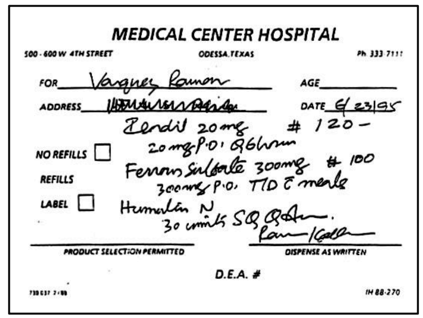
EMORY UNIVERSITY

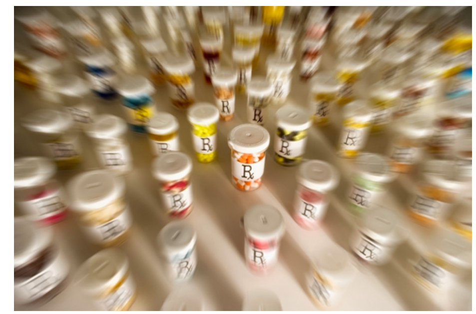
A Deep Deterministic Policy Gradient Approach to Medication Dosing and Surveillance in the ICU

Rongmei Lin, Mohammad Ghassemi, Shamim Nemati

Introduction

- Medication dosing is error prone
  - Complicated situations (e.g., intensive care)
  - Accidental human errors



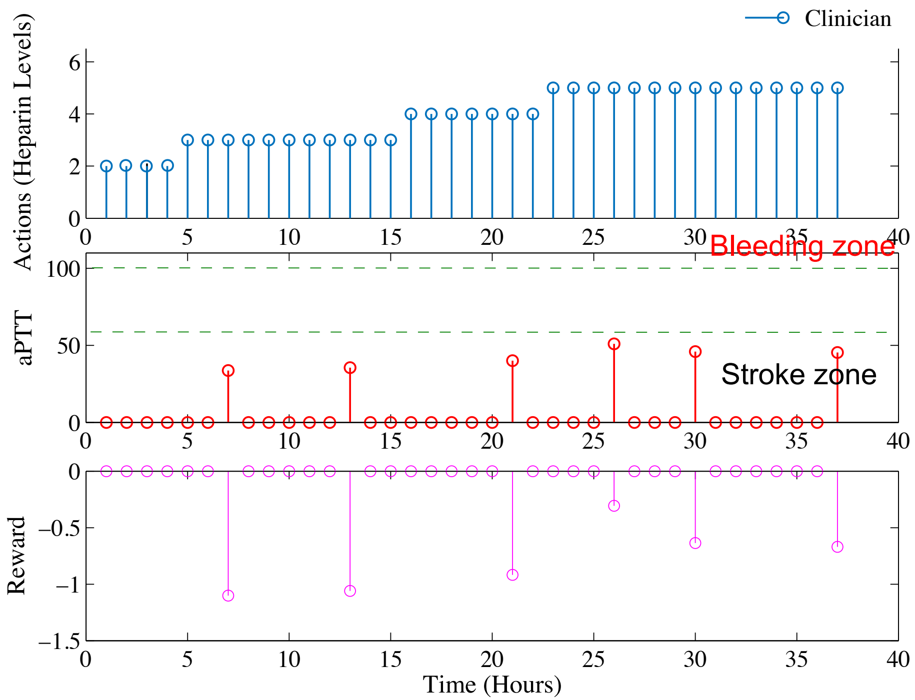


- Preventable Adverse Event (PAE)
  - A lower limit of 210,000 premature deaths associated with PAEs per y ear in the United States [1].
  - The number of severe harm cases are even as many as 10 to 20 time s of these fatal harm cases.

[1] James JT. A new, evidence-based estimate of patient harms associated with hospital care. Journal of patient safety. 2013 Sep 1;9(3):122-8.

An Example of Medication Dosing

- Heparin: An anticoagulant with sensitive therapeutic windows
  - Misdosing heparin can place patients at unnecessary risk and increase length of hospital stay
- aPTT: Activated Partial thromboplastin time (units of seconds)
  - Therapeutic range [60 100]
- ↑Heparin, ↑aPTT, Longer time to form clots



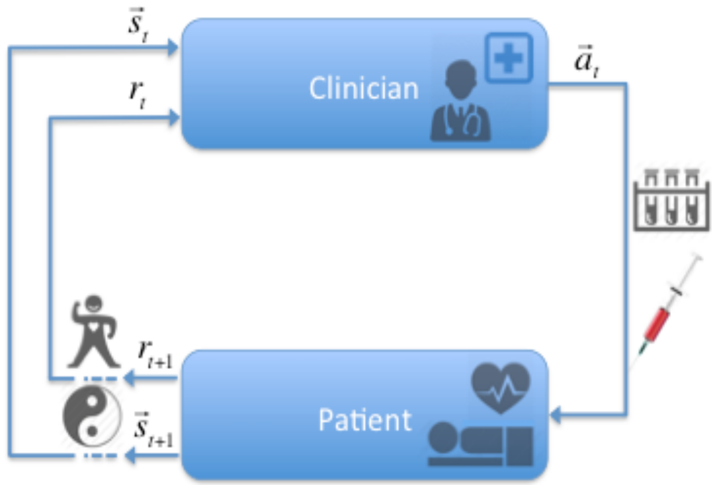
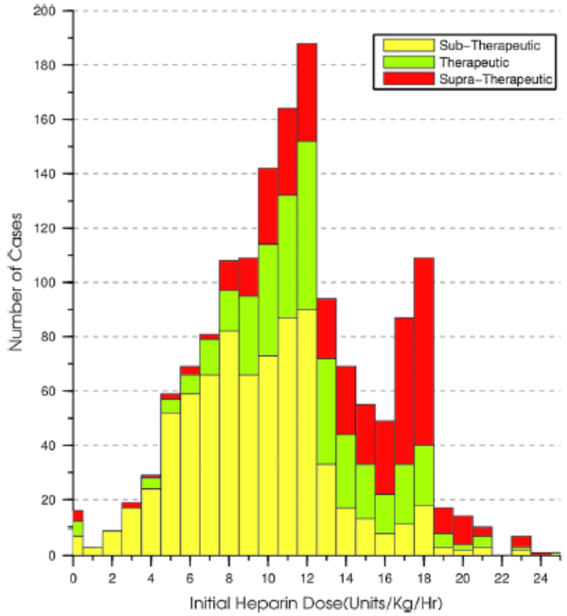


Figure. Generic Sequential Decision Making in medicine

Problem and the Proposed Solution

Mohammad M. Ghassemi et al., Intensive Care Med (2014)



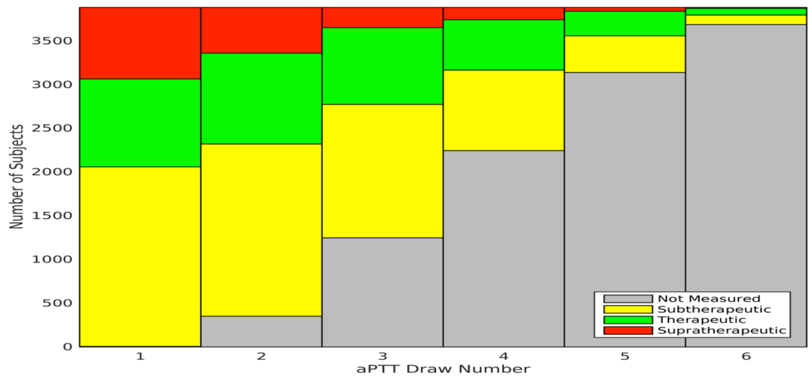


Figure. The distribution of our MIMIC II cohort's aPTT measures over a 48 hours period.

- Utilize the framework of reinforcement learning in continuous state -action spaces to learn a better policy for heparin dosing from obs ervational data.
- Statistically assess if the learned policy is in fact better than the exi sting hospital protocols.

Brief Introduction to Reinforcement Learning

- State (s_t)
- Action (a_t)
- Reward/Reinforcement (r_t)
- Policy (π: s_t -> a_t)

Objective: Find an optimal policy that maximizes the expected (discounted) total reward

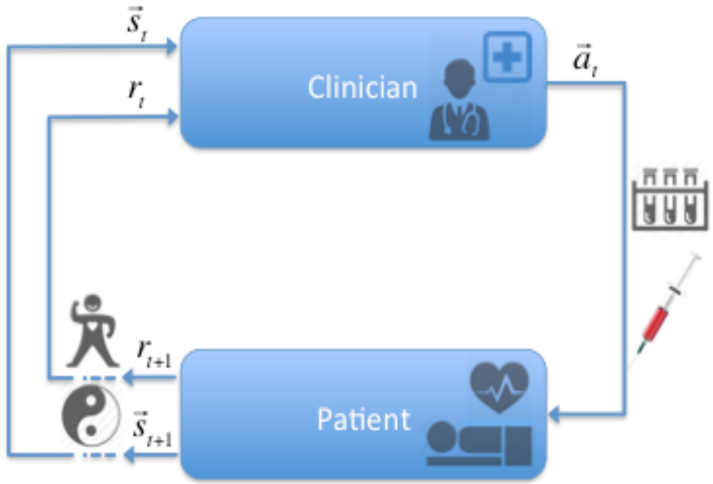


Figure. Generic Sequential Decision Making in medicine

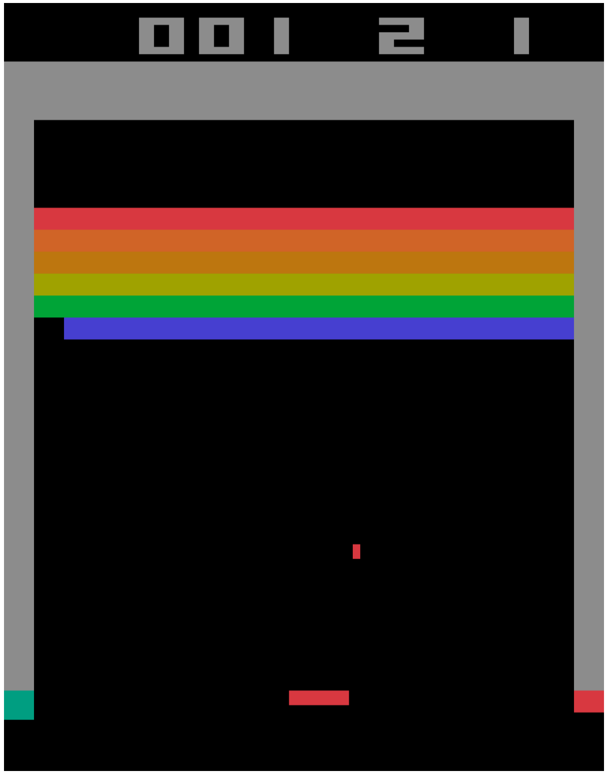
- **Optimal state-action value function:**

$$Q^*(s,a) = \max_\pi E[r_t + \gamma r_{t+1} + \gamma^2 r_{t+2} + \cdots \mid s_t = s, a_t = a, \pi]$$

Brief Introduction to Reinforcement Learning

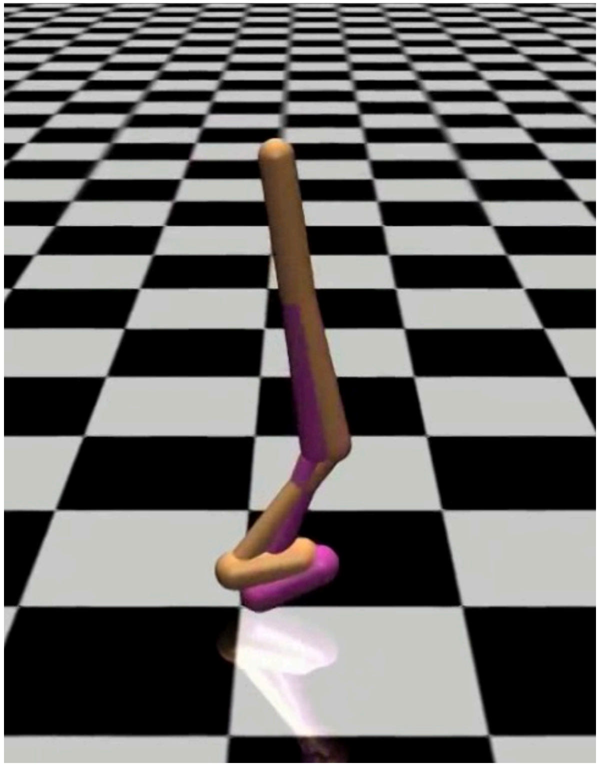
- Value based algorithms (Q-learning, SARSA, etc.)

  discrete action domain



- Policy based algorithm (Reinforce, Actor-Critique, etc.)

  continuous action domain



Data preprocessing

- Heparin problem setting
  - Actions (heparin dosing) occur in continuous domain.
  - Heparin dose can be changed every hour.
  - Monitoring with activated partial thromboplastin time (aPTT).
  - aPTT is measured sparsely (e.g. every 4 hours) → delayed reward
  - Our agent takes the reward and the features (or state) at each hour an d determines the dosing for the next hour.
  - It behaves like a "aPTT GPS" → only makes recommendations
- Data used in this project
  - MIMIC-II (Multiparameter Intelligent Monitoring in Intensive Care- II) da tabase.
  - Emory Hospital Intensive Care Unit clinical data.

Data preprocessing

- **MIMIC-II database description**
  - 25,328 ICU stays between 2001 and 2007.
  - Collected from Beth Israel Deaconess Medical Center in Boston by MIT l ab.
  - 4470 patients with Heparin.
  - 2598 patients with complete Heparin information.

- **Emory ICU data description**
  - Over 30,000 ICU stays between 2013 and 2015.
  - Collected from Emory University Hospital in Atlanta.
  - ICD-9 codes are extracted to provide extra information.
  - 2310 patients with complete Heparin information (no less than 8 hours an d no more than 20 days).

Results and evaluation

➢ Emory ICU data results: example of dosing

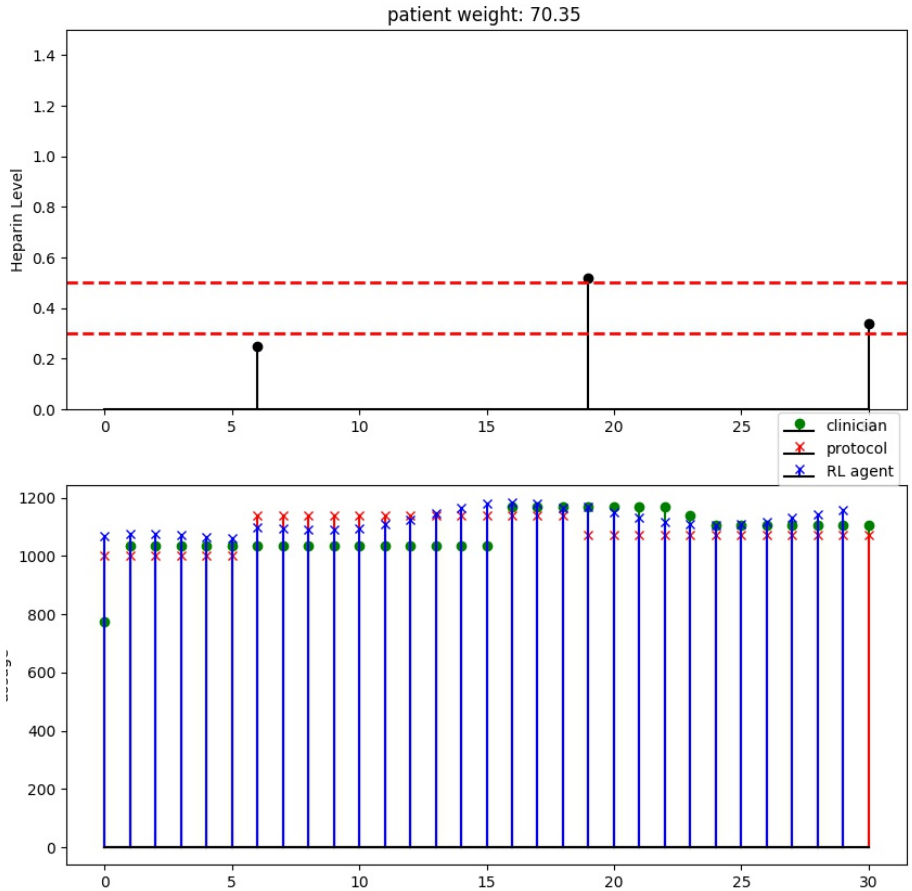
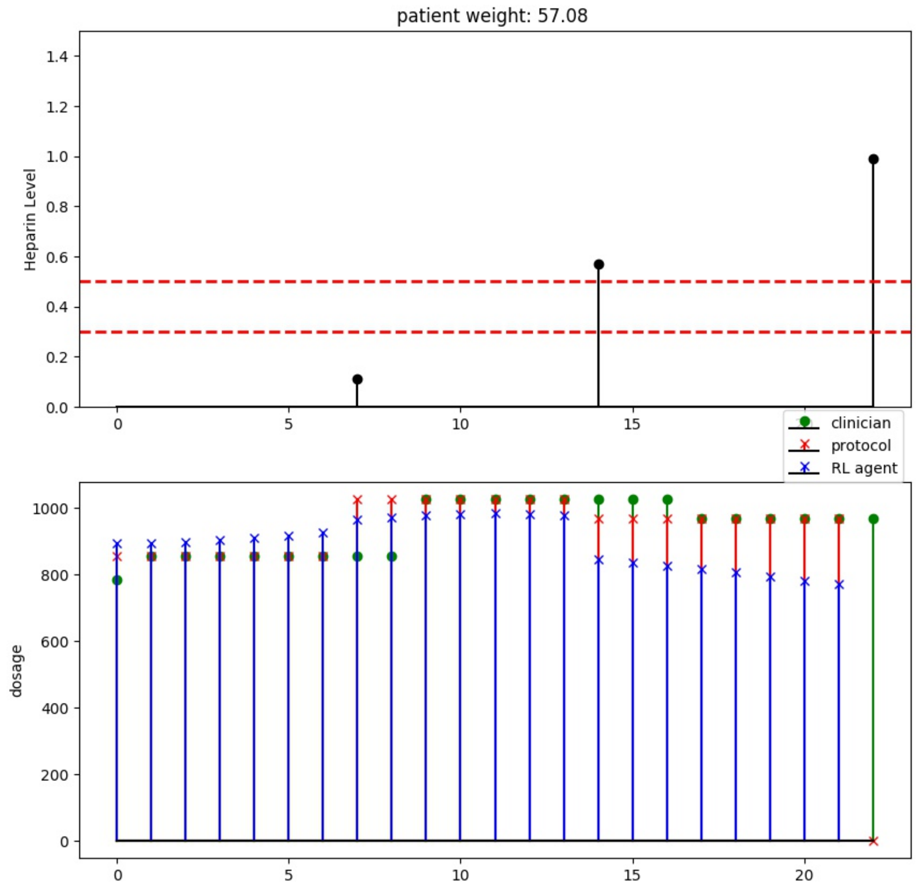


Figure. Results on normal heparin level

Results and evaluation

- Emory ICU data results: example of dosing

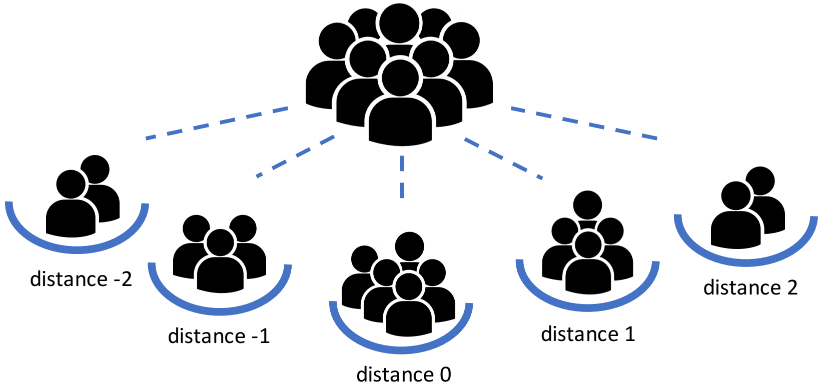


distance from RL policy

Figure. Results on abnormal heparin level

Results and evaluation

- Emory ICU data evaluations

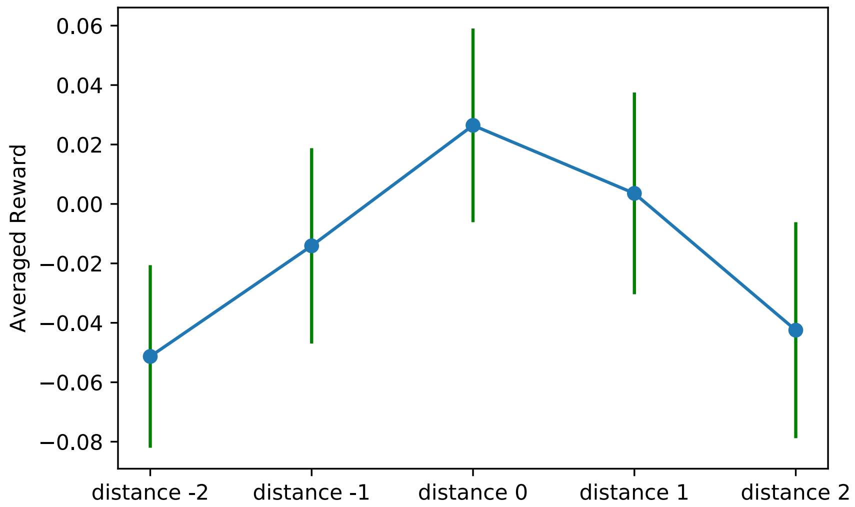


$$Distance = \mathbb{E}_t[Recommendations - Clinicians]$$

Is deviation from optimal RL policy associated with adverse outcomes?

Association with Average Reward

- Emory ICU data evaluations



In the Emory ICU data, It can be seen from figure that the distance 0 class achieved the highest reward. The reward will decrease with the increase of absolute distance.

$$Distance = \mathbb{E}_t[Recommendations - Clinicians]$$

Association with Average Reward

- Emory ICU data evaluations

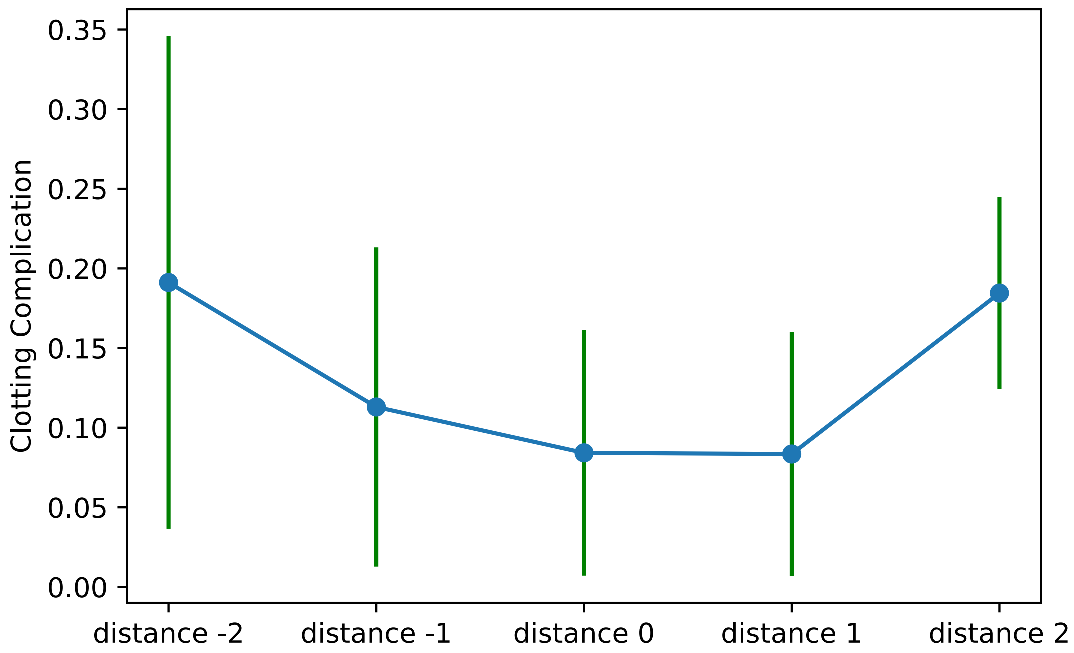
OLS Regression Results

	coef	std err	t	$P > \|t\|$	[95% Conf. Interval]
const	0.0032	0.004	0.847	0.397	[-0.004 , 0.011]
distance	-0.0198	0.006	-3.397	0.001	[-0.030 , -0.008]
hi_clot	0.0074	0.006	1.217	0.224	[-0.005 , 0.019]
hi_blood	0.0048	0.006	0.788	0.431	[-0.007 , 0.017]
weight	0.0004	0.006	0.060	0.952	[-0.011 , 0.012]
age	0.0059	0.006	1.024	0.306	[-0.005 , 0.017]
SOFA	-0.0029	0.006	-0.517	0.605	[-0.014 , 0.008]

- We extract the history of clotting complication (hi_clot) and bleeding complication (hi_blood) from daily ICD-codes of patients.
- Only the distance has significant p-value.
- It is negatively associated with rewards. In other words, the closer a dosing compared with the recommendation, the higher reward it will achieve.

Association with Clotting Complications

- Emory ICU data evaluations



$$Distance = \mathbb{E}_t[Recommendations - Clinicians]$$

Association with Clotting Complications

- Emory ICU data evaluations

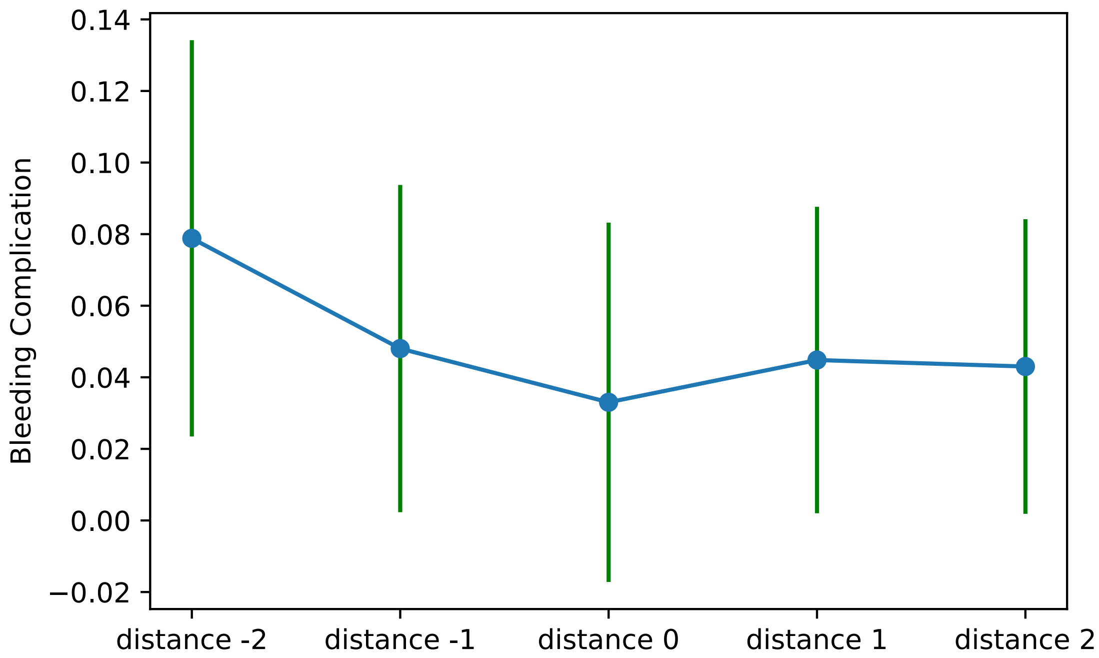
Logit Regression Results

	coef	std err	z	$P > \|z\|$	[95% Conf. Interval]
const	-2.3724	0.076	-31.179	0.000	[-2.522 , -2.223]
distance	0.0146	0.004	3.784	0.001	[0.007 , 0.022]
hi_clot	0.1156	0.064	1.795	0.073	[-0.011 , 0.242]
weight	0.0740	0.069	1.077	0.282	[-0.061 , 0.209]
age	-0.1416	0.073	-1.937	0.053	[-0.285 , 0.002]
SOFA	-0.1059	0.083	-1.275	0.202	[-0.269 , 0.057]

- Distance is the only significant variable with p-value smaller than 0.05.
- With the increase of distance between recommendation and clinicians dosing, the patient might not received enough heparin dosing. As a con sequence, the probability of clotting complication will be higher.

Association with Bleeding Complications

- Emory ICU data evaluations



$$Distance = \mathbb{E}_t[Recommendations - Clinicians]$$

Association with Bleeding Complications

➢ Emory ICU data evaluations

Logit Regression Results

	coef	std err	z	$P > \|z\|$	[95% Conf. Interval]
const	-3.0050	0.099	-30.220	0.000	[-3.200 , -2.810]
distance	-0.0282	0.004	-7.198	0.000	[-0.036 , -0.021]
hi_bleed	-0.1086	0.105	-1.029	0.303	[-0.315 , 0.098]
weight	-0.0112	0.101	-0.111	0.912	[-0.210 , 0.187]
age	0.0027	0.098	0.028	0.978	[-0.190 , 0.196]
SOFA	0.2492	0.074	3.353	0.001	[0.104 , 0.395]

- The first significant variable is distance. A decrease of distance is asso ciated with an increase of bleeding probability.
- The second variable is the coagulation SOFA scores. Low platelets co unts results in a high SOFA score.

Conclusions

- Results and Evaluation
  - We showed that an RL agent can learn reasonable medicatio n dosing policies from observational data (two separate datas ets)
  - After adjusting for confounding factors, deviation from RL pol icy is associated with adverse outcomes
- Limitation on Learning Ability
  - Some useful strategies are not learned by the RL agent, such as rapid turning off of Heparin drip
- Ongoing work
  - Interpretability, via relevance score/relevance propagation
  - Clinical Implementation and prospective validation

Thank you!
Questions?

Brief Introduction to Reinforcement Learning

- Stochastic policy gradient

Classic stochastic policy gradient will consider objective function $J(\theta)$ for state density $\rho^{\pi}(s)$ as follows:

$$J(\theta) = \mathbb{E}_s\Big[\int_a \pi_\theta(s,a)R(s,a)da\Big]$$

The gradient can be calculated by the policy gradient theorem:

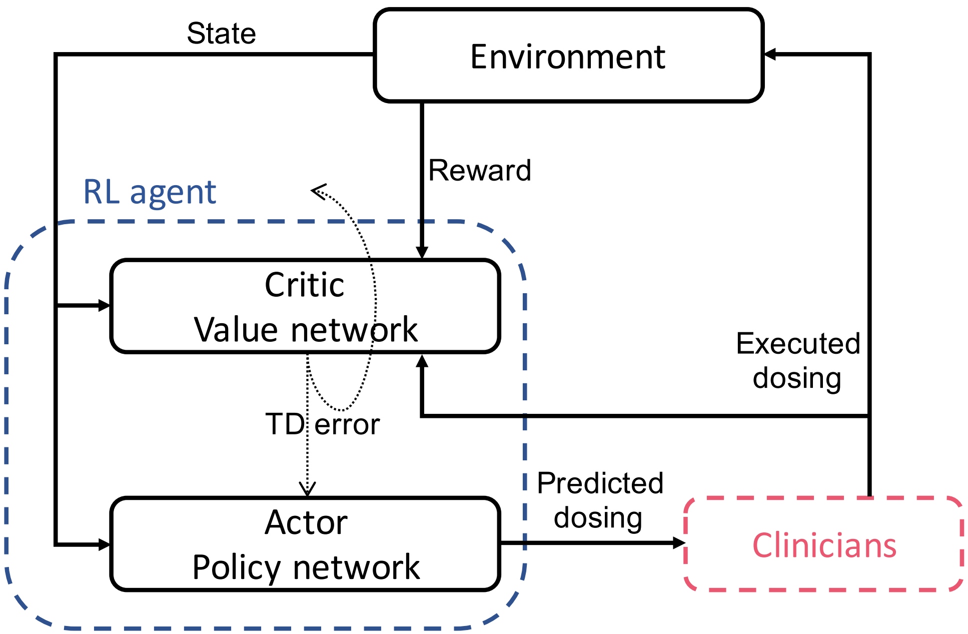
$$\nabla_\theta J(\theta) = \mathbb{E}_s[\int_a \nabla_\theta \pi_\theta(s,a)Q^{\pi}(s,a)da]$$

$$= \mathbb{E}_{s,a}[\nabla_\theta \log \pi_\theta(s,a)Q^{\pi}(s,a)]$$

Experiments

- Proposed framework (clinician-in-the-loop)

1. Instead of generating new episode by interacting with environment, the feasible w ay to implement RL algorith m is analyzing real episode from retrospective clinical d ata.
2. In the sequential decision making process, the agent will predict a action accordi ng to the current state, but the executed action is deter mined by clinicians.



Introduction

- Actor-Critic architecture

  Based on the fundamental theorem, the actor-critic architecture is widely used to represent the components inside policy gradient.

  Actor: adjust parameter of policy

  $$\pi_\theta(s, a)$$

  Critic: estimate action-value

  $$Q^\omega(s, a) \approx Q^\pi(s, a)$$

  The policy gradient update will be:

  $$\Delta\theta = \alpha \nabla_\theta \log \pi_\theta(s, a) Q^\omega(s, a)$$