Designing a Multi-agent RL Algorithm for Improving Post-HCT Medication Adherence via a Digital Intervention

INFORMS 2024, Seattle
October 23rd

Ziping Xu

Postdoctoral Research Fellow
Department of Statistics, Harvard University

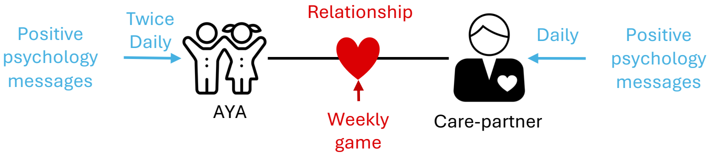
A Mobile Health Clinical Trial

- **Target population**:
  - Adolescents and young adults (AYA) with blood cancer
  - Received hematopoietic stem cell transplantation (HCT)
- **Severe complication**:
  - graft-versus-host disease (GVHD)
  - must take medication twice-daily
- Low medication adherence (60%)!

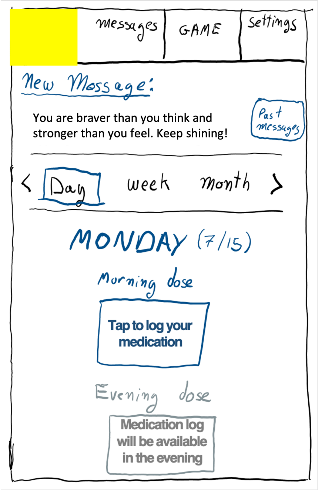
- **ADAPTS-HCT** mobile health clinical trial
  - Deliver digital interventions to improve AYA medication adherence

Dyadic Structure and Intervention Package

- **Dyadic structure**
  - AYAs are vulnerable groups (very sick!)
  - 73% of care-partners (often parents) manage AYA medication
- **Intervention package**
  - Daily positive psychology messages (mitigate psychological distress)
  - Weekly collaborative word-guessing game (improve relationship quality)



Message View and Game View



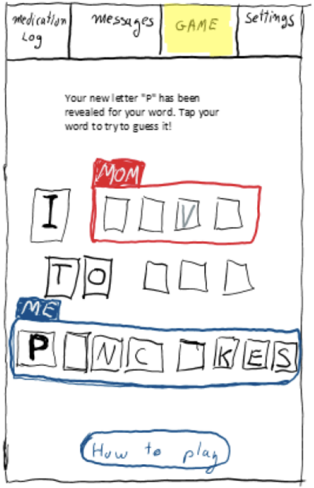
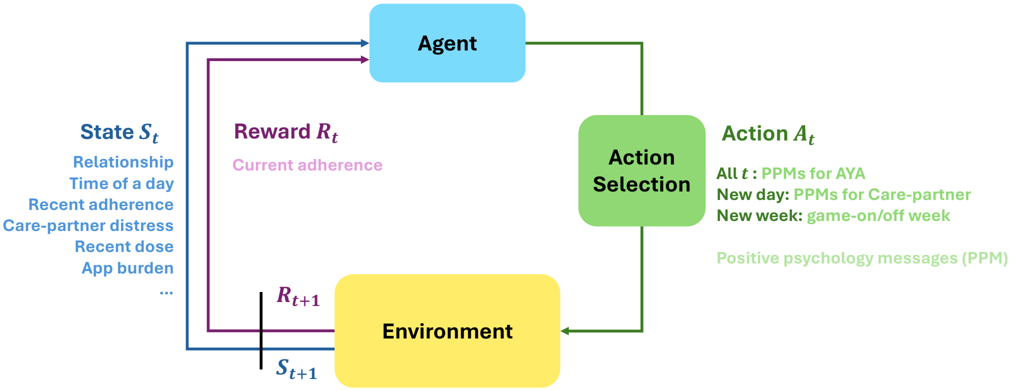


Figure: An example app view used during focus group interviews

Environment Formulation

Each dyad stays for 100 days with $t = 1, \ldots, 200$ (twice-daily) decision times



Heterogeneity in action spaces at different t!

A Hierarchical Multi-Agent Algorithm

Three agents:

- AYA agent (twice-daily): $A_t^{\mathtt{AYA}}$ for all t
- Care-partner agent (daily): $A_d^{\mathtt{CARE}}$ for day d
- Game agent (weekly): $A_w^{\mathtt{GAME}}$ for week w

- Lower level agents include higher level agents' action in their state

Advantages:

- Flexible feature constructions
- Flexible reward designs
- Flexible algorithm designs
- Decentralization
  - One agent does not model other agents' behavior

Challenges

Inherited challenges from the mHealth environment

- Low signal-to-noise ratio
- Low sample size (25 dyads)
- High non-stationarity within each dyad: increasing app burden

Challenges from multi-agent RL:

- Non-stationarity due to the learning of other agents

Leveraging environment structure (or domain knowledge)!

Knowledge on the mechanism

Learning $A_d^{\texttt{CARE}}$ through primary outcomes (adherence) is extremely difficult

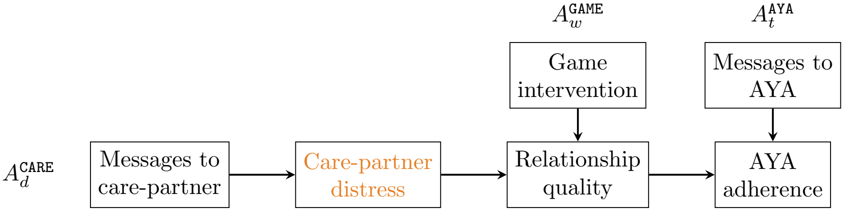


Figure: Causal DAG based on domain knowledge

- The effect from $A_d^{\texttt{CARE}}$ to future AYA adherence is distal
- Other agents' action creates non-stationarity
  - Care-partner agent does not predict what AYA agent will do in the future

Tackle Distal Effect

Solution: construct surrogate rewards through mediators

- $R_d^{\texttt{CARE}}$: negative next day care-partner psychological distress
- $R_w^{\texttt{GAME}}$: next week relationship quality
- $R_t^{\texttt{AYA}}$: time t medication adherence

Evaluation and Base Algorithm

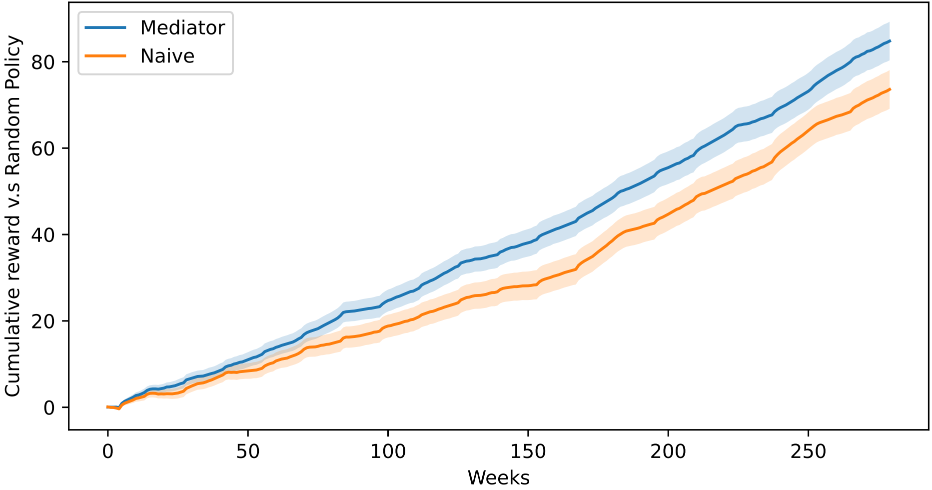
Results evaluation: build a "digital twin" of the target population

- Based on available data + health domain expertise
- Replicate the expected noise structure

Base algorithm:

- Infinite horizon RLSVI for all three agents
- Action centering (or orthogonal estimation) [1, 2]
  - Mititgate non-stationarity

Results



Mediator
Naive
Cumulative reward v.s Random Policy
0
20
40
60
80
0
50
100
150
200
250
Weeks

Theory in Surrogate Rewards

Questions:

- Does surrogate rewards induce the same optimal policy as true rewards?
- What is the benefit of using surrogate reward?

Theory in Surrogate Rewards

Consider linear MDPs (Markov Decision Process) with mediators

- State $S_t \in \mathcal{S}$, action $A_t \in \mathcal{A}$, mediator $M_t \in \mathbb{R}^{d_M}$
- Feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$

Transition dynamic:

$$S_{t+1} \sim \langle \phi(S_t, A_t), \mu_S(\cdot) \rangle$$
$$M_t \sim \Theta\phi(S_t, A_t) + \eta_t \quad \text{and} \quad R_t = \langle M_t, \theta_R \rangle + \epsilon_t$$

- $\Theta \in \mathbb{R}^{d_M \times d}$; $\eta_t \in \mathbb{R}^{d_M}$ and $\epsilon_t \in \mathbb{R}$ are noise
- Property: linear Q-value function
  - $Q^\pi(s, a) = \langle \phi(s, a), \omega^\pi \rangle$ for some $\omega \in \mathbb{R}^d$

MDP Variance Quantity

Variance quantity:

$$\mathbb{V} := \sup_{s,a,\pi} \mathbb{V}^{\pi}(s,a) := \sup_{s,a} \mathrm{Var}\left(R_t + \gamma V^{\pi}(S_{t+1}) \mid S_t = s, A_t = a\right).$$

There exists online algorithm with sample complexity linear in $\sqrt{\mathbb{V}}$ [3]

Reduction in Variance Quantity

Surrogate reward through mediator (if know θ_R):

$$\bar{R}_t = \mathbb{E}[R_t \mid M_t] = M_t^\top \theta_R$$

- Same Q-function: $\bar{Q}^\pi = Q^\pi$
- Constant reduction in variance quantity:

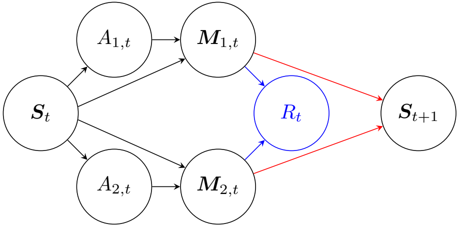
$$\mathbb{V}^\pi(s,a) - \bar{\mathbb{V}}^\pi(s,a) = \mathrm{Var}(\epsilon_t)$$

The reduction is significant if

$$\mathrm{Var}(\epsilon_t) \gg \mathrm{Var}(\eta_t^\top \theta_R)$$

Does the same reward design ($\mathbb{E}[R_t \mid M_t]$) work in the multi-agent setting?

Extension to Multi-agent RL (MARL)



Multi-agent linear MDPs with mediators:

$$M_{i,t} \sim \langle \phi_i(S_t, A_{i,t}), \mu_i(\cdot) \rangle, \tag{1}$$

$$S_{t+1} \sim \sum_i \langle M_{i,t}, \nu_i(\cdot) \rangle \quad \text{and} \quad R_t = \sum_i \langle M_{i,t}, \boldsymbol{\theta}_i \rangle + \epsilon_t \tag{2}$$

- Each agent has their own mediator $M_{i,t}$
- Effects of different mediators are additive

Failure of $\bar{R}_{i,t} = M_{i,t}^\top \theta_i$

The reward design of $\bar{R}_{i,t} = M_{i,t}^\top \theta_i$ is no longer valid

- Think about $\theta_i = 0$: all policies π_i are optimal for reward $\bar{R}_{i,t}$
- However, $A_{i,t} \to S_{t+1} \to M_{j,t+1} \to R_{t+1}$ for $j \neq i$ with $\theta_j \neq 0$

This is the case in ADAPTS-HCT

- Care-partner psychological distress ($M_{2,t}$) has no direct arrow to R_t
- The above design design will give $\bar{R}_{i,t} \equiv 0$ $(\times)$

We must predict the delayed effects of mediators!

Decompose Q-value Function

The surrogate reward must account for the delayed effect onto other mediators

We first show that the value function can indeed be decomposed

Proposition (Decomposing Q-value function)

For any joint policy $\bar{\pi} : \mathcal{S} \mapsto \mathcal{A}^N$*, there exists functions* $f_i^{\bar{\pi}} : \mathcal{S} \times \mathcal{A}_i \mapsto \mathbb{R}$ *such that*

$$Q^{\bar{\pi}}(s, \boldsymbol{a}) = \sum_i f_i^{\bar{\pi}}(s, a_i)$$

A valid Design

Define $\beta_{i,j}^{\bar{\pi}} = \int_{s'} f_j^{\bar{\pi}}(s', \bar{\pi}(s')_j)\nu_i(s')\mathrm{d}s'$: effects of $M_{i,t}$ onto agent j's next-step value

Theorem (A valid design)

Choose the following reward design

$$R_{i,t} = M_{i,t}^\top \left(\theta_i + \gamma \sum_{j \neq i} \beta_{i,j}^{\bar{\pi}^*} \right).$$

The advantage function is consistent

$$f_i^{\bar{\pi}^*}(s, a_i') - f_i^{\bar{\pi}^*}(s, a_i) \equiv$$

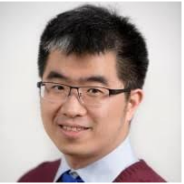
$$\mathbb{E}^{\bar{\pi}^*}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_{i,t} \mid S_t = s, A_{i,t} = a_i'\right] - \mathbb{E}^{\bar{\pi}^*}\left[\sum_{t=1}^{\infty} \gamma^{t-1} R_{i,t} \mid S_t = s, A_{i,t} = a_i\right]$$

Discussion in ADAPTS-HCT

In ADAPTS-HCT, let $i = 1, 2, 3$ be AYA, care-partner, and game agent, respectively

- Care-partner mediator $M_{2,t}$, psychological distress, a scalar
  - $M_{2,t}$ has no direct impact on adherence $\theta_2 = 0$
  - $M_{2,t}$ has a negative impact onto relationship: $\beta_{2,3}^{\bar{\pi}^*} < 0$
  - $M_{2,t}$ has no direct impact onto AYA: $\beta_{2,1}^{\bar{\pi}^*} = 0$
- Thus, $R_{2,t} = -M_{2,t}$ will induce the correct optimal policy

Collaborators



Dr. Pei-yao Hung
University of Michigan

Prof. Susan A. Murphy
Harvard University

Prof. Sung Won Choi
University of Michigan

Dr. Guy Shani
Michigan State University

Prof. Inbal Nahum-Shani
University of Michigan

Prof. Alexandra M Psihogios
Northwestern University

K. Greenewald, A. Tewari, S. Murphy, and P. Klasnja.
Action centered contextual bandits.
Advances in neural information processing systems, 30, 2017.

A. Zhou.
Orthogonalized estimation of difference of q-functions.
arXiv preprint arXiv:2406.08697, 2024.

R. Zhou, Z. Zihan, and S. S. Du.
Sharp variance-dependent bounds in reinforcement learning: Best of both worlds in stochastic and deterministic environments.
In *International Conference on Machine Learning*, pages 42878–42914. PMLR, 2023.