}^{(1)} \times \text{selfefficacy} \right) \times \text{source} \\ & + \left(\psi_{10}^{(2)} + \psi_{11}^{(2)} \times \text{education} \right) \times \text{story} + \text{error}.\end{aligned}$$

Table 3.1 Regression coefficients and 95 % bootstrap confidence intervals at stage 1 (significant effects are in bold)

Variable	Coefficient	95 % CI
motivation	0.04	(−0.00, 0.08)
selfefficacy	0.03	(0.00, 0.06)
education	−0.01	(−0.07, 0.06)
source	−0.15	(−0.35, 0.06)
source × selfefficacy	0.03	(0.00, 0.06)
story	0.05	(−0.01, 0.11)
story × education	− 0.07	(−0.13, −0.01)

[3.4.3 from Statistical Methods for
Dynamic Treatment Regimes]

Agenda for today

- ~~1. Housekeeping~~
- ~~2. Review lecture material [15 mins]~~
- ~~3. Smoking cessation two-stage example [15 mins]~~
4. Broader discussion [15 mins]

Evaluating Reinforcement Learning Algorithms in Observational Health Settings

Omer Gottesman¹, Fredrik Johansson², Joshua Meier¹, Jack Dent¹,
Donghun Lee¹, Srivatsan Srinivasan¹, Linying Zhang³, Yi Ding³, David
Wihl¹, Xuefeng Peng¹, Jiayu Yao¹, Isaac Lage¹, Christopher Mosch⁴, Li-wei
H. Lehman², Matthieu Komorowski^{5,6}, Aldo Faisal⁷, Leo Anthony Celi^{5,8,9},
David Sontag², and Finale Doshi-Velez¹

¹Paulson School of Engineering and Applied Sciences, Harvard University

²Institute for Medical Engineering and Science, MIT

³T.H. Chan School of Public Health, Harvard University

⁴Department of Statistics, Harvard University

⁵Laboratory for Computational Physiology, Harvard-MIT Health Sciences &
Technology, MIT

⁶Department of Surgery and Cancer, Faculty of Medicine, Imperial College
London

⁷Department of Bioengineering, Imperial College London

⁸Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel
Deaconess Medical Center

⁹MIT Critical Data

<https://arxiv.org/pdf/1805.12298.pdf>

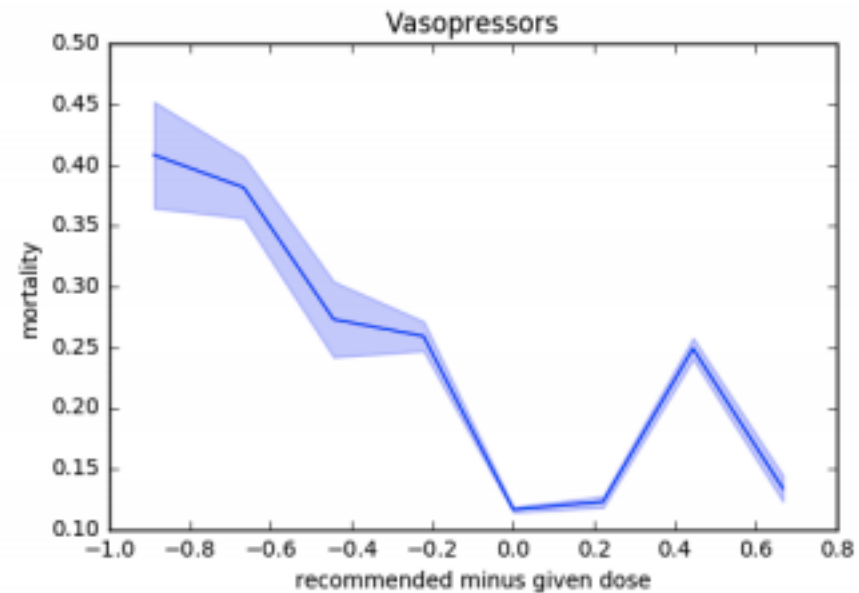
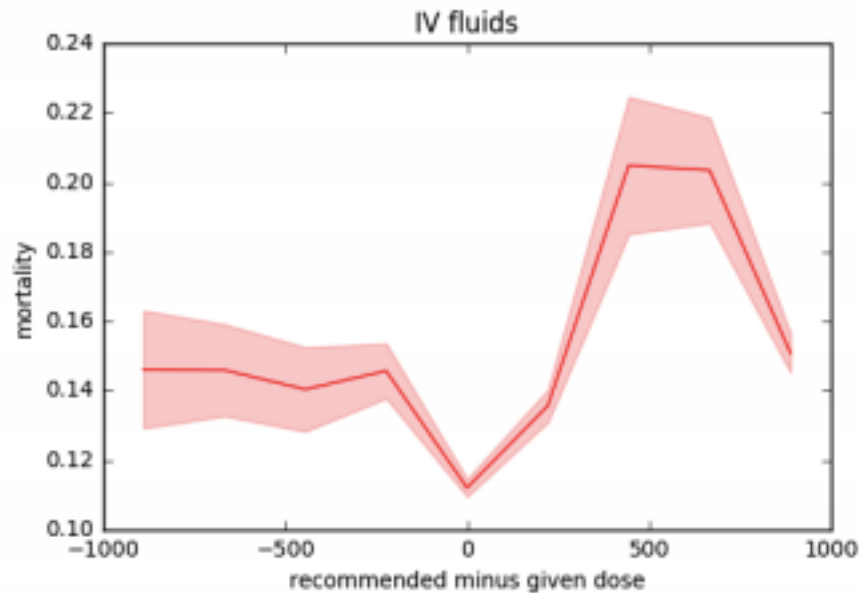
What is the correct representation?

- We want to include all confounders in patient history
- Large feature spaces may make reinforcement learning intractable, so how do we learn a succinct but comprehensive representation?
- Mini experiment:
 - K-means cluster into 100 or 200 “patient types”
 - Find optimal sepsis treatment based on patient type as covariate
 - Repeat 5 times with different clustering initializations
- When 100 types, agreement on optimal treatment was 26%; when 200 types, agreement was 14%

[Gottesman et al, 2018
Nature Medicine]

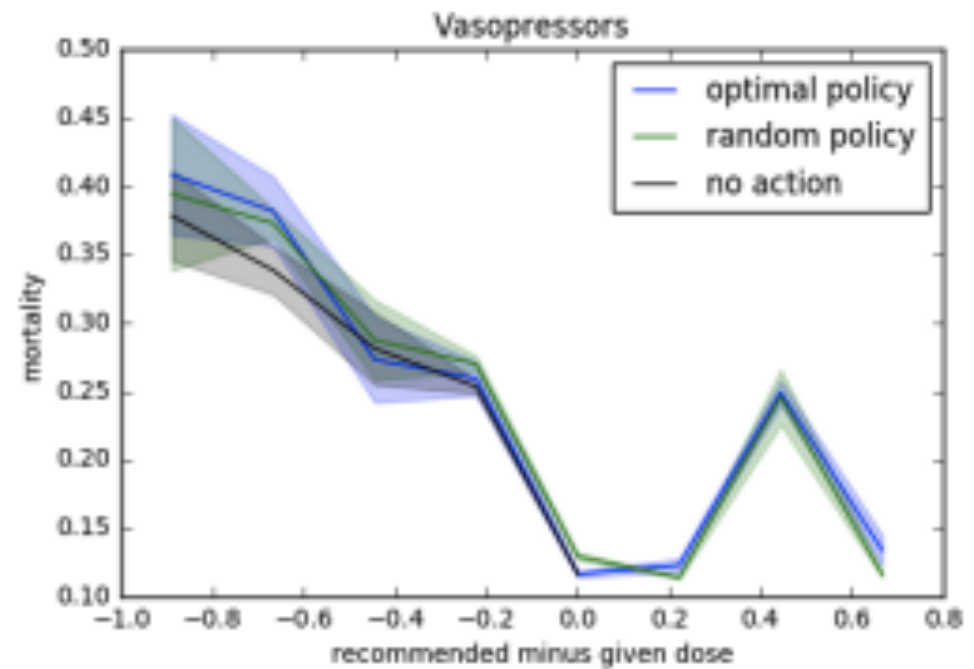
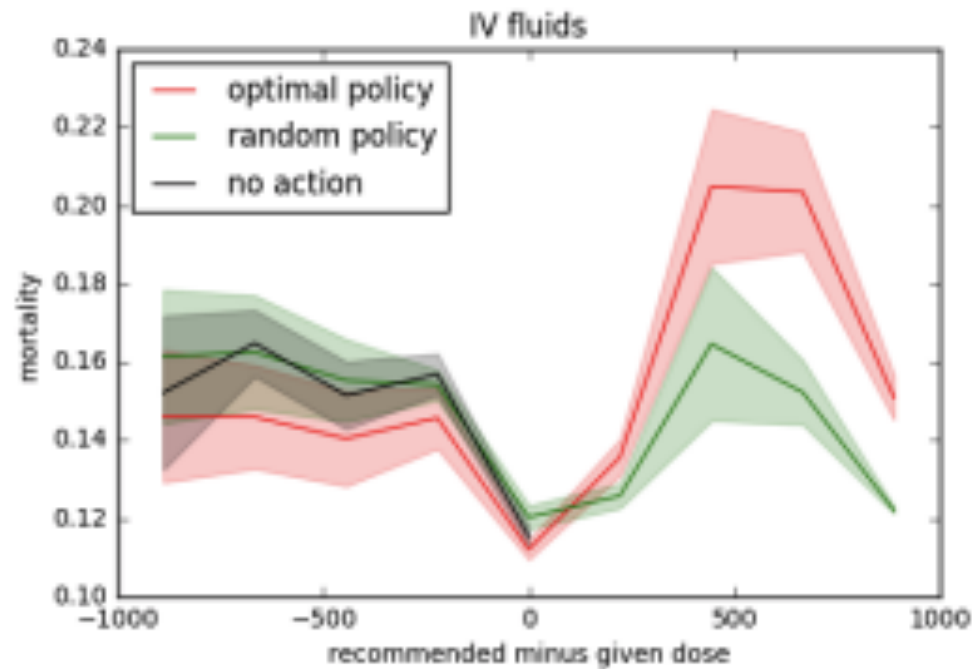
What about ad-hoc evaluation methods?

- U-curve evaluation: Difference between clinician policy and evaluation policy should be correlated with outcome like mortality



[Gottesman et al, 2018
Nature Medicine]

What about ad-hoc evaluation methods?



[Gottesman et al, 2018
Nature Medicine]

Other considerations and recommendations

- Design data collection and representations to make causal conclusions
- Limit yourself to actions and policies similar to physicians
- Be cognizant of effective sample size
- Clearly explain limitations

[Gottesman et al, 2018
Nature Medicine]

Takeaways: Clinical RL

- In theory, RL fits well into **existing clinical workflow**
- **Model-based and value-based** learning work well on observational healthcare data – with well defined actions and states
- Current bleeding edge research works through very **few steps**
- Design analysis to mimic **clinician perspective** and test **sensitivity and robustness**

Have a great weekend!