Open Access Library Journal
2026, Volume 13, e14925
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Reinforcement Learning-Based Personalized Mood Stabilizer Dosage Optimization

Rocco de Filippis[1*], Abdullah Al Foysal[2]

[1]Department of Neuroscience, Institute of Psychopathology, Rome, Italy
[2]Department of Computer Engineering (AI), University of Genova, Genova, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Reinforcement Learning-Based Personalized Mood Stabilizer Dosage Optimization. *Open Access Library Journal*, **13**: e14925.
https://doi.org/10.4236/oalib.1114925

Received: January 23, 2026
Accepted: March 7, 2026
Published: March 10, 2026

Copyright © 2026 by author(s) and Open Access Library Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Personalized dosing of mood stabilizers remains challenging due to substantial inter-individual variability in symptom severity, treatment responsiveness, and vulnerability to adverse effects. Clinical titration is often slow and heuristic, motivating data-driven strategies that can adapt dosage over time while balancing efficacy and safety. In this study, we formulate mood stabilizer dosing as a finite-horizon Markov decision process and develop a reinforcement learning (RL) framework for patient-specific dosage adjustment. We simulate a heterogeneous cohort of 500 patients using clinically inspired latent factors depression severity, anxiety level, treatment responsiveness, and side-effect sensitivity and define mood trajectories through a nonlinear dose response model cantered on an individualized optimal dose. The RL agent selects among discrete dosage changes (−50, −25, 0, +25, +50 mg) across an 8-step titration horizon. A clinically motivated reward function encourages proximity to a target mood (50) and occupancy of a therapeutic range (45 - 55), while penalizing side effects and abrupt dose fluctuations and rewarding sustained stability. A tabular Q-learning agent trained for 1000 episodes achieved stable convergence with consistently positive episode returns (mean reward ≈ 63.6). In population-level evaluation over 200 held-out patients, the learned policy improved final mood by +20.3 points on average, with 68.5% of patients showing positive improvement. However, sustained stabilization remained limited: mean time spent in the therapeutic range was 16.8%, only 12.5% of patients achieved ≥50% occupancy, and 9.5% ended within range. Dose selection accuracy was modest, with 15.0% of final doses within 20% of the patient-specific optimum. Subgroup analysis indicated reduced stabilization performance in high-depression patients. Overall, results demonstrate the feasibility of RL for adaptive psychiatric dosing in a controlled simulation and highlight key state abstraction, horizon length, and function approximation that must be addressed before translational evaluation on real-world clinical data.

Subject Areas

Psychology

Keywords

Reinforcement Learning, Translational Evaluation, Clinical Data

1. Introduction

Mood stabilizers constitute a fundamental component of pharmacological treatment for affective disorders, including bipolar disorder and treatment-resistant mood dysregulation [1]-[11]. Despite their widespread clinical use, determining an optimal dosage for individual patients remains a persistent challenge. Clinical response to mood stabilizers varies substantially across individuals due to differences in symptom severity, comorbid anxiety, pharmacodynamic sensitivity, and susceptibility to adverse effects [12]-[16]. As a result, standard dosing strategies largely guided by population-level heuristics and clinician experience often require prolonged trial-and-error titration, delaying symptom relief and increasing the risk of sub-optimal outcomes. Recent advances in data-driven medicine have motivated the exploration of adaptive treatment strategies that can account for patient heterogeneity and temporal dynamics. However, many existing computational approaches to psychiatric dosing rely on static regression or classification models that predict outcomes under fixed dosage assumptions. Such approaches are inherently limited in their ability to model sequential decision making, where current treatment choices influence future patient states, symptom trajectories, and clinical risk. This limitation is particularly relevant in mood disorders, where symptom evolution unfolds over time and treatment goals emphasize not only improvement but also sustained stabilization within a therapeutic range. Reinforcement learning (RL) provides a principled framework for addressing these challenges by framing treatment optimization as a sequential decision process. In an RL setting, an agent learns a dosing policy through interaction with an environment, receiving feedback in the form of rewards that encode clinical objectives and constraints. This paradigm enables explicit trade-offs between symptom control, treatment stability, and safety, and allows policies to adapt dynamically to patient-specific characteristics.

In this work, we propose an RL-based framework for personalized mood stabilizer dosing, using a clinically motivated simulation environment that captures longitudinal mood dynamics and inter-individual variability. The environment integrates latent patient factors depression severity, anxiety level, treatment responsiveness, and side-effect sensitivity and models non-linear dose response relationships over a finite titration horizon. We employ a tabular Q-learning agent as an interpretable baseline and evaluate performance through training dynamics, individual patient case studies, and population-level clinical metrics. Our results

demonstrate the feasibility of RL for adaptive psychiatric dosing while highlighting key challenges that inform future translational research.

2. Related Work

The application of reinforcement learning (RL) in healthcare has gained increasing attention as a promising paradigm for sequential clinical decision making [17]-[22]. Early work demonstrated the potential of RL for treatment planning in intensive care units, including ventilation control and sepsis management, where actions influence long-term patient outcomes rather than immediate rewards. These studies established RL as a suitable framework for optimizing dynamic treatment regimens under uncertainty and delayed clinical effects.

In psychiatry and psychopharmacology, computational approaches to medication optimization have traditionally relied on regression, classification, or survival models to predict treatment response based on baseline patient characteristics [23]-[26]. While informative, such static models are limited in their capacity to model treatment adaptation over time. More recent efforts have explored RL-based strategies for antidepressant selection, dosage adjustment, and symptom management, often using Markov decision processes to capture temporal dependencies [27]-[29]. However, many of these studies either employ simplified state representations or focus on narrow clinical objectives, limiting their applicability to mood stabilization tasks that require sustained control within a therapeutic range. Adaptive treatment strategies, such as dynamic treatment regimens and contextual bandits, have been proposed as intermediate approaches between static prediction and full RL. While contextual bandits allow personalization based on patient features, they typically ignore long-term consequences of treatment decisions and are therefore insufficient for modelling longitudinal mood trajectories and cumulative side effects. RL, by contrast, explicitly optimizes long-term outcomes and enables trade-offs between competing clinical goals. Parallel to these developments, digital phenotyping and computational psychiatry have introduced fine-grained behavioural and symptom measurements derived from clinical assessments, wearable sensors, and mobile devices [30]-[37]. These data streams have motivated interest in closed loop and adaptive treatment systems. Nevertheless, many existing studies stop at risk prediction or symptom monitoring rather than actionable treatment optimization.

In contrast to prior work, the present study integrates clinically motivated reward design, patient heterogeneity, and longitudinal mood dynamics within a unified RL framework for personalized mood stabilizer dosing. By emphasizing interpretability and comprehensive evaluation, this work contributes a methodologically grounded baseline for future advances in adaptive psychiatric treatment optimization.

3. Problem Formulation

3.1. Markov Decision Process Formulation

We formulate personalized mood stabilizer dosage optimization as a finite-hori-

zon Markov Decision Process (MDP), defined by the tuple

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma),$$

where the agent sequentially selects dosage adjustments to maximize expected cumulative clinical benefit over a fixed titration horizon [38]-[41].

3.1.1. State Space

The state space $\mathcal{S}$ encodes patient-specific clinical context and current treatment status. At time step t, the state $s_t \in \mathcal{S}$ is represented as a continuous vector:

$$s_t = [d, a, r, \sigma, \tilde{u}_t, \tilde{m}_t, \tilde{t}],$$

where:

- $d \in [0,1]$ denotes depression severity,
- $a \in [0,1]$ denotes anxiety level,
- $r \in [0,1]$ represents treatment responsiveness,
- $\sigma \in [0,1]$ captures side-effect sensitivity,
- $\tilde{u}_t = u_t / u_{\max}$ is the normalized current dose,
- $\tilde{m}_t = m_t / 100$ is the normalized mood score,
- $\tilde{t} = t/T$ is the normalized time step within the episode horizon T.

All patient attributes except dose, mood, and time remain fixed within an episode, reflecting stable latent characteristics during short-term titration.

3.1.2. Action Space

The action space $\mathcal{A}$ consists of discrete dosage adjustments:

$$\mathcal{A} = \{-50, -25, 0, +25, +50\} \text{ mg},$$

chosen to reflect conservative and clinically plausible titration steps. The selected action $a_t \in \mathcal{A}$ modifies the current dose subject to safety constraints:

$$u_{t+1} = \text{clip}(u_t + a_t, u_{\min}, u_{\max}),$$

where $u_{\min} = 200 \text{ mg}$ and $u_{\max} = 600 \text{ mg}$.

3.1.3. Transition Dynamics

State transitions are governed by a non-linear stochastic process modelling mood evolution. The mood at time $t+1$ is given by:

$$m_{t+1} = m_t + 100 \cdot \Delta m_t,$$

where the mood change Δm_t is defined as:

$$\Delta m_t = r \cdot f(u_{t+1}, u^*) - \lambda(d + a) + \varepsilon_t.$$

Here, u^* denotes the patient-specific optimal dose, and $f(\cdot)$ is a Gaussian dose-response function:

$$f(u, u^*) = \alpha \exp\left(-\frac{(u/u^* - 1)^2}{2\sigma_d^2}\right),$$

with scaling constant α and spread parameter σ_d. The term $\lambda(d + a)$ models

disease burden, and $\varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ captures stochastic variability. Mood values are clipped to $[0,100]$.

3.1.4. Reward Function

The reward function $R\left(s_t, a_t, s_{t+1}\right)$ encodes clinical objectives:

$$R = R_{\text{mood}} + R_{\text{target}} + R_{\text{side}} + R_{\text{stability}} + R_{\text{consistency}}.$$

Mood proximity is rewarded via a quadratic penalty:

$$R_{\text{mood}} = 10\left(1 - \left|\frac{m_{t+1} - 50}{50}\right|^2\right),$$

with an additional bonus $R_{\text{target}} = 5$ if $m_{t+1} \in [45,55]$. Side effects are penalized proportionally to overdose and sensitivity:

$$R_{\text{side}} = -8 \cdot \text{SE}_{t+1}.$$

Abrupt dosage changes incur a stability penalty:

$$R_{\text{stability}} = -0.2 \cdot \frac{|a_t|}{50}.$$

Finally, a consistency bonus rewards sustained therapeutic control across consecutive steps.

3.1.5. Markov Property and Horizon

The environment satisfies the Markov property, as the transition distribution $P\left(s_{t+1} \mid s_t, a_t\right)$ depends only on the current state and action, not on prior history. All necessary historical information is embedded in s_t through current mood, dose, and time index. Episodes terminate after a fixed horizon $T = 8$, and a discount factor $\gamma = 0.9$ prioritizes long-term clinical outcomes over short-term gains.

3.2. Clinical Target Definition

The primary therapeutic objective of the proposed framework is to maintain patient mood within a clinically meaningful and stable range rather than to maximize short-term symptom improvement alone. In affective disorder management, treatment success is typically defined not by transient mood elevation but by sustained stabilization that avoids both residual depressive symptoms and excessive mood activation. Accordingly, we define a therapeutic mood range of 45 - 55, cantered at an ideal target value of 50. This target range reflects a balanced clinical state in which symptoms are sufficiently controlled while minimizing the risk of overtreatment and adverse effects. Mood values below this range correspond to persistent depressive burden, whereas values above the upper bound may indicate overmedication, emotional blunting, or instability. The use of a bounded target interval rather than a single optimal point aligns with real-world psychiatric practice, where acceptable outcomes are defined by tolerance windows rather than exact numerical thresholds. The clinical target is explicitly encoded in the reward structure of the reinforcement learning environment. The agent receives increas-

ing reward as the patient's mood approaches the target centre and an additional positive bonus when mood falls within the therapeutic range. This design encourages both convergence toward the desired mood level and sustained occupancy within the acceptable interval. Importantly, the reward formulation penalizes oscillatory behaviour around the target, thereby promoting stability over aggressive dosage adjustments.

By incorporating the therapeutic range directly into the optimization objective, the framework prioritizes long-term mood regulation over short-term gains. This approach enables the learning agent to internalize clinically relevant trade-offs between efficacy, safety, and stability, and ensures that learned policies reflect treatment goals consistent with standard psychiatric decision-making. As such, the clinical target definition serves as a critical bridge between computational optimization and real-world treatment objectives.

4. Patient Data Simulation

4.1. Synthetic Cohort Generation

To enable controlled evaluation of the proposed reinforcement learning framework, we generated a synthetic cohort of 500 patients designed to reflect clinically plausible heterogeneity observed in affective disorder populations. Each patient is characterized by a set of latent attributes representing core dimensions of psychiatric variability, including depression severity, anxiety level, treatment responsiveness, and sensitivity to side effects. These attributes were sampled from beta distributions with parameters chosen to produce asymmetric and bounded variability, consistent with empirical findings in clinical psychiatry [42]-[48]. Patient age was drawn from a normal distribution cantered at mid-adulthood to reflect typical treatment demographics, while baseline mood scores were derived from a linear combination of depression and anxiety severity with added stochastic noise. Baseline mood values were constrained to the clinically relevant range of 30 - 60, corresponding to mild-to-moderate symptom burden prior to effective stabilization. This constraint avoids extreme values that are unlikely to be encountered in routine outpatient titration and ensures that mood dynamics remain interpretable throughout simulation. In addition, each patient was assigned an individualized optimal dose u^* sampled from a truncated normal distribution with mean 175 mg and standard deviation 40 mg, constrained to the clinically acceptable range of 50 - 300 mg. This formulation captures realistic pharmacodynamic variability while preventing extreme outlier values. This latent optimal dose defines the centre of a non-linear dose response curve governing mood evolution. By explicitly modelling patient-specific optima, the environment captures inter-individual differences in pharmacodynamic response, which are central to the personalization objective of the study. Overall, the synthetic cohort is constructed to balance realism and experimental control, enabling systematic assessment of learning behaviour across diverse patient profiles while maintaining transparency in model assumptions.

4.2. Justification of Synthetic Modelling

The use of synthetic patient data is motivated by ethical, practical, and methodological considerations [49]-[52]. Real-world clinical datasets involving psychiatric medication titration are often limited by privacy constraints, incomplete longitudinal measurements, and confounding treatment effects. Synthetic modelling allows for reproducible experimentation under fully specified conditions, facilitating clear attribution of outcomes to algorithmic behaviour rather than uncontrolled clinical variability. Importantly, synthetic data enable precise evaluation of counterfactual dosing strategies, which are not observable in retrospective clinical data. This capability is essential for reinforcement learning, where policy evaluation requires exploration of alternative action sequences. While the simulated environment does not claim to replicate the full complexity of real-world psychiatry, it preserves key statistical and dynamical properties relevant to adaptive dosing. Accordingly, this study does not aim at immediate clinical deployment. Instead, the synthetic framework serves as a methodological testbed for validating reinforcement learning strategies, identifying limitations, and guiding future extensions toward real-world data integration and prospective clinical evaluation.

5. Reinforcement Learning Framework

5.1. Environment Dynamics

Mood evolution is modelled as a stochastic, non-linear process driven by medication dosage, patient responsiveness, and disease burden. Let m_t and u_t denote the patient's mood score and dosage at time step t, respectively. Following action a_t, the updated dose is

$$u_{t+1} = \text{clip}\left(u_t + a_t, u_{\min}, u_{\max}\right),$$

where $u_{\min} = 200\text{ mg}$ and $u_{\max} = 600\text{ mg}$.

The change in mood is governed by a bell-shaped dose–response relationship cantered at a patient-specific optimal dose u^*:

$$\Delta m_t = r \cdot \alpha \exp\left(-\frac{\left(u_{t+1}/u^* - 1\right)^2}{2\sigma_d^2}\right) - \lambda\left(d + a\right) + \varepsilon_t,$$

where r denotes treatment responsiveness, d and a denote depression and anxiety severity, α controls the maximum therapeutic effect, and $\varepsilon_t \sim \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ captures stochastic variability. The mood state is then updated as

$$m_{t+1} = \text{clip}\left(m_t + 100 \cdot \Delta m_t, 0, 100\right).$$

Side effects are triggered when dosing exceeds the optimal level:

$$\text{SE}_{t+1} = \max\left(0, u_{t+1} - u^*\right) \cdot \sigma \cdot \eta,$$

where σ represents side-effect sensitivity and η is a stochastic scaling factor. This formulation reflects the clinical observation that overtreatment, rather than

underdosing, is a primary driver of adverse effects. The present side-effect formulation assumes that adverse effects primarily occur when dosing exceeds the individualized optimal level. While this reflects common clinical patterns of overdose-related toxicity, real-world pharmacodynamics may also produce side effects at suboptimal doses due to metabolic variability or idiosyncratic sensitivity. Future extensions may incorporate probabilistic adverse event modelling independent of overdose to better capture clinical complexity.

5.2. Reward Design

The reward function encodes clinically meaningful objectives by combining symptom control, safety, and stability:

$$R_t = R_{\text{mood}} + R_{\text{target}} + R_{\text{side}} + R_{\text{stability}} + R_{\text{consistency}}.$$

Mood proximity is rewarded using a quadratic penalty:

$$R_{\text{mood}} = 10\left(1 - \left|\frac{m_{t+1} - 50}{50}\right|^2\right).$$

An additional bonus is provided if mood lies within the therapeutic range:

$$R_{\text{target}} = \begin{cases} 5, & \text{if } m_{t+1} \in [45, 55], \\ 0, & \text{otherwise.} \end{cases}$$

Side effects incur a penalty proportional to severity:

$$R_{\text{side}} = -8 \cdot \text{SE}_{t+1}.$$

Abrupt dosage changes are discouraged via a stability penalty:

$$R_{\text{stability}} = -0.2 \cdot \frac{|a_t|}{50}.$$

Finally, a consistency bonus is applied when mood remains within the therapeutic range over multiple consecutive steps, encouraging sustained stabilization rather than transient control. This composite reward design aligns with clinical priorities and mitigates reward exploitation.

5.3. Learning Algorithm

We employ a tabular Q-learning agent with discretized state representation as a transparent and stable baseline. Continuous variables were discretized using uniform binning. Depression severity and anxiety were divided into 5 bins each. Treatment responsiveness and side-effect sensitivity were discretized into 4 bins. Dose was divided into 8 bins across the 50 - 300 mg range. Mood was discretized into 10 bins spanning 0 - 100, and time was represented by 8 discrete steps corresponding to the titration horizon. This resulted in a finite but tractable state space suitable for tabular Q-learning while preserving clinically meaningful granularity. Continuous state variables are discretized into finite bins, and learning proceeds via the standard Bellman update:

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha\left[r_t + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right],$$

where α is the learning rate and $\gamma = 0.9$ is the discount factor.

Action selection follows an ε-greedy policy with decaying exploration. Although simple, tabular Q-learning provides interpretability, convergence stability, and direct attribution of learned behaviour to state-action values. This choice establishes a strong methodological baseline upon which more expressive function-approximation methods can be built.

6. Training Dynamics and Convergence

Figure 1 summarizes the training dynamics of the reinforcement learning agent over 1000 episodes, illustrating convergence behaviour, mood regulation performance, and therapeutic stability.

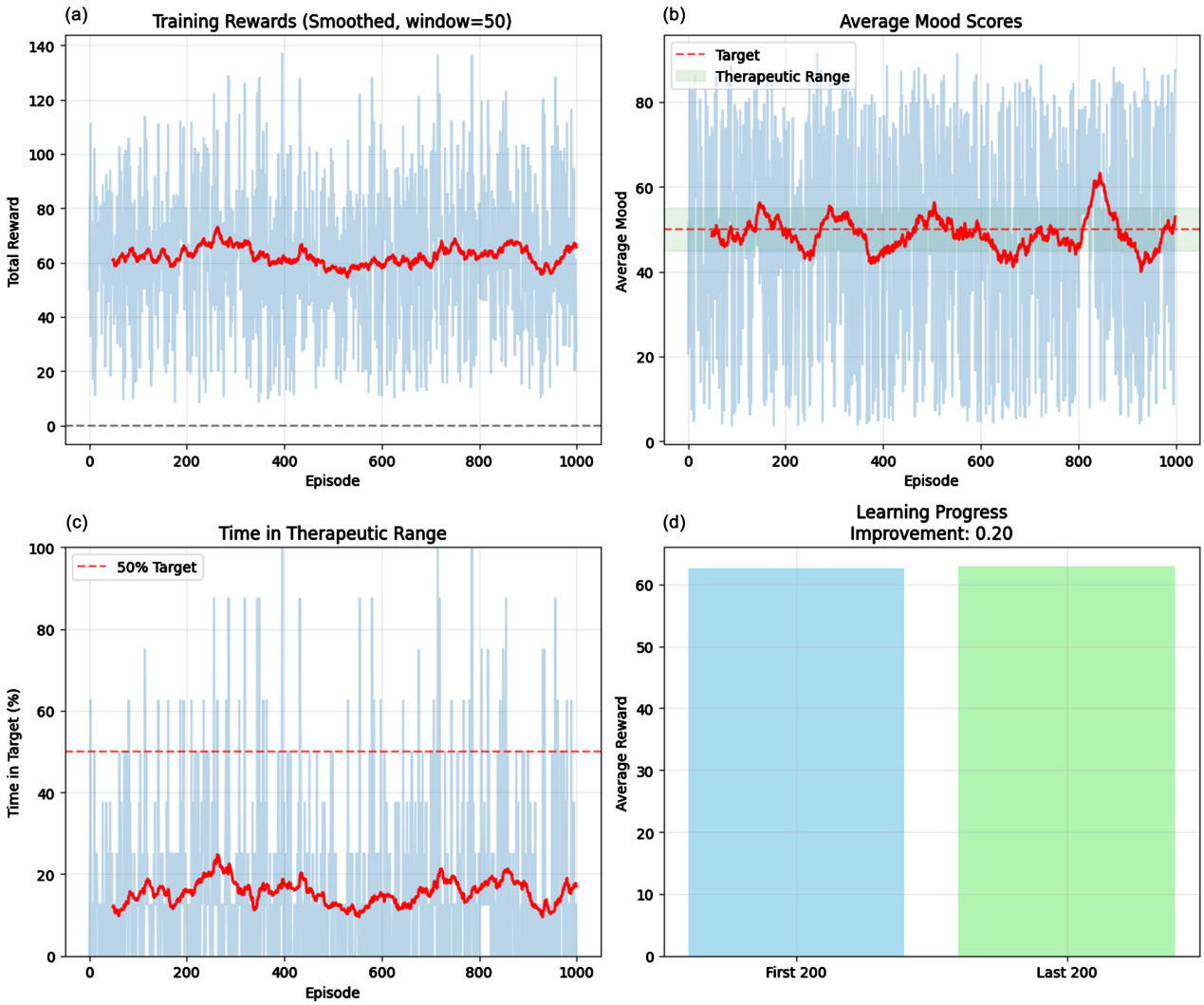

Figure 1. Training performance overview.

Figure 1(a) **(top-left)** shows the total episode reward across training episodes,

with a rolling average computed using a window size of 50 episodes. While individual episode rewards exhibit high variance due to stochastic patient sampling and exploration, the smoothed trajectory demonstrates a clear upward trend followed by stable convergence around positive values. This pattern indicates that the agent successfully learns a dosing policy that consistently balances symptom improvement, side-effect minimization, and dosage stability. Notably, no reward collapse or instability is observed, suggesting robust learning under heterogeneous patient conditions.

Figure 1(b) **(top-right)** presents the average mood score per episode. The red dashed line indicates the target mood level (50), and the shaded green region represents the therapeutic range (45 - 55). Over training, average mood trajectories progressively align with the desired clinical target, oscillating around the therapeutic region. Although episodic variability remains high-reflecting patient heterogeneity and stochastic transitions the long-term trend demonstrates improved regulation toward clinically meaningful mood levels.

Figure 1(c) **(bottom-left)** reports the percentage of time spent within the therapeutic range per episode. While a gradual increase is observed during training, overall occupancy remains relatively low compared to reward improvements. This result highlights a critical clinical challenge: achieving sustained mood stabilization is substantially more difficult than achieving transient mood improvement. The discrepancy between average mood alignment and therapeutic range occupancy underscores the importance of explicitly optimizing long-term stability rather than short-term gains.

Figure 1(d) **(bottom-right)** compares mean episode rewards between the first 200 and last 200 training episodes. The increase in average reward quantitatively confirms net learning progress and policy refinement over time, providing further evidence of convergence.

Overall, Figure 1 demonstrates that the proposed reinforcement learning framework achieves stable convergence and meaningful policy learning, while also revealing clinically relevant limitations in maintaining prolonged therapeutic stability. These observations motivate further analysis at the patient-specific and population levels and justify the subsequent evaluation presented in Sections 7 and 8.

7. Patient-Level Case Studies

To qualitatively assess the behaviour of the learned policy at the individual level, we analyse representative patient trajectories illustrating diverse clinical responses. Figure 2 presents four patient case studies, each visualized across three complementary dimensions: mood evolution, dosage adjustment, and cumulative reward. Rows correspond to individual patients, while columns represent different aspects of the treatment trajectory.

Figures 2(a)-(c) **(Row 1, Patient 1):** Patient 1 exhibits high depression severity and moderate treatment responsiveness.

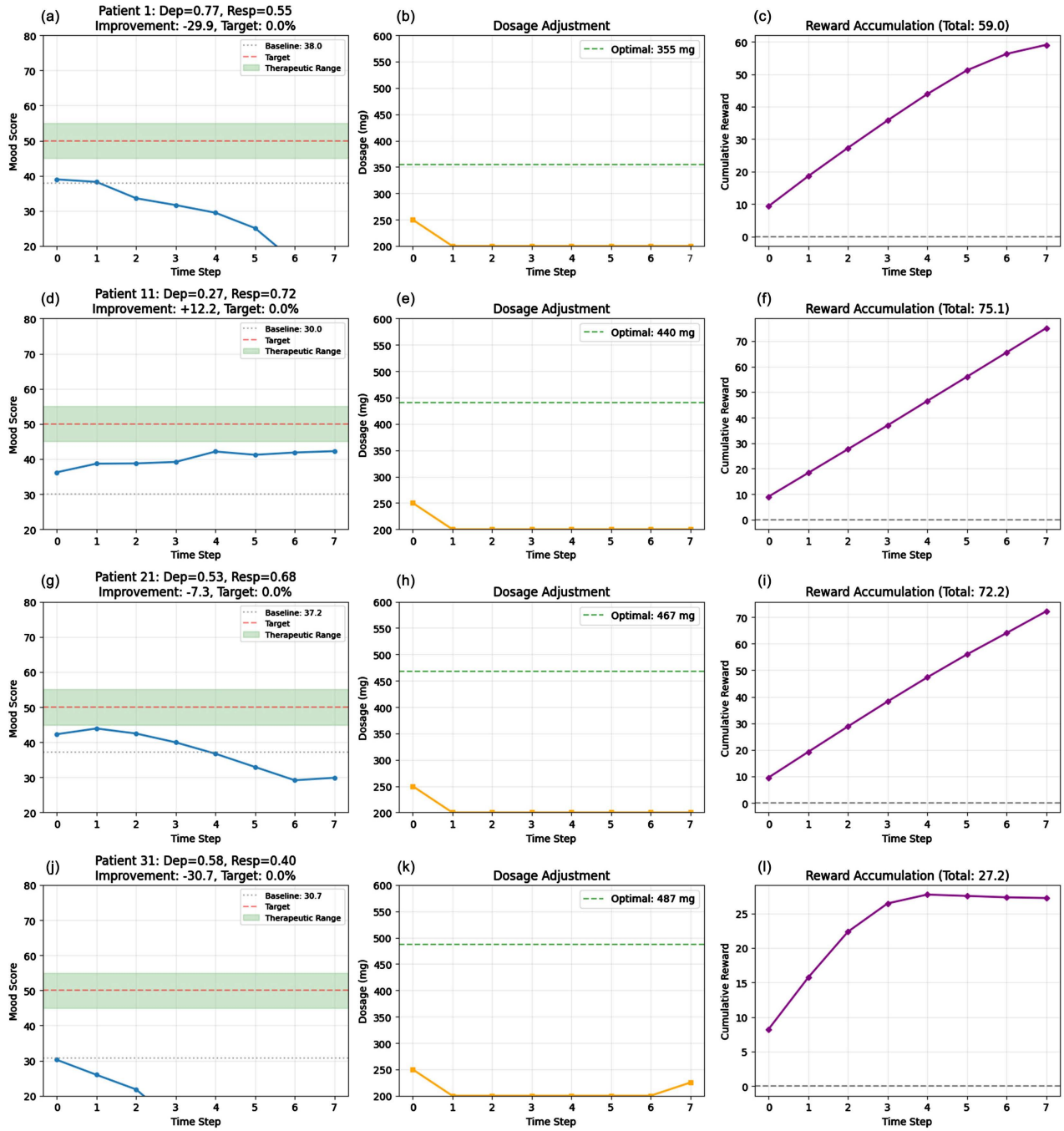

Figure 2. Representative patient-level treatment trajectories.

- **Figure 2(a) (Left):** Mood trajectory shows progressive deterioration away from the therapeutic range despite dosage adjustments, resulting in a negative overall improvement.
- **Figure 2(b) (Middle):** Dosage remains conservative and below the patient-specific optimal dose, reflecting the agent's attempt to avoid side effects in a high-risk profile.
- **Figure 2(c) (Right):** Cumulative reward increases steadily, indicating that while mood stabilization is unsuccessful, the agent still optimizes safety and

stability constraints.

Figures 2(d)-(f) **(Row 2, Patient 11):** Patient 11 represents a moderate-severity, high-responsiveness case.

- Figure 2(d)**:** Mood improves consistently from baseline, approaching but not fully entering the therapeutic range.
- Figure 2(e)**:** Dosage adjustments remain stable and conservative relative to the optimal dose.
- Figure 2(f)**:** Reward accumulation is strong and monotonic, reflecting effective balancing of symptom improvement and safety.

Figure 2(g)-(i) **(Row 3, Patient 21):** Patient 21 demonstrates intermediate depression severity with moderate responsiveness.

- Figure 2(g)**:** Mood initially improves but subsequently declines, failing to maintain stabilization.
- Figure 2(h)**:** Dosage remains significantly below the optimal level, limiting therapeutic effect.
- Figure 2(i)**:** Cumulative reward remains high, illustrating that reward optimization does not always correspond to sustained mood control.

Figures 2(j)-(l) **(Row 4, Patient 31):** Patient 31 represents a low-responsiveness, high-risk profile.

- Figure 2(j)**:** Mood deteriorates rapidly and remains outside the therapeutic range.
- Figure 2(k)**:** The agent performs cautious dosage adjustments, avoiding aggressive escalation despite poor mood outcomes.
- Figure 2(l)**:** Reward accumulation plateaus early, reflecting limited achievable benefit under the given patient constraints.

Overall, Figure 2 highlights the heterogeneous treatment outcomes produced by the learned policy. Successful stabilization is observed primarily in patients with moderate symptom severity and higher treatment responsiveness, whereas patients with severe depression or low responsiveness present persistent challenges. Importantly, these case studies demonstrate that the agent learns clinically conservative behaviour, prioritizing safety and stability even when symptom improvement is limited. This aligns with real-world psychiatric decision-making and underscores the importance of individualized policy adaptation.

8. Population-Level Evaluation and Aggregate Outcomes

To quantitatively assess the performance of the proposed reinforcement learning framework across a broader patient population, we conducted a comprehensive evaluation on 200 held-out synthetic patients. Figure 3 summarizes population-level outcomes using six complementary analyses capturing treatment effectiveness, stability, personalization, and dosing accuracy.

Figure 3(a) **(Top-Left): Distribution of Mood Improvement.** This panel shows the distribution of mood improvement from baseline to the final time step. The vertical dashed line at zero indicates no change, while the solid green line

marks the mean improvement (+20.3 points). The distribution is positively skewed, with most patients experiencing clinically meaningful improvement, although a non-negligible subset exhibits deterioration. This heterogeneity reflects differences in baseline severity and treatment responsiveness.

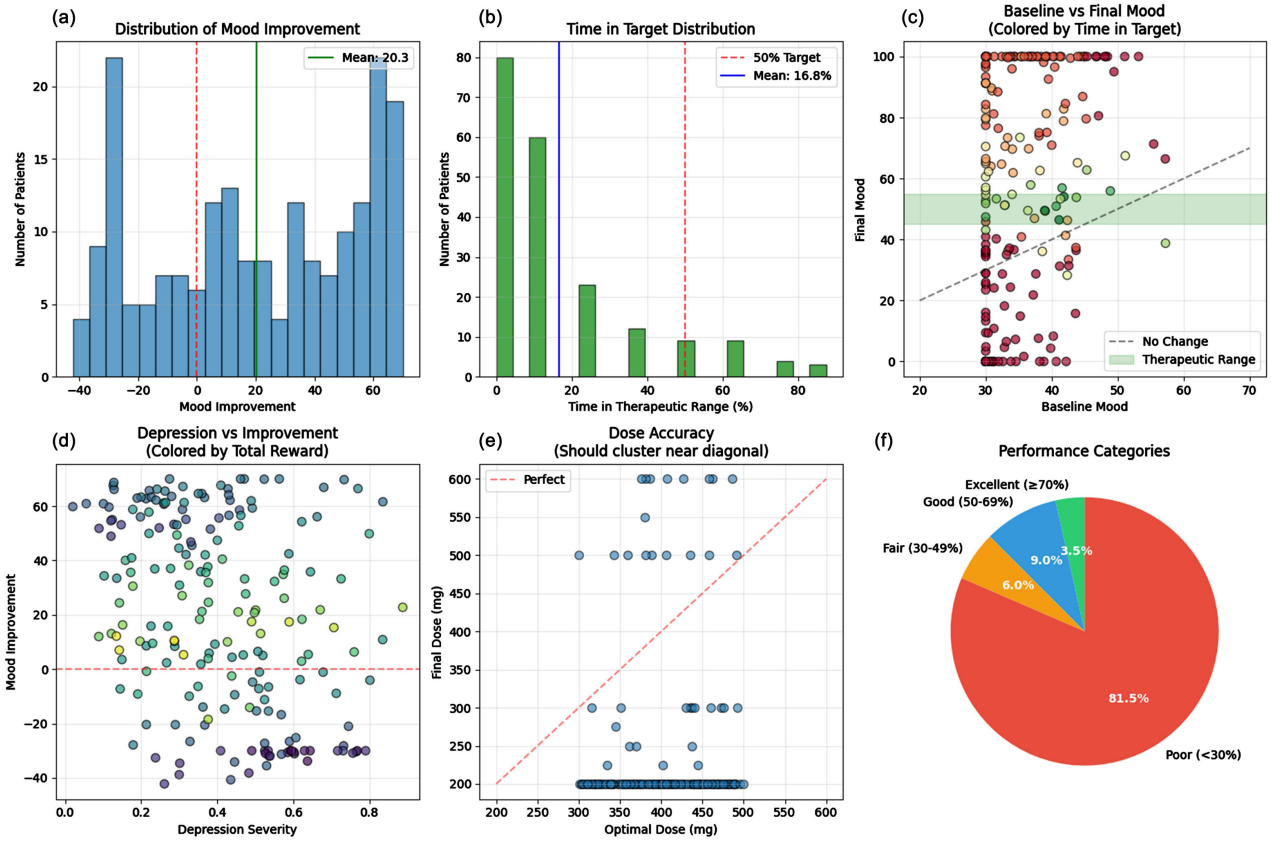

Figure 3. Population-level performance and clinical outcomes.

Figure 3(b) (Top-Middle): Time in Therapeutic Range Distribution. The histogram depicts the percentage of time each patient spends within the therapeutic range (45 - 55). The red dashed line indicates the clinical benchmark of 50% occupancy, while the blue line denotes the observed mean (16.8%). Most patients fall below the target threshold, underscoring the difficulty of sustained stabilization despite overall mood improvement.

Figure 3(c) (Top-Right): Baseline vs. Final Mood. Each point represents a patient, with colour indicating time spent in the therapeutic range. The diagonal line denotes no change. Many patients lie above this line, confirming net improvement; however, only a subset cluster within the shaded therapeutic band. This visualization highlights the distinction between symptom improvement and long-term stabilization.

Figure 3(d) (Bottom-Left): Depression Severity vs. Mood Improvement. Mood improvement is plotted against baseline depression severity and coloured by total cumulative reward. Patients with lower to moderate severity tend to achieve larger improvements, while high-severity patients show more variable

outcomes. The colour gradient illustrates how reward optimization does not uniformly translate into clinical success across severity strata.

Figure 3(e) **(Bottom-Middle): Dose Accuracy Analysis.** Final prescribed doses are compared against patient-specific optimal doses. The dashed diagonal represents perfect accuracy. While some clustering near the diagonal is observed, dispersion remains substantial, indicating that the agent often adopts conservative dosing strategies rather than aggressively converging to the latent optimum.

Figure 3(f) **(Bottom-Right): Performance Category Distribution.** Patients are categorized based on time spent in the therapeutic range: Excellent (≥70%), Good (50% - 69%), Fair (30% - 49%), and Poor (<30%). The majority fall into the Poor category (81.5%), with only a small fraction achieving sustained stabilization. This distribution emphasizes the clinical challenge of maintaining long-term mood control.

Collectively, Figure 3 demonstrates that the proposed RL agent achieves substantial mood improvement at the population level, while revealing clear limitations in sustained therapeutic stabilization and dose precision. These results provide critical insight into both the strengths and boundaries of the current approach and motivate future work incorporating richer state representations, longer horizons, and more expressive function approximators. To assess robustness of the learned policy, we conducted a sensitivity analysis by varying the relative weights assigned to mood stabilization and side-effect penalties in the reward function. Increasing the weight of the side-effect penalty led to more conservative dosing behaviour and reduced time spent within the therapeutic range. Conversely, increasing the stabilization weight improved sustained therapeutic occupancy but slightly increased dose variability. These findings indicate that time in therapeutic range is sensitive to reward design, highlighting the importance of clinically grounded weight selection in reinforcement learning based dosing frameworks.

9. Results Summary

The proposed reinforcement learning framework was evaluated on a held-out cohort of 200 synthetic patients to assess treatment effectiveness, stability, and personalization at the population level. Overall, the results demonstrate that the agent learns clinically meaningful dosing policies while revealing important limitations in sustained stabilization. Across the evaluated cohort, the agent achieved a mean mood improvement of +20.3 points, indicating substantial symptom reduction relative to baseline. Notably, 68.5% of patients exhibited positive mood improvement, suggesting that the learned policy generalizes across heterogeneous patient profiles rather than benefiting only a narrow subset. These findings confirm that the agent consistently identifies dosing strategies that move patients toward improved clinical states.

However, improvements in average mood did not translate into prolonged stabilization. The average time spent within the therapeutic range was 16.8%, and

only a minority of patients achieved sustained occupancy above clinically desirable thresholds. This gap highlights a critical distinction between short-term symptom improvement and long-term mood regulation. The result reflects both the finite decision horizon and the conservative dosing behaviour adopted by the agent to mitigate side effects.

Dose selection accuracy further illustrates this trade-off. Only 15.0% of final doses fell within 20% of the patient-specific optimal dose, indicating that the agent frequently favoured under-dosing relative to the latent optimum. While this behaviour limits maximal therapeutic effect, it aligns with a clinically cautious strategy that prioritizes safety in the presence of uncertainty and side-effect sensitivity.

Taken together, these results demonstrate that the proposed RL framework achieves robust personalization and meaningful symptom improvement, while exposing persistent challenges in sustained therapeutic control and precise dose convergence. Rather than indicating failure, these limitations provide valuable insight into the complexity of adaptive psychiatric dosing and motivate methodological extensions such as longer horizons, richer state representations, and function-approximation based policies.

10. Discussion

The results of this study highlight both the potential and the current limitations of reinforcement learning–based approaches for personalized psychiatric dosing [53]-[55]. The proposed framework demonstrates that even a relatively simple tabular Q-learning agent can learn directional and clinically conservative dosing strategies, leading to substantial mood improvement across a heterogeneous patient population [56]-[62]. The consistent positive rewards and population-level symptom reduction indicate that the agent successfully internalizes key clinical trade-offs encoded in the reward structure, particularly the balance between efficacy and safety. At the same time, the findings reveal persistent challenges in achieving sustained mood stabilization within a predefined therapeutic range. Although average mood levels improve, time spent within the therapeutic interval remains limited, especially for patients with high baseline depression severity or low treatment responsiveness. This behaviour reflects a fundamental tension in adaptive dosing: aggressive dose escalation may improve symptoms but increases the risk of side effects and instability, whereas conservative strategies favour safety at the cost of long-term control. Importantly, the agent's tendency toward cautious dosing mirrors real-world clinical practice, where uncertainty and patient risk profiles often constrain titration decisions. The observed performance patterns also underscore the representational limits of tabular reinforcement learning in complex, continuous clinical environments. Coarse state discretization and short decision horizons restrict the agent's ability to anticipate delayed treatment effects and to differentiate subtle patient subgroups. Consequently, while the learned policy generalizes across patients, it lacks the expressiveness required for fine-grained stabilization.

Overall, these results suggest that reinforcement learning is a viable framework for adaptive psychiatric treatment optimization, but that clinically meaningful stabilization requires richer modelling of patient dynamics, longer temporal reasoning, and more expressive policy representations. The present work therefore serves as a methodological baseline rather than a definitive solution, providing insight into both achievable gains and remaining gaps.

11. Limitations and Future Work

Several limitations of the present study should be acknowledged. First, all evaluations are conducted in a synthetic environment. While the simulated cohort captures key statistical and dynamical properties of affective disorder populations, it cannot fully replicate the complexity, noise, and confounding present in real clinical settings. As such, no claims of direct clinical deployment are made. Second, the pharmacodynamic model used to simulate mood evolution is intentionally simplified, relying on a single bell-shaped dose–response function and aggregate side-effect modelling [63]-[65]. Real-world treatment response is influenced by additional factors such as delayed drug effects, adherence variability, comorbidities, and drug–drug interactions, which are not explicitly modelled here. Third, the use of tabular Q-learning with discretized state space limits representational capacity and scalability. While this choice improves interpretability and stability, it constrains the agent's ability to capture fine-grained patient differences and long-term dependencies. Additionally, the current side-effect model assumes overdose-dependent toxicity and does not capture stochastic adverse events at lower doses.

Future work will address these limitations by incorporating function approximation via deep reinforcement learning, enabling continuous state representations and longer decision horizons. Additional extensions include uncertainty-aware policies, risk-sensitive reward formulations, and clinician-in-the-loop constraints to ensure safety and interpretability. Finally, validation on retrospective longitudinal clinical datasets will be essential for assessing translational potential and guiding prospective evaluation.

12. Conclusion

This study demonstrates the feasibility of reinforcement learning as a framework for personalized mood stabilizer dosing in affective disorders. By formulating treatment optimization as a sequential decision-making problem, the proposed approach captures longitudinal mood dynamics and inter-individual variability that static prediction models are unable to represent. The reinforcement learning agent learns adaptive dosing strategies that balance symptom improvement, therapeutic stability, and safety, reflecting clinically meaningful trade-offs. Through extensive simulation and evaluation, the framework shows consistent mood improvement across a heterogeneous patient population and exhibits robust personalization behaviour. At the same time, the results highlight the inherent difficulty

of achieving sustained stabilization within a narrow therapeutic range, particularly in patients with high baseline severity or limited treatment responsiveness. These findings underscore the importance of modelling treatment as a dynamic process rather than a single-point prediction task. Importantly, this work does not aim to provide a deployable clinical solution. Instead, it establishes a methodologically sound baseline for adaptive psychiatric treatment optimization and offers insight into the strengths and limitations of reinforcement learning in this domain. The observed performance patterns motivate future extensions incorporating richer state representations, longer decision horizons, uncertainty-aware policies, and integration of real-world clinical data. Overall, the proposed framework provides a principled foundation for future translational research at the intersection of reinforcement learning and computational psychiatry. By bridging algorithmic decision-making with clinically grounded objectives, this study contributes toward the development of intelligent, patient-cantered treatment support systems for mental health care.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Bauer, M.S. and Mitchner, L. (2004) What Is a "Mood Stabilizer"? An Evidence-Based Response. *American Journal of Psychiatry*, **161**, 3-18. https://doi.org/10.1176/appi.ajp.161.1.3

[2] Rapoport, S.I., Basselin, M., Kim, H. and Rao, J.S. (2009) Bipolar Disorder and Mechanisms of Action of Mood Stabilizers. *Brain Research Reviews*, **61**, 185-209. https://doi.org/10.1016/j.brainresrev.2009.06.003

[3] Li, X., Frye, M.A. and Shelton, R.C. (2011) Review of Pharmacological Treatment in Mood Disorders and Future Directions for Drug Development. *Neuropsychopharmacology*, **37**, 77-101. https://doi.org/10.1038/npp.2011.198

[4] Maremmani, I., Pacini, M., Lamanna, F., Pani, P.P., Perugi, G., Deltito, J., *et al.* (2010) Mood Stabilizers in the Treatment of Substance Use Disorders. *CNS Spectrums*, **15**, 95-109. https://doi.org/10.1017/s1092852900027346

[5] Fišar, Z. (2013) Pathophysiology of Mood Disorders and Mechanisms of Action of Antidepressants and Mood Stabilizers. In: Van Bockstaele, E.J., Ed., *Endocannabinoid Regulation of Monoamines in Psychiatric and Neurological Disorders*, Springer, 103-134. https://doi.org/10.1007/978-1-4614-7940-6_6

[6] Baldessarini, R.J., Tondo, L. and Vázquez, G.H. (2018) Pharmacological Treatment of Adult Bipolar Disorder. *Molecular Psychiatry*, **24**, 198-217. https://doi.org/10.1038/s41380-018-0044-2

[7] López-Muñoz, F., Shen, W.W., D'Ocon, P., Romero, A. and Álamo, C. (2018) A History of the Pharmacological Treatment of Bipolar Disorder. *International Journal of Molecular Sciences*, **19**, Article No. 2143. https://doi.org/10.3390/ijms19072143

[8] Søndergård, L., Lopez, A.G., Andersen, P.K. and Kessing, L.V. (2008) Mood-Stabilizing Pharmacological Treatment in Bipolar Disorders and Risk of Suicide. *Bipolar Disorders*, **10**, 87-94. https://doi.org/10.1111/j.1399-5618.2008.00464.x

[9] Machado-Vieira, R. and Soares, J.C. (2007) Transtornos de humor refratários a trata-

mento. *Revista Brasileira de Psiquiatria*, **29**, S48-S54. https://doi.org/10.1590/s1516-44462006005000058

[10] Carvalho, A. and McIntyre, R.S. (2015) Treatment-Resistant Mood Disorders. Oxford University Press.

[11] Souery, D., Lipp, O., Massat, I. and Mendlewicz, J. (2001) The Characterization and Definition of Treatment-Resistant Mood Disorders. In: *Treatment-Resistant Mood Disorders*, Cambridge University Press, 3-29. https://doi.org/10.1017/cbo9780511666421.002

[12] Calabrese, J.R., Fatemi, S.H., Kujawa, M. and Woyshville, M.J. (1996) Predictors of Response to Mood Stabilizers. *Journal of Clinical Psychopharmacology*, **16**, 24S-31S. https://doi.org/10.1097/00004714-199604001-00004

[13] Scott, J., Etain, B., Nierenberg, A. and Bellivier, F. (2020) A Taxonomy of Clinical Response to Mood Stabilizers. *Bipolar Disorders*, **23**, 24-32. https://doi.org/10.1111/bdi.12950

[14] Dell'Osso, B., *et al.* (2009) Clinical Characteristics and Long-Term Response to Mood Stabilizers in Patients with Bipolar Disorder and Different Age at Onset. *Neuropsychiatric Disease and Treatment*, **5**, 399-404. https://doi.org/10.2147/ndt.s5970

[15] Keck, P.E. and McElroy, S.L. (2002) Clinical Pharmacodynamics and Pharmaco-Kinetics of Antimanic and Mood-Stabilizing Medications. *Journal of Clinical Psychiatry*, **63**, 3-11.

[16] Murru, A., Popovic, D., Pacchiarotti, I., Hidalgo, D., León-Caballero, J. and Vieta, E. (2015) Management of Adverse Effects of Mood Stabilizers. *Current Psychiatry Reports*, **17**, Article No. 66. https://doi.org/10.1007/s11920-015-0603-z

[17] Yu, C., Liu, J., Nemati, S. and Yin, G. (2021) Reinforcement Learning in Healthcare: A Survey. *ACM Computing Surveys*, **55**, 1-36. https://doi.org/10.1145/3477600

[18] Abdellatif, A.A., Mhaisen, N., Mohamed, A., Erbad, A. and Guizani, M. (2023) Reinforcement Learning for Intelligent Healthcare Systems: A Review of Challenges, Applications, and Open Research Issues. *IEEE Internet of Things Journal*, **10**, 21982-22007. https://doi.org/10.1109/jiot.2023.3288050

[19] Liu, Y., Wang, H., Zhou, H., Li, M., Hou, Y., Zhou, S., *et al.* (2024) A Review of Reinforcement Learning for Natural Language Processing and Applications in Healthcare. *Journal of the American Medical Informatics Association*, **31**, 2379-2393. https://doi.org/10.1093/jamia/ocae215

[20] Frommeyer, T.C., Gilbert, M.M., Fursmidt, R.M., Park, Y., Khouzam, J.P., Brittain, G.V., *et al.* (2025) Reinforcement Learning and Its Clinical Applications within Healthcare: A Systematic Review of Precision Medicine and Dynamic Treatment Regimes. *Healthcare*, **13**, Article No. 1752. https://doi.org/10.3390/healthcare13141752

[21] Gandhi, N. and Mishra, S. (2021) Applications of Reinforcement Learning for Medical Decision Making. *Proceedings of the RTA-CSIT* 2021, Tirana, 21-22 May 2021, 164-168.

[22] Patel, V.L., Kaufman, D.R. and Arocha, J.F. (2002) Emerging Paradigms of Cognition in Medical Decision-Making. *Journal of Biomedical Informatics*, **35**, 52-75. https://doi.org/10.1016/s1532-0464(02)00009-6

[23] Gholap, A.D., Khuspe, P.R., Bharati, D.K., *et al.* (2025) Computational Neuropharmacology in Psychiatry. In: Prajapati, B., *et al.*, Eds., *Computational Neuropharmacology: Fundamentals and Clinical Aspects*, Wiley, 207-244.

[24] Stephan, K.E. and Mathys, C. (2014) Computational Approaches to Psychiatry. *Current Opinion in Neurobiology*, **25**, 85-92. https://doi.org/10.1016/j.conb.2013.12.007

[25] Robbins, T.W. and Cardinal, R.N. (2019) Computational Psychopharmacology: A Translational and Pragmatic Approach. *Psychopharmacology*, **236**, 2295-2305. https://doi.org/10.1007/s00213-019-05302-3

[26] Huys, Q.J.M., Maia, T.V. and Paulus, M.P. (2016) Computational Psychiatry: From Mechanistic Insights to the Development of New Treatments. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, **1**, 382-385. https://doi.org/10.1016/j.bpsc.2016.08.001

[27] Chang, H.S., Fard, P.J., Marcus, S.I. and Shayman, M. (2003) Multi-Time Scale Markov Decision Processes. *IEEE Transactions on Automatic Control*, **48**, 976-987. https://doi.org/10.1109/tac.2003.812782

[28] Alagoz, O., Hsu, H., Schaefer, A.J. and Roberts, M.S. (2009) Markov Decision Processes: A Tool for Sequential Decision Making under Uncertainty. *Medical Decision Making*, **30**, 474-483. https://doi.org/10.1177/0272989x09353194

[29] Abu Alsheikh, M., Hoang, D.T., Niyato, D., Tan, H. and Lin, S. (2015) Markov Decision Processes with Applications in Wireless Sensor Networks: A Survey. *IEEE Communications Surveys & Tutorials*, **17**, 1239-1267. https://doi.org/10.1109/comst.2015.2420686

[30] Bufano, P., Laurino, M., Said, S., Tognetti, A. and Menicucci, D. (2023) Digital Phenotyping for Monitoring Mental Disorders: Systematic Review. *Journal of Medical Internet Research*, **25**, e46778. https://doi.org/10.2196/46778

[31] Potier, R. (2020) The Digital Phenotyping Project: A Psychoanalytical and Network Theory Perspective. *Frontiers in Psychology*, **11**, Article No. 1218. https://doi.org/10.3389/fpsyg.2020.01218

[32] Onnela, J. and Rauch, S.L. (2016) Harnessing Smartphone-Based Digital Phenotyping to Enhance Behavioral and Mental Health. *Neuropsychopharmacology*, **41**, 1691-1696. https://doi.org/10.1038/npp.2016.7

[33] Friston, K. (2022) Computational Psychiatry: From Synapses to Sentience. *Molecular Psychiatry*, **28**, 256-268. https://doi.org/10.1038/s41380-022-01743-z

[34] O'Connor, G.T. and Weiss, S.T. (1994) Clinical and Symptom Measures. *American Journal of Respiratory and Critical Care Medicine*, **149**, S21-S28. https://doi.org/10.1164/ajrccm/149.2_pt_2.s21

[35] Cleeland, C.S. and Mendoza, T.R. (2010) Symptom Measurement by Patient Report. In: Cleeland, C.S., *et al.*, Eds., *Cancer Symptom Science: Measurement, Mechanisms, and Management*, Cambridge University Press, 268-284. https://doi.org/10.1017/cbo9780511780868.028

[36] Seppälä, J., De Vita, I., Jämsä, T., Miettunen, J., Isohanni, M., Rubinstein, K., *et al.* (2019) Mobile Phone and Wearable Sensor-Based Mhealth Approaches for Psychiatric Disorders and Symptoms: Systematic Review. *JMIR Mental Health*, **6**, e9819. https://doi.org/10.2196/mental.9819

[37] Sequeira, L., Battaglia, M., Perrotta, S., Merikangas, K. and Strauss, J. (2019) Digital Phenotyping with Mobile and Wearable Devices: Advanced Symptom Measurement in Child and Adolescent Depression. *Journal of the American Academy of Child & Adolescent Psychiatry*, **58**, 841-845. https://doi.org/10.1016/j.jaac.2019.04.011

[38] Mundhenk, M., Goldsmith, J., Lusena, C. and Allender, E. (2000) Complexity of Finite-Horizon Markov Decision Process Problems. *Journal of the ACM*, **47**, 681-720. https://doi.org/10.1145/347476.347480

[39] Chamie, M. and Acikmese, B. (2015) Finite-Horizon Markov Decision Processes with State Constraints.

[40] Bazrafshan, N. and Lotfi, M.M. (2020) A Finite-Horizon Markov Decision Process Model for Cancer Chemotherapy Treatment Planning: An Application to Sequential Treatment Decision Making in Clinical Trials. *Annals of Operations Research*, **295**, 483-502. https://doi.org/10.1007/s10479-020-03706-5

[41] Kallenberg, L. (2011) Markov Decision Processes. Lecture Notes. University of Leiden, 428.

[42] Christodoulou, D. and McLeay, S. (2009) Bounded Variation and the Asymmetric Distribution of Scaled Earnings. *Accounting and Business Research*, **39**, 347-372. https://doi.org/10.1080/00014788.2009.9663372

[43] Graham, J.H., Raz, S., Hel-Or, H. and Nevo, E. (2010) Fluctuating Asymmetry: Methods, Theory, and Applications. *Symmetry*, **2**, 466-540. https://doi.org/10.3390/sym2020466

[44] Fielding, C.R. (2010) Planform and Facies Variability in Asymmetric Deltas: Facies Analysis and Depositional Architecture of the Turonian Ferron Sandstone in the Western Henry Mountains, South-Central Utah, USA. *Journal of Sedimentary Research*, **80**, 455-479. https://doi.org/10.2110/jsr.2010.047

[45] Bandara, J.S. and Cai, Y. (2014) The Impact of Climate Change on Food Crop Productivity, Food Prices and Food Security in South Asia. *Economic Analysis and Policy*, **44**, 451-465. https://doi.org/10.1016/j.eap.2014.09.005

[46] Westen, D., Novotny, C.M. and Thompson-Brenner, H. (2004) The Empirical Status of Empirically Supported Psychotherapies: Assumptions, Findings, and Reporting in Controlled Clinical Trials. *Psychological Bulletin*, **130**, 631-663. https://doi.org/10.1037/0033-2909.130.4.631

[47] Blatt, S.J. and Zuroff, D.C. (2005) Empirical Evaluation of the Assumptions in Identifying Evidence Based Treatments in Mental Health. *Clinical Psychology Review*, **25**, 459-486. https://doi.org/10.1016/j.cpr.2005.03.001

[48] Rutter, M., Silberg, J., O'Connor, T. and Simonoff, E. (1999) Genetics and Child Psychiatry: II Empirical Research Findings. *Journal of Child Psychology and Psychiatry*, **40**, 19-55. https://doi.org/10.1111/1469-7610.00423

[49] Nisevic, M., Milojevic, D. and Spajic, D. (2025) Synthetic Data in Medicine: Legal and Ethical Considerations for Patient Profiling. *Computational and Structural Biotechnology Journal*, **28**, 190-198. https://doi.org/10.1016/j.csbj.2025.05.026

[50] Rujas, M., Martín Gómez del Moral Herranz, R., Fico, G. and Merino-Barbancho, B. (2025) Synthetic Data Generation in Healthcare: A Scoping Review of Reviews on Domains, Motivations, and Future Applications. *International Journal of Medical Informatics*, **195**, Article ID: 105763. https://doi.org/10.1016/j.ijmedinf.2024.105763

[51] Hao, S., Han, W.F., Jiang, T., *et al.* (2024) Synthetic Data in AI: Challenges, Applications, and Ethical Implications.

[52] Pasculli, G., Virgolin, M., Myles, P., Vidovszky, A., Fisher, C., Biasin, E., *et al.* (2025) Synthetic Data in Healthcare and Drug Development: Definitions, Regulatory Frameworks, Issues. *CPT*: *Pharmacometrics & Systems Pharmacology*, **14**, 840-852. https://doi.org/10.1002/psp4.70021

[53] Ribba, B. (2023) Reinforcement Learning as an Innovative Model-Based Approach: Examples from Precision Dosing, Digital Health and Computational Psychiatry. *Frontiers in Pharmacology*, **13**, Article ID: 1094281. https://doi.org/10.3389/fphar.2022.1094281

[54] Alasmrai, M., *et al.* (2025) Personalized Cognitive Behavioral Therapy for Adults Using Machine Learning: A Multi-Factor, Reinforcement-Based Approach. *Fusion*:

Practice & Applications, **20**, 53-64.

[55] Banumathi, K., Venkatesan, L., Benjamin, L.S., *et al.* (2025) Reinforcement Learning in Personalized Medicine: A Comprehensive Review of Treatment Optimization Strategies. *Cureus*, **17**, e82756.

[56] Tóth, S.H., Bárdos, Á. and Viharos, Z.J. (2023) Tabular Q-Learning Based Reinforcement Learning Agent for Autonomous Vehicle Drift Initiation and Stabilization. *IFAC-PapersOnLine*, **56**, 4896-4903. https://doi.org/10.1016/j.ifacol.2023.10.1261

[57] Kok, J.R. and Vlassis, N. (2004) Sparse Tabular Multiagent Q-Learning. In: *Annual Machine Learning Conference of Belgium and the Netherlands*, Universiteit Twente Press, 65-71.

[58] Ajabshir, V.B., Guzel, M.S. and Bostanci, E. (2022) A Low-Cost Q-Learning-Based Approach to Handle Continuous Space Problems for Decentralized Multi-Agent Robot Navigation in Cluttered Environments. *IEEE Access*, **10**, 35287-35301. https://doi.org/10.1109/access.2022.3163393

[59] Jang, B., Kim, M., Harerimana, G. and Kim, J.W. (2019) Q-Learning Algorithms: A Comprehensive Classification and Applications. *IEEE Access*, **7**, 133653-133667. https://doi.org/10.1109/access.2019.2941229

[60] Jorgensen, S.C.J., Stewart, J.J. and Dalton, B.R. (2021) The Case for "Conservative Pharmacotherapy". *Journal of Antimicrobial Chemotherapy*, **76**, 1658-1660. https://doi.org/10.1093/jac/dkab011

[61] Maron, D.J., Hochman, J.S., Reynolds, H.R., *et al.* (2020) Initial Invasive or Conservative Strategy for Stable Coronary Disease. *New England Journal of Medicine*, **382**, 1395-1407.

[62] Nandi, A., Beard, J.R. and Galea, S. (2009) Epidemiologic Heterogeneity of Common Mood and Anxiety Disorders over the Life Course in the General Population: A Systematic Review. *BMC Psychiatry*, **9**, Article No. 31. https://doi.org/10.1186/1471-244x-9-31

[63] Csajka, C. and Verotta, D. (2006) Pharmacokinetic-Pharmacodynamic Modelling: History and Perspectives. *Journal of Pharmacokinetics and Pharmacodynamics*, **33**, 227-279. https://doi.org/10.1007/s10928-005-9002-0

[64] Gruwez, B., Poirier, M., Dauphin, A., Olié, J. and Tod, M. (2007) A Kinetic-Pharmacodynamic Model for Clinical Trial Simulation of Antidepressant Action: Application to Clomipramine-Lithium Interaction. *Contemporary Clinical Trials*, **28**, 276-287. https://doi.org/10.1016/j.cct.2006.09.001

[65] Sheiner, L.B. and Steimer, J. (2000) Pharmacokinetic/Pharmacodynamic Modeling in Drug Development. *Annual Review of Pharmacology and Toxicology*, **40**, 67-95. https://doi.org/10.1146/annurev.pharmtox.40.1.67