A multi-agent reinforcement learning algorithm for personalized recommendations in bipolar disorder

Sidian Lin,[a, b] Soroush Saghafian,[b,*] Jessica M Lipschitz[c,d] and Katherine E Burdick[c,d]

[a]The Kenneth C. Griffin Graduate School of Arts and Sciences, Harvard University, Cambridge, 02138, MA, USA, [b] Harvard Kennedy School, Cambridge, 02138, MA, USA, [c]Department of Psychiatry, Brigham and Women's Hospital, Boston, 02115, MA, USA and [d]Department of Psychiatry, Harvard Medical School, Boston, 02115, MA, USA

[*]To whom correspondence should be addressed: soroush_saghafian@hks.harvard.edu

FOR PUBLISHER ONLY Received on Date Month Year; accepted on Date Month Year

Abstract

This study introduces a novel multi-agent reinforcement learning (MARL) algorithm designed for identifying and optimizing personalized recommendations in bipolar disorder. The algorithm leverages longitudinal offline data from wearables to recommend self-care strategies tailored to individual patients. We focus on self-care strategies involving physical activity (measured by steps), sleep duration, and bedtime consistency, aiming to reduce the periods of mood exacerbations. Findings suggest that following our algorithm's self-care recommendations could significantly reduce periods of elevated mood symptoms, resulting in improved overall well-being. In addition, the algorithm offers important clinical insights for treating bipolar patients, and shows promising theoretical properties showcasing its potential for use in other chronic diseases.

Key words: Multi-agent reinforcement learning, dynamic treatment regime, bipolar disorder, offline reinforcement learning

Significance statement

Our multi-agent reinforcement learning algorithm provides an innovative approach to personalizing self-care recommendations for bipolar disorder patients. By integrating longitudinal data from wearable devices and self-reports, the algorithm dynamically tailors self-care recommendations to individual patient's profiles. Findings suggest that the self-care recommendations generated by our algorithm could reduce the occurrence of clinically significant mood symptoms. The algorithm, therefore, has the potential to guide behavior changes that might have the biggest impact on patients' symptoms, allowing them to better manage their chronic disease.

Introduction

In the conventional reinforcement learning (RL) paradigm, a single agent interacts with the environment to form a policy. This policy, essentially a function connecting states to actions, aims to maximize a given cumulative reward, with the main goal of learning the optimal policy that results in the maximum possible reward. Multi-Agent RL (MARL, [73], [42]) is a subfield of RL where multiple agents learn simultaneously. Under MARL, the actions of all the agents collectively define the state of the environment, making the learning problem more complex. However, many real-world problems naturally involve multi-agent systems. Thus, MARL is needed for these situations as it allows multiple agents to learn and adapt to their environment and each other simultaneously. This is crucial for complex tasks, where the collective behavior of multiple entities can outperform any individual entity.

Across different fields, MARL has numerous applications. For example, in Robotics, MARL is used to train multiple robots to work together to achieve a common goal ([63]). In traffic scenarios, MARL can be used to train multiple vehicles to navigate in a shared space ([70], [10]). MARL can

Table 1. Eligibility Criteria [33]

Inclusion criteria	Exclusion criteria
(1) Age 18-68	(1) History of central nervous system trauma (including concussion with known loss of consciousness $>$ 1 minute)
(2) BD I or II diagnosis per the Structured Clinical Interview for DSM-5	(2) Any diagnosed neurological disorder
	(3) Attention deficit hyperactivity disorder that was diagnosed and treated in childhood (prior to onset of BD) or known learning disability
	(4) Current diagnosis of mild cognitive impairment or dementia
	(5) Substance abuse disorder (per SCID-5) within 3 months
	(6) Active, unstable medical problem that may interfere with cognition
	(7) Electroconvulsive therapy in the past year

Abbreviation: DSM-5 = Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition.

also be applied to optimize the use of shared resources in scenarios like power grid management ([5]) or packet routing ([37]), where multiple entities have to make decisions on how to use shared resources. MARL is still not common in the healthcare domain, though there is some previous work exploring using multiple-agent systems. Roche et al. ([48]) developed a multi-agent system model for vector-borne disease transmission in a realistic spatial environment, which took the natural complexity observed in both natural and man-made ecosystems into account. For personalized medicine in chronic disease management, different treatments can be seen as agents, and these agents can learn to coordinate their actions over time to provide the most effective overall treatment recommendation for individual patients.

However, there are several challenges when applying MARL, especially in the offline setting, which is common in healthcare applications using observational datasets. First, multi-agent offline RL suffers from issues like distributional shift and extrapolation error more severely than single-agent offline reinforcement learning (RL) ([26], [30]), partially due to a larger action and state space and complex inter-dependencies among agents ([77]). Second, to the best of our knowledge, there is no multi-agent offline RL algorithm that explicitly takes the dependency among agents into consideration, which is crucial in various healthcare applications, because agents are not totally independent entities and their actions need to be coordinated. We will discuss these challenges in detail in Section 1.2.

Experimental evidence from our research indicates that our algorithm is adept at learning from observational datasets and making personalized and effective recommendations for bipolar disorder patients. While our performance metrics demonstrate an improvement over the observed patient behaviors in the case of bipolar disorder, our MARL approach is general and offers a step forward in developing adaptive, data-informed self-care recommendations that could significantly improve the quality of life for individuals with other chronic health conditions.

Data and Setting

Our goal in this paper is to identify data-informed behavioral recommendations for bipolar disorder (BD) patients, focusing on activity and sleep behaviors. To this end, we collaborated with medical professionals from Brigham and Women's Hospital and Harvard Medical School. This was a secondary data analysis of a large longitudinal study on cognitive and psychosocial functioning in BD (see also [33]). The study was approved by Brigham and Women's Hospital's Institutional Review Board. Participants for the longitudinal study were recruited via hospital listservs, patient registries, and other ongoing studies of BD. All participants enrolled in the longitudinal study between November 1, 2020 and June 12, 2023 were invited to participate in the digital phenotyping investigation and the study team obtained informed consent from interested participants. Detailed eligibility criteria are provided in Table 1 (see also [33]).

Data collected include Fitbit data and biweekly self-reports. Fitbit data included data of steps per day/per hour/per minute, daily floors total, daily activity, calories, sleep condition and duration, heart rate per second/minute/15 minutes. Patients were asked to complete the Patient Health Questionnaire-8 (PHQ-8) ([25]) and Altman Self-Rating Mania Scale (ASRM) ([1]) via RedCAP every other week, so the second part of the data is biweekly self-reports including PHQ8 scores, on which there are cutoffs that define clinical depression (see Table 2 for a general data summary and Table 3 for a summary of Fitbit data features). Cutoff values are PHQ-8 $\geq$ 10 (indicating probable depressive episode at that timepoint) ([71]) and ASRM $\geq$ 6 (indicating probable manic or hypomanic episode at that timepoint, hereafter referred to as (hypo)mania) ([1]).

In this study, we introduce a novel MARL algorithm designed for offline use, tailored to capture the complex interactions among agents which mirror the dynamic nature of patient behaviors. This algorithm stands out for its unique approach to significantly reducing extrapolation errors, thereby enhancing its ability to learn from available data. We then leverage this algorithm to identify personalized self-care recommendations for individuals with bipolar disorder.

Our primary objective is to reduce the frequency of weeks with clinically-significant mood symptoms in BD patients. Throughout the paper, "mood episode" is used to refer to two-week periods of clinically-significant depressive or (hypo)manic symptoms based on established clinical cutoffs for the PHQ-8 and Altman Mania Rating Scale. As used in this paper, therefore, the term "mood episodes" does not reflect major depressive, manic, or hypomanic episodes as defined in the Diagnostic and Statistical Manual of Mental Illness (DSM-5-TR). To achieve our primary objective, we consider treatment recommendations in three distinct categories: step-based activity interventions, sleep duration adjustments, and bedtime consistency enhancements. Each category is managed by a separate "agent" within our algorithm, representing the various aspects of daily life that could be helpful if adjusted in BD patients.

Table 2. Description of Dataset

Variable	Description	Availability
Self-Report		
PHQ-8	Biweekly self-report rating of severity of depression symptoms with a clinical cutoff of ≥ 10	✓ cleaned
ASRM	Biweekly self-report rating of mania/hypomania symptoms with a clinical cutoff of ≥ 6	✓ cleaned
Fitbit Data		
Steps	Broken down by day (or smaller increments if desired). Can also be broken down by milage traveled.	✓ messy
Sleep	Total minutes in bed; Total minutes asleep; Total minutes light sleep, deep sleep, and REM; Total minutes of nighttime awakenings.	✓ messy
Heart Rate	Broken down by 15 minutes, 5 minutes, 1 minute, or seconds. Daily Resting HR and Daily HR Zones (below, above, fat burn, cardio, peak, out of range, custom zone) also available.	✓ messy
Intensity	Distance traveled and minutes in the following categories – Very Active, Moderately Active, Lightly Active, and Sedentary. Can be broken down by day or by hour.	✓ messy
Activity	Auto-detected types of exercise (ex. Sport, Run, Kayaking, Aerobic workout, etc.) and associated steps, distance, calories, Lightly/Fairly/Very Active Minutes, average heart rate, out of range heart rate minutes, fat burn heart rate minutes, cardio heart rate minutes, peak heart rate minutes	✓ messy

Table 3. Summary of Fitbit data features used in our MARL approach (For a full summary of our Fitbit data, see [33].)

Feature Label (Abbreviation)	Description	Mean (SD)
Step count (Steps)	Average number of steps taken per day during the observation window.	6631.75 (3585.15)
Heart rate (HR)	Daily heart rate value averaged over the observation window.	78.42 (7.62)
Resting heart rate (Resting)	Daily resting heart rate value averaged over the observation window.	69.38 (7.96)
Total sleep time (TST)	Total number of minutes classified as being asleep per night averaged over the observation window excluding nights without sleep. This metric does not include nights without any sleep.	430.95 (81.10)
Sleep efficiency score (Effi)	100 * [minutes asleep / (time in bed – minutes after waking up before user changes their device out of sleep mode)] per night averaged over the observation window.	92.27 (7.13)
Awakenings duration (Awake)	Number of minutes classified as being awake during the sleep stages record each night averaged over the observation window.	57.11 (17.15)
Median bedtime	Median bedtime during the observation window. Bedtime was recoded such that 6 pm = 0 and 10 am the next morning = 16. Thus a value of 5 indicates an 11 pm bedtime.	5.53 (1.62)

Challenges in Applying RL in to Our Setting

In the evolving field of RL, the integration of multi-agent dynamics presents a complex challenge, particularly in specialized domains such as healthcare. This paper seeks to explore and address the unique difficulties posed by applying multi-agent RL techniques to healthcare scenarios, where data limitations and ethical considerations significantly shape the research approach. Below, we outline several key issues that our current study aims to tackle:

(C1) In the domain we are examining, several critical aspects highlight the limitations of traditional single-agent RL approaches, which struggle to address the complexity of problems where interactions among multiple agents are significant. Furthermore, in the context of multi-agent RL, there appears to be a lack of algorithms that have been effectively adapted to personalized self-care treatment recommendations, highlighting a gap in the current research landscape.

(C2) In settings where learning can only occur from observational data, multi-agent offline RL encounters significant challenges. These include more pronounced issues of distributional shift and extrapolation error compared to single-agent offline RL. Such difficulties stem from the enlarged action and state spaces, as well as the intricate inter-dependencies among agents, which are crucial in accurately modeling the dynamics of the environment (patient conditions in our setting).

(C3) To the best of our knowledge, no existing multi-agent offline RL algorithm currently incorporates a mechanism to explicitly consider dependencies among agents. This is a crucial oversight, particularly because most available decomposition approaches fail to reflect the complex realities of agent interactions ([69], [64], [45]).

(C4) Although bipolar disorder is a lifelong condition, existing data encapsulates only a finite period of observation with no real "termination" point, posing significant challenges in learning a stationary policy ([14], [79]) from such partial episodic data. The temporal limitations of data necessitate careful consideration of how policies derived based on data that is not long-term might perform over the extended course of a lifelong disorder.

(C5) The direct implementation, testing, or validation of the learned policy in a real-world setting is precluded due to the ethical and sensitive nature of healthcare interventions. This limitation requires us to adopt off-policy evaluation (OPE) methods ([66], [22], [4]) to ascertain the effectiveness and safety of the learned policies, ensuring that we do not directly impact real patients without thorough validation of the algorithm-based recommendations' potential outcomes.

Literature Review

Our study intersects with three main streams of literature: (1) MARL algorithms and their applications; (2) mobile health and the integration of RL; (3) research on the severity and treatment of bipolar disorder.

MARL is an extension of RL that considers environments with multiple agents, aimed at addressing the sequential decision-making problem that involves several learners. MARL has recently achieved notable advancements. For example, Shalev-Shwartz et al. ([57]) applied deep MARL to the problem of strategic long-term driving planning. Multi-agent systems have applications in many other domains including finance ([28], [29]), communication networks ([78], [10]), and social sciences ([7], [8]). For example, in robotics, control is naturally distributed among multiple robots ([63]). And for adaptive traffic signal control in intricate urban networks, MARL is considered a natural solution for overcoming scalability issues by distributing the global control to individual RL agents ([40]).

Offline MARL, an emerging subarea, has attracted significant attention because real-world applications often have limited opportunities for collecting online data, while ample offline data may be available. In single-agent domains, offline RL has been extensively studied, identifying distributional shift as a major hurdle ([17]). Consequently, modern offline algorithms like Batch-Constrained deep Q-learning (BCQ, [17]) are designed to ensure that the learned policy remains close to the behavior policy or to suppress the Q-value outright. MARL scenarios, however, confront a substantially larger action space that expands exponentially with the number of agents. This increase leads to a growth in unseen state-action pairs, potentially resulting in accumulating extrapolation errors. To address this, [74] introduced an offline RL algorithm, Implicit Constraint Q-learning (ICQ), that relies solely on the dataset's state-action pairs for value estimation. ICQ has been extended to multi-agent tasks by decomposing the joint policy under an implicit constraint. Despite the necessity to train agents for tasks involving physical, social, and team dynamics, most existing works, including ICQ, typically assume agents make decisions independently, overlooking the intricate interdependencies among agents. Wang et al. ([68]) propose using copulas to model agent correlations and coordination explicitly within multi-agent imitation learning. In this paper, we incorporate copulas into the ICQ framework to more accurately capture agents' behaviors and interactions.

With the widespread application of the Internet, and connected sensors that talk to each other via Internet of Things ([55]), data collection is often combined with wireless transmission. Data can be uploaded to the network and used to generate a database, thereby providing long-term disease monitoring. This technology, when combined with approaches such as human-algorithm hybrid models ([41]) and dynamic resource assignment strategies ([53]), can create important opportunities to scale and personalize healthcare. This technology can also be used to guide individualized recommendations in the context of entirely digital mHealth interventions that are much more accessible than clinician-provided treatment ([54]).

Although various studies have proposed mHealth recommender systems ([54]), applications of MARL in healthcare, especially in offline settings and the multi-agent domain, are sparse. For instance, [9] developed IntelliCare, a suite of apps for managing depression and anxiety with an app recommender system aimed at enhancing engagement, and [72] introduced emHealth, an intelligent health recommendation system for emotional well-being. These systems use predictive algorithms like decision trees and support vector machines. A wide array of RL algorithms has been proposed to support medical decision-making and to optimize treatment recommendations ([39], [76], [27], [61], [34]). In applications with online data, contextual multi-armed bandits have been utilized in order to combat declining patient engagement over time ([65]). For offline applications, [20] suggests a generalized linear mixed-model framework under a contextual bandit to achieve personalized interventions. Ameko et al. ([2]) presents a novel treatment recommender system for emotion regulation. In a comprehensive RL setting, [16] reported on a weight loss study employing interventions selected bi-weekly over a 12-week period. Liao et al. ([32]) developed an RL algorithm that iteratively learns and refines treatment policies in just-in-time adaptive interventions (JITAI) based on ongoing data collection. Saghafian ([50]) extended Dynamic Treatment Regimes (DTRs) to Ambiguous Dynamic Treatment Regimes (ADTRs) and developed new RL methods using an Ambiguous Partially Observable Markov Decision Process (APOMDP, [49]) to improve personalized and dynamic treatment plans. Our work focuses on developing an offline MARL framework for bipolar disorder (BD) patients, suitable for recommending timely and personalized self-care strategies that could easily be conveyed in an mHealth interventions.

BD is a chronic and severe psychiatric condition characterized by extreme mood swings, ranging from depression to hypomania or mania. Patients with BD typically experience an average of three mood polarity changes each year ([23]). Treatment strategies for BD focus on reducing the occurrence and intensity of these mood episodes to alleviate emotional distress and mitigate the mood episodes' adverse effects on patients' personal and professional lives ([75], [62]). Timely identification of mood episodes is crucial, as it enables healthcare providers to identify patients at risk of more severe outcomes, such as cognitive decline ([56]) and increased suicidality ([21]). Furthermore, rapid cycling, which is characterized by frequent mood episodes of depression or mania within a year, is often underrecognized ([6]) but is associated with a more challenging disease course and poorer prognosis ([13]). Given the consensus on the importance of reducing mood episodes in BD, we pursue the objective of recommending self-care strategies that can minimize the frequency of mood episodes given each individual patient's profile.

Methods

In this section, we discuss the design of our MARL algorithm. We first address how our algorithm tackles the challenges outlined in previous sections. Then, we provide a detailed description of our algorithm and the objective function we utilize during its training process.

Addressing the Challenges (C1 - C4)

We now discuss how our approach addresses the four challenges mentioned earlier (C1 - C4). To address challenges (C1) and (C2), we adopt the foundational framework of the ICQ algorithm ([74]), a MARL algorithm originally developed for optimizing policies in the game Starcraft. We implement substantial modifications to tailor this algorithm to our specific healthcare setting, ensuring it can handle the unique complexities and constraints of medical data.

To address challenge (C3), we integrate copulas into the characterization of the joint policy, which allows for a more nuanced representation of dependencies among agents. Additionally, we revise the computation of joint Q-values to avoid the simplistic approaches commonly used in Q-value decomposition ([36], [31]), thus enhancing the model's ability to capture complex agent interactions.

To overcome the limitations described in (C4), we modify the "reward" associated with the last time point in our dataset. Specifically, we use the data from all time points except the last for training purposes, while the final state information from each patient's trajectory is reserved to estimate a potential future reward. This approach is designed to facilitate the learning of a stationary policy despite the episodic truncation of the data.

Regarding (C5), we evaluate our derived policies using an off-policy evaluation method similar to [59], which is tailored for multi-agent settings and accounts for interference effects among agents. This method aligns well with our use of copulas, providing a robust framework for assessing policy effectiveness in a complex, interconnected environment. However, as we will discuss, this method necessitates knowledge of the behavior policy, which poses a significant challenge in our context due to the difficulty of accurately estimating this policy from observational healthcare data.

The Proposed Multi-Agent Reinforcement Learning (MARL) Framework

Notation. We model our problem as a fully cooperative multi-agent task, which is usually modeled as a Decentralized Partially Observable Markov Decision Process (Dec-POMDP) consisting of the tuple $G = \langle S, \mathbf{A}, P, r, \Omega, O, n, \gamma, \rho_0 \rangle$. In this setting, S is the common state space characterized by "leading" features selected from a prior study. We denote by A_i the action space for agent $i \in N \equiv \{1, ..., n\}$. All the agents form a joint action $\boldsymbol{a} \in \boldsymbol{A} \equiv A_1 \times ... \times A_n$. Of note, we have in total three agents in our problem, executing actions corresponding to steps, sleep duration, and sleep hygiene. $P(s'|s, \boldsymbol{A}) : S \times \boldsymbol{A} \times S \to [0, 1]$ is the state transition function, and $r(s, \boldsymbol{a}) : S \times \boldsymbol{A} \to \mathbb{R}$ indicates the team reward function. $\gamma \in [0, 1)$ is the discount factor, and $\rho_0 : S \to \mathbb{R}$ is the distribution of the initial state S_0.

In the POMDP setting, each agent generates individual observation $o^i \in \Omega$ based on the observation function $O(s, a) : S \times \mathbf{A} \to \Omega$. Each agent maintains an action-observation history $\tau^i \in \mathbf{T} \equiv (\Omega \times \mathbf{A})^t$, which serves as conditioning for a stochastic policy $\pi^i(a^i \mid \tau^i)$: $\mathbf{T} \times \mathbf{A} \to [0, 1]$, which we parameterize by θ_i. The joint action-value function is defined as $Q^{\boldsymbol{\pi}}(\boldsymbol{\tau}, \boldsymbol{a}) \triangleq \mathbb{E}_{s_{0:\infty}, \boldsymbol{a}_{0:\infty}} \left[\sum_{t=0}^{\infty} \gamma^t r_t \mid s_0 = s, \boldsymbol{a}_0 = \boldsymbol{a}, \boldsymbol{\pi}\right]$, where $\boldsymbol{\pi}$ represents the joint policy with parameters $\theta = \langle \theta_1, \ldots, \theta_n \rangle$. The joint policy $\boldsymbol{\pi}$ with parameters θ, also written as $\boldsymbol{\pi}_{\boldsymbol{\theta}}$, the transition kernel P, and the initial state distribution ρ_0, induce a marginal state distribution at time t, $d_\theta^t(s)$, which is a probability mass when S is discrete. The offline dataset $\mathcal{B}$ contains trajectories generated by a behavior policy μ, and can be used to obtain a joint optimal policy $\boldsymbol{\pi}^*$ that maximizes the discounted sum of rewards.

We make substantial modifications on the offline actor-critic RL algorithm proposed by [74], termed *Implicit Constrained Q-Learning* (ICQ), to make it suitable for our setting. In offline tasks, we do not have an environment for agents to interact with, and the standard Bellman operator would suffer from the out-of-distribution (OOD) issue. By adopting the importance sampling method, the OOD issue can be avoided, but in most scenarios, it is hard to estimate the exact behavior policy to calculate the importance sampling weight $\rho(\tau', a') \triangleq \frac{\pi(a'|\tau')}{\mu(a'|\tau')}$. The core idea of ICQ is that only policies similar to the behavior policy are preferred while maximizing the accumulated reward $Q^\pi(\tau, a)$, which is ensured via the constraint, $D_{KL}(\pi \mid\mid \mu)[\tau] \leq \epsilon$. In other words, ICQ utilizes a constraint on the KL divergence between the learned and behavior policy to alleviate extrapolation errors. Specifically, the optimization problem in ICQ is:

$$\begin{aligned} \pi_{k+1} = \arg\max_{\pi} \mathbb{E}_{a \sim \pi(\cdot|\tau)}[Q^{\pi_k}(\tau, a)] \\ \text{s.t.} \quad D_{KL}(\pi||\mu)[\tau] \leq \epsilon, \\ \sum_a \pi(a|\tau) = 1, \end{aligned} \tag{1}$$

where the subscript k is used to denote the iteration step. Using the KKT condition, the optimal policy π^*_{k+1} can be obtained as

$$\pi^*_{k+1}(a|\tau) = \frac{1}{Z(\tau)} \mu(a|\tau) \ \exp\left(\frac{Q^{\pi_k}(\tau, a)}{\alpha}\right), \tag{2}$$

where $\alpha > 0$ is the Lagrangian coefficient and $Z(\tau) = \sum_{\tilde{a}} \mu(\tilde{a}|\tau) \exp(Q^{\pi_k}(\tau, \tilde{a})/\alpha)$ is the normalizing partition function. Dividing Eq. (2) by μ, we can obtain the importance sampling weight $\rho(\tau, a)$, which is

$$\rho(\tau, a) = \frac{\pi^*_{k+1}(a|\tau)}{\mu(a|\tau)} = \frac{1}{Z(\tau)} \ \exp\left(\frac{Q^{\pi_k}(\tau, a)}{\alpha}\right).$$

Thus, the Implicit Constraint Q-learning operator is defined as

$$\mathcal{T}_{ICQ} Q(\tau, a) = r + \gamma \mathbb{E}_{a' \sim \mu}\left[\frac{1}{Z(\tau')} \ \exp\left(\frac{Q(\tau', a')}{\alpha}\right) Q(\tau', a')\right].$$

In the single-agent case, based on the operator $\mathcal{T}_{ICQ}$, ICQ learns $Q(\tau, a; \phi)$, where ϕ is used to parameterize the function, by minimizing

$$\begin{aligned} \mathcal{J}_Q(\phi) = \mathbb{E}_{\tau, a, \tau', a' \sim \mathcal{B}}\Bigg[r + \gamma \frac{1}{Z(\tau')} \exp\left(\frac{Q\left(\tau', a'; \phi'\right)}{\alpha}\right) \cdot \\ Q(\tau', a'; \phi') - Q\left(\tau, a; \phi\right)\Big]^2, \end{aligned} \tag{3}$$

which is the critic loss.

As for the policy learning, ICQ aims to minimize the following KL distance parameterized by θ,

$$\begin{aligned} \mathcal{J}_\pi(\theta) &= \mathbb{E}_{\tau \sim \mathcal{B}}\left[D_{\mathrm{KL}}\left(\pi^*_{k+1} \| \pi_\theta\right)[\tau]\right] \\ &= \mathbb{E}_{\tau \sim \mathcal{B}}\left[-\sum_a \pi^*_{k+1}(a \mid \tau) \log \frac{\pi_\theta(a \mid \tau)}{\pi^*_{k+1}(a \mid \tau)}\right] \\ &\overset{(a)}{=} \mathbb{E}_{\tau \sim \mathcal{B}}\left[\sum_a \frac{\pi^*_{k+1}(a \mid \tau)}{\mu(a \mid \tau)} \mu(a \mid \tau)\left(-\log \pi_\theta(a \mid \tau)\right)\right] \\ &\overset{(b)}{=} \mathbb{E}_{\tau, a \sim \mathcal{B}}\left[-\frac{1}{Z(\tau)} \log(\pi(a \mid \tau; \theta)) \exp\left(\frac{Q(\tau, a)}{\alpha}\right)\right], \end{aligned}$$

where equivalence (a) is obtained by ignoring one term that is constant in θ, and equality (b) is obtained by applying the

importance sampling weight under the KL constraint. Equality (b) is the actor loss in the single-agent setting; we will derive its multi-agent version below.

In the *centralized training and decentralized execution* (CTDE, [60]) framework designed for multi-agent RL, the algorithm is provided with the true state s and the action-observation history τ_i of each agent. It also has the flexibility to share all information among agents. However, in the execution step, each agent can only access its own action-observation history.

We will train separate policies for different agents to execute under this CTDE framework. It should be noted that calculating $\mathbb{E}_{\boldsymbol{\mu}}[\rho(\boldsymbol{\tau}', \mathbf{a}')Q^{\boldsymbol{\pi}}(\boldsymbol{\tau}', \boldsymbol{a}')]$ in multi-agent policy evaluation is difficult due to its computational complexity of $O(|A|^n)$. Additionally and importantly, the majority of current studies focusing on multi-agent RL (including ICQ) assume that agents independently make decisions based on their observations, disregarding the intricate interdependency among agents. In what follows, we first introduce the utilization of copula c, a rigorous statistical quantity for capturing correlation among random variables, to explicitly model the dependency (and hence, coordination) among agents. Specifically, we first consider the joint policy as:

$$\boldsymbol{\pi}(\boldsymbol{a} \mid \boldsymbol{\tau}) \triangleq \Pi_{i \in N} \pi^i(a^i \mid \tau^i) \cdot c(F^1(a^1|\tau^1), ..., F^N(a^N|\tau^N)|\boldsymbol{\tau}),$$

where $\pi^i(a^i \mid \tau^i)$ is the marginal policy of agent i, and F^i is the corresponding cumulative distribution function; the function c is the density of the copula on the transformed actions $u^i = F^i(a^i \mid \tau^i)$ obtained by processing original actions with probability integral transform. There are many ways to learn function $c(\cdot)$ from the observed data. We make use of Algorithm 2 (presented in the supplementary material) proposed by [68] to learn this function from our data.

Since the learned $c(\cdot)$ is the dependency among agents in the observational dataset, and not necessarily the optimal policy, we introduce a decision parameter $\beta \in [0, 1]$ and incorporate $c^\beta(\cdot) \triangleq [c(\cdot)]^\beta$ in deriving the optimal policy. When $\beta = 0$, $c^\beta(\cdot)$ equals 1, meaning agents must act independently with no consideration of others. Conversely, when $\beta = 1$, the influence of β vanishes, meaning that agents must exhibit the same dependencies as observed in the data. Thus, we modify our original optimization problem as follows:

$$\begin{aligned} \pi_{k+1} = \arg\max_{\pi, \beta} \mathbb{E}_{a \sim \pi(\cdot|\tau)}[Q^{\pi_k}(\tau, a)] \\ \text{s.t.} \quad D_{KL}(\pi||\mu)[\tau] \leq \epsilon, \\ \sum_a \pi(a|\tau) = 1, \\ 0 \leq \beta \leq 1. \end{aligned}$$

That is, we keep the KL divergence in the same format as before but consider β as a decision variable and impose constraints on it. The Lagrangian function for the optimization problem is

$$\begin{aligned} \mathcal{L}(\pi, \alpha, \lambda, \beta, \gamma_1, \gamma_2) =& \mathbb{E}_{a \sim \pi}[Q^{\pi_k}(\tau, a)] + \alpha(\epsilon - D_{KL}(\pi||\mu)[\tau]) \\ &+ \lambda \left(1 - \sum_a \pi(a|\tau)\right) + \gamma_1(1 - \beta) + \gamma_2 \beta, \end{aligned}$$

where $\alpha, \gamma_1, \gamma_2 > 0$ denote the Lagrangian coefficients. To use the KKT condition, we need to take the partial derivative of $\mathcal{L}$ with respect to π and set it to zero.

Doing so, it can be seen that the optimal policy can be represented as

$$\boldsymbol{\pi}^*(\boldsymbol{a} \mid \boldsymbol{\tau}) \triangleq \Pi_{i \in N} \pi^{*i}(a^i \mid \tau^i) \cdot c^\beta(F^1(a^1|\tau^1), ..., F^N(a^N|\tau^N)|\boldsymbol{\tau}), \quad \beta \in [0, 1],$$

where π^{*i} is the individual optimal policy of agent i. To obtain the joint and individual Q values, we then introduce a value-decomposition assumption. The intuition is that for agent i, the action a^i is preferred when for some other agent j, its action a^j is 1) positively correlated with a^i, and 2) has a high Q value.

$$\begin{aligned} Q^\pi(\boldsymbol{\tau}, a) &= \sum_{i=1}^N w^i(\boldsymbol{\tau}) \hat{Q}^i(\boldsymbol{\tau}^i, a^i) + b(\boldsymbol{\tau}). \\ \hat{Q}^i(\boldsymbol{\tau}^i, a^i) &= Q^i(\boldsymbol{\tau}^i, a^i) + \sum_{j \neq i} \sum_{a'} \eta^i(a') Q^j(\boldsymbol{\tau}^j, a') \\ \eta^i(a) &= \log c(a^i, a^{-i}) \times [\sum_{j \neq i} N(a^i; 1)(a) - \frac{1}{\|A\|}] \end{aligned} \tag{4}$$

Here, the values of $w^i(\boldsymbol{\tau}) \geq 0$ and $b(\boldsymbol{\tau})$ are obtained from the Mixer Network ([74]), which takes the global observation-action history as inputs. The normal distribution is used to characterize the distance between two actions. As long as the action is ordinal or continuous, as is the case with our study, this distance can be combined with the copula to serve as an indicator of the correlation between the actions of two different agents. The subtraction of $\frac{1}{\|A\|}$ removes the baseline effect of a uniform distribution, ensuring that $\eta^i(a)$ only captures meaningful deviations from random behavior in the action space. While $\frac{1}{\|A\|}]$ vanishes in the limit of large $\|A\|$, it provides an important baseline adjustment for smaller or moderate action spaces, helping to stabilize and properly scale the correlation terms during learning. We will hereinafter use $\hat{Q}^i$ to denote the copula-adjusted individual Q value.

Making use of these assumptions, we propose the decomposed multi-agent joint policy under an implicit constraint with copulas in Theorem 1 in the next section.

Besides considering dependency among agents, another modification we make, to address (C4), is to allow the algorithm to learn a stationary policy from finite horizon data, which is the case we encounter in our BD dataset. Although due to the nature of the data the agents face time constraints in their iterations with the environment, our objective remains to learn a long-run policy. This is important because managing moods for bipolar patients is a "time-unlimited" task, which requires learning a stationary policy. To this end, we follow the procedure in [43], which involves bootstrapping at states where termination is due to time constraints. Specifically, we use only the first T-1 periods in each trajectory for training and reserve the state information of the last period (s_T) to construct a potential future reward based on the critic estimator within the algorithm. By denoting an estimate of the critic value as $\hat{v}_\pi$, the target y for a one-step Temporal Difference (TD) update at the T-1 period, after transitioning to s_T and receiving a reward r, is formulated as $y = r + \gamma \hat{v}_\pi(s_T)$.

Results

Theoretical Results

In this section, we develop the theoretical framework for our copula-adjusted multi-agent reinforcement learning (MARL) algorithm. Using copulas, we model inter-agent dependencies

and decompose the joint policy into individual policies, as shown in Theorem 1. This leads to the Copula-Adjusted MARL algorithm (Algorithm 1) within the actor-critic framework. We further derive two gradient estimators—one accounting for agent dependencies and the other for independent agents—and analyze their variance in Theorem 2. The bounded variance difference between the estimators confirms that our approach maintains comparable convergence properties to the independent case.

Theorem 1 *Assuming the joint action-value function to be the form of Eq. (4), we can decompose the multi-agent joint policy under implicit constraint with copulas as follows:*

$$\begin{aligned}
&\boldsymbol{\pi}^* \\
&= \underset{\pi^1,\ldots,\pi^n,\beta}{\arg\min} \sum_i \underset{\boldsymbol{\tau},a^i\sim\boldsymbol{\mathcal{B}}}{\mathbb{E}} \left[-\frac{\log(\pi^i(a^i \mid \tau^i))}{Z^i(\tau^i)} \exp\left(\frac{w^i(\boldsymbol{\tau})\hat{Q}^i(\tau^i,a^i)}{\alpha}\right)\right] \\
&\quad + \sum_i \mathbb{E}_{\boldsymbol{\tau},a^i\sim\boldsymbol{\mathcal{B}}} \left[-\frac{\beta}{\hat{Z}^i(\tau^i)} \log(c(\cdot)) \exp\left(\frac{w^i(\boldsymbol{\tau})\hat{Q}^i(\tau^i,a^i)}{\alpha}\right)\right],
\end{aligned}$$

where $\hat{Z}^i(\tau^i) = \sum_{\tilde{a}^i} \mu^i(\tilde{a}^i \mid \tau^i) \exp\left(\frac{1}{\alpha} w^i(\boldsymbol{\tau})\hat{Q}^i(\tau^i,\tilde{a}^i)\right)$ *is the normalizing partition function.*

The decomposed multi-agent joint policy can be expressed concisely. Furthermore, we can train individual policies π^i *by minimizing the following expression:*

$$\begin{aligned}
&\mathcal{J}_{\boldsymbol{\pi}}(\theta) \\
&= \sum_i \mathbb{E}_{\boldsymbol{\tau},a^i\sim\boldsymbol{\mathcal{B}}} \left[-\frac{\log(\pi^i(a^i \mid \tau^i;\theta_i))}{\hat{Z}^i(\tau^i)} \exp\left(\frac{w^i(\boldsymbol{\tau})\hat{Q}^i(\tau^i,a^i)}{\alpha}\right)\right] \\
&\quad + \sum_i \mathbb{E}_{\boldsymbol{\tau},a^i\sim\boldsymbol{\mathcal{B}}} \left[-\frac{\beta}{\hat{Z}^i(\tau^i)} \log(c(\cdot)) \exp\left(\frac{w^i(\boldsymbol{\tau})\hat{Q}^i(\tau^i,a^i)}{\alpha}\right)\right].
\end{aligned} \tag{5}$$

Proof : See the supplementary material. □

Using Theorem 1, we present Algorithm 1 within the actor-critic framework.

We next derive two estimators from the gradient of $\mathcal{J}_\pi(\theta)$ presented in Theorem 1, one for copula-adjusted (denoted by $\mathbf{g}^i_C$), where agent dependencies are taken into account ($\beta \neq 0$), and one for the original ICQ (denoted by $\mathbf{g}^i_O$), where agents are modeled as independent entities ($\beta = 0$). These estimators are, respectively, given by $\mathbf{g}^i_C = -\frac{1}{\hat{Z}^i(\tau^i)} \exp\left(\frac{w^i(\tau)\hat{Q}^i(\tau^i,a^i)}{\alpha}\right) \nabla_{\theta^i} \log \pi^i_\theta(a^i|s)$, and $\mathbf{g}^i_O = -\frac{1}{Z^i(\tau^i)} \exp\left(\frac{w^i(\tau)Q^i(\tau^i,a^i)}{\alpha}\right) \nabla_{\theta^i} \log \pi^i_\theta(a^i|s)$.

We then introduce the following two assumptions:

Assumption 1. *The state space* S *and every agent* i*'s action space* A^i *is either discrete and finite, or continuous and compact.*

Assumption 2. *For all* $i \in N$, $s \in S$, *and* $a^i \in A^i$, *the map* $\theta^i \mapsto \pi^i_\theta(a^i|s)$ *is continuously differentiable.*

Based on the assumptions above, we can prove the boundedness of the difference in the variance of the two estimators as stated in Theorem 2.

Algorithm 1: Copula-Adjusted MARL algorithm

Input: Offline buffer $\mathcal{B}$, copula density values c, target network update rate d.

Initialize critic networks $Q^i(\cdot;\phi_i)$, actor networks $\pi^i(\cdot;\theta_i)$ and Mixer network $M(\cdot;\psi)$ with random parameters; initialize Lagrangian coefficient α and decision parameter β.

Initialize target networks: $\phi' = \phi$, $\theta' = \theta$, $\psi' = \psi$.

for $t = 1$ **to** $T-1$ **do**

Sample trajectories from $\mathcal{B}$.

Train individual policy according to Eq. (5) where $\hat{Q}^i$ is calculated by Eq. (4).

Train critic according to

$$\begin{aligned}
&\mathcal{J}_Q(\phi,\psi) \\
&= \mathbb{E}_{\mathcal{B}} \left\{ \sum_{t\geq 0} (\gamma\lambda)^t \left[r_t + \gamma \frac{\exp\left(\frac{1}{\alpha} Q(\boldsymbol{\tau}_{t+1},\boldsymbol{a}_{t+1};\phi',\psi')\right)}{Z(\boldsymbol{\tau}_{t+1};\phi',\psi')} \cdot \right.\right. \\
&\qquad \left.\left. Q(\boldsymbol{\tau}_{t+1},\boldsymbol{a}_{t+1};\phi',\psi') - Q(\boldsymbol{\tau}_t,\boldsymbol{a}_t;\phi,\psi)\right]\right\}^2.
\end{aligned}$$

if $t \bmod d = 0$ **then**

Update target networks: $\phi' = \phi$, $\theta' = \theta$, $\psi' = \psi$.

Update α and β.

end

end

Theorem 2 *Under Assumption 1 and 2, the copula-adjusted and the original ICQ estimators of MAPG (multi-agent policy gradient) satisfy*

$$\mathbf{Var}_{s_{0:\infty}\sim d^{0:\infty}_\theta, \mathbf{a}_{0:\infty}\sim\pi_\theta}[\mathbf{g}^i_C] - \mathbf{Var}_{s_{0:\infty}\sim d^{0:\infty}_\theta, \mathbf{a}_{0:\infty}\sim\pi_\theta}[\mathbf{g}^i_O] \leq B_i,$$

where $B_i = \sup_{s,\mathbf{a}} \|\nabla_{\theta^i} \log \pi^i_\theta(a^i|s)\|$.

Proof : See the supplementary material. □

Theorem 2 shows that in any scenario where the original ICQ algorithm is effective, our copula-adjusted MARL algorithm (Algorithm 1) can achieve a desired convergence rate. This is because the variance difference remains within a reasonable range, ensuring the convergence property of the copula-adjusted algorithm is comparable to that of the original ICQ algorithm.

Numerical Results

In our longitudinal data, we have 52 complete trajectories from bipolar disorder patients participating in a study by wearing a Fitbit device for a total of T=9 consecutive time periods, where each time period represents two weeks (see also [33]).

Among those patients, first we imputed values for missing Fitbit data. Missingness was defined by the presence of missing intervals in the 1-minute interval heart rate data. Missing data were replaced using the random forest imputation method, which has advantages over parametric multivariate imputation by chained equations (MICE) insofar as it accommodates nonlinearities and interactions [33]. Where minute-level data were available, imputations were done at the minute level, and the corresponding 15-minute/hourly/daily data were aggregated across intervals of missingness (details of imputation can be found in the supplementary material).

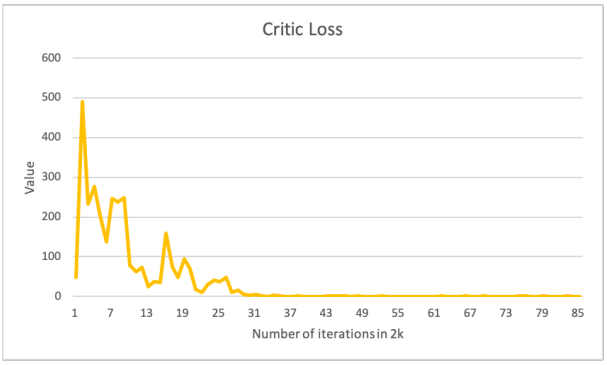

Fig. 1. Critic loss value

Then, we introduced three "imaginary" agents: the steps-controlling agent, the total sleep time (TST)-controlling agent, and the bedtime consistency (BC)-controlling agent. Bedtime consistency (BC) in our setting is characterized by the standard deviation of bedtime over a two-week period. Each agent can choose one of four actions based on a comparison with the last period: 1) change this period's value by a small amount; 2) change this period's value by a medium amount; 3) change this period's value by a large amount[1]; 4) do nothing. Here, "change" indicates an increase in steps and TST but a decrease in the standard deviation of bedtime.

Based on our prior study of mood episode detection in bipolar disorder ([33]), we used the following seven variables to represent the state for each patient: 1) sleep efficiency score (Effi), 2) awakenings duration (Awake), 3) heart rate (HR), 4) resting heart rate (Resting), 5) steps (Steps), 6) total sleep time excluding nights without sleep (TST), and 7) standard deviation of the bedtime (BC). Of note, all these variables were measured using the mean value during the two-week period. Furthermore, we categorize all the state variables into three levels: stable around the individual mean, lower than the individual mean, and higher than the individual mean. These categories were represented by 0, 1, and 2, respectively. The cutoffs used for these categories for each state variable can be found in the supplementary material.

Based on our goal of minimizing the number of mood episodes, we assigned the following reward values:

- R = {0, if not in an mood episode; - 10, if in a single mood episode; -20, if in a mixed mood episode, i.e., observations of both depression and (hypo)mania above the clinical cutoff in a two-week period}

These reward values assign depression and hypo(mania) equal weights consistent with the goal of minimizing the number of mood episodes during the monitoring periods.

We ran our Copula-Adjusted MARL algorithm (Algorithm 1) on our BD dataset for 17,000 iterations to learn the treatment recommendations. Figure 1 shows the critic loss over all iterations. We observe a temporary increase in the critic loss after adding the potential future value in the 5,000th (corresponding to 25 on the x-axis) iteration, but it then quickly converges, demonstrating the validity of our one-step future reward modification. Once convergence is obtained, we extracted the policy corresponding to the minimum critic loss as the final learned (i.e., optimal) recommendation policy.

To assess the potential improvement in (reduction in) the number of mood episodes for BD patients, we evaluated our learned policies using the off-policy evaluation (OPE) method proposed by [59]. This method, specifically introduced for multi-agent reinforcement learning, accounts for the interference effects among agents, aligning well with our approach of incorporating copulas into the setting. However, the algorithm necessitates having access to a known (or fully estimated) behavior policy in the observational data. To overcome this challenge, we set $\alpha = \infty$ and $\beta = 1$ in the modified ICQ algorithm (discouraging actions not observed in the data and preserving the full dependence structure as in the data), thereby performing behavior cloning to obtain the behavior policy.

Figure 2 shows the OPE results for each patient, depicting the individual impact of the counterfactual policy offered by our algorithm (Algorithm 1). It is important to highlight that the OPE estimator we adopted is doubly robust, meaning that it provides reliable estimates even if either the behavior policy or the learned model is misspecified. This gives additional confidence in the accuracy of the results shown. On average, the number of mood episodes decreased by 15.7%. Furthermore, our analysis shows that only nine out of fifty-two patients exhibited an increased number of mood episodes, as shown by negative bars in Fig. 2.

Next, to gain deeper insights, we randomly selected patients from the cohort and investigated their counterfactual trajectories under the policy learned from our algorithm versus that under the behavior policy. Figure 3 shows a comparison for a representative patient who experienced five mood episodes over the course of 16 weeks. For example, during the third period, this patients is recommended to take a lot more steps than in the second period, while the learned policy recommends also increasing the total sleep time.

Since our state space encompasses $3^7 = 2187$ different combinations, we use our algorithm to output suggested treatments for each combination. We then run a decision-tree model using states as input and treatment suggestions as output to 1) generate an interpretable representation of the policies our algorithm suggests, and 2) obtain clinically relevant insights. The tree is shown in Figure 4. There are 55 unique treatments with the detailed corresponding relations shown in the table in the supplementary material. Among the 55 unique treatments aimed at minimizing the number of mood episodes, 13 did not appear in the behavioral treatment data from the dataset. Figure 5 demonstrates the feature importance obtained from the decision tree. Bedtime consistency, resting heart rate, and steps are all very important in providing corresponding treatment suggestions. Additional plots based on SHAP (Shapley Additive exPlanations, [35]) values can be found in the supplementary material. These plots also show directional relationships in addition to feature importance values, offering additional insights with important clinical relevance.

[1] "small" corresponds to a 5% - 10% increase in steps or total sleep time, a 0 - 5% decrease in BC, "medium" corresponds to a 10% - 20% increase in steps or total sleep time, a 5 - 10% decrease in BC, and "large" corresponds to an over 20% increase in steps or total sleep time, an over 10% decrease in BC. Those cutoffs are chosen to make actions of different agents roughly evenly distributed in the dataset.

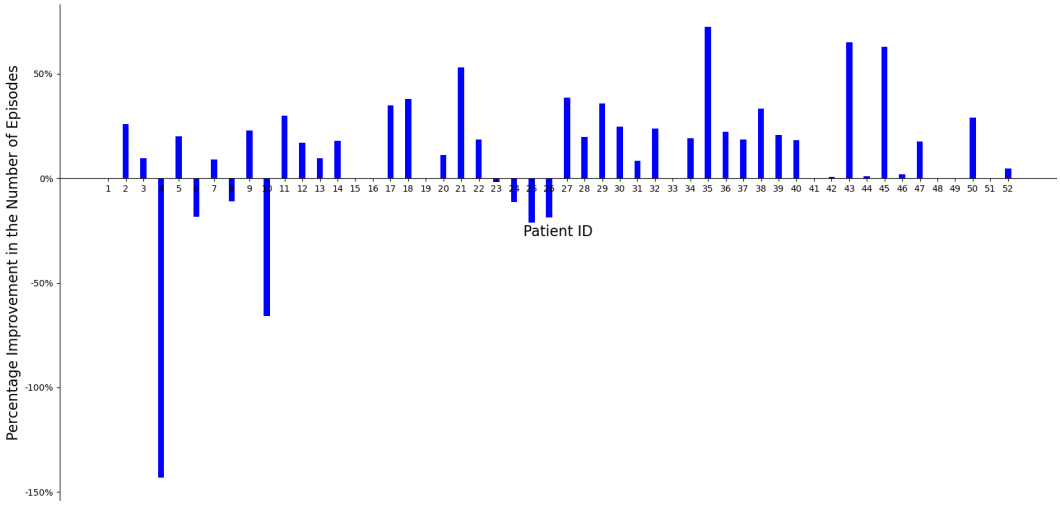

Fig. 2. Off-policy evaluation (OPE) results showing the impact of the learned policy on each patient's number of mood episodes. The average improvement across all patients is 15.7 %.

Heterogeneity Analysis

In a comparative analysis, 25 participants were identified based on their higher gain from our algorithm. We then compare the characteristics of these selected patients against the full BD sample (N = 52).

The high-gain subset has an average age of approximately 38.5 years, compared to 40.2 years in the full BD sample (p = 0.31). Although this difference is not statistically significant, it points toward a slightly younger demographic in the high-gain group. In terms of sex distribution, about 60% of the high-gain participants are female versus 72% in the full sample (p = 0.45). A similar trend emerges in race, with about 68% of high-gain participants identifying as White, compared to roughly 75% in the broader sample. While none of these comparisons rise to the level of statistical significance, they hint that the high-gain subset may show modest demographic variations relative to the full BD cohort.

Looking at disease-specific characteristics, the age of BD onset is slightly later among the high-gain group (18.2 years) compared to the broader BD sample (17.5 years), with a p-value of 0.64 indicating no statistically significant difference. The duration of illness follows suit, as the high-gain subgroup reports an average of 20.3 years since onset, versus 22.7 years for the full sample (p = 0.29). Although these p-values exceed conventional thresholds for significance, the mild trends observed may still represent meaningful differences in clinical trajectory and illness course. Ultimately, these comparisons underscore the potential subtleties in demographic and onset features that distinguish high-gain participants from the wider BD population, warranting further investigation with larger or more granular datasets.

Using K-means clustering, we identified three distinct subgroups based on features such as age, age of onset, and duration[2]. This method partitions individuals into clusters such that those within the same cluster share a greater similarity in their characteristics compared to those in other clusters. The results presented in Figure 6 reveal meaningful patterns. Notably, one cluster is composed of younger individuals with shorter illness durations, and another is composed of older individuals with extended durations of illness. These results highlight the heterogeneity in the dataset and provide insights into potential subgroup-specific trajectories or interventions. Specifically, the accompanying pairplot in Figure 6, which visualizes the clustering results, offers a detailed view of how the features interact across clusters. Diagonal histograms display the distribution of each feature within clusters, while scatterplots depict relationships between feature pairs. As can be seen, clear distinctions in age and duration emerge between the clusters, reflecting distinct clinical profiles within the data.

Discussion

To help patients with bipolar disorder, we proposed an algorithm that demonstrates significant potential to reduce weeks with clinically significant symptoms. By leveraging patient states to tailor self-care recommendations, our algorithm suggests behavior changes associated with an average 15.7% decrease in weeks with clinically significant symptoms. These findings align with existing research highlighting the critical role of physical activity and sleep in managing bipolar disorder. For instance, [15] has shown that regular physical activity is associated with improved mood and reduced symptom severity in bipolar disorder patients. Similarly, [18]

[2] Given 1) recent studies showing that the benefit of including race in models is often less than expected [12] and 2) the fact that the majority of patients in our sample are white, we did not include race in our analysis.

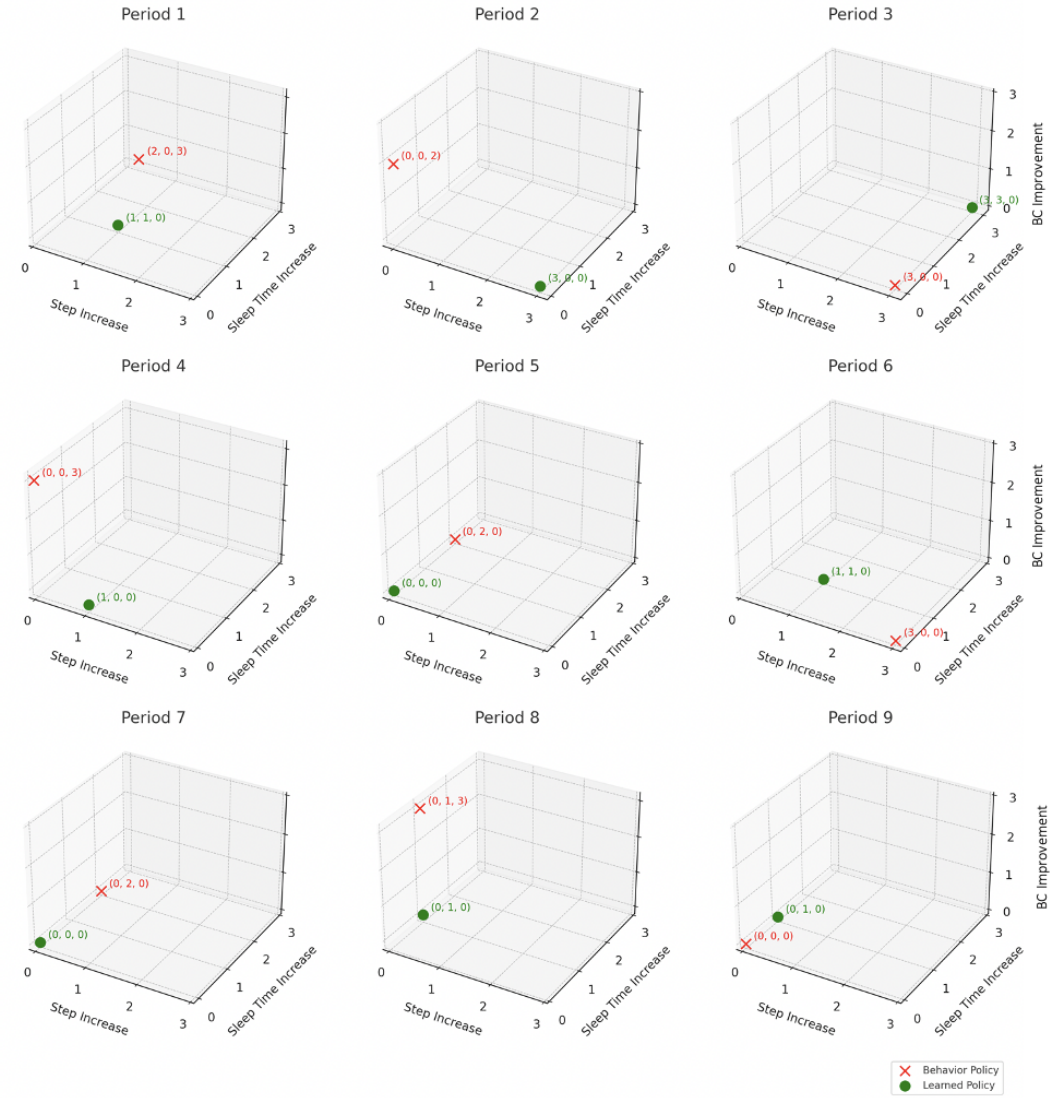

Fig. 3. 3D plot for policy comparison for a representative patient during the 9 periods. The number on each axis indicates the extent of change, where 0 = no change, 1 = small change, 2 = medium change, and 3 = large change compared with the previous period (see also footnote 1).

emphasizes the importance of sleep hygiene in stabilizing mood and preventing relapse.

Bipolar disorder profoundly affects individuals, causing significant disruptions in mood, daily functioning, and overall well-being. It is estimated to impact approximately 2.8% of U.S. adults—around 7 million individuals (National Institute of Mental Health, 2021). Globally, the disorder contributed to 9.9 million Disability-Adjusted Life Years (DALYs) in 2013, underscoring its serious effects on quality of life and productivity ([19]). By reducing the number of weeks with clinically significant mood symptoms, our algorithm has the potential to alleviate much of this personal suffering and improve overall well-being.

In addition to personal benefits, the economic impact of bipolar disorder is substantial. In the United States alone, the disorder generates an estimated $202.1 billion in direct and indirect costs annually ([11]). Personalized management recommendations that help reduce mood symptoms could decrease the need for emergency interventions, hospitalizations, and other costly treatments, resulting in significant savings. Additionally, reducing the personal and clinical burden may mitigate indirect costs associated with lost productivity and long-term disability ([19]).

Furthermore, reducing the potential risk of suicide among bipolar disorder patients is a critical public health goal. Suicidal ideation is highly prevalent during depressive phases, occurring in 79% of patients, and is closely linked to mixed mood episodes ([38]). More precisely, suicide attempts or completions mostly occur during severe depressive or mixed mood episodes (78–89%) and less frequently during dysphoric mania (11–20%) or euphoric mania and euthymia (0–7%) ([3], [24], [44], [47], [58], [67]). By reducing the frequency of mood episodes through tailored self-care recommendations, our algorithm could reduce suicide attempts in this high-risk population. This not only has the potential to save lives but also to reduce the

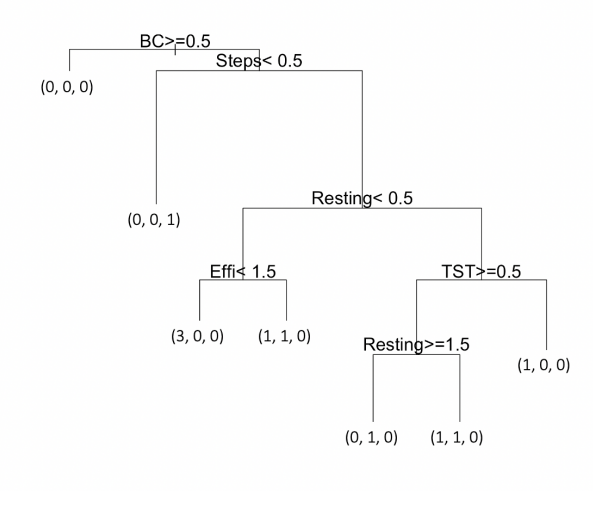

Fig. 4. The Decision Tree structure

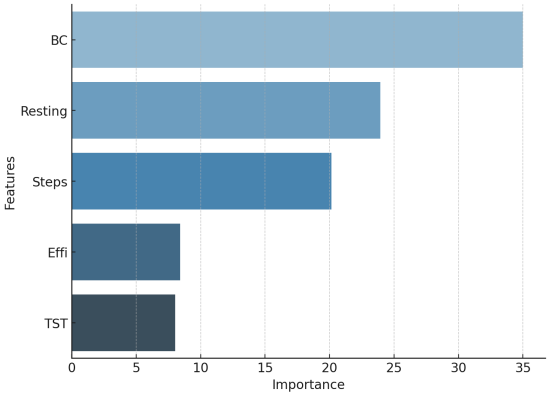

Fig. 5. Feature importance based on the Decision Tree analyses

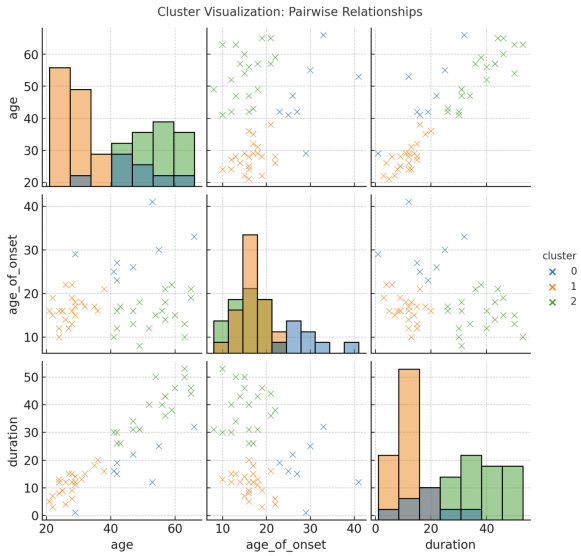

Fig. 6. Pairwise relationships among features (age, age of onset, and duration) across clusters identified using K-means clustering. Diagonal histograms show the distribution of each feature within clusters, while scatterplots illustrate interactions between feature pairs.

associated healthcare costs, which are significant due to high hospitalization and emergency intervention rates ([19]).

The quality-adjusted life years (QALYs) and health-related quality of life (HRQOL) measures further indicate that improvements in managing depressive episodes significantly enhance patient well-being while reducing the long-term burden on healthcare systems ([46]).

Our algorithm's ability to offer personalized self-care recommendations and provide actionable goals for patients makes it a promising tool for managing bipolar disorder. By decreasing the frequency of mood episodes, it could lead to fewer hospitalizations and reduced healthcare costs, improving resource allocation and mental health outcomes on a broader scale. In light of the increased public reporting efforts of hospitals' outcomes ([52], [51]), this can in turn help hospitals improve their performance measures. Furthermore, integrating such an algorithm into routine clinical practice has the potential to transform the standard of care for bipolar disorder, making it more responsive to patient needs and more effective in mitigating the severe impacts of this debilitating illness.

In addition to its public health impact, our algorithm is distinguished by several innovative features. The use of copulas to model dependencies among agents is a key advancement, allowing the algorithm to capture complex interactions that are typically challenging in multi-agent settings. This approach mitigates the limitations of traditional Q-value decomposition methods, ensuring more accurate treatment recommendations. Furthermore, the algorithm's offline operation is particularly advantageous in healthcare applications, where data privacy and patient safety are paramount. Finally, despite the episodic nature of the data, the algorithm's stationary policy learning underscores its robustness and adaptability in addressing the long-term management of bipolar disorder. These features collectively make our algorithm a novel and powerful tool for personalization in healthcare, offering notable advantages for improving care across many chronic conditions well beyond the bipolar disorder focus of this study.

Supplementary material

Supplementary material is available at PNAS Nexus online.

Funding

This research was supported by a Young Investigator Grant from the Brian & Behavior Research Foundation (#28537; to JML), a grant from the Harvard Brain Science Initiative Bipolar Disorder Seed Grant Program, and a Pathways Research Award from Alkermes, Inc. The data collection for the longitudinal study was supported in part by the Baszucki Brain Research Fund (to KEB) and the Harvard Brain Science Initiative Bipolar Disorder Seed Grant Program (to KEB). Additionally, Dr. Lipschitz's time was partially supported by the National Institute of Mental Health (NIMH) Grant MH120324. Dr. Saghafian's time was partially supported through a grant from Harvard's Middle East Initiative Kuwait Science Program, which is aimed at improving population health via machine learning and AI-enabled mobile health interventions. The funding organizations played no role in study design, data collection, analysis and interpretation of data, or the writing of this manuscript.

Author contributions statement

SL and SS were primarily responsible for data analysis and model building, whereas JML and KEB were primarily responsible for data acquisition. SL conducted the experiments and was primarily responsible for drafting the manuscript. All other authors—SS, JML and KEB—were involved in the interpretation of findings and critical revision of the manuscript and approval of the final submission.

Data availability

We are not currently able to share data because data collection is still ongoing (findings presented are based on interim analyses) and the study is not yet federally funded. The informed consent used in this study allows for data sharing and we do expect to share our data once data collection is complete and a data repository is in place.

References

1. Edward G Altman, Donald Hedeker, James L Peterson, and John M Davis. The altman self-rating mania scale. *Biological psychiatry*, 42(10):948–955, 1997.
2. Mawulolo K Ameko, Miranda L Beltzer, Lihua Cai, Mehdi Boukhechba, Bethany A Teachman, and Laura E Barnes. Offline contextual multi-armed bandits for mobile health interventions: A case study on emotion regulation. In *Fourteenth ACM Conference on Recommender Systems*, pages 249–258, 2020.
3. Ross J Baldessarini, Maurizio Pompili, and Leonardo Tondo. Suicidal risk in antidepressant drug trials. *Archives of general psychiatry*, 63(3):246–248, 2006.
4. Andrew Bennett and Nathan Kallus. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *Operations Research*, 72(3):1071–1086, 2024.
5. David Biagioni, Xiangyu Zhang, Dylan Wald, Deepthi Vaidhynathan, Rohit Chintala, Jennifer King, and Ahmed S Zamzam. Powergridworld: A framework for multi-agent reinforcement learning in power systems. In *Proceedings of the thirteenth ACM international conference on future energy systems*, pages 565–570, 2022.
6. André F Carvalho, Dimos Dimellis, Xenia Gonda, Eduard Vieta, Roger S McIntyre, and Konstantinos N Fountoulakis. Rapid cycling in bipolar disorder: a systematic review. *The Journal of clinical psychiatry*, 75(6):16864, 2014.
7. Cristiano Castelfranchi. The theory of social functions: challenges for computational social science and multi-agent learning. *Cognitive Systems Research*, 2(1):5–38, 2001.
8. Paul Chelarescu. Deception in social learning: A multi-agent reinforcement learning perspective. *arXiv preprint arXiv:2106.05402*, 2021.
9. Ken Cheung, Wodan Ling, Chris J Karr, Kenneth Weingardt, Stephen M Schueller, and David C Mohr. Evaluation of a recommender app for apps for the treatment of depression and anxiety: an analysis of longitudinal user engagement. *Journal of the American Medical Informatics Association*, 25(8):955–962, 2018.
10. Tianshu Chu, Sandeep Chinchali, and Sachin Katti. Multi-agent reinforcement learning for networked system control. *arXiv preprint arXiv:2004.01339*, 2020.
11. Martin Cloutier, Mallik Greene, Annie Guerin, Maelys Touya, and Eric Wu. The economic burden of bipolar i disorder in the united states in 2015. *Journal of affective disorders*, 226:45–51, 2018.
12. Madison Coots, Soroush Saghafian, David M Kent, and Sharad Goel. A framework for considering the value of race and ethnicity in estimating disease risk. *Annals of Internal Medicine*, 2024.
13. N Cruz, E Vieta, M Comes, Josep Maria Haro, C Reed, J Bertsch, et al. Rapid-cycling bipolar i disorder: course and treatment outcome of a large sample across europe. *Journal of psychiatric research*, 42(13):1068–1075, 2008.
14. Damien Ernst and Arthur Louette. Introduction to reinforcement learning. *Feuerriegel, S., Hartmann, J., Janiesch, C., and Zschech, P.(2024). Generative ai. Business & Information Systems Engineering*, 66(1):111–126, 2024.
15. Joseph Firth, Josh A Firth, Brendon Stubbs, Davy Vancampfort, Felipe B Schuch, Mats Hallgren, Nicola Veronese, Alison R Yung, and Jerome Sarris. Association between muscular strength and cognition in people with major depression or bipolar disorder and healthy controls. *JAMA psychiatry*, 75(7):740–746, 2018.
16. Evan M Forman, Stephanie G Kerrigan, Meghan L Butryn, Adrienne S Juarascio, Stephanie M Manasse, Santiago Ontañón, Diane H Dallal, Rebecca J Crochiere, and Danielle Moskow. Can the artificial intelligence technique of reinforcement learning use continuously-monitored digital data to optimize treatment for weight loss? *Journal of behavioral medicine*, 42:276–290, 2019.
17. Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International conference on machine learning*, pages 2052–2062. PMLR, 2019.
18. Philip D Harvey, Kelly Posner, Nallakkandi Rajeevan, Kseniya V Yershova, Mihaela Aslan, and John Concato. Suicidal ideation and behavior in us veterans with schizophrenia or bipolar disorder. *Journal of psychiatric research*, 102:216–222, 2018.
19. Hairong He, Chuanyu Hu, Zhenhu Ren, Ling Bai, Fan Gao, and Jun Lyu. Trends in the incidence and dalys of bipolar disorder at global, regional, and national levels: results from the global burden of disease study 2017. *Journal of psychiatric research*, 125:96–105, 2020.
20. Xinyu Hu, Min Qian, Bin Cheng, and Ying Kuen Cheung. Personalized policy learning using longitudinal mobile health data. *Journal of the american statistical association*, 116(533):410–420, 2021.
21. Kay Redfield Jamison. Suicide and bipolar disorder. *Bipolar Disorder*, pages 115–119, 2019.
22. Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. In *International conference on machine learning*, pages 652–661. PMLR, 2016.
23. Lewis L Judd, Hagop S Akiskal, Pamela J Schettler, Jean Endicott, Jack Maser, David A Solomon, Andrew C Leon, John A Rice, and Martin B Keller. The long-term natural history of the weekly symptomatic status of bipolar i disorder. *Archives of general psychiatry*, 59(6):530–537, 2002.
24. LV Kessing and PK Andersen. Does the risk of developing dementia increase with the number of episodes in patients with depressive disorder and in patients with bipolar disorder? *Journal of Neurology, Neurosurgery & Psychiatry*, 75(12):1662–1666, 2004.

25. Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. The phq-8 as a measure of current depression in the general population. *Journal of affective disorders*, 114(1-3):163–173, 2009.
26. Aviral Kumar, Aurick Zhou, George Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *Advances in Neural Information Processing Systems*, 33:1179–1191, 2020.
27. Eric B Laber, Kristin A Linn, and Leonard A Stefanski. Interactive model building for q-learning. *Biometrika*, 101(4):831–847, 2014.
28. Jae Won Lee and Jangmin O. A multi-agent q-learning framework for optimizing stock trading systems. In *International Conference on Database and Expert Systems Applications*, pages 153–162. Springer, 2002.
29. Jae Won Lee, Jonghun Park, O Jangmin, Jongwoo Lee, and Euyseok Hong. A multiagent approach to q-learning for daily stock trading. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 37(6):864–877, 2007.
30. Sergey Levine, Aviral Kumar, George Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *arXiv preprint arXiv:2005.01643*, 2020.
31. Chenghao Li, Tonghan Wang, Chengjie Wu, Qianchuan Zhao, Jun Yang, and Chongjie Zhang. Celebrating diversity in shared multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:3991–4002, 2021.
32. 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.
33. Jessica M Lipschitz, Sidian Lin, Soroush Saghafian, Chelsea K Pike, and Katherine E Burdick. Digital phenotyping in bipolar disorder: Using longitudinal fitbit data and personalized machine learning to predict mood symptomatology. *Acta Psychiatrica Scandinavica*, 2024.
34. Daniel J Luckett, Eric B Laber, Anna R Kahkoska, David M Maahs, Elizabeth Mayer-Davis, and Michael R Kosorok. Estimating dynamic treatment regimes in mobile health using v-learning. *Journal of the American Statistical Association*, 2019.
35. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.
36. Xiaoteng Ma, Yiqin Yang, Chenghao Li, Yiwen Lu, Qianchuan Zhao, and Yang Jun. Modeling the interaction between agents in cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2102.06042*, 2021.
37. Xuan Mai, Quanzhi Fu, and Yi Chen. Packet routing with graph attention multi-agent reinforcement learning. In *2021 IEEE Global Communications Conference (GLOBECOM)*, pages 1–6. IEEE, 2021.
38. Roger S McIntyre, Sara Higa, Quan V Doan, Diana Amari, David Oliveri, Patrick Gillard, and Amanda Harrington. Place of care and costs associated with acute episodes and remission in bipolar i disorder. *Journal of Medical Economics*, 25(1):1110–1117, 2022.
39. Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.
40. Mohammad Noaeen, Atharva Naik, Liana Goodman, Jared Crebo, Taimoor Abrar, Zahra Shakeri Hossein Abad, Ana LC Bazzan, and Behrouz Far. Reinforcement learning in urban network traffic signal control: A systematic literature review. *Expert Systems with Applications*, 199:116830, 2022.
41. Agni Orfanoudaki, Soroush Saghafian, Karen Song, Harini A Chakkera, and Curtiss Cook. Algorithm, human, or the centaur: How to enhance clinical care? 2022.
42. Afshin Oroojlooy and Davood Hajinezhad. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.
43. Fabio Pardo, Arash Tavakoli, Vitaly Levdik, and Petar Kormushev. Time limits in reinforcement learning. In *International Conference on Machine Learning*, pages 4045–4054. PMLR, 2018.
44. Maurizio Pompili, Maria Masocco, Monica Vichi, David Lester, Marco Innamorati, Roberto Tatarelli, and Nicola Vanacore. Suicide among italian adolescents: 1970–2002. *European child & adolescent psychiatry*, 18:525–533, 2009.
45. Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51, 2020.
46. Taeho Greg Rhee, Emily S Gillissie, Andrew A Nierenberg, and Roger S McIntyre. Association of current and remitted bipolar disorders with health-related quality of life: Findings from a nationally representative sample in the us. *Journal of Affective Disorders*, 321:33–40, 2023.
47. Zoltán Rihmer. Suicide risk in mood disorders. *Current opinion in psychiatry*, 20(1):17–22, 2007.
48. Benjamin Roche, Jean-François Guégan, and François Bousquet. Multi-agent systems in epidemiology: a first step for computational biology in the study of vector-borne disease transmission. *BMC bioinformatics*, 9:1–9, 2008.
49. Soroush Saghafian. Ambiguous partially observable markov decision processes: Structural results and applications. *Journal of Economic Theory*, 178:1–35, 2018.
50. Soroush Saghafian. Ambiguous dynamic treatment regimes: A reinforcement learning approach. *Management Science*, 2023.
51. Soroush Saghafian and Wallace J Hopp. The role of quality transparency in health care: Challenges and potential solutions. *NAM perspectives*, 2019, 2019.
52. Soroush Saghafian and Wallace J Hopp. Can public reporting cure healthcare? the role of quality transparency in improving patient–provider alignment. *Operations Research*, 68(1):71–92, 2020.
53. Soroush Saghafian, Derya Kilinc, and Stephen J Traub. Dynamic assignment of patients to primary and secondary inpatient units: Is patience a virtue? *HKS Faculty Research Working Paper Series*, 2022.
54. Soroush Saghafian and Susan A Murphy. Innovative health care delivery: The scientific and regulatory challenges in designing mhealth interventions. *NAM perspectives*, 2021, 2021.
55. Soroush Saghafian, Brian Tomlin, and Stephan Biller. The internet of things and information fusion: who talks to who? *Manufacturing & Service Operations Management*, 24(1):333–351, 2022.
56. Eva María Sánchez-Morla, Ana López-Villarreal, Estela Jiménez-López, Ana Isabel Aparicio, Vicente Martínez-Vizcaíno, Rodriguez-Jimenez Roberto, Eduard Vieta, and

José-Luis Santos. Impact of number of episodes on neurocognitive trajectory in bipolar disorder patients: a 5-year follow-up study. *Psychological medicine*, 49(8):1299–1307, 2019.

57. Shai Shalev-Shwartz, Shaked Shammah, and Amnon Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.
58. Leo Sher, Michael F Grunebaum, Gregory M Sullivan, Ainsley K Burke, Thomas B Cooper, J John Mann, and Maria A Oquendo. Testosterone levels in suicide attempters with bipolar disorder. *Journal of psychiatric research*, 46(10):1267–1271, 2012.
59. Chengchun Shi, Runzhe Wan, Ge Song, Shikai Luo, Rui Song, and Hongtu Zhu. A multi-agent reinforcement learning framework for off-policy evaluation in two-sided markets. *arXiv preprint arXiv:2202.10574*, 2022.
60. Kyunghwan Son, Daewoo Kim, Wan Ju Kang, David Earl Hostallero, and Yung Yi. Qtran: Learning to factorize with transformation for cooperative multi-agent reinforcement learning. In *International conference on machine learning*, pages 5887–5896. PMLR, 2019.
61. Rui Song, Weiwei Wang, Donglin Zeng, and Michael R Kosorok. Penalized q-learning for dynamic treatment regimens. *Statistica Sinica*, 25(3):901, 2015.
62. Paul Stang, Cathy Frank, M Ulcickas Yood, Karen Wells, and Steven Burch. Impact of bipolar disorder: results from a screening study. *Primary care companion to the Journal of clinical psychiatry*, 9(1):42, 2007.
63. Peter Stone and Manuela Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8:345–383, 2000.
64. Peter Sunehag, Guy Lever, Audrunas Gruslys, Wojciech Marian Czarnecki, Vinicius Zambaldi, Max Jaderberg, Marc Lanctot, Nicolas Sonnerat, Joel Z Leibo, Karl Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
65. Ambuj Tewari and Susan A Murphy. From ads to interventions: Contextual bandits in mobile health. In *Mobile Health*, pages 495–517. Springer, 2017.
66. Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148. PMLR, 2016.
67. Hanna M Valtonen, Kirsi Suominen, Outi Mantere, Sami Leppämäki, Petri Arvilommi, and Erkki Isometsä. Suicidal behaviour during different phases of bipolar disorder. *Journal of affective disorders*, 97(1-3):101–107, 2007.
68. Hongwei Wang, Lantao Yu, Zhangjie Cao, and Stefano Ermon. Multi-agent imitation learning with copulas. In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part I 21*, pages 139–156. Springer, 2021.
69. Jianhao Wang, Zhizhou Ren, Terry Liu, Yang Yu, and Chongjie Zhang. Qplex: Duplex dueling multi-agent q-learning. *arXiv preprint arXiv:2008.01062*, 2020.
70. Tong Wu, Pan Zhou, Kai Liu, Yali Yuan, Xiumin Wang, Huawei Huang, and Dapeng Oliver Wu. Multi-agent deep reinforcement learning for urban traffic light control in vehicular networks. *IEEE Transactions on Vehicular Technology*, 69(8):8243–8256, 2020.
71. Yin Wu, Brooke Levis, Kira E Riehm, Nazanin Saadat, Alexander W Levis, Marleine Azar, Danielle B Rice, Jill Boruff, Pim Cuijpers, Simon Gilbody, et al. Equivalency of the diagnostic accuracy of the phq-8 and phq-9: a systematic review and individual participant data meta-analysis. *Psychological medicine*, 50(8):1368–1380, 2020.
72. Shiqi Yang, Ping Zhou, Kui Duan, M Shamim Hossain, and Mohammed F Alhamid. emhealth: towards emotion health through depression prediction and intelligent health recommender system. *Mobile Networks and Applications*, 23(2):216–226, 2018.
73. Yaodong Yang and Jun Wang. An overview of multi-agent reinforcement learning from game theoretical perspective. *arXiv preprint arXiv:2011.00583*, 2020.
74. Yiqin Yang, Xiaoteng Ma, Chenghao Li, Zewu Zheng, Qiyuan Zhang, Gao Huang, Jun Yang, and Qianchuan Zhao. Believe what you see: Implicit constraint approach for offline multi-agent reinforcement learning. *Advances in Neural Information Processing Systems*, 34:10299–10312, 2021.
75. Lakshmi N Yatham, Sidney H Kennedy, Sagar V Parikh, Ayal Schaffer, David J Bond, Benicio N Frey, Verinder Sharma, Benjamin I Goldstein, Soham Rej, Serge Beaulieu, et al. Canadian network for mood and anxiety treatments (canmat) and international society for bipolar disorders (isbd) 2018 guidelines for the management of patients with bipolar disorder. *Bipolar disorders*, 20(2):97–170, 2018.
76. Baqun Zhang, Anastasios A Tsiatis, Eric B Laber, and Marie Davidian. A robust method for estimating optimal treatment regimes. *Biometrics*, 68(4):1010–1018, 2012.
77. Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.
78. Sai Qian Zhang, Qi Zhang, and Jieyu Lin. Succinct and robust multi-agent communication with temporal message control. *Advances in Neural Information Processing Systems*, 33:17271–17282, 2020.
79. Zihan Zhang, Xiangyang Ji, and Simon Du. Horizon-free reinforcement learning in polynomial time: the power of stationary policies. In *Conference on Learning Theory*, pages 3858–3904. PMLR, 2022.