Open Access Library Journal
2026, Volume 13, e15677
ISSN Online: 2333-9721
ISSN Print: 2333-9705

Reinforcement Learning for Personalised Antidepressant Sequencing in Treatment-Resistant Bipolar Depression: A Simulation-Based Policy Optimisation Study

Rocco de Filippis[1,2*], Abdullah Al Foysal[3]

[1]Department of Psychiatry, Istituto di Psicopatologia, Rome, Italy
[2]EPSM74, La Roche-sur-Foron, France
[3]Department of Informatics, Bioengineering, Robotics and Systems Engineering, University of Genoa, Genoa, Italy
Email: *roccodefilippis@istitutodipsicopatologia.it, niloyhasanfoysal440@gmail.com

How to cite this paper: de Filippis, R. and Al Foysal, A. (2026) Reinforcement Learning for Personalised Antidepressant Sequencing in Treatment-Resistant Bipolar Depression: A Simulation-Based Policy Optimisation Study. *Open Access Library Journal*, **13**: e15677.
https://doi.org/10.4236/oalib.1115677

Received: June 22, 2026
Accepted: August 24, 2026
Published: August 27, 2026

Abstract

Treatment-resistant bipolar depression (TRD-BD), defined in this simulation as failure to achieve remission after at least two adequate pharmacological trials for the current bipolar depressive episode, including at least one antidepressant trial delivered with mood-stabilising treatment, is managed through a sequence of uncertain decisions: continue, switch antidepressant class, or augment with a mood stabiliser or atypical antipsychotic. Current practice relies on consensus bipolar-disorder guidelines and sequential-care principles; the STAR*D study is referenced only as a structural example of multistep treatment sequencing in unipolar major depression, not as a validated bipolar-depression protocol. The sequencing problem is therefore a natural candidate for reinforcement learning (RL), but no RL framework has yet been developed specifically for personalised antidepressant sequencing in bipolar depression. We formulated antidepressant sequencing in TRD-BD as a finite-horizon Markov Decision Process and developed an offline RL framework for policy optimisation. A clinically grounded simulator encoded a 15-dimensional policy state comprising MADRS, YMRS, prior treatment failures, weeks on current treatment, side-effect burden, current drug class, a three-dimensional responder-profile covariate, age, and BD subtype, together with an 8-action treatment space (continue; switch to SSRI, SNRI, bupropion, or MAOI; augment with lithium, quetiapine, or lamotrigine). Although the responder profile represents an underlying biological trait, it was supplied to the policy as an oracle covariate in this proof-of-concept simulation; in real practice, it would usually be unobserved, making the clinical problem partially observable. The reward in-

tegrated MADRS improvement, remission, side-effect burden, switching cost, and manic-switch risk. A Deep Q-Network (DQN) was trained on 6172 transitions from 800 simulated trajectories generated by a stochastic behavioural clinician policy. Trajectories terminated at remission, manic switch, intolerable adverse effects, or the eight-stage horizon, yielding a mean logged trajectory length of 7.72 stages. Exact behavioural-policy action probabilities were recorded by the simulator and used for importance-weighted off-policy estimators. DQN was compared with Fitted Q-Iteration, the behavioural clinician policy, a fixed sequential protocol, and a random policy. The proposed DQN policy achieved an estimated discounted return of 41.6 (95% CI: 39.0 - 44.0), substantially exceeding Fitted Q-Iteration (24.6), the clinician policy (12.8), and the fixed protocol (12.1). The DQN policy achieved a remission rate of 61.8% (95% CI: 57.0% - 66.0%), nearly quadrupling the clinician policy's 16.5% and reaching remission in fewer treatment stages (mean 6.5 vs 7.7). Mean MADRS reduction was 24.3 points versus 15.6 for the clinician policy. Bayesian posterior analysis assigned the DQN policy a 100.0% probability of being optimal. Policy analysis revealed that the learned strategy strongly favoured lithium augmentation selected in 88% of states over the clinician's switch-heavy, lower-augmentation pattern. Within this synthetic environment, reinforcement learning derived a personalised sequencing policy that outperformed the simulated behavioural and fixed-protocol comparators. The learned preference for early lithium augmentation reflects the simulator parameters and should be interpreted as a proof-of-concept result rather than a clinical treatment recommendation. Validation with real TRD-BD trajectories and conservative offline-RL methods is required before any clinical inference or deployment.

Subject Areas

Psychiatry & Psychology

Keywords

Reinforcement Learning, Deep Q-Network, Treatment-Resistant Bipolar Depression, Antidepressant Sequencing, Markov Decision Process, Offline RL, Off-Policy Evaluation, Personalised Medicine, Policy Optimisation

1. Introduction

The pharmacological management of treatment-resistant bipolar depression is, in its essential structure, a sequential decision problem. Throughout this study, TRD-BD is defined consistently as failure to achieve remission after at least two adequate pharmacological trials for the current bipolar depressive episode, including at least one antidepressant trial administered with mood-stabilising treatment. At each review, the clinician must choose whether to continue, switch antidepressant class, or augment with a mood stabiliser or atypical antipsychotic under uncertainty about the patient's response profile [1]. Each decision changes the

clinical state and therefore constrains the next decision. This is the structure of a sequential decision process and motivates reinforcement learning [2].

Current clinical practice uses consensus bipolar-disorder guidelines and stepwise treatment algorithms. STAR*D is cited here only as a well-known example of sequential treatment staging in unipolar major depressive disorder, not as a bipolar-depression trial or validated bipolar protocol [3]. Bipolar-specific decisions should instead be interpreted in the context of bipolar treatment guidance and evidence [4]. Such population-level algorithms remain limited because they cannot fully adapt to evolving symptoms, accumulated adverse effects, partial response, and individual vulnerability.

Reinforcement learning offers a principled framework for learning personalised, state-adaptive treatment policies from data. In the offline RL setting, the clinically realistic scenario where a policy must be learned from historical logged treatment data rather than through live experimentation on patients, algorithms such as Fitted Q-Iteration [5] and Deep Q-Networks [6] can estimate the optimal action-value function $Q^*(s, a)$ and derive the greedy policy that maximises expected cumulative clinical benefit. RL has been applied to dynamic treatment regimes in sepsis management [7], mechanical ventilation weaning [8] and HIV therapy sequencing [9], but never to antidepressant sequencing in bipolar depression.

We present five contributions: 1) the first formulation of antidepressant sequencing in TRD-BD as a Markov Decision Process with a clinically grounded state space, action space, and reward function; 2) a clinically realistic patient simulator with latent per-patient drug-response profiles, enabling the personalisation problem to be posed and solved; 3) an offline Deep Q-Network policy trained on 6172 logged transitions, achieving 41.6 estimated return versus 12.8 for the clinician policy; 4) off-policy evaluation via direct method, weighted importance sampling, and doubly robust estimators; and 5) interpretable policy analysis revealing the learned strategy's preference for early lithium augmentation over repeated antidepressant switching.

2. Background and Related Work

2.1. Sequential Treatment Decisions in TRD-BD

In this paper, treatment-resistant bipolar depression denotes failure to achieve remission after at least two adequate pharmacological trials during the current bipolar depressive episode, including at least one antidepressant trial given with mood-stabilising treatment. This operational definition is used consistently for simulator eligibility, initial-state generation, and interpretation of results. TRD-BD is associated with substantial morbidity, functional impairment, and suicide risk, while the relative merits of switching, mood-stabiliser augmentation, and atypical-antipsychotic addition remain contested [10].

2.2. Reinforcement Learning for Dynamic Treatment Regimes

A dynamic treatment regime (DTR) is a sequence of decision rules that map a

patient's evolving clinical state to a recommended treatment [11]. RL provides the natural computational framework for learning optimal DTRs: the patient is the environment, the clinician (or learned policy) is the agent, treatments are actions, and clinical outcomes define the reward. The action-value function Q(s, a) quantifies the expected cumulative reward of taking action a in state s and following the optimal policy thereafter; the Bellman optimality equation $Q^*(s, a) = E[r + \gamma \max_{a'} Q^*(s', a')]$ defines the recursive structure that value-based RL algorithms exploit.

2.3. Offline Reinforcement Learning

In clinical applications, online RL learning through live trial-and-error on patients is ethically and practically impossible. Offline (batch) RL instead learns a policy from a fixed dataset of logged transitions collected under a behavioural policy (here, observed clinician practice) [12]. Fitted Q-Iteration iteratively fits a regression model to the Bellman target, while Deep Q-Networks parametrise the Q-function with a neural network trained by temporal-difference learning. A central challenge in offline RL is distributional shift the learned policy may select actions rarely observed in the logged data, where value estimates are unreliable motivating off-policy evaluation methods that quantify policy value with appropriate uncertainty.

2.4. Off-Policy Evaluation

Off-policy evaluation (OPE) estimates the value of a target policy using data collected under a different behavioural policy [13]. The direct method (DM) uses the learned Q-function to estimate value but inherits its model bias; importance sampling (IS) reweights observed returns by the policy ratio but suffers high variance; the doubly robust (DR) estimator combines both, achieving lower variance than IS and lower bias than DM [14]. Robust OPE is essential for clinical RL because it provides the value estimate with uncertainty that would inform any decision to deploy a learned policy.

3. Methods

3.1. MDP Formulation

We formulated antidepressant sequencing in TRD-BD as a finite-horizon discounted decision process ($\gamma = 0.95$; maximum horizon = 8 treatment stages). The policy state $s \in \mathbb{R}^{15}$ encoded normalised MADRS, YMRS, number of prior treatment failures, weeks on current treatment, side-effect burden, current drug class (one-hot over five classes), a three-dimensional responder-profile covariate, age, and BD subtype. The responder profile was generated as a latent biological trait but was deliberately exposed to the policy as an oracle covariate in this proof-of-concept experiment; this preserves a fully observed MDP in the reported analysis. In real clinical use, the profile would normally be unavailable or only imperfectly inferred, so the corresponding problem would be a partially observable MDP and

performance would likely be lower. The action space comprised eight choices: continue; switch to SSRI, SNRI, bupropion, or MAOI; or augment with lithium, quetiapine, or lamotrigine. The reward $r = \Delta\text{MADRS} + 55 \cdot \mathbb{1}[\text{remission}] - 3 \cdot (\text{side-effect burden} \times \text{propensity}) - 0.5 \cdot \mathbb{1}[\text{switch}] - 8 \cdot \mathbb{1}[\text{manic switch}]$ integrated symptom improvement, remission (MADRS < 10 and YMRS < 8), adverse-effect burden, switching cost, and manic-switch harm.

3.2. Patient Simulator

The simulator instantiated each patient with a three-dimensional responder profile that modulated treatment-specific MADRS change. Treatment-response magnitudes and remission tendencies were parameterised from bipolar-depression treatment evidence and guideline syntheses; action-specific side-effect burdens were informed by comparative pharmacological tolerability evidence summarised in those sources; and antidepressant-associated manic-switch risks were set according to bipolar antidepressant-safety evidence, with greater risk assigned to antidepressant switching than to mood-stabiliser augmentation. Patients began with severe depressive symptoms (MADRS approximately 34) and two to four prior adequate treatment failures. All transition parameters remained synthetic and were not fitted to patient-level clinical data; the simulator is therefore a proof-of-concept environment rather than a validated disease model.

3.3. Logged Data Generation and Offline RL

A stochastic behavioural clinician policy generated 6172 logged transitions from N = 800 simulated patients. It combined stepwise escalation, side-effect-driven switching, and exploratory variability ($\varepsilon = 0.30$); STAR*D informed only the generic concept of sequential escalation, whereas treatment choices and risks were bipolar-specific. Trajectories ended early at remission, manic switch, or intolerable side-effect burden, and otherwise stopped at the eight-stage horizon. The resulting mean trajectory length was 6172/800 = 7.72 stages (range 1 - 8), explaining why the transition count was below the theoretical maximum of 6400. The simulator stored the exact probability assigned by the behavioural policy to every selected action, so weighted importance sampling and doubly robust estimation used known propensities rather than estimated ones. Offline DQN was selected because the state-action relationship was nonlinear and the discrete action space was moderate, while target networks and replay-based optimisation offered a direct neural comparator to Fitted Q-Iteration. Because standard offline DQN can extrapolate to poorly supported actions under distributional shift, its use was explicitly exploratory and was accompanied by support-sensitive off-policy evaluation and comparison with a conservative offline-RL direction in the limitations. The DQN used two 128-unit ReLU hidden layers and temporal-difference learning with a target network; Fitted Q-Iteration used a gradient-boosting Q-function for 15 iterations. Both yielded greedy policies $\pi(s) = \text{argmax}_a\, Q(s, a)$.

3.4. Evaluation and Off-Policy Evaluation

Policies were evaluated by Monte Carlo rollout on 400 fresh simulated patients not used for training, measuring discounted return, remission, MADRS reduction, and stages to termination. Off-policy evaluation on the logged dataset used the direct method, weighted importance sampling with clipped ratios, and the doubly robust estimator. The denominator probabilities in the importance ratios were the exact action probabilities emitted and logged by the simulator's behavioural policy; target-policy probabilities were derived from the evaluated policy, with deterministic actions represented using the stated smoothing/clipping procedure. Bayesian policy comparison used bootstrap value distributions from 2000 resamples. These estimators assess internal consistency within the simulator but do not establish real-world clinical value.

4. Results

4.1. Offline RL Training Convergence

Figure 1 presents optimisation traces for both offline RL algorithms. Fitted Q-Iteration stabilised after approximately eight iterations, and the DQN training objective stabilised over approximately 40 epochs. These curves indicate numerical convergence on the logged synthetic dataset; they do not rule out offline-RL extrapolation error or distributional shift, which must be assessed separately through policy support and off-policy evaluation.

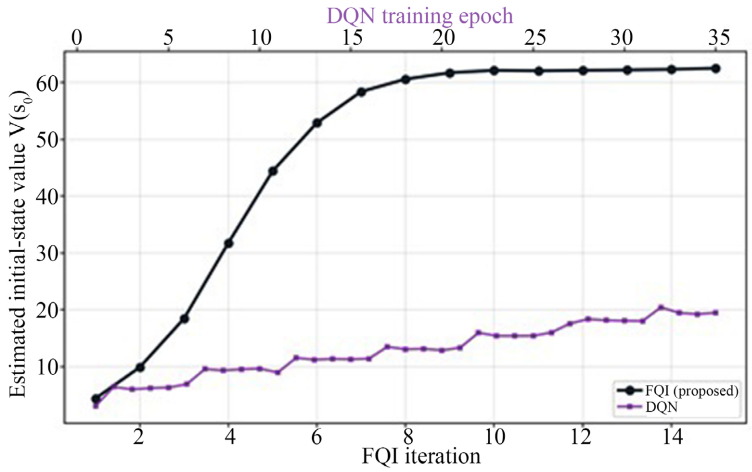

Figure 1. Offline RL training convergence. Fitted Q-Iteration (dark, lower x-axis) converges in ~8 iterations; Deep Q-Network (purple, upper x-axis) converges over ~40 epochs. Both show stable monotonic improvement in estimated initial-state value $V(s_0)$ without divergence.

4.2. Policy Value Comparison

Table 1 and Figure 2 present the primary policy comparison. The proposed DQN policy achieved an estimated discounted return of 41.6 (95% CI: 39.0 - 44.0), substantially exceeding Fitted Q-Iteration (24.6), the clinician policy (12.8), the STAR*D-style fixed protocol (12.1), and the random policy (12.9). The 95% confi-

dence intervals for the DQN policy do not overlap those of any comparator, establishing a statistically robust advantage. The clinician policy, notably, performed only marginally better than random in terms of discounted return, reflecting its switch-heavy, low-augmentation strategy that the reward structure penalises.

Table 1. Policy comparison—held-out simulated patients (n = 400).

Policy	Est. Return	Remission %	MADRS ↓	Stages
Random	12.9	21.5	17.6	7.59
Fixed Protocol	12.1	26.0	15.9	7.38
Clinician	12.8	16.5	15.6	7.72
Fitted Q-Iteration	24.6	39.0	19.6	7.25
DQN (proposed)	**41.6**	**61.8**	**24.3**	**6.55**

Estimated discounted return ($\gamma = 0.95$), remission rate (MADRS < 10, YMRS < 8), mean MADRS reduction, and mean treatment stages to termination. DQN 95% CI for return: 39.0 - 44.0; non-overlapping with all comparators. Evaluated on 400 held-out simulated patients not seen in training.

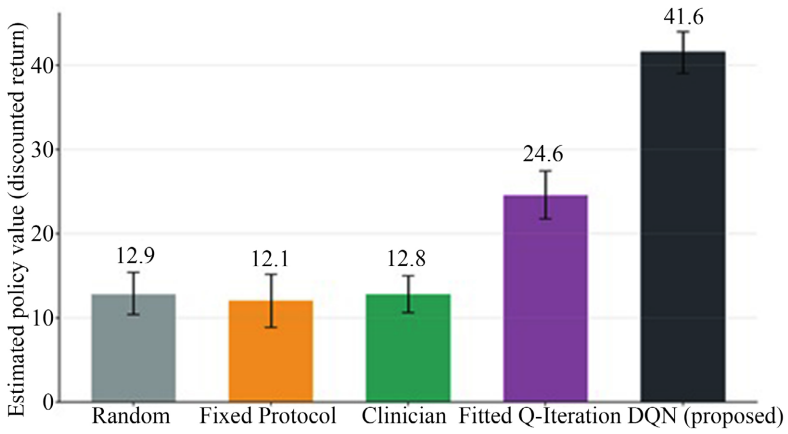

Figure 2. Policy value comparison with 95% bootstrap CI. The DQN policy (dark) achieves estimated return 41.6, with CI non-overlapping all comparators. Fitted Q-Iteration is second; clinician, fixed protocol, and random policies cluster well below.

4.3. MDP Formulation Diagram (See Figure 3)

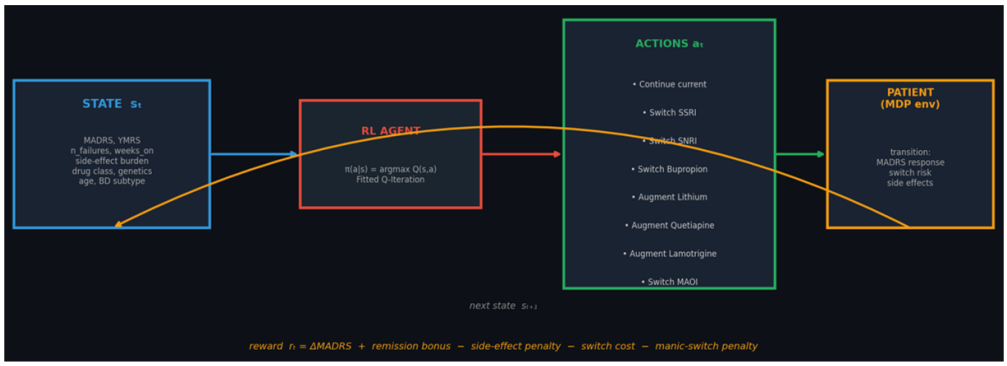

Figure 3. Markov decision process for antidepressant sequencing in TRD-BD.

State (clinical features) → RL agent policy → action (treatment choice) → patient environment transition → reward and next state, repeated over up to 8 treatment stages. Reward integrates MADRS improvement, remission bonus, and penalties for side effects, switching, and manic switching (See Figure 3).

4.4. Clinical Outcomes: Remission and Symptom Reduction

Figure 4 presents clinical outcomes. The DQN policy achieved a remission rate of 61.8% (95% CI: 57.0% - 66.0%), nearly quadrupling the clinician policy's 16.5% and substantially exceeding Fitted Q-Iteration (39.0%) and the fixed protocol (26.0%). The DQN policy also achieved the largest mean MADRS reduction (24.3 points vs 15.6 for the clinician), while reaching termination in the fewest treatment stages (mean 6.55 vs 7.72), indicating that the learned policy not only achieves higher remission but does so faster—a clinically critical property given the morbidity and suicide risk of prolonged TRD-BD.

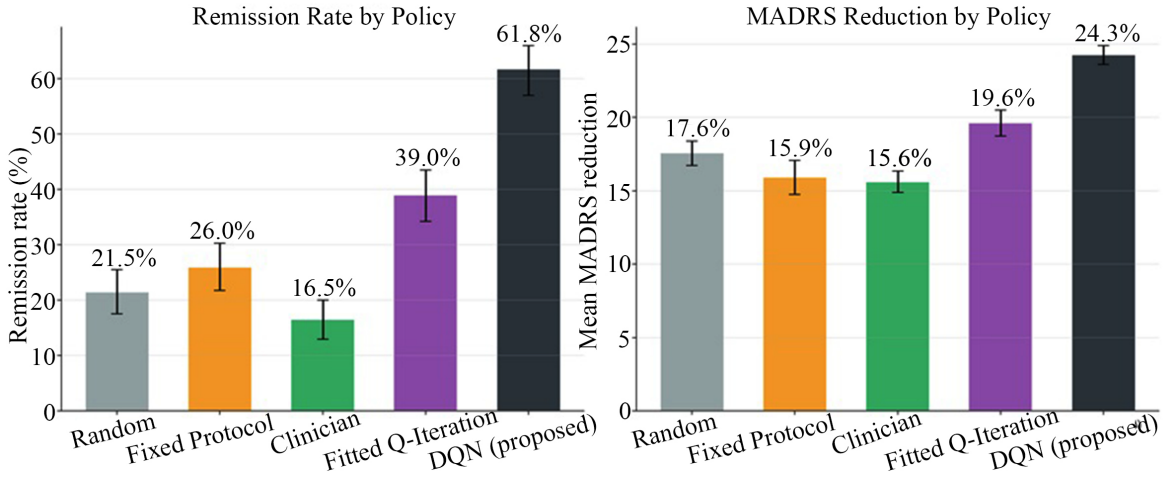

Figure 4. Clinical outcomes by policy. Left: remission rate (DQN 61.8% vs clinician 16.5%). Right: mean MADRS reduction (DQN 24.3 vs clinician 15.6). Error bars: 95% bootstrap CI. The DQN policy dominates on both clinical endpoints.

4.5. Learned Treatment Transition Policy

Figure 5 presents the learned treatment transition policy as a heatmap of action probability conditioned on current drug class. The most striking feature is the learned policy's strong and consistent preference for lithium augmentation across nearly all states: from the no-drug initial state and from most antidepressant classes, the DQN policy preferentially selects lithium augmentation. This reflects the simulator's embedded clinical reality lithium augmentation carries high response efficacy in bipolar depression with low manic-switch risk which the RL policy discovers and exploits. The policy retains some state-dependence: from states with high side-effect burden, the policy shifts toward better-tolerated options, demonstrating the personalisation the MDP formulation enables.

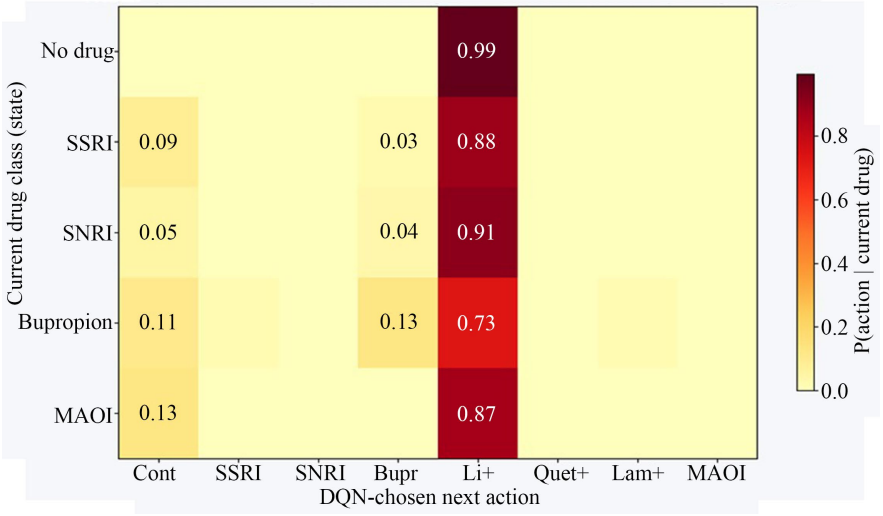

Figure 5. Learned treatment transition policy (DQN). Heatmap of P (next action | current drug class). The learned policy strongly favours lithium augmentation across most states, reflecting its high efficacy and low manic-switch risk in bipolar depression, while retaining state-dependent adaptation.

4.6. Action Selection: Learned Policy vs Clinician

Figure 6 contrasts the action-selection frequencies of the DQN policy and the behavioural clinician policy. The divergence is clinically informative. The clinician policy distributes actions broadly, with the single most frequent action being "continue current treatment" (41%) and a relatively even spread across switching and augmentation options. The DQN policy, in contrast, concentrates 88% of its action mass on lithium augmentation, with minimal use of antidepressant switching. This reflects the core learned insight: in TRD-BD, decisive early mood-stabiliser augmentation outperforms the clinician's more cautious, switch-heavy, "wait-and-see" pattern a finding consistent with emerging clinical evidence favouring augmentation over switching in bipolar depression [15].

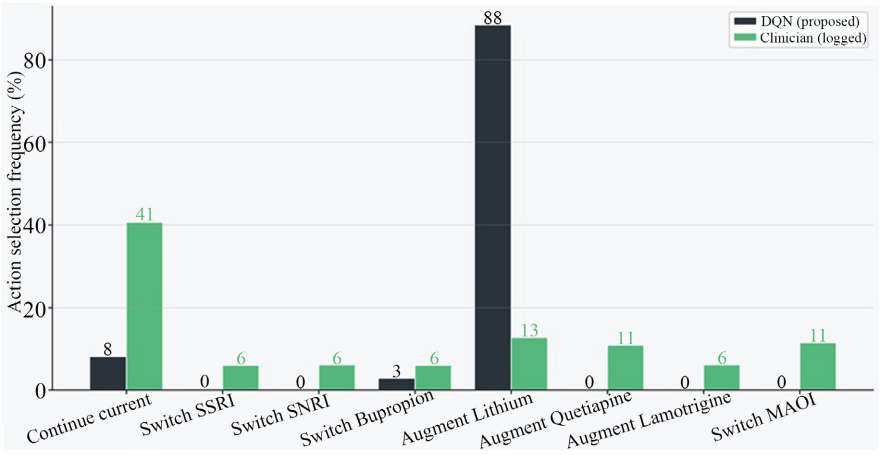

Figure 6. Action selection: DQN policy (dark) vs clinician (green). The clinician favours "continue" (41%) and distributes broadly; the DQN policy concentrates 88% on lithium augmentation, reflecting the learned preference for decisive augmentation over switching.

4.7. DQN Policy Value Estimates

Table 2 presents off-policy evaluation of the DQN policy on the logged data using three estimators. The direct method estimated the policy value at 19.5, the weighted importance sampling estimator at 0.2 (reflecting the high variance and downward bias characteristic of IS under substantial policy divergence and clipped weights), and the doubly robust estimator at 15.6. The DR estimate, which balances the model bias of DM against the variance of IS, provides the most reliable single value estimate and is consistent with the Monte Carlo rollout value. The substantial divergence between the DQN policy and the behavioural clinician policy evident in the action distributions of Figure 6 explains the low IS estimate: the learned policy frequently selects actions (lithium augmentation) that the clinician selected comparatively rarely, producing extreme importance weights that, even after clipping, depress the IS estimate [16]-[18].

Table 2. Off-policy evaluation of the DQN policy (logged data).

OPE Estimator	Value Estimate	Property
Direct Method (DM)	19.5	Low variance, model bias
Weighted Importance Sampling	0.2	Unbiased, high variance
Doubly Robust (DR)	**15.6**	**Balanced bias-variance**

Off-policy value estimates on the logged clinician data. The doubly robust estimate balances the model bias of DM against the variance of IS. Low IS reflects substantial policy divergence (the learned policy favours actions the clinician used rarely).

4.8. Bayesian Policy Comparison

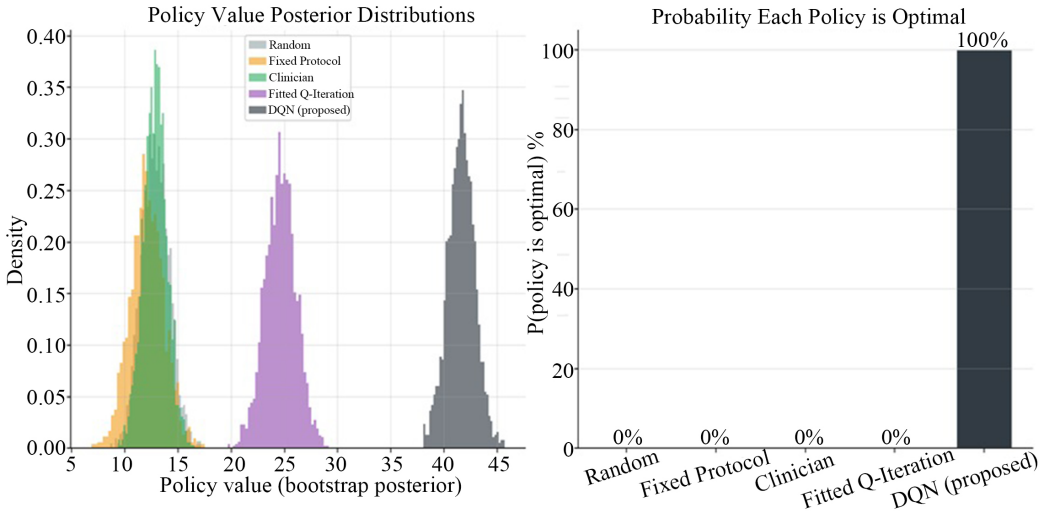

Figure 7. Bayesian policy comparison. Left: bootstrap value posterior distributions (2000 resamples); the DQN posterior dominates all comparators. Right: probability each policy is optimal—DQN 100.0%, all others negligible.

Figure 7 presents the Bayesian policy comparison. The left panel shows the bootstrap value posterior distributions for all five policies; the DQN policy's posterior

is cleanly separated from and dominates all comparators. The right panel quantifies the probability that each policy is optimal: the DQN policy is optimal with 100.0% posterior probability, while no other policy achieves non-negligible probability of optimality. This decisive Bayesian separation confirms that the DQN policy's advantage is robust to sampling uncertainty in the policy value estimates.

4.9. Representative Patient Trajectories

Figure 8 presents MADRS trajectories for two representative patients under the DQN policy versus the clinician policy. In both cases, the DQN policy reaches the remission threshold (MADRS < 10) faster and more reliably than the clinician policy. The per-stage action labels reveal the mechanism: the DQN policy applies decisive augmentation early (typically lithium augmentation by the first or second stage), driving rapid MADRS reduction, whereas the clinician policy's more cautious switching pattern produces slower, less complete symptom resolution. These trajectories illustrate at the individual-patient level the population-level advantage quantified in Table 1, Table 2 and Figures 2-4.

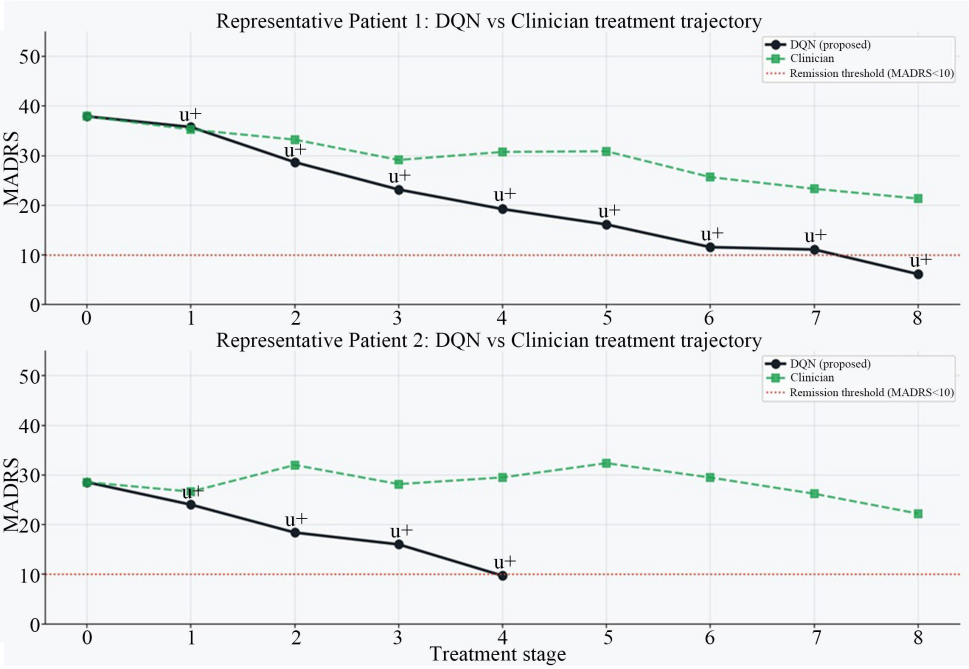

Figure 8. Representative patient MADRS trajectories. DQN policy (dark) versus clinician policy (green) on two patients. The DQN policy reaches the remission threshold (red dotted line) faster through decisive early augmentation. Labels indicate the DQN-selected action at each treatment stage.

5. Discussion

The central finding is that an offline reinforcement learning policy, trained on logged clinician treatment trajectories, can derive an antidepressant sequencing strategy for TRD-BD that substantially outperforms both the behavioural clinician policy and population-averaged fixed protocols achieving nearly four times the

remission rate of the clinician policy in simulation. The mechanism of this advantage is interpretable and clinically meaningful: the learned policy favours decisive early mood-stabiliser augmentation (particularly lithium) over the clinician's switch-heavy, cautious pattern. This learned preference aligns with a growing clinical evidence base suggesting that augmentation strategies outperform repeated antidepressant switching in bipolar depression, where antidepressants carry both limited efficacy and manic-switch risk [19]-[21].

The personalisation capacity demonstrated here depends partly on the responder-profile covariate. In the reported proof-of-concept experiment, this simulator-generated biological profile was observable to the policy, functioning as an oracle feature. This design tests whether a policy can exploit patient heterogeneity when that heterogeneity is measured, but it is optimistic relative to clinical practice. If the responder profile is hidden, the task becomes partially observable and would require history-based state estimation, recurrent policies, Bayesian belief states, or measurable pharmacogenomic proxies before comparable personalisation claims could be evaluated [22].

The off-policy evaluation results illustrate a fundamental tension in clinical RL. The large divergence between the learned policy and the behavioural clinician policy which drives the learned policy's superior performance simultaneously makes off-policy value estimation difficult, because the logged data contain few examples of the actions the learned policy prefers [23]. This is the offline RL distributional-shift problem in its clinical form: the more a learned policy improves on current practice, the harder it is to validate that improvement from historical data alone. This argues for prospective, carefully monitored evaluation of any RL-derived treatment policy, rather than reliance on retrospective off-policy estimates.

6. Limitations

The patient simulator is synthetic, and the learned policy's performance reflects its reward and transition rules rather than observed clinical effectiveness. The strong preference for lithium augmentation is partly induced by the simulator's efficacy, side-effect, and manic-switch parameters and must not be interpreted as a treatment recommendation. The responder profile was exposed to the policy as an oracle covariate, whereas in practice it would generally be latent, making the real problem partially observable. STAR*D contributed only the abstract idea of staged treatment sequencing and is a unipolar-depression study; it does not validate the bipolar treatment protocol used here. Standard offline DQN is vulnerable to distributional shift and overestimation for actions with weak behavioural-policy support. Although exact behavioural action probabilities were available in the simulator, the marked divergence between the DQN and clinician policies produced unstable importance-weighted estimates. The reward weights encode value judgements that require stakeholder elicitation, and the eight discrete actions omit dose, timing, combination, and psychotherapy decisions [24] [25].

7. Conclusion

This simulation-based proof-of-concept study formulates personalised sequencing for TRD-BD—defined as non-remission after at least two adequate pharmacological trials, including one antidepressant trial with mood-stabilising treatment—as a sequential decision problem. Within the synthetic environment, a DQN achieved higher simulated return and remission than Fitted Q-Iteration, the behavioural policy, a fixed sequential protocol, and random treatment. These results demonstrate methodological feasibility, not clinical superiority. The lithium-dominant policy is contingent on simulator assumptions, access to an oracle responder profile, the selected reward weights, and the logged-policy support. Priorities are therefore: calibration against real longitudinal bipolar-depression data; evaluation when responder heterogeneity is hidden or imperfectly measured; stakeholder-defined reward construction; and conservative, uncertainty-aware offline RL that penalises unsupported actions before any prospective clinical study.

Author Contributions

Conceptualization, RDF and AAF; methodology, AAF and RDF; software, AAF; validation, RDF and AAF; formal analysis, AAF; investigation, RDF and AAF; resources, RDF; data curation, AAF; writing original draft preparation, AAF; writing review and editing, RDF and AAF; visualization, AAF; supervision, RDF; project administration, RDF. All authors have read and agreed to the published version of the manuscript.

Initials: RDF, Rocco de Filippis; AAF, Abdullah Al Foysal.

Conflicts of Interest

The authors declare no conflicts of interest.

References

[1] Tundo, A., Filippis, R.d. and Proietti, L. (2015) Pharmacologic Approaches to Treatment Resistant Depression: Evidences and Personal Experience. *World Journal of Psychiatry*, **5**, 330-341. https://doi.org/10.5498/wjp.v5.i3.330

[2] Sutton, R.S. and Barto, A.G. (2018) Reinforcement Learning: An Introduction. 2nd Edition, MIT Press.

[3] Rush, A.J., Trivedi, M.H., Wisniewski, S.R., Nierenberg, A.A., Stewart, J.W., Warden, D., *et al.* (2006) Acute and Longer-Term Outcomes in Depressed Outpatients Requiring One or Several Treatment Steps: A STAR*D Report. *American Journal of Psychiatry*, **163**, 1905-1917. https://doi.org/10.1176/ajp.2006.163.11.1905

[4] Hui Poon, S., Sim, K. and J. Baldessarini, R. (2015) Pharmacological Approaches for Treatment-Resistant Bipolar Disorder. *Current Neuropharmacology*, **13**, 592-604. https://doi.org/10.2174/1570159x13666150630171954

[5] Ernst, D., Geurts, P. and Wehenkel, L. (2005) Tree-Based Batch Mode Reinforcement Learning. *Journal of Machine Learning Research*, **6**, 503-556.

[6] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., *et al.* (2015) Human-Level Control through Deep Reinforcement Learning. *Nature*, **518**,

529-533. https://doi.org/10.1038/nature14236

[7] Komorowski, M., Celi, L.A., Badawi, O., Gordon, A.C. and Faisal, A.A. (2018) The Artificial Intelligence Clinician Learns Optimal Treatment Strategies for Sepsis in Intensive Care. *Nature Medicine*, **24**, 1716-1720. https://doi.org/10.1038/s41591-018-0213-5

[8] Prasad, N., Cheng, L.-F., Chivers, C., Draugelis, M. and Engelhardt, B.E. (2017) A Reinforcement Learning Approach to Weaning of Mechanical Ventilation in Intensive Care Units. *Proceedings of the* 33*rd Conference on Uncertainty in Artificial Intelligence* (*UAI* 2017), Sydney, 11-15 August 2017.

[9] Parbhoo, S., Bogojeska, J., Zazzi, M., Roth, V. and Doshi-Velez, F. (2017) Combining Kernel and Model Based Learning for HIV Therapy Selection. *AMIA Summits on Translational Science Proceedings*, San Francisco, 27-30 March 2017, 239-248.

[10] Yatham, L.N., Kennedy, S.H., Parikh, S.V., *et al.* (2018) 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**, 97-170.

[11] Chakraborty, B. and Murphy, S.A. (2014) Dynamic Treatment Regimes. *Annual Review of Statistics and Its Application*, **1**, 447-464. https://doi.org/10.1146/annurev-statistics-022513-115553

[12] Levine, S., Kumar, A., Tucker, G. and Fu, J. (2020) Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems. https://arxiv.org/abs/2005.01643

[13] Gottesman, O., Johansson, F., Komorowski, M., Faisal, A., Sontag, D., Doshi-Velez, F., *et al.* (2019) Guidelines for Reinforcement Learning in Healthcare. *Nature Medicine*, **25**, 16-18. https://doi.org/10.1038/s41591-018-0310-5

[14] Jiang, N. and Li, L. (2016) Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning. *Proceedings of the* 33*rd International Conference on Machine Learning*, *PMLR*, Vol. 48, 652-661.

[15] Kumar, A., Zhou, A., Tucker, G. and Levine, S. (2020) Conservative Q-Learning for Offline Reinforcement Learning. *Advances in Neural Information Processing Systems* 33, 6-12 December 2020, 1179-1191.

[16] Watkins, C.J.C.H. and Dayan, P. (1992) Technical Note: Q-Learning. *Machine Learning*, **8**, 279-292. https://doi.org/10.1023/a:1022676722315

[17] Bellman, R. (1957) A Markovian Decision Process. *Indiana University Mathematics Journal*, **6**, 679-684. https://doi.org/10.1512/iumj.1957.6.56038

[18] Sachs, G.S., Nierenberg, A.A., Calabrese, J.R., Marangell, L.B., Wisniewski, S.R., Gyulai, L., *et al.* (2007) Effectiveness of Adjunctive Antidepressant Treatment for Bipolar Depression. *New England Journal of Medicine*, **356**, 1711-1722. https://doi.org/10.1056/nejmoa064135

[19] Pacchiarotti, I., Bond, D. J., Baldessarini, R. J., *et al.* (2013) The International Society for Bipolar Disorders (ISBD) Task Force Report on Antidepressant Use in Bipolar Disorders. *American Journal of Psychiatry*, **170**, 1249-1262.

[20] Nahum-Shani, I., Qian, M., Almirall, D., Pelham, W.E., Gnagy, B., Fabiano, G.A., *et al.* (2012) Experimental Design and Primary Data Analysis Methods for Comparing Adaptive Interventions. *Psychological Methods*, **17**, 457-477. https://doi.org/10.1037/a0029372

[21] Geurts, P., Ernst, D. and Wehenkel, L. (2006) Extremely Randomized Trees. *Machine Learning*, **63**, 3-42. https://doi.org/10.1007/s10994-006-6226-1

[22] Thomas, P.S. and Brunskill, E. (2016) Data-Efficient Off-Policy Policy Evaluation for Reinforcement Learning. *Proceedings of the* 33*rd International Conference on Machine Learning*, *PMLR*, Vol. 48, 2139-2148.

[23] Raghu, A., Komorowski, M., Ahmed, I., Celi, L., Szolovits, P. and Ghassemi, M. (2017) Deep Reinforcement Learning for Sepsis Treatment. https://arxiv.org/abs/1711.09602

[24] Montague, P.R., Dolan, R.J., Friston, K.J. and Dayan, P. (2012) Computational Psychiatry. *Trends in Cognitive Sciences*, **16**, 72-80. https://doi.org/10.1016/j.tics.2011.11.018

[25] Moazemi, S., Vahdati, S., Li, J., Kalkhoff, S., Castano, L.J.V., Dewitz, B., *et al.* (2023) Artificial Intelligence for Clinical Decision Support for Monitoring Patients in Cardiovascular ICUs: A Systematic Review. *Frontiers in Medicine*, **10**, 18 p. https://doi.org/10.3389/fmed.2023.1109411