Reinforcement Learning on AYA Dyads to Enhance Medication Adherence

Ziping Xu[1], Hinal Jajal[1], Sung Won Choi[2], Inbal Nahum-Shani[2], Guy Shani[2], Alexandra M. Psihogios[3], Pei-Yao Hung[2], and Susan Murphy[1]

[1]Harvard University, Cambridge, MA, USA
[2]University of Michigan, Ann Arbor, MI, USA
[3]Northwestern University, Feinberg School of Medicine, Chicago, IL, USA

Abstract. Medication adherence is critical for the recovery of adolescents and young adults (AYAs) who have undergone hematopoietic cell transplantation (HCT). However, maintaining adherence is challenging for AYAs after hospital discharge, who experience both individual (e.g. physical and emotional symptoms) and interpersonal barriers (e.g., relational difficulties with their care partner, who is often involved in medication management). To optimize the effectiveness of a three-component digital intervention targeting both members of the dyad as well as their relationship, we propose a novel Multi-Agent Reinforcement Learning (MARL) approach to personalize the delivery of interventions. By incorporating the domain knowledge, the MARL framework, where each agent is responsible for the delivery of one intervention component, allows for faster learning compared with a flattened agent. Evaluation using a dyadic simulator environment, based on real clinical data, shows a significant improvement in medication adherence (approximately 3%) compared to purely random intervention delivery. The effectiveness of this approach will be further evaluated in an upcoming trial.

Keywords: Reinforcement Learning · Dyadic Relationships · Medication Adherence · Digital Health

1 Introduction

For patients who have undergone allogeneic **hematopoietic stem cell transplantation (HCT)**, strict adherence to medication regimens, such as prophylactic immunosuppressant therapy (i.e., calcineurin inhibitors, such as tacrolimus or cyclosporine, taken twice-daily), is crucial for mitigating the risk of acute graft-versus host disease (GVHD) [5]. Acute GVHD occurs in 50-70% of patients following HCT. A lower medication adherence (60%) rate is shown to associate with higher severity of GVHD [9].

The challenges of adherence management are amplified among **adolescents and young adults (AYAs)**, who often demonstrate poorer medication adherence [23,19,14]. For AYAs with cancer, self-management rarely involves the individual alone. Instead, up to 73% of family care partners bear the primary responsibility for managing cancer-related medications for AYAs [22].

Many of these dyads express a desire to move toward sharing these responsibilities with each other [22]. Indeed, for AYAs with chronic health conditions, this developmental period often marks a shift from relying solely on a caregiver to taking more personal responsibility for health care. While shifts in autonomy versus dependence and navigating the ensuing family conflict that can arise from these new dynamics are normative parts of AYA development, difficult family interactions can have a detrimental impact on medication adherence. For example, in a meta-analysis [20], higher level of family conflict and lower levels of family cohesion were significantly associated with worse medication adherence across pediatric illnesses and age groups.

After being discharged from the hospital, both individuals in the dyad face significant emotional and physical challenges as they adjust to managing medication regimens *outside the hospital environment*. For AYAs, the daily challenges of managing complex medication regimens, coping with treatment side effects, coping with stress, and maintaining normal activities in the context of a complex medical regimen can create distress in their home environment. Similarly, care partners must balance caregiving responsibilities with their personal obligations. Those who shoulder heavy caregiving responsibilities at home face higher physical and emotional stressors, which can impede their ability to provide effective care, make sound decisions, and support their AYA's self-care [24].

This need for support outside the inpatient environment motivates the development of interventions that leverage **digital technologies** such as mobile devices [31]. Digital interventions are promising for supporting both AYAs and care partners *at home* on a daily basis, compared to traditional clinical support delivered with limited frequency (e.g., weekly clinical visits for post-HCT AYAs). There is strong heterogeneity across dyads and the users' context are constantly changing, which makes it important to personalize the intervention delivery to optimize the effectiveness of digital interventions. Reinforcement Learning (RL), a machine learning technique that adaptively learns the optimal behavior in an unknown environment to maximize cumulative rewards, is a promising approach for achieving this personalization. RL has been successfully applied in a variety of digital interventions [13,1,28,4].

In this paper, we describe our work in developing an RL algorithm for ADAPTS-HCT [26]. ADAPTS-HCT is a digital intervention for improving medication adherence by AYAs over 100 days after receiving HCT. ADAPTS-HCT integrates three components: (1) twice-daily messages promoting positive emotions for the AYA, (2) daily messages focusing on coping and self-care strategies for the care partner, and (3) a weekly collaborative game for improving their relationship [26]. *We call the three components AYA, care partner, and relationship component*, respectively. Table 1 summarizes these components. The fully developed intervention package will be evaluated in the upcoming clinical trial.

Goals. Our goal is to design an RL algorithm that can personalize the delivery of these interventions to optimize their effectiveness. Given the complexity of the dyadic structure, we identify the following two key challenges:

Table 1: Intervention components in ADAPTS-HCT

Component	Intervention
AYA	Twice-daily positive psychology messages
Care partner	Daily positive psychology messages
Relationship	Weekly collaborative game designed to facilitate positive dyadic interpersonal relationship

1. **Managing multiple intervention decisions across different multi-scales.** There are three intervention components, each requiring decisions to be made at a different time scales. The decision-making occurs twice daily for AYAs, daily for care partners, and weekly for the relationship component. Making decisions on multiple timescales complicates the algorithm design.
2. **Accelerating learning in noisy, data-limited settings.** Observed data in digital intervention deployment is quite noisy [29]. Furthermore, limited data will be available to support in decision making for dyads recruited early in the clinical trial. Additionally, less data is available for learning decisions that occur at slower timescales. These factors necessitate a sample-efficient algorithm that learns faster given limited data.

Contribution. Our contribution is a novel multi-agent RL (MARL) framework involving three RL agents, where each agent is responsible for making decisions for one specific intervention component and operates at the timescale corresponding to its intervention component timescale, which directly addresses challenge (1) about multi-scale decision-making.

The use of MARL decouples the decision processes of different intervention components, thus improving interpretability of the agent model design. This improved interpretability allows us to incorporate domain knowledge into the agent-specific algorithm designs to address challenge (2). To further accelerate learning, we propose a novel **reward engineering** method that learns a less noisy surrogate reward function for each component. Through evaluation in a carefully designed dyadic environment, we demonstrate both the superior performance of our proposed algorithm and strong collaborative behavior among the three agents. Lack of collaboration is often a critical issue in MARL [16].

2 RL Framework and Domain Knowledge

We start with formulating the intervention decision making as an RL problem, where we underscore the challenge in the multiple time scales. HCT treatment is followed by an outpatient 14-weeks twice-daily medication regimen. Decision times within the 14 weeks are denoted by (w, d, t) where $w \in \{1, \dots, 14\}$ is the week index, $d \in \{1, \dots, 7\}$ is the day index, and $t \in \{1, 2\}$ is the decision window within a day.

Primary goal. The primary goal is to make decisions at each decision time t to maximize cumulative sum of medication adherence $\sum_{w=1}^{14} \sum_{d=1}^{7} \sum_{t=1}^{2} R^{\texttt{AYA}}_{w,d,t}$,

Table 2: Summary of variables about each target component

Target	Variable	Type	Description
AYA	$R^{\texttt{AYA}}_{w,d,t}$	binary	Medication adherence at time t on day d in week w
	$A^{\texttt{AYA}}_{w,d,t}$	binary	Intervention at time t on day d in week w
	$B^{\texttt{AYA}}_{w,d,t}$	continuous	App burden at time t on day d in week w
Care partner	$Y^{\texttt{CARE}}_{w,d}$	continuous	Psychological distress on day d in week w
	$A^{\texttt{CARE}}_{w,d}$	binary	Intervention on day d in week w
	$B^{\texttt{CARE}}_{w,d}$	continuous	App burden on day d in week w
Relationship	$Y^{\texttt{REL}}_{w}$	binary	Relationship quality at the end of week w
	$A^{\texttt{REL}}_{w}$	binary	Game intervention at the beginning of week w

where $R^{\texttt{AYA}}_{w,d,t}$ is medication adherence at window t on day d in week w. See Table 2, for selected information that will be collected on the dyad.

Action space. All actions are binary (deliver versus do not deliver intervention content); see Tables 1,2. When the current time $(d = 1, t = 1)$ is the first decision time on the first day of the week, the agent chooses a three-dimensional action corresponding to all three interventions components. If the current time is the first time on a day after the first day of the week $(d > 1, t = 1)$, the agent chooses a two-dimensional action corresponding to only the AYA intervention and the care partner components. At the second time on each day $(t = 2)$ the agent chooses a one-dimensional action corresponding to only the AYA intervention component.

Observation space. Apart from the dynamic action space, we collect observations about different components at different time scales as well; see Table 2. At each time (w, d, t), we collect the current medication adherence and digital intervention burden from the AYA component. In the end of each day d, we collect the psychological distress and digital intervention burden from the care partner component. In the end of a week w, we collect the relationship quality from questionnaires from both the AYA and the care partner.

3 Domain Knowledge through Causal Diagram

Our algorithm design is guided by domain knowledge encoded as the causal diagram in Fig. 1. This diagram describes the scientific team's understanding of the primary causal relationships between the variables in each component listed in Table 2. Note that the causal relationships are likely more complex and direct paths may exist between any two variables. However the scientific team believes that these other paths are likely to be less detectable given the noise in digital intervention data. We summarize the primary pathways that interventions can take to effect the AYA's adherence in the following.

1. **AYA intervention.** The AYA interventions $A^{\texttt{AYA}}_{w,d,t}$ should directly influence the immediate AYA's adherence $R^{\texttt{AYA}}_{w,d,t}$ (black arrows).

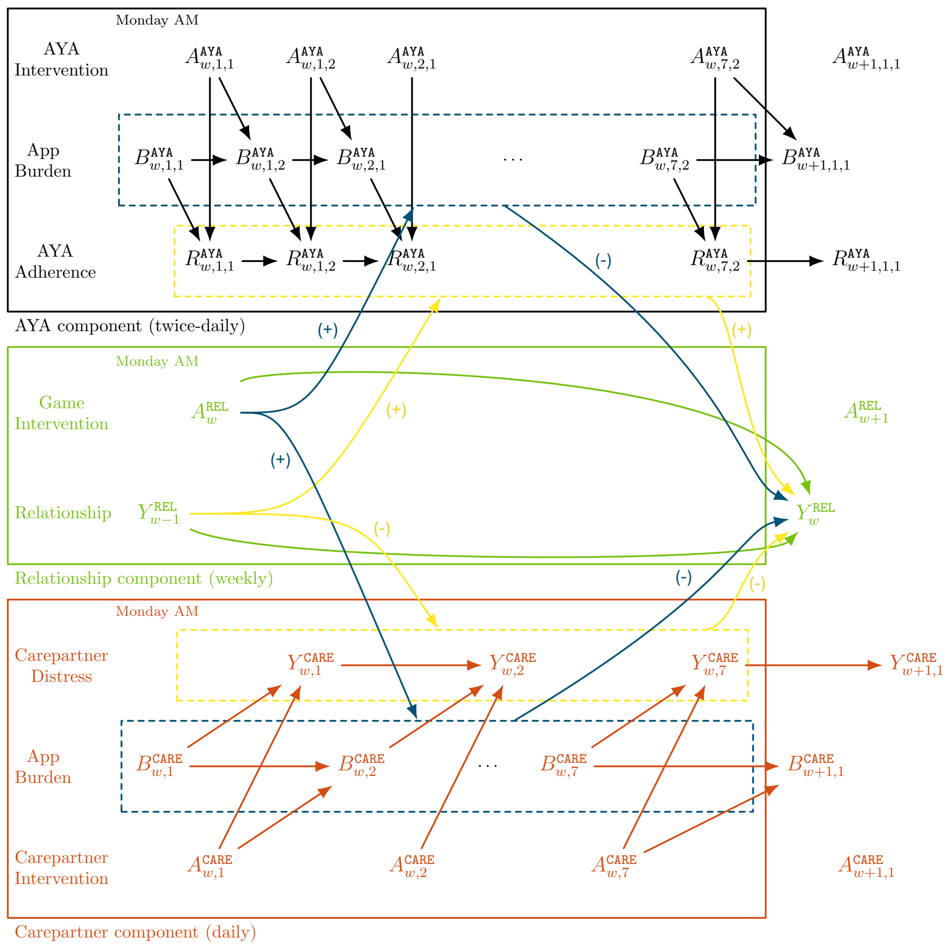


Fig. 1: Causal diagram for ADAPTS-HCT intervention [1]. We categorize the variables into three components: AYA component (marked in **black**), care partner component (marked in red), and relationship component (marked in green). Each component operates at different time scales. Variables in the AYA component evolve on a twice-daily basis, while the care partner component operates on a daily basis. The relationship component operates on a weekly basis. The arrows indicate the direct causal effects.

[1] In the causal inference literature, this is called a causal Directed Acyclic Graph (DAG), a graphical representation of causal relationships among a set of variables [18].

2. **Game intervention.** The game intervention $A_w^{\texttt{REL}}$ has two pathways by which it is expected to effect AYA's adherence. First, $A_w^{\texttt{REL}}$ is expected to increase the AYA's burden $B_{w,d,t}^{\texttt{AYA}}$ throughout the week w. And AYA's burden $B_{w,d,t}^{\texttt{AYA}}$ is expected to decrease the AYA's adherence $R_{w,d,t}^{\texttt{AYA}}$ (blue arrows). Second, the game intervention $A_w^{\texttt{REL}}$ is expected to effect next week AYA's adherence $R_{w+1,d,t}^{\texttt{AYA}}$ by improving the end of the week relationship quality $Y_w^{\texttt{REL}}$ (green arrows).
3. **Care partner intervention.** The care-partner intervention $A_{w,d}^{\texttt{CARE}}$ is expected to effect the AYA's adherence indirectly. First, $A_{w,d}^{\texttt{CARE}}$ should decrease the care partner's psychological distress $Y_{w,d}^{\texttt{CARE}}$, which should increase the end of week relationship quality $Y_w^{\texttt{REL}}$ (yellow arrows). Second, $A_{w,d}^{\texttt{CARE}}$ should increase the care partner's burden $B_{w,d}^{\texttt{CARE}}$, which should decrease the the end of week relationship quality $Y_w^{\texttt{REL}}$ (blue arrows).

We further note that the variables from different components are generally independent conditioned on the bottleneck variables, e.g., the relationship quality that blocks all the paths from the care partner variables to the AYA's adherence. This forms the basis of our multi-agent RL design.

4 Proposed Multi-Agent RL Approach

The conditional independence property observed from Fig. 1 motivates us to design a multi-agent RL (MARL) comprising three agents: the AYA agent, the care partner agent, and the relationship agent. Each makes decisions at different time scales for their own component.

The MARL approach allows us to tailor the agent design choices for each agent to optimize the learning speed. Our base RL algorithm for each agent is Randomized Least Square Value Iteration (RLSVI) [17], which has been proven as stable in deployment of mobile health applications [28,4]. Additionally, we use linear models, which helps in discussions of the algorithm and its parameters with domain scientists.

We construct agent-specific features based on Fig. 1. Specifically, the AYA agent's model uses its own variables ($B_{w,d,t}^{\texttt{AYA}}, R_{w,d,t-1}^{\texttt{AYA}}$) and the variables in the relationship component ($Y_{w-1}^{\texttt{REL}}, A_w^{\texttt{REL}}$). Similarly, the care partner agent uses its own variables, as well as the variables in the relationship component. The relationship agent's model uses $Y_{w-1}^{\texttt{REL}}$, and previous weeks' $B_{w-1,7,2}^{\texttt{AYA}}$, $B_{w-1,7}^{\texttt{CARE}}$, as well as a weighted average of AYA adherence and care partner distress in the past week.

4.1 Surrogate Reward Function Design Through Domain Knowledge

Typical MARL [16] with independent learners considers agents making decisions without communication. In our study, the lack of communication is due to the different time scales–the relationship agent that makes decisions in the beginning of a week may not predict the AYA and care partner agents' decisions throughout the week. This may prevent the agents from **collaborating**. For example, the relationship agent may choose to always intervene so as to improve the relationship quality (the primary goal of the game intervention), which may not be optimal for the AYA's adherence.

Furthermore, the effects of care partner intervention and the game intervention are highly delayed. The game intervention improves end of week relationship quality with a significant delayed effect onto the adherence in the next week. The care partner intervention (positive messages for the care partner) is designed to mitigate the care partner's psychological distress, which only has indirect and delayed effects on the AYA's adherence.

To address the above two issues, we engineer the reward function to account for the delayed effects and across-component effects of each intervention component to promot collaboration. Similar reward engineering in the context of digital interventions is discussed in [30]. Our approach is distinct in that we explore the principles for incorporating domain knowledge to guide the reward function design.

Domain knowledge informed surrogate reward functions. We introduce the surrogate reward functions for the relationship agent and the care partner agent. As informed by Fig. 1, the delayed effect of the game intervention is through the relationship quality and the AYA burden. This motivates us to fit a linear model to predict the sum of medication adherence within week w, $\sum_{d=1}^{7}\sum_{t=1}^{2} R^{\texttt{AYA}}_{w,d,t}$, using $(1, Y^{\texttt{REL}}_{w-1}, B^{\texttt{AYA}}_{w,1,1}, A^{\texttt{REL}}_{w}, A^{\texttt{REL}}_{w} \cdot Y^{\texttt{REL}}_{w-1})$ as the covariates. To account for the delayed effect, we engineer the surrogate reward function for the relationship agent as: $r^{\texttt{REL}}_{w} = (1, Y^{\texttt{REL}}_{w-1}, B^{\texttt{AYA}}_{w,1,1}, A^{\texttt{REL}}_{w}, A^{\texttt{REL}}_{w} \cdot Y^{\texttt{REL}}_{w-1})\boldsymbol{\beta}^{\texttt{REL}} + \max_a(1, Y^{\texttt{REL}}_{w}, B^{\texttt{AYA}}_{w+1,1,1}, a, a \cdot Y^{\texttt{REL}}_{w})\boldsymbol{\beta}^{\texttt{REL}}$, where $\boldsymbol{\beta}^{\texttt{REL}} \in \mathbb{R}^5$ are Bayesian linear regression estimates. [1] The above reward yields a two-step greedy policy, which is a good enough approximation for the total sum of the medication adherence. We opt for a simple, linear model here because the bias trade-off is justified by the faster learning and reduction in noise.

The design of the care partner agent is similar. A key observation is that the end of the week relationship quality blocks all the paths from the care partner variables to the AYA's adherence. Thus, we fit a linear model to predict the end of week relationship quality $Y^{\texttt{REL}}_{w+1}$ using $(1, Y^{\texttt{CARE}}_{w,d}, B^{\texttt{CARE}}_{w,d+1}, Y^{\texttt{REL}}_{w-1}, A^{\texttt{CARE}}_{w,d})$ as covariates. The surrogate reward function is: $r^{\texttt{CARE}}_{w,d} = (1, Y^{\texttt{CARE}}_{w,d}, B^{\texttt{CARE}}_{w,d+1}, Y^{\texttt{REL}}_{w-1}, A^{\texttt{CARE}}_{w,d})\boldsymbol{\beta}^{\texttt{CARE}}$, where $\boldsymbol{\beta}^{\texttt{CARE}} \in \mathbb{R}^5$ are Bayesian linear regression estimates.

5 Results

We simulate a *dyadic environment* to evaluate the performance of the proposed framework. The environment design should replicate the noise level and structure that we expect to encounter in the forthcoming ADAPTS-HCT clinical trial.

Our environment is based on Roadmap 2.0 dataset involving 171 dyads, each consisting of a patient undergone HCT (target person) and a care partner. Roadmap 2.0 provides daily positive psychology interventions to the care partner

[1] We choose the prior mean to reflect our guesses on the sign the coefficients. The prior variance is chosen so the prior mean dominates until around the 5th dyads. The complete prior is provided in Appendix.

only. Roadmap 2.0 collects wearable devices data, for example, physical activity, and self-report data, for example, mood score.

We build upon the environment design in [11], which also uses the Roadmap 2.0 data, but primarily focuses on AYA and relationship intervention component. We extend the environment to include the care partner intervention component. Specifically, we fit a separate multi-variate linear model for each component's outcome (ie., $R^{\texttt{AYA}}_{w,d,t}, Y^{\texttt{CARE}}_{w,d}, Y^{\texttt{REL}}_{w}$) in the dataset. These models simulate the user trajectories under no intervention.

To simulate outcomes under treatments, we impute the treatment effects of the interventions and the effects of app burden, so the induced standard treatment effects (STE) [2] are around 0.15, 0.3, and 0.5. These STEs are commonly seen in behavioral science studies [3]. A complete description and code of the dyadic environment is provided in supplementary material [3].

5.1 Cumulative Adherence Improvement

We simulate 25 dyads, the planned sample size in the upcoming pilot study, by sampling dyads sequentially with replacement from Roadmap 2.0 dataset. Each dyad is simulated for 14 weeks. We implement the following three algorithms: `SingleAgent`, `MultiAgent`, and `MultiAgent+SurrogateRwd`. Here the `SingleAgent` is the algorithm that trains a single agent that outputs all the three types of actions. The `MultiAgent` is the proposed MARL algorithm using the adherence as the learning reward signal for all three agents. The `MultiAgent+SurrogateRwd` is the proposed MARL algorithm using the surrogate reward functions. The full details of the algorithms are described in supplementary material.

We compare the proposed RL algorithms with a random policy, where $P(A^{\texttt{AYA}}_{w,d,t} = 1) = P(A^{\texttt{CARE}}_{w,d} = 1) = P(A^{\texttt{REL}}_{w} = 1) \equiv 0.5$ in terms of the cumulative adherence improvement.

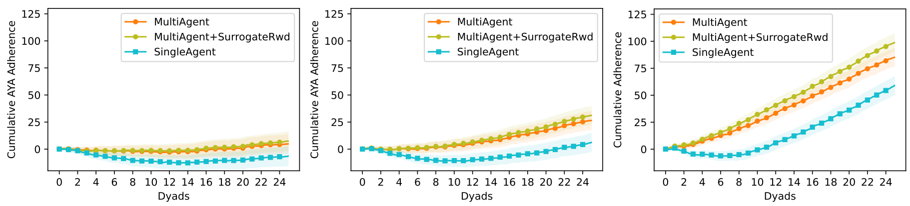
We also observe that all the algorithms can make more significant improvement over the random policy under a higher STE. `SingleAgent` takes longer to learn due to the larger number of parameters compared to `MultiAgent`. We also see an advantage of using surrogate rewards through an increased cumulative adherence at all levels of STE. Notice that for a low STE, the learning is slow, which is intuitive given that the the signal-to-noise ratio is low in such an environment. Additional ablation studies and analysis on the collaborating behavior is provided in the supplementary material.

6 Discussion

In this paper, we propose an MARL algorithm that effectively learns to optimize delivery of the ADAPTS-HCT digital interventions. While this presents a

[2] STE here is defined as the difference in the mean of the primary outcomes under the proposed intervention package and these under no intervention, which is further standardized by the standard deviation under no intervention.

[3] https://github.com/StatisticalReinforcementLearningLab/ADAPTS-HCT-AIME



(a) STE 0.15 (b) STE 0.3 (c) STE 0.5

Fig. 2: Cumulative adherence improvement over the uniform random policy for all three components under dyadic environments with different STEs. The confidence interval is the standard deviation based on 1000 independent runs.

significant step towards preparing for the ADAPTS-HCT clinical trial, several challenges remain to be addressed. First, in the real-clinical trial, the participants are recruited incrementally with significant overlaps, whereas our dyadic environment assumes a simple sequential recruitment. Second, the clinical trial study emphasizes the need for after-study analysis, such as causal inference on treatment effects, which often requires smooth allocation functions [36]. Additionally, there is room to further improve algorithm performance. For example, our proposed algorithm pools data across dyads to reduce learning variance but does not account for heterogeneity across dyads. The algorithm may benefit from a more flexible pooling, e.g., a random effect model.

References

1. Samuel L Battalio, David E Conroy, Walter Dempsey, and et al. Sense2stop: a micro-randomized trial using wearable sensors to optimize a just-in-time-adaptive stress management intervention for smoking relapse prevention. *Contemporary Clinical Trials*, 109:106534, 2021.
2. George EP Box and Norman R Draper. *Empirical model-building and response surfaces*. John Wiley & Sons, 1987.
3. Jacob Cohen. *Statistical power analysis for the behavioral sciences*. routledge, 2013.
4. Susobhan Ghosh, Yongyi Guo, Pei-Yao Hung, Lara Coughlin, Erin Bonar, Inbal Nahum-Shani, Maureen Walton, and Susan Murphy. Miwaves reinforcement learning algorithm. *arXiv preprint arXiv:2408.15076*, 2024.
5. B Gresch, M Kirsch, K Fierz, JP Halter, G Nair, K Denhaerynck, and S De Geest. Medication nonadherence to immunosuppressants after adult allogeneic haematopoietic stem cell transplantation: a multicentre cross-sectional study. *Bone marrow transplantation*, 52(2):304–306, 2017.
6. Qiang He, Xingwei Wang, Yong Zhao, Bo Yi, Xijia Lu, Mingzhou Yang, and Min Huang. Reinforcement-learning-based competitive opinion maximization approach in signed social networks. *IEEE Transactions on Computational Social Systems*, 9(5):1505–1514, 2021.
7. Søren Højsgaard, Ulrich Halekoh, and Jun Yan. The r package geepack for generalized estimating equations. *Journal of statistical software*, 15:1–11, 2006.
8. Chi Jin, Qinghua Liu, Yuanhao Wang, and Tiancheng Yu. V-learning–a simple, efficient, decentralized algorithm for multiagent rl. *arXiv preprint arXiv:2110.14555*, 2021.

9. M Kirsch, A Gotz, JP Halter, U Schanz, G Stussi, F Dobbels, and S De Geest. Differences in health behaviour between recipients of allogeneic haematopoietic sct and the general population: a matched control study. *Bone marrow transplantation*, 49(9):1223–1230, 2014.
10. Barbora Kostalova, Janette Ribaut, Fabienne Dobbels, Sabine Gerull, Katerina Mala-Ladova, Leah L Zullig, and Sabina De Geest. Medication adherence interventions in transplantation lack information on how to implement findings from randomized controlled trials in real-world settings: a systematic review. *Transplantation Reviews*, 36(1):100671, 2022.
11. Shuangning Li, Lluis Salvat Niell, Sung Won Choi, Inbal Nahum-Shani, Guy Shani, and Susan Murphy. Dyadic reinforcement learning. *arXiv preprint arXiv:2308.07843*, 2023.
12. ZhiRu Li, FangYan Lu, JingYun Wu, RuiJie Bao, YuXin Rao, Yun Yang, and Huafen Wang. Usability and effectiveness of ehealth and mhealth interventions that support self-management and health care transition in adolescents and young adults with chronic disease: Systematic review. *Journal of Medical Internet Research*, 26:e56556, 2024.
13. Peng Liao, Kristjan Greenewald, Predrag Klasnja, and Susan Murphy. Personalized heartsteps: A reinforcement learning algorithm for optimizing physical activity. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 4(1):1–22, 2020.
14. Karen S Lyons and Christopher S Lee. The theory of dyadic illness management. *Journal of family nursing*, 24(1):8–28, 2018.
15. Christine J McPherson and Alanna Devereaux. A dyadic examination of patients' and caregivers' attachment orientations and mutually supportive care in cancer caregiving. *Psycho-Oncology*, 33(1):e6256, 2024.
16. Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.
17. Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. In *International Conference on Machine Learning*, pages 2377–2386. PMLR, 2016.
18. Judea Pearl et al. Models, reasoning and inference. *Cambridge, UK: CambridgeUniversityPress*, 19(2):3, 2000.
19. Alexandra M Psihogios, Annisa M Ahmed, Elise R McKelvey, Toto, and et al. Social media to promote treatment adherence among adolescents and young adults with chronic health conditions: A topical review and tiktok application. *Clinical Practice in Pediatric Psychology*, 10(4):440, 2022.
20. Alexandra M Psihogios, Heather Fellmeth, Lisa A Schwartz, and Lamia P Barakat. Family functioning and medical adherence across children and adolescents with chronic health conditions: a meta-analysis. *Journal of Pediatric Psychology*, 44(1):84–97, 2019.
21. Alexandra M Psihogios, Mashfiqui Rabbi, Annisa Ahmed, Elise R McKelvey, Yimei Li, Jean-Philippe Laurenceau, Stephen P Hunger, Linda Fleisher, Ahna LH Pai, Lisa A Schwartz, et al. Understanding adolescent and young adult 6-mercaptopurine adherence and mhealth engagement during cancer treatment: protocol for ecological momentary assessment. *JMIR Research Protocols*, 10(10):e32789, 2021.
22. Alexandra M Psihogios, Lisa A Schwartz, Kylie B Ewing, Bryn Czerniecki, Leslie S Kersun, Ahna LH Pai, Janet A Deatrick, and Lamia P Barakat. Adherence to multiple treatment recommendations in adolescents and young adults with cancer: a mixed methods, multi-informant investigation. *Journal of adolescent and young adult oncology*, 9(6):651–661, 2020.

23. Alexandra M Psihogios, Colleen Stiles-Shields, and Martha Neary. The needle in the haystack: identifying credible mobile health apps for pediatric populations during a pandemic and beyond. *Journal of Pediatric Psychology*, 45(10):1106–1113, 2020.
24. Susan C Reinhard, Barbara Given, Nirvana Huhtala Petlick, and Ann Bemis. Supporting family caregivers in providing care. *Patient safety and quality: An evidence-based handbook for nurses*, 2008.
25. Daniel J Russo, Benjamin Van Roy, Abbas Kazerouni, Ian Osband, Zheng Wen, et al. A tutorial on thompson sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96, 2018.
26. Guy Shani, Sung Won Choi, Susan Murphy, Inbal Billie Nahum-Shani, and et al. Designing a dyadic just-in-time adaptive intervention for medication adherence post-hematopoietic cell transplantation. In *8th Annual Technology in Psychiatry Summit (TIPS)*, Phoenix, AZ, December 2024.
27. Ji Youn Shin, Tammy I Kang, Robert B Noll, and Sung Won Choi. Supporting caregivers of patients with cancer: A summary of technology-mediated interventions and future directions. In *American Society of Clinical Oncology Educational Book. American Society of Clinical Oncology. Annual Meeting*, volume 38, pages 838–849, 2018.
28. Anna L Trella, Kelly W Zhang, Hinal Jajal, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. A deployed online reinforcement learning algorithm in an oral health clinical trial. *arXiv preprint arXiv:2409.02069*, 2024.
29. Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Designing reinforcement learning algorithms for digital interventions: pre-implementation guidelines. *Algorithms*, 15(8):255, 2022.
30. Anna L Trella, Kelly W Zhang, Inbal Nahum-Shani, Vivek Shetty, Finale Doshi-Velez, and Susan A Murphy. Reward design for an online reinforcement learning algorithm supporting oral self-care. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 15724–15730, 2023.
31. Fabio Alexis Rincón Uribe, Maria Fernanda Monteiro Favacho, Paula Marília Nascimento Moura, and et al. Effectiveness of an app-based intervention to improve well-being through cultivating positive thinking and positive emotions in an adult sample: study protocol for a randomized controlled trial. *Frontiers in Psychology*, 14:1200960, 2023.
32. Chiara Visintini, Irene Mansutti, and Alvisa Palese. Medication adherence among allogeneic haematopoietic stem cell transplant recipients: a systematic review. *Cancers*, 15(9):2452, 2023.
33. Chao Wang, Yiming Liu, Xiaofeng Gao, and Guihai Chen. A reinforcement learning model for influence maximization in social networks. In *International Conference on Database Systems for Advanced Applications*, pages 701–709. Springer, 2021.
34. Jianrui Wang, Yitian Hong, Jiali Wang, Jiapeng Xu, Yang Tang, Qing-Long Han, and Jürgen Kurths. Cooperative and competitive multi-agent systems: From optimization to games. *IEEE/CAA Journal of Automatica Sinica*, 9(5):763–783, 2022.
35. Shuxin Yang, Quanming Du, Guixiang Zhu, Jie Cao, Lei Chen, Weiping Qin, and Youquan Wang. Balanced influence maximization in social networks based on deep reinforcement learning. *Neural Networks*, 169:334–351, 2024.
36. Kelly W Zhang, Nowell Closser, Anna L Trella, and Susan A Murphy. Replicable bandits for digital health interventions. *arXiv preprint arXiv:2407.15377*, 2024.
37. A Ziegler and M Vens. Generalized estimating equations. *Methods of information in medicine*, 49(05):421–425, 2010.

A Related Work

Below we summarize the most relevant literature from both the medical lens and the algorithm lens.

RL on social networks. We design and implement RL on dyads that are small social networks in this paper. Existing works on RL on social networks are mostly focused on maximizing social influence or opinion spreading [33,6,35] with large scale social networks in mind. These problems are usually formulated as a constrained Markov Decision Process (CMDP) [35], where the goal is to allocate incentive to maximize the social influence or opinion spreading. Our focuses are on the challenges in the multi-scale decision making and the design of the RL algorithms that incorporate domain knowledge about the social networks. These differences make our algorithm designs unique contributions to the literature.

Dyadic structure in health care. Social relationships between patients and carepartners are proven to be important in many critical health outcomes. Studies have shown that the patient-caregiver dyad functions as a unit, with the well-being and coping strategies of one member significantly impacting the other [27,15]. The quality of this relationship can affect treatment outcomes such as medication adherence [21,10,5], and chronic disease management [32,12].

Multi-agent RL (MARL). Our proposed approach falls into the range of the independent learners in the MARL literature [16]. Previous literature on MARL in a collaborative game focuses on finding the (approximate) Nash equilibrium of the game through interacting with an unknown environment [34,8]. However, in our paper, we emphasize the advantage of MARL in terms of its strong interpretability and being able to make decisions in multiple time-scales.

B Algorithm Details

We provide the complete details of the proposed `MultiAgent+SurrogateRwd` algorithm as well as the baseline `SingleAgent` algorithm.

We first introduce the infinite horizon RLSVI (Randomized Least Squares Value Iteration) algorithm in Alg. 1 [25]. This algorithm is a model-free posterior sampling approach that samples a random value function from its posterior distribution, and the agent acts greedily with respect to the sampled value function. We use the infinite horizon variant of RLSVI, which perturbs the Bayesian regression parameters with a random noise ω' (line 4). We introduce temporal correlation between the current noise w' and the previous noise w to introduce persistence in exploration.

Algorithm 1 Infinite Horizon RLSVI (Inf-RLSVI)

1: Input: discount factor $\gamma \in \mathbb{R}$, previous dataset $\mathcal{D} = (s_i, a_i, r_i)_{i=1}^{n-1} \cup \{s_n\}$, previous perturbation $w \in \mathbb{R}^d$, feature mapping $\phi : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$, previous parameter $\theta \in \mathbb{R}^d$
2: Generate regression matrix and vector

$$X \leftarrow \begin{bmatrix} \phi\left(s_1, a_1\right) \\ \vdots \\ \phi\left(s_{n-1}, a_{n-1}\right) \end{bmatrix} \quad y \leftarrow \begin{bmatrix} r_1 + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_2, \alpha), \theta \rangle \\ \vdots \\ r_{n-1} + \gamma \max_{\alpha \in \mathcal{A}} \langle \phi(s_n, \alpha), \theta \rangle \end{bmatrix}$$

3: Estimate value function

$$\bar{\theta} \leftarrow \frac{1}{\sigma^2} \left(\frac{1}{\sigma^2} X^\top X + \lambda I \right)^{-1} X^\top y \quad \mathbf{\Sigma} \leftarrow \left(\frac{1}{\sigma^2} X^\top X + \lambda I \right)^{-1}$$

4: Sample $w' \sim \mathcal{N}(\gamma w, (1 - \gamma^2)\mathbf{\Sigma})$ and set $\theta' = \bar{\theta} + w'$
5: **Output:** θ' and w'

We use the same hyperparameters $\lambda = 0.75$ and $\sigma = 0.5$ for all the algorithms, which achieves an overall good performance for all the algorithms.

Additional notation. We use w, d, t to denote the week, day, and time of the decision. When we increment the time, we use $w, d, t + 1$ to denote the next decisioin time right after w, d, t, and $w, d, t - 1$ to denote the previous decision time right before w, d, t. Note that if $t = 1$, then $w, d, t - 1$ is the evening decision time of the previous day.

B.1 Single Agent Algorithm

Our `SingleAgent` algorithm runs the RLSVI algorithm in Alg. 1 using the all the obervations available at time w, d, t as the state variable:

$$S_{w,d,t} = \left(Y^{\texttt{CARE}}_{w,d-1}, Y^{\texttt{REL}}_{w-1}, R^{\texttt{AYA}}_{w,d,t-1}, \bar{Y}^{\texttt{AYA}}_{w-1}, \bar{Y}^{\texttt{CARE}}_{w-1}, B^{\texttt{AYA}}_{w,d,t}, B^{\texttt{CARE}}_{w,d,t}, A^{\texttt{CARE}}_{w,d}, A^{\texttt{REL}}_{w}\right) \in \mathbb{R}^9.$$

Here we slightly abuse the notation by using $R^{\texttt{AYA}}_{w,d,t-1}$ to represent the AYA adherence at half-day decision time prior to the current decision time w, d, t. This means that if $t = 1$, a morning decision time, then $R^{\texttt{AYA}}_{w,d,t-1}$ is the AYA adherence at the previous night.

The `SingleAgent` algorithm has the three dimensional action space $\boldsymbol{a} = (a_1, a_2, a_3)^\top \in \{0, 1\}^3$, each entry corresponding to one of the three interventions. The second action a_2 will only be effective on a new day and the third action a_3 will only be effective on a new week. The feature mapping ϕ for the single agent algorithm is defined as

$$\phi(s, \boldsymbol{a}) = (1, s, a_1, a_2, a_3, s \cdot a_1, s \cdot a_2, s \cdot a_3) \in \mathbb{R}^{40}.$$

Algorithm 2 `SingleAgent` Algorithm

Input: discount factor $\gamma = 0.5$
Initialize: $\theta_{1,1,1} = \mathbf{0} \in \mathbb{R}^{40}$; dataset $\mathcal{D}_{1,1,1} = \emptyset$
for $w = 1, 2, \ldots, 14$ **do**
  for $d = 1, 2, \ldots, 7$ **do**
    for $t = 1, 2$ **do**
      Call Algorithm 1 and update $\theta_{w,d,t}$
      $\boldsymbol{a} = \arg\max_{\alpha} \langle \phi(S_{w,d,t}, \alpha), \theta_{w,d,t} \rangle$
      if $t = 1$ and $d = 1$ (New Week) **then**
        Set $A_w^{\texttt{REL}} = \boldsymbol{a}_3$
      end if
      if $t = 1$ (New Day) **then**
        Set $A_{w,d}^{\texttt{CARE}} = \boldsymbol{a}_2$
      end if
      Set $A_{w,d,t}^{\texttt{AYA}} = \boldsymbol{a}_1$
      Environment generates $R_{w,d,t}^{\texttt{AYA}}$ and next state $S_{w,d,t+1}$
      Update $\mathcal{D}_{w,d,t} = \mathcal{D}_{w,d,t-1} \cup \{(S_{w,d,t}, \boldsymbol{a}, R_{w,d,t}^{\texttt{AYA}})\}$
    end for
  end for
end for

B.2 MultiAgent Algorithm

The `MultiAgent` algorithm runs an RLSVI agent for each of the three interventions. We use agent-specific feature mapping $\phi^{\texttt{AYA}}, \phi^{\texttt{CARE}}, \phi^{\texttt{REL}}$ for the AYA, carepartner, and relationship agents, respectively. The state construction and the feature mapping for Q-value function are given by Table 3. The `MultiAgent` algorithm is described in Alg. 3, where the carepartner and the relationship agents learns based on the naive rewards that are the sum of the AYA rewards over the day, and over the week, respectively (line 15 and line 18).

Table 3: State and feature construction for the Q-value function by agent.

Agent	State or Feature Mapping
AYA State	$S_{w,d,t}^{\texttt{AYA}} = \left(R_{w,d,t-1}^{\texttt{AYA}}, B_{w,d,t}^{\texttt{AYA}}, Y_w^{\texttt{REL}}, A_w^{\texttt{REL}}\right) \in \mathbb{R}^4$
AYA Feature	$\phi^{\texttt{AYA}}(s, a) = (1, s, a, s \cdot a) \in \mathbb{R}^{10}$
Carepartner State	$S_{w,d}^{\texttt{CARE}} = \left(Y_{w,d-1}^{\texttt{CARE}}, B_{w,d}^{\texttt{CARE}}, Y_w^{\texttt{REL}}, A_w^{\texttt{REL}}\right) \in \mathbb{R}^4$
Carepartner Feature	$\phi^{\texttt{CARE}}(s, a) = (1, s, a, s \cdot a) \in \mathbb{R}^{10}$
Relationship State	$S_w^{\texttt{REL}} = \left(Y_{w-1}^{\texttt{REL}}, B_{w,1,1}^{\texttt{AYA}}, B_{w,1}^{\texttt{CARE}}, \bar{Y}_{w-1}^{\texttt{AYA}}, \bar{Y}_{w-1}^{\texttt{CARE}}\right) \in \mathbb{R}^5$
Relationship Feature	$\phi^{\texttt{REL}}(s, a) = (1, s, a, s \cdot a) \in \mathbb{R}^{12}$

The `MultiAgent+SurrogateRwd` algorithm is described in Alg. 4. The only difference between the `MultiAgent` and `MultiAgent+SurrogateRwd` is that the later agent optimizes the surrogate reward functions, defined in Equ. (1) and

Algorithm 3 `MultiAgent` Algorithm

Input: discount factor $\gamma^{\texttt{AYA}} = 0.5$, $\gamma^{\texttt{CARE}} = 0.5$, $\gamma^{\texttt{REL}} = 0$
Initialize: $\theta^{\texttt{AYA}}_{1,1,1} = \mathbf{0} \in \mathbb{R}^{10}$; $\theta^{\texttt{CARE}}_{1,1} = \mathbf{0} \in \mathbb{R}^{10}$; $\theta^{\texttt{REL}}_{1} = \mathbf{0} \in \mathbb{R}^{12}$; dataset $\mathcal{D}^{\texttt{AYA}}_{1,1,1} = \emptyset$; $\mathcal{D}^{\texttt{CARE}}_{1,1} = \emptyset$; $\mathcal{D}^{\texttt{REL}}_{1} = \emptyset$
for $w = 1, 2, \ldots, 14$ **do**
  Call Algorithm 1 using $\mathcal{D}^{\texttt{REL}}_{w}, \gamma^{\texttt{REL}}$, and update $\theta^{\texttt{REL}}_{w}$
  Set $A^{\texttt{REL}}_{w} = \arg\max_{\alpha} \langle \phi^{\texttt{REL}}(S^{\texttt{REL}}_{w}, \alpha), \theta^{\texttt{REL}}_{w} \rangle$
  for $d = 1, 2, \ldots, 7$ **do**
    Call Algorithm 1 using $\mathcal{D}^{\texttt{CARE}}_{w,d}, \gamma^{\texttt{CARE}}$, and update $\theta^{\texttt{CARE}}_{w,d}$
    Set $A^{\texttt{CARE}}_{w,d} = \arg\max_{\alpha} \langle \phi^{\texttt{CARE}}(S^{\texttt{CARE}}_{w,d}, \alpha), \theta^{\texttt{CARE}}_{w,d} \rangle$
    for $t = 1, 2$ **do**
      Call Algorithm 1 using $\mathcal{D}^{\texttt{AYA}}_{w,d,t}, \gamma^{\texttt{AYA}}$, and update $\theta^{\texttt{AYA}}_{w,d,t}$
      $A^{\texttt{AYA}}_{w,d,t} = \arg\max_{\alpha} \langle \phi^{\texttt{AYA}}(S^{\texttt{AYA}}_{w,d,t}, \alpha), \theta^{\texttt{AYA}}_{w,d,t} \rangle$
      Environment generates $R^{\texttt{AYA}}_{w,d,t}$ and next state $S_{w,d,t+1}$
      Update $\mathcal{D}^{\texttt{AYA}}_{w,d,t} = \mathcal{D}^{\texttt{AYA}}_{w,d,t-1} \cup \{(S^{\texttt{AYA}}_{w,d,t}, A^{\texttt{AYA}}_{w,d,t}, R^{\texttt{AYA}}_{w,d,t})\}$
    end for
    Compute care-partner reward $R^{\texttt{CARE}}_{w,d} = \sum_{t=1}^{2} R^{\texttt{AYA}}_{w,d,t}/2$
    Update $\mathcal{D}^{\texttt{CARE}}_{w,d} = \mathcal{D}^{\texttt{CARE}}_{w,d-1} \cup \{(S^{\texttt{CARE}}_{w,d}, A^{\texttt{CARE}}_{w,d}, R^{\texttt{CARE}}_{w,d})\}$
  end for
  Compute relationship reward $R^{\texttt{REL}}_{w} = \sum_{d=1}^{7} R^{\texttt{CARE}}_{w,d}/7$
  Update $\mathcal{D}^{\texttt{REL}}_{w} = \mathcal{D}^{\texttt{REL}}_{w-1} \cup \{(S^{\texttt{REL}}_{w}, A^{\texttt{REL}}_{w}, R^{\texttt{REL}}_{w})\}$
end for

Equ. (2), where the coefficients are estimated using Bayesian Ridge Regression, with the prior mean given in Table 4.

$$\begin{aligned} r^{\texttt{REL}}_{w} =& (1, Y^{\texttt{REL}}_{w-1}, B^{\texttt{AYA}}_{w,1,1}, A^{\texttt{REL}}_{w}, A^{\texttt{REL}}_{w} \cdot Y^{\texttt{REL}}_{w-1})\boldsymbol{\beta}^{\texttt{REL}} \\ &+ \max_{a \in \{0,1\}} (1, Y^{\texttt{REL}}_{w}, B^{\texttt{AYA}}_{w+1,1,1}, a, a \cdot Y^{\texttt{REL}}_{w})\boldsymbol{\beta}^{\texttt{REL}}, \end{aligned} \tag{1}$$

$$r^{\texttt{CARE}}_{w,d} = (1, Y^{\texttt{CARE}}_{w,d}, B^{\texttt{CARE}}_{w,d+1}, Y^{\texttt{REL}}_{w-1}, A^{\texttt{CARE}}_{w,d})\boldsymbol{\beta}^{\texttt{CARE}}, \tag{2}$$

Table 4: Prior mean for coefficients in the surrogate reward functions.

Agent	Intercept	$Y^{\texttt{REL}}_{w}$	$B^{\texttt{AYA}}_{w}$	$A^{\texttt{REL}}_{w}$	$A^{\texttt{REL}}_{w} \cdot Y^{\texttt{REL}}_{w}$
$\boldsymbol{\beta}^{\texttt{REL}}$	1	1	-1	-1	0.5
Agent	Intercept	$Y^{\texttt{CARE}}_{w,d}$	$B^{\texttt{CARE}}_{w,d}$	$Y^{\texttt{REL}}_{w-1}$	$A^{\texttt{CARE}}_{w,d}$
$\boldsymbol{\beta}^{\texttt{CARE}}$	1	-1	-1	1	-0.5

Algorithm 4 `MultiAgent+SurrogateRwd` Algorithm

Input: discount factor $\gamma^{\texttt{AYA}} = 0.5$, $\gamma^{\texttt{CARE}} = 0.5$, $\gamma^{\texttt{REL}} = 0$
Initialize: $\theta^{\texttt{AYA}}_{1,1,1} = \mathbf{0} \in \mathbb{R}^{10}$; $\theta^{\texttt{CARE}}_{1,1} = \mathbf{0} \in \mathbb{R}^{10}$; $\theta^{\texttt{REL}}_{1} = \mathbf{0} \in \mathbb{R}^{12}$; dataset $\mathcal{D}^{\texttt{AYA}}_{1,1,1} = \emptyset$; $\mathcal{D}^{\texttt{CARE}}_{1,1} = \emptyset$; $\mathcal{D}^{\texttt{REL}}_{1} = \emptyset$
for $w = 1, 2, \ldots, 14$ **do**
  Call Algorithm 1 using $\mathcal{D}^{\texttt{REL}}_{w}, \gamma^{\texttt{REL}}$, and update $\theta^{\texttt{REL}}_{w}$
  Set $A^{\texttt{REL}}_{w} = \arg\max_{\alpha} \langle \phi^{\texttt{REL}}(S^{\texttt{REL}}_{w}, \alpha), \theta^{\texttt{REL}}_{w} \rangle$
  for $d = 1, 2, \ldots, 7$ **do**
    Call Algorithm 1 using $\mathcal{D}^{\texttt{CARE}}_{w,d}, \gamma^{\texttt{CARE}}$, and update $\theta^{\texttt{CARE}}_{w,d}$
    Set $A^{\texttt{CARE}}_{w,d} = \arg\max_{\alpha} \langle \phi^{\texttt{CARE}}(S^{\texttt{CARE}}_{w,d}, \alpha), \theta^{\texttt{CARE}}_{w,d} \rangle$
    for $t = 1, 2$ **do**
      Call Algorithm 1 using $\mathcal{D}^{\texttt{AYA}}_{w,d,t}, \gamma^{\texttt{AYA}}$, and update $\theta^{\texttt{AYA}}_{w,d,t}$
      $A^{\texttt{AYA}}_{w,d,t} = \arg\max_{\alpha} \langle \phi^{\texttt{AYA}}(S^{\texttt{AYA}}_{w,d,t}, \alpha), \theta^{\texttt{AYA}}_{w,d,t} \rangle$
      Environment generates $R^{\texttt{AYA}}_{w,d,t}$ and next state $S_{w,d,t+1}$
      Update $\mathcal{D}^{\texttt{AYA}}_{w,d,t} = \mathcal{D}^{\texttt{AYA}}_{w,d,t-1} \cup \{(S^{\texttt{AYA}}_{w,d,t}, A^{\texttt{AYA}}_{w,d,t}, R^{\texttt{AYA}}_{w,d,t})\}$
    end for
    Compute care-partner reward $\tilde{R}^{\texttt{CARE}}_{w,d}$ based on Equ. (2)
    Update $\mathcal{D}^{\texttt{CARE}}_{w,d} = \mathcal{D}^{\texttt{CARE}}_{w,d-1} \cup \{(S^{\texttt{CARE}}_{w,d}, A^{\texttt{CARE}}_{w,d}, \tilde{R}^{\texttt{CARE}}_{w,d})\}$
  end for
  Compute relationship reward $\tilde{R}^{\texttt{REL}}_{w}$ based on Equ. (1)
  Update $\mathcal{D}^{\texttt{REL}}_{w} = \mathcal{D}^{\texttt{REL}}_{w-1} \cup \{(S^{\texttt{REL}}_{w}, A^{\texttt{REL}}_{w}, \tilde{R}^{\texttt{REL}}_{w})\}$
end for

C The Dyadic Environment

C.1 Overview of the Simulated Dyadic Environment

We construct a dyadic simulation environment to evaluate the performance of the proposed algorithm. The 1st order goal of the environment design is to replicate the noise level and structure that we expect to encounter in the forthcoming ADAPTS-HCT clinical trial. This noise often encompasses the stochasticity in the state transition of each participant and the heterogeneity across participants.

The environment is based on Roadmap 2.0, a care partner-facing mobile health application that provides daily positive psychology interventions to the care partner only. Roadmap 2.0 involves 171 dyads, each consisting of a patient undergone HCT (target person) and a care partner. Each participant in the dyad had the Roadmap mobile app on their smartphone and wore a Fitbit wrist tracker. The Fitbit wrist tracker recorded physical activity, heart rate, and sleep patterns. Furthermore, each participant was asked to self-report their mood via the Roadmap app every evening. A list of variables in Roadmap 2.0 is reported in Table 6.

Roadmap 2.0 data is suitable for constructing a dyadic environment for developing the RL algorithm for ADAPTS-HCT in that Roadmap 2.0 has the same dyadic structure about the participants–post-HCT cancer patients and their care partner. Moreover, Roadmap 2.0 encompasses some context variables that align with those to be collected in ADAPTS-HCT, for example, the daily self-reported mood score.

Overcoming impoverishment. From the viewpoint of evaluating dyadic RL algorithms, this data is impoverished [29] mainly in two aspects. First, Roadmap 2.0 does not include micro-randomized daily or weekly intervention actions (i.e., whether to send a positive psychology message to the patient/care partner and whether to engage the dyad into a weekly game). Second, it does not include observations on the adherence to the medication–the primary reward signal, as well as other important measurements such as the strength of relationship quality.

To overcome this impoverishment, we construct surrogate variables from the Roadmap 2.0 data to represent the variables intended to be collected in ADAPTS-HCT. A list of substitutes is reported in Table 5. Worthnoting, the substitute for the AYA medication adherence is based on the step count. There is evidence on the association between the step count and the adherence.

We further impute the treatment effects of the intervention actions so the marginal effects after normalization, which we call the standardized treatment effects (STE), are around 0.15, 0.3, and 0.5, corresponding to small, medium, and large effect sizes in typical behavioral science studies.

Constructing the dyadic environment. We follow the environment design in [11], which also uses the Roadmap 2.0 data, but primarily focuses on AYA intervention and relationship intervention. We extend the environment to include the care partner intervention. Specifically, we fit a separate multi-variate linear model for each participant in the dataset with the AR(1) working correlation using the generalized estimating equation (GEE) approach [37,7]. We impute the treatment effects of the intervention actions based on the typical STE around 0.15, 0.3, and 0.5, which completes a generative model for the state transitions. The environment simulates a trial by randomly sampling dyads from the dataset, and simulate their trajectories based on the actions selected by the RL algorithm. The environment details are described in Appendix C. Our experiments primarily focus on the three vanilla testbeds corresponding to the three STEs.

C.2 Using the Roadmap 2.0 Dataset

This section outlines our approach to addressing the limitations of the Roadmap 2.0 dataset, specifically its absence of micro-randomized interventions and reward signals.

To circumvent the lack of interventions, we impute treatment effects that represent the burden of the digital interventions, assuming that frequent notifications diminish both weekly and the daily treatment effects. Based on prior literature, we choose the scale of the treatment effect to be smaller than the baseline effect of features [2].

To address the missing reward signals, we use directly measurable variables in Roadmap 2.0 dataset as proxies to the outcomes we will observe in the real clinical trial. We approximate AYA adherence, $R^{\texttt{AYA}}_{w,d,t}$, using the 12-hourly step count from Roadmap 2.0. Previous work has found the two values to be strongly correlated [Hinal:TODO cite]. Since adherence is a binary signal in the ADAPTS-HCT trial, we discretize step count into a binary variable. Furthermore, we approximate the carepartner's daily psychological distress, $Y_d^{\texttt{CARE}}$, using the daily length of their sleep. Finally, the weekly relationship between the AYA and their carepartner is estimated using the self-reported mood as a surrogate. Specifically, we let $Y_w^{\texttt{REL}} = \mathbb{1}\{\sum_{d=1}^{7} \texttt{Mood}_{w,d}^{\texttt{AYA}} \geq \texttt{Mood}^{\texttt{AYA}}\}\mathbb{1}\{\sum_{d=1}^{7} \texttt{Mood}_{w,d}^{\texttt{CARE}} \geq \texttt{Mood}^{\texttt{CARE}}\}$. Here, $\texttt{Mood}_{w,d}^{\texttt{AYA}}$ is the daily self-reported mood on week w and day d, and $\texttt{Mood}^{\texttt{AYA}}$ is the q-th quantile of the the weekly summed mood across all AYA observations. We choose the quantile level q such that approximately 50% of the dataset satisfies $Y_w^{\texttt{REL}} = 1$.

Table 5 summarizes the main variables and their replacements from the Roadmap 2.0 dataset.

Table 5: Substitutes of the main variables from Roadmap 2.0 dataset.

Variables	Substitutes
AYA adherence	Binary step count $\mathbb{1}\{\texttt{Step}_{w,d,t}^{\texttt{AYA}} \geq \texttt{Step}^{\texttt{AYA}}\}$
Carepartner distress	Carepartner daily length of sleep $\texttt{Sleep}_{w,d}^{\texttt{CARE}}$
Weekly relationship quality	Mood indicator: $\mathbb{1}_{\{\sum_d \texttt{Mood}_{w,d}^{\texttt{CARE}} \geq \texttt{Mood}^{\texttt{CARE}}\}} \mathbb{1}_{\{\sum_d \texttt{Mood}_{w,d}^{\texttt{AYA}} \geq \texttt{Mood}^{\texttt{AYA}}\}}$
Effects of interventions $A_{w,d,t}^{\texttt{AYA}}, A_{w,d}^{\texttt{CARE}}, A_w^{\texttt{REL}}$	Imputed based on domain knowledge
Effects of digital interventions burden $B_{w,d,t}^{\texttt{AYA}}, B_{w,d}^{\texttt{CARE}}$	Imputed based on domain knowledge

Table 6: List of variables in Roadmap 2.0 and the measuring frequencies.

Variables[4]
$\texttt{Step}_{w,d,t}$: twice-daily cumulative step count
$\texttt{Heart}_{w,d,t}$: twice-daily average heart rate
$\texttt{Sleep}_{w,d}$: daily length of sleep
$\texttt{Mood}_{w,d}$: daily self-report mood measurement

C.3 Environment Model Design

We now describe how these surrogate variables are used to build the full environment model. Our approach involves fitting two state transition models for digital intervention burden (AYA and carepartner) and three models for rewards (AYA adherence, carepartner stress, and relationship quality).

For all transition models, we fit the baseline parameters – which represent system dynamics under no intervention – for each dyad using its respective dataset and a generalized estimating equation [7] approach. We impute the remaining parameters using domain knowledge. Further detail on the choice of the coefficients is in Appendix C.4.

Transition models for the AYA component: The digital intervention burden transition for AYA follows a linear model with covariates $(B^{\texttt{AYA}}_{w,d,t}, A^{\texttt{AYA}}_{w,d,t}, A^{\texttt{REL}}_{w})$.

$$B^{\texttt{AYA}}_{w,d,t+1} \sim \theta^{\texttt{AYA}}_0 + \theta^{\texttt{AYA}}_1 B^{\texttt{AYA}}_{w,d,t} + \theta^{\texttt{AYA}}_2 A^{\texttt{AYA}}_{w,d,t} + \theta^{\texttt{AYA}}_3 A^{\texttt{REL}}_{w} + \eta^{\texttt{AYA}}_{w,d,t},$$
$$\text{where } \eta^{\texttt{AYA}}_{w,d,t} \sim \mathcal{N}(0, (\omega^{\texttt{AYA}})^2). \tag{3}$$

For the primary outcome, AYA adherence, we fit a generalized linear model with a sigmoid link function:

$$\begin{aligned}
R^{\texttt{AYA}}_{w,d,t} &\sim \text{Bernoulli}(\text{sigmoid}(P^{\texttt{AYA}}_{w,d,t})),\\
P^{\texttt{AYA}}_{w,d,t} &= (1-M_t)\big(\beta^{\texttt{AYA}}_{0,\texttt{AM}} + \beta^{\texttt{AYA}}_{1,\texttt{AM}} R^{\texttt{AYA}}_{w,d,t-1} + \beta^{\texttt{AYA}}_{2,\texttt{AM}} Y^{\texttt{REL}}_{w-1} + \beta^{\texttt{AYA}}_{3,\texttt{AM}} Y^{\texttt{CARE}}_{w,d-1} + \beta^{\texttt{AYA}}_{4,\texttt{AM}} B^{\texttt{AYA}}_{w,d,t}\\
&\quad + \tau^{\texttt{AYA}}_{0,\texttt{AM}} A^{\texttt{AYA}}_{w,d,t} + \tau^{\texttt{AYA}}_{1,\texttt{AM}} A^{\texttt{AYA}}_{w,d,t} Y^{\texttt{REL}}_{w-1} + \tau^{\texttt{AYA}}_{2,\texttt{AM}} A^{\texttt{AYA}}_{w,d,t} B^{\texttt{AYA}}_{w,d,t}\big)\\
&\quad + M_t\big(\beta^{\texttt{AYA}}_{0,\texttt{PM}} + \beta^{\texttt{AYA}}_{1,\texttt{PM}} R^{\texttt{AYA}}_{w,d,t-1} + \beta^{\texttt{AYA}}_{2,\texttt{PM}} Y^{\texttt{REL}}_{w-1} + \beta^{\texttt{AYA}}_{3,\texttt{PM}} Y^{\texttt{CARE}}_{w,d-1,t} + \beta^{\texttt{AYA}}_{4,\texttt{PM}} B^{\texttt{AYA}}_{w,d,t}\\
&\quad + \tau^{\texttt{AYA}}_{0,\texttt{PM}} A^{\texttt{AYA}}_{w,d,t} + \tau^{\texttt{AYA}}_{1,\texttt{PM}} A^{\texttt{AYA}}_{w,d,t} Y^{\texttt{REL}}_{w-1} + \tau^{\texttt{AYA}}_{2,\texttt{PM}} A^{\texttt{AYA}}_{w,d,t} B^{\texttt{AYA}}_{w,d,t}\big)
\end{aligned} \tag{4}$$

where M_t is a decision window indicator defined as:

$$M_t = \begin{cases} 0 & \text{if } t = 2k-1 \text{ (AM decision window)} \quad \text{for } k = 1, 2, \ldots \\ 1 & \text{if } t = 2k \qquad\quad \text{(PM decision window)} \quad \text{for } k = 1, 2, \ldots \end{cases},$$

Note that we exclude any effect of relationship interventions on AYA adherence as the game is designed without reinforcements and, thus, is not supposed to directly improve adherence.

Transition models for the carepartner component: The digital intervention burden transition for the carepartner is a linear model:

$$B^{\texttt{CARE}}_{w,d+1} = \theta^{\texttt{CARE}}_0 + \theta^{\texttt{CARE}}_1 B^{\texttt{CARE}}_{w,d} + \theta^{\texttt{CARE}}_2 A^{\texttt{CARE}}_{w,d} + \theta^{\texttt{CARE}}_3 A^{\texttt{REL}}_{w} + \eta^{\texttt{CARE}}_{w,d},$$
$$\text{where } \eta^{\texttt{CARE}}_{w,d} \sim \mathcal{N}(0, (\omega^{\texttt{CARE}})^2). \tag{5}$$

For the carepartner's psychological distress level, $R^{\texttt{CARE}}_{d}$, we fit another linear model:

$$\begin{aligned}
Y^{\texttt{CARE}}_{w,d} =& \beta^{\texttt{CARE}}_0 + \beta^{\texttt{CARE}}_1 Y^{\texttt{CARE}}_{w,d-1} + \beta^{\texttt{CARE}}_2 R^{\texttt{AYA}}_{w,d,t-1} + \beta^{\texttt{CARE}}_3 Y^{\texttt{REL}}_{w-1} + \beta^{\texttt{CARE}}_4 B^{\texttt{CARE}}_{w,d} +\\
&\tau^{\texttt{CARE}}_0 A^{\texttt{CARE}}_{w,d} + \tau^{\texttt{CARE}}_1 A^{\texttt{CARE}}_{w,d} Y^{\texttt{REL}}_{w-1} + \tau^{\texttt{CARE}}_2 A^{\texttt{CARE}}_{w,d} B^{\texttt{CARE}}_{w,d} + \epsilon^{\texttt{CARE}}_{w,d}
\end{aligned} \tag{6}$$

where $\epsilon^{\texttt{CARE}}_{w,d} \sim \mathcal{N}(0, (\sigma^{\texttt{CARE}})^2)$. Similar to (4), we do not include relationship intervention $A^{\texttt{REL}}_{w-1}$.

Transition model for the weekly relationship: For the shared component, we only fit a transition model for the reward, which is the weekly relationship quality. Specifically, we fit a generalized linear model with a sigmoid link function:

$$\begin{aligned}
Y^{\texttt{REL}}_{w+1} \sim \text{Bernoulli}(\text{sigmoid}\big(&\beta^{\texttt{REL}}_0 + \beta^{\texttt{REL}}_1 Y^{\texttt{REL}}_{w} + \beta^{\texttt{REL}}_2 \bar{R}^{\texttt{AYA}}_{w} + \beta^{\texttt{REL}}_3 \bar{R}^{\texttt{CARE}}_{w}\\
&+\tau^{\texttt{REL}}_0 A^{\texttt{REL}}_{w} + \tau^{\texttt{REL}}_1 A^{\texttt{REL}}_{w}(B^{\texttt{CARE}}_{w,d} + B^{\texttt{AYA}}_{w,d,t}))\big)
\end{aligned} \tag{7}$$

where $\bar{R}_w^{\texttt{AYA}} = \sum_{d=1}^{7} \sum_{t=1}^{2} \gamma^{14-(7(w-1)+d)+2(t-1)} R_{w,d,t}^{\texttt{AYA}}$ is the exponentially weighted average of adherence within week w, and $\bar{R}_w^{\texttt{CARE}} = \sum_{d=1}^{7} \gamma^{7-d} Y_{w,d}^{\texttt{CARE}}$ is the exponentially weighted average of carepartner distress within week w.

C.4 Selecting Environment Model Parameters

We list all the parameters that must be either imputed based on domain knowledge or estimated from the existing dataset.

1. The baseline transition parameters β's can be estimated directly from the dataset:
   (a) AYA state transition: $\boldsymbol{\beta}_{\texttt{AM}}^{\texttt{AYA}} = (\beta_{0,\texttt{AM}}^{\texttt{AYA}}, \beta_{1,\texttt{AM}}^{\texttt{AYA}}, \beta_{2,\texttt{AM}}^{\texttt{AYA}}, \beta_{3,\texttt{AM}}^{\texttt{AYA}}, \beta_{4,\texttt{AM}}^{\texttt{AYA}})$ and $\boldsymbol{\beta}_{\texttt{PM}}^{\texttt{AYA}} = (\beta_{0,\texttt{PM}}^{\texttt{AYA}}, \beta_{1,\texttt{PM}}^{\texttt{AYA}}, \beta_{2,\texttt{PM}}^{\texttt{AYA}}, \beta_{3,\texttt{PM}}^{\texttt{AYA}}, \beta_{4,\texttt{PM}}^{\texttt{AYA}})$.
   (b) Carepartner state transition: $\boldsymbol{\beta}^{\texttt{CARE}} = (\beta_0^{\texttt{CARE}}, \beta_1^{\texttt{CARE}})$.
   (c) Relationship transition: $\boldsymbol{\beta}^{\texttt{REL}} = (\beta_0^{\texttt{REL}}, \beta_1^{\texttt{REL}}, \beta_2^{\texttt{REL}}, \beta_3^{\texttt{REL}})$.
2. Imputed or tuned based on domain knowledge:
   (a) Burden transitions: coefficients $\boldsymbol{\theta}^{\texttt{AYA}} = (\theta_0^{\texttt{AYA}}, \theta_1^{\texttt{AYA}}, \theta_2^{\texttt{AYA}}, \theta_3^{\texttt{AYA}})$, $\boldsymbol{\theta}^{\texttt{CARE}} = (\theta_0^{\texttt{CARE}}, \theta_1^{\texttt{CARE}}, \theta_2^{\texttt{CARE}}, \theta_3^{\texttt{CARE}})$; burden noise variance $\omega^{\texttt{AYA}}$ and $\omega^{\texttt{CARE}}$.
   (b) Main effects of burden: $\beta_{4,\texttt{AM}}^{\texttt{AYA}}, \beta_{4,\texttt{PM}}^{\texttt{AYA}}$, and $\beta_4^{\texttt{CARE}}$.
   (c) AYA treatment effects: $\{\tau_{i,\texttt{AM}}^{\texttt{AYA}}\}_{i=0}^{2}$, $\{\tau_{i,\texttt{PM}}^{\texttt{AYA}}\}_{i=0}^{2}$ and $\{\sigma_{i,\texttt{AM}}^{\texttt{AYA}}\}_{0=1}^{2}$, $\{\sigma_{i,\texttt{PM}}^{\texttt{AYA}}\}_{i=0}^{2}$.
   (d) Carepartner treatment effects: $\{\tau_i^{\texttt{CARE}}\}_{i=0}^{2}$ and $\{\sigma_i^{\texttt{CARE}}\}_{i=0}^{2}$.
   (e) Relationship treatment effects: $\tau^{\texttt{REL}}$ and $\sigma^{\texttt{REL}}$.

Fitting parameters (1a-d): We estimate the baseline transition parameters under no intervention directly from the Roadmap 2.0 dataset. For the parameters in Equation (4), we have the correspondences $\beta_{i,\texttt{AM}}^{\texttt{AYA}} = \hat{\beta}_{i,\texttt{AM}}^{\texttt{AYA}}$ and $\beta_{i,\texttt{PM}}^{\texttt{AYA}} = \hat{\beta}_{i,\texttt{PM}}^{\texttt{AYA}}$ for $i = 0, 1, \ldots, 3$, where $\hat{\beta}_{i,\texttt{AM}}^{\texttt{AYA}}$ and $\hat{\beta}_{i,\texttt{PM}}^{\texttt{AYA}}$ are fitted coefficients obtained using the generalized estimating equation (GEE) approach. Since we assume that app burden only moderates the effects of AYA interventions without directly influencing adherence, we set $\beta_{4,\texttt{AM}}^{\texttt{AYA}} = \beta_{4,\texttt{PM}}^{\texttt{AYA}} = 0$. Similarly, for parameters in Equation (6), the correspondence is $\beta_i^{\texttt{CARE}} = \hat{\beta}_i^{\texttt{CARE}}$ for $i = 0, \ldots, 3$, and we set $\beta_4^{\texttt{CARE}} = 0$ under the same assumption for carepartner distress. For the relationship quality model in Equation (7), the correspondence is $\beta_i^{\texttt{REL}} = \hat{\beta}_i^{\texttt{REL}}$ for $i = 0, \ldots, 3$. Based on domain knowlege, we also truncate the parameters as follows: $\beta_{2,*}^{\texttt{AYA}} = \max\{0, \hat{\beta}_{2,*}^{\texttt{AYA}}\}$, reflecting the assumption that weekly relationship quality non-negatively influences AYA adherence, $\beta_{3,*}^{\texttt{AYA}} = \min\{0, \hat{\beta}_{3,*}^{\texttt{AYA}}\}$, as carepartner distress is expected to negatively influence adherence, and $\beta_3^{\texttt{REL}} = \min\{0, \hat{\beta}_3^{\texttt{REL}}\}$ as carepartner distress is expected negatively impact relationship quality.

Imputing Burden Transitions (2a): We set $\theta_1^{\texttt{AYA}} = \frac{13}{14}, \quad \theta_1^{\texttt{CARE}} = \frac{6}{7}$, so that the memory of digital burden spans roughly one week for both AYA and carepartner. We choose $\theta_2^{\texttt{AYA}} = 5\,\theta_3^{\texttt{AYA}} = 1, \quad \theta_2^{\texttt{CARE}} = 5\,\theta_3^{\texttt{CARE}} = 1$ so that daily interventions exert five times more burden than the weekly relationship intervention. The intercepts are $\theta_0^{\texttt{AYA}} = 0.2, \quad \theta_0^{\texttt{CARE}} = 0.2$, and chosen so that participants have around a 20% baseline burden even without an intervention. We set $\omega^{\texttt{AYA}} = \omega^{\texttt{CARE}} = 2.4$ to obtain a moderate noise-to-signal ratio, set so that

$(\theta_1^{\texttt{AYA}} + \theta_2^{\texttt{AYA}})/\omega^{\texttt{AYA}} \approx 0.5$. We then truncate burdens at zero and standardize them separately for AYA and carepartner by simulating 10,000 steps with random interventions.

Imputing main effects of app burden (2b). We set $\beta_{4,\texttt{AM}}^{\texttt{AYA}} = \beta_{4,\texttt{PM}}^{\texttt{AYA}} = \beta_4^{\texttt{CARE}} = 0$ based on the assumption that digital app intervention burden does not directly affect AYA adherence or carepartner distress, unless through moderating the digital interventions.

Imputing treatment Effects (2c–2f): Since digital health environments are noisy, treatment terms likely have a lower effect on transitions than the baseline transitions under no intervention. Hence, we scale all intervention effects relative to the baseline effects using a single, global hyperparameter C_{treat}.

For each time of the day (AM or PM), the AYA intervention increases adherence by $\tau_{0,*}^{\texttt{AYA}} = C_{\text{treat}} \left|\beta_{1,*}^{\texttt{AYA}}\right|$, where $* \in \{\text{AM}, \text{PM}\}$ and $\beta_{1,*}$ is the corresponding baseline coefficient estimated from Roadmap 2.0.

We further define $\left|\beta_{1,*}^{\texttt{AYA}}\right|$ and $\tau_{\text{burden},*}^{\texttt{AYA}} = -C_{\text{treat}} \left|\beta_{1,*}^{\texttt{AYA}}\right|$ because the AYA intervention's effectiveness can be increased by good relationship quality and decreased by high digital-intervention burden.

To account for individual heterogeneity across dyads, each treatment-effect coefficient has an associated random effect with variance $\sigma_{0,*}^{\texttt{AYA}} = C_{\text{treat}} \sigma_{\beta_{1,*}^{\texttt{AYA}}}$, where $\sigma_{\beta_{1,*}^{\texttt{AYA}}}$ is the empirical standard deviation across dyads of the baseline coefficient $\beta_{1,*}^{\texttt{AYA}}$.

For carepartner interventions, the main effect on distress is scaled as $\tau_0^{\texttt{CARE}} = -C_{\text{treat}} \left|\beta_1^{\texttt{CARE}}\right|$, where the negative sign is due to the intervention reducing distress. Lastly, the effect of the weekly relationship intervention on improving relationship quality is given by $\tau^{\texttt{REL}} = C_{\text{treat}} \left|\beta_1^{\texttt{REL}}\right|$.

We summarize the imputation design in Table 7.

Tuning $C_{\texttt{Treat}}$: We tune the hyperparameter $C_{\texttt{Treat}}$ such that the standardized treatment effects (STE) are around 0.15, 0.3, and 0.5, where STE is defined as:

$$\text{STE}(C_{\texttt{Treat}}) = \frac{\mathbb{E}\left[\mathbb{E}[\text{CR}(\pi_e^*) \mid e] - \mathbb{E}[\text{CR}(\pi_0, e) \mid e]\right]}{\sqrt{\text{Var}(\mathbb{E}[\text{CR}(\pi_0, e) \mid e])}}, \tag{8}$$

Here, e corresponds to the resulting environment model for dyad e when the hyperparameter is set to be $C_{\texttt{Treat}}$, and π_e^* is the optimal policy for dyad e. $\text{CR}(\pi, e)$ is the cumulative rewards earned by running policy π on dyad e, and π_0 is the reference policy that always chooses action 0 for all components.

Figure 3 plots the value of the hyperparameter versus the STE computed using the optimal policy in the environment defined by the hyperparameter. We outline our approximation of the optimal policy in Appendix C.5. By default, we choose an environment with mediator effect = 1. This results in three dyadic environments, which we summarize in Table 8.

C.5 Optimal Policy Approximation

To approximate the optimal policy, we generate a dataset under a random policy with $P(A_{w,d,t}^{\texttt{AYA}} = 1) = P(A_{w,d}^{\texttt{CARE}} = 1) = P(A_w^{\texttt{REL}} = 1) = 0.5$ and apply offline

Table 7: Summary of burden transition design and treatment effects design.

Burden transition	
Intercept $\theta_0^{\texttt{AYA}}$	Based on domain knowledge $\theta_0^{\texttt{AYA}} = 0.2$
Intervention coefficients $\theta_2^{\texttt{AYA}}, \theta_3^{\texttt{AYA}}$	$\theta_2^{\texttt{AYA}} = 5\theta_3^{\texttt{AYA}} = 1$ (Because relationship intervention produces lower burden)
Noise standard deviation $\omega^{\texttt{AYA}}$	Based on the typical noise-to-signal ratio $\omega^{\texttt{AYA}} = 2.4$
Treatment effect for twice-daily adherence transition (* stands for AM or PM)	
Main effect of AYA intervention $\tau_{0,*}^{\texttt{AYA}}$	Hyper-parameter $\tau_{0,*}^{\texttt{AYA}} = C_{\texttt{Treat}}\lvert\beta_{1,*}^{\texttt{AYA}}\rvert$
Rel. and AYA int. interaction $\tau_{2,*}^{\texttt{AYA}}$	Hyper-parameter $\tau_{1,*}^{\texttt{AYA}} = C_{\texttt{Treat}}\lvert\beta_{1,*}^{\texttt{AYA}}\rvert$
Burden and AYA int. interaction $\tau_{4,*}^{\texttt{AYA}}$	Hyper-parameter $\tau_{2,*}^{\texttt{AYA}} = C_{\texttt{Treat}}\lvert\beta_{1,*}^{\texttt{AYA}}\rvert$
Random treatment variance $\{\sigma_{i,*}^{\texttt{AYA}}\}_{i=0}^{5}$	Scales with the variance of $\beta_{1,*}^{\texttt{AYA}}$: $\sigma_{i,*}^{\texttt{AYA}} = \tau_{i,*}^{\texttt{AYA}} \cdot \sigma_{\beta_{1,*}^{\texttt{AYA}}} / \lvert\beta_{1,*}^{\texttt{AYA}}\rvert$
Treatment effect for weekly relationship transition	
Main effect of relationship int. $\tau^{\texttt{REL}}$	Hyper-parameter $\tau^{\texttt{REL}} = C_{\texttt{Treat}}\lvert\beta_1^{\texttt{REL}}\rvert$

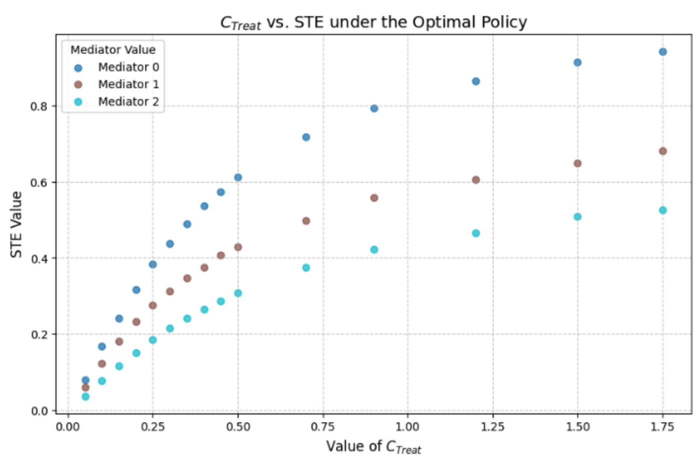


Fig. 3: Relationship between the hyperparameters and the STE, categorized by the mediator effect value.

Table 8: Summary of all testbeds

Treatment effect size	Value of $C_{\mathtt{Treat}}$
0.15 (Small)	0.2
0.3 (Medium)	0.3
0.5 (Large)	0.5

Q-learning on this dataset. To make the computation tractable, we discretize and subset the features. Specifically, we use six features: the intercept, AYA adherence, carepartner distress, AYA burden, carepartner burden, and relationship quality. Numerical features (carepartner distress, AYA burden, and carepartner burden) are discretized into 10 bins.

Finally, we evaluate the performance of this approximation against other baseline policies, including micro-randomized actions with fixed probabilities of 0.5, 0.6, 0.7, 0.8, and 0.9. Our approximation consistently outperforms these baselines.

C.6 Evidence of the Need for Collaboration in the Dyadic Environment

To show that each agent impacts the performance of other agents, we consider the following toy setting. We fix the care partner agent's randomization probability at 0.5 and vary the AYA agent's probability to be 0.25 and 0.75. Then, for each fixed AYA agent's probability, we identify the value of the relationship agent's probability that maximizes average weekly adherence. We find that this relationship probability changes from 1.0 to 0.0 when we change AYA agent's probability from 0.25 to 0.75.

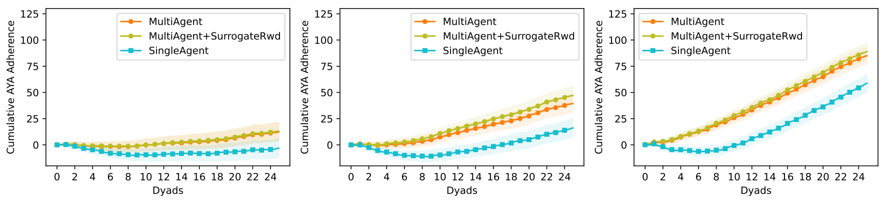
We repeat this experiment for the care partner agent by fixing the AYA agent's probability at 0.5 and varying the relationship agent's probability to be 0.25 and 0.75. Similarly, we find that the care partner agent's probability that maximizes adherence changes from 0.6 to 0.5 when we vary the relationship probability from 0.25 to 0.75.

These results indicate that the agents must change their behavior to account for the other agents' behavior.

D Additional Results

D.1 Ablation Study

No Mediator Effect The improvement from using a surrogate reward is through the effects of the mediator variables. For example, the relationship intervention $A_w^{\mathtt{REL}}$ improves the mediator relationship, which may improve the primary outcome, medication adherence. The care-partner intervention $A_{w,d}^{\mathtt{CARE}}$ mitigates the distress, which may improve relationship. In Fig. 4, we run the all three algorithms under a testbed variant for which we force the above two mediator effects to be 0, i.e., no effect from relationship to adherence or effect from distress to relationship. In this testbed variant, `MutiAgent+SurrogateRwd` performs the same as `MutiAgent`–there is no cost of reward learning under no mediator effect.



(a) STE 0.15 (b) STE 0.3 (c) STE 0.5

Fig. 4: Cumulative rewards improvement over the uniform random policy for all three components under the testbed without the effect of care-partner distress onto relationship quality or the effect of relationship quality onto AYA's adherence.

Other Testbed Variants. To further violate the assumptions made from the causal diagram, we made the following two changes to test the robustness of our proposed algorithm: 1) we add a direct effect from care-partner psychological distress to AYA medication adherence; 2) we generate random mediator effects, effect from relationship to adherence and effect from distress to relationship. This later one violates the monotonicity assumptions learned from principles.

D.2 Collaboration of Multi-Agent RL

We train each individual agent in the `MultiAgent+SurrogateRwd` algorithm over 1000 dyads under the STE 0.5 environment, while fixing the randomization probability of the other agents. We denote the randomization probability of the AYA agent, care partner agent, and relationship agent as $p^{\texttt{AYA}}$, $p^{\texttt{CARE}}$, and $p^{\texttt{REL}}$ respectively.

We first train the relationship agent while fixing $p^{\texttt{CARE}} = 0.5$. We see that the average probability of sending an intervention for the relationship agent is 0.57 and 0.42 under $p^{\texttt{AYA}} = 0.25$ and 0.75, respectively. This indicates that the relationship agent learns to *reduce* the intervention probability when the AYA agent is more likely to send an intervention.

Similarly, we train the care partner agent while fixing $p^{\texttt{AYA}} = 0.5$. We see that the average probability of sending an intervention for the care partner agent is 0.61 and 0.45 under $p^{\texttt{REL}} = 0.25$ and 0.75, respectively. This indicates that the care partner agent learns to *reduce* the intervention probability when the relationship agent is more likely to send an intervention.