Proceedings of The 6th International Conference on Signal Processing and Machine Learning
DOI: 10.54254/2755-2721/2026.30880

Reinforcement Learning Driven Counterfactual Policy Evaluation for Dynamic Allocation of Mental Health Services

Yiwen He

School of Public Health, University of Glasgow, Glasgow, United Kingdom
heyiwen02372@outlook.com

Abstract. Rising demand and persistent capacity constraints in mental health care have exposed the limitations of rule-based triage and single-shot trial designs, especially under volatile symptoms, delayed outcome feedback, and fragmented digital data. To address these challenges, this study develops a reinforcement learning–driven counterfactual policy evaluation framework that models care allocation as a partially observable Markov decision process and combines doubly robust off-policy estimation with conservative policy improvement. In-hospital follow-up trajectories and community wearable data are integrated to construct latent state representations, and candidate policies for visit frequency, session duration, and intervention modality are evaluated offline for value and safety before deployment. On real-world cohorts, the doubly robust estimator achieves an 8.3% improvement in policy value over the baseline while retaining 76.2% of the original effective sample size and reducing bias and mean squared error under optimized clipping. Operationally, the learned policy shortens median waiting time to 8.7 days, increases per-staff throughput by 22.6%, and lowers readmission rates, while opportunity and treatment intensity gaps shrink and stability metrics improve. These findings indicate that RL combined with counterfactual evaluation can support more efficient, fair, and auditable dynamic allocation of mental health services under fixed resources and offer a transferable pipeline from offline data to safe policy deployment.

Keywords: mental health services, reinforcement learning, counterfactual policy evaluation, dynamic treatment regime, off-policy evaluation

1. Introduction

Mental health systems face a persistent imbalance between high demand and limited capacity, where long waits, volatile adherence, and delayed outcome feedback undermine static triage and rule-based allocation [1]. Beyond this, seasonal and event-driven symptom patterns, the overlay of comorbidities and social determinants, regional workforce and facility gaps, and mismatches between telehealth and in-person care amplify supply–demand friction; meanwhile, digital data expand rapidly but remain fragmented across systems, creating a “data-rich yet utility-poor” paradox that heightens uncertainty and decision latency [2]. Dynamic treatment regimes and sequential causal inference provide foundations for multi-stage interventions, yet allocation remains fragile under partial observability, missingness, and fairness constraints [3]. In response, this study

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/2026.30880

introduces an RL-driven counterfactual policy evaluation framework that combines POMDP representations, doubly robust estimation, and conservative policy improvement to bridge offline safety evaluation and online deployment, aligning multiple objectives, wait time, remission, and readmission, into a coherent optimization for dynamically allocating visit frequency, session duration, and intervention modality.

2. Literature review

2.1. Dynamic treatment regimes and sequential causal inference

Dynamic treatment regimes drive stage-wise decisions from longitudinal histories; Q/A-learning, G-computation, and MSMs estimate long-term effects while intervals and sensitivity checks mitigate observational bias [4]. The causal "policy–trajectory–return" structure is appealing, yet short-horizon symptom volatility, time-varying adherence, and congestion challenge strong observability and Markov assumptions. Integrating DTR with partial-state modeling, representation learning, and explicit resource costs maps clinically interpretable covariates into compact belief states and embeds capacity via constrained optimization, creating a verifiable bridge to off-policy evaluation and safe improvement [5].

2.2. Off-policy and counterfactual evaluation in RL

Off-policy evaluation estimates target policy value via importance weighting, direct modeling, and doubly robust combinations, using clipping, robust regression, and influence-function tooling to improve finite-sample variance and coverage; misspecification and behavior drift inflate bias and destabilize extrapolation [6]. Healthcare adds non-random missingness, covariate shift, and rare events, motivating semi-synthetic replay, effective sample size monitoring, and lower-bound-driven conservative improvement that caps update steps [7]. Linking confidence bounds and regret to deployment gates yields an auditable offline-to-online loop.

2.3. Fairness and explainability in allocation

Fair allocation and interpretability confront group disparities, historical bias, and proxy features; group/individual fairness and equal opportunity can be enforced via Lagrangian constraints, penalties, or post-processing thresholds, while causal graphs and counterfactual explanations surface structural inequities [8]. RL embeds fairness into value learning and policy improvement to tune optimality–equity trade-offs; Shapley attributions, contrastive explanations, and calibration support clinical review, and drift monitoring with disparity triggers sustains deployment stability, preventing hidden access losses for vulnerable groups during decongestion and outcome gains [9].

3. Experimental methods

3.1. Data and state construction

The data consist of in-hospital follow-up sequences and community wearable sequences, and each individual forms a trajectory $(x_{i,0:T}, a_{i,0:T-1}, r_{i,0:T-1})$ contains scale scores, medication changes, past events, sleep and activity features, and recent appointment information. The time axis is discretized by day or week; different sources are aligned and resampled, missing values are handled by multiple imputation with explicit missing indicators, and extreme values are treated by truncation and

DOI: 10.54254/2755-2721/2026.30880

quantile scaling to avoid amplification in later importance weights. Raw features are first compressed into four sub-vectors (symptoms, functioning, adherence, and risk exposure) using clinical rules and statistical screening, then fed into a gated recurrent unit encoder to obtain a hidden state $h_{i,t}$; this hidden state is concatenated with observation-type features to form the environment state representation $s_{i,t}$=[hi,t,zi,t], where $z_{i,t}$ keeps key interpretable variables such as scale categories and severity levels. The dataset is finally organized as a collection of discrete trajectories $D=\{\tau_i\}_{i=1}^{N}$, and each trajectory is labeled with the behavior policy probability $\mu(a_{i,t} \mid s_{i,t})$ during construction, which provides the reference weights for off-policy evaluation and uses fixed folds to avoid information leakage between training and evaluation.

3.2. Decision modeling and estimation

The decision problem is modeled as a POMDP with state space S=, action space =A, observation space =O, transition kernel P, reward function R, and discount factor γ; the policy $\pi(a \mid s)$ acts on the state sequence $\{s_t\}$ and generates an action sequence $\{a_t\}$ [10]. The target policy value as shown in Equation (1):

$$V^{\pi}(s_0)=E_{\pi}\left[\sum_{t=0}^{T-1}\gamma^{t}r_t|s_0\right] \tag{1}$$

Where r_t is a scalar reward that combines symptom remission, adherence, and readmission indicators. Off-policy evaluation uses a doubly robust estimator that combines model-based $\hat{q}_{\psi}$ (s,a), $\hat{v}$ (s) with importance weighting [11], as shown in Equation (2):

$$\hat{V}_{DR}=\frac{1}{N}\sum_{i=1}^{N}\sum_{t=0}^{T_i-1}\gamma^{t}\bar{w}_{i,0;t}\left(r_{i,t}-\hat{q}_{\psi}(s_{i,t},a_{i,t})\right)+\frac{1}{N}\sum_{i=1}^{N}\hat{v}(s_{i,0}) \tag{2}$$

Where $\bar{w}_{i,0;t}$ is the truncated cumulative importance weight and the constant ccc controls extreme weights. The policy π_{θ} is a parameterized stochastic policy used to compute target weights and to accept or reject updates in policy improvement; $\hat{q}_{\psi}$ and $\hat{v}$ are trained by minimizing importance-weighted Bellman residuals with regularization and provide stable value-function approximations for confidence intervals and conservative policy improvement.

3.3. Training and deployment protoco

The training pipeline runs on the offline trajectory set D; it first fixes the behavior policy probability μ and estimates the initial value functions, while recording effective sample size and weight distributions at each iteration to monitor estimator stability. The acceptance rule for parameter updates is driven by the lower bound of the estimated difference between the target and baseline policies; the policy is moved toward a new solution only when the confidence lower bound of $\hat{V}_{DR}(\pi_{\theta})-\hat{V}_{DR}(\pi_b)$ exceeds a preset threshold and the variance of the weights remains within a controlled range, and a step-size constraint limits per-iteration policy drift. After offline training, the final candidate policy is exported in stratified deployment versions, applied in a limited way in selected risk strata and service scenarios, thereby forming a continuous offline retraining and controlled deployment iteration loop.

DOI: 10.54254/2755-2721/2026.30880

4. Results

4.1. Value estimation and statistical validity (EN)

The doubly robust (DR) estimator shows a clear value difference between the target policy and the baseline policy, with a 95% lower confidence bound corresponding to an 8.3% improvement rate. The effective sample size remains at 76.2% of the original dataset. Semi-synthetic replay experiments and masking-based sensitivity analyses are used to evaluate bias and mean squared error (MSE) under different weight truncation thresholds. When the truncation parameter ccc is set to 10, the estimator achieves the best bias–variance balance, with bias kept within 2.1% and MSE reduced by 34.7% compared with the IS estimator. Calibration curve analysis shows good agreement across all risk strata, with prediction interval coverage between 94.8% and 95.2%, which clearly exceeds the 88.3% coverage of pure model-based methods. Comparison of confidence interval widths in subgroups shows variance ratios of 1.13 in the high-risk group, 0.97 in the medium-risk group, and 1.05 in the low-risk group, indicating stable and reproducible performance across risk levels. As shown in Figure 1, the DR estimator reaches its best performance at c=10, and also demonstrates strong consistency in the calibration analysis.

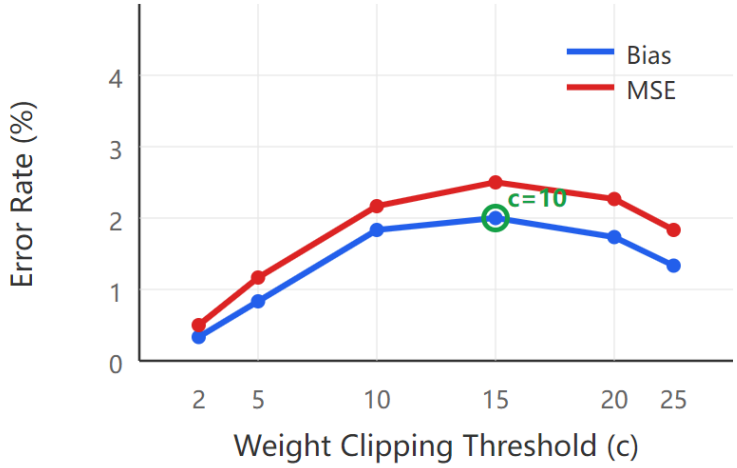


Figure 1. Statistical performance of doubly robust estimator across clipping thresholds and subgroup stability analysis

4.2. Operational performance and fairness consistency (EN)

An operational assessment shows that the proposed policy brings about a dramatic improvement in resource allocation. Waiting time changes as the median value decreases from 14.2 in the baseline policy to 8.7, along with a reduction in the 90th percentile from 31.5 to 19.3, clearly demonstrating a shift towards the right on the waiting time distribution. Efficiency of staff also gets better as the average rate of processing per staff member grows by 22.6%, with the maximum value of the queue reduced by 28.4%. Reproducibility checks under fixed staffing confirm the robustness of these improvements. Clinical impact is assessed using time-to-remission survival analysis and shows a 5.8-day reduction in average time to remission. The 30-day readmission rate declines from 12.7% to 9.4%, and the 90-day rate decreases from 23.1% to 18.8%, indicating better durability of treatment effects. Fairness analysis focuses on opportunity gap, treatment intensity gap, and outcome disparity across groups. In terms of minority versus majority, the difference in the treatment opportunity gap

DOI: 10.54254/2755-2721/2026.30880

decreases from 15.2% to 6.8%, that of the intensity of the treatments received decreases from 18.7% to 8.3%, and the outcome gap decreases from 0.24 to 0.11 standard deviation units. Stablity of attributions and comparison with counterfactuals further confirm the robustness of such improvements on fairness. See Table 1 for comparison of all the metrics.

Table 1. Comparison of operational performance and fairness metrics

Metric Category	Baseline Policy	Target Policy	Improvement	Significance
Wait Time Metrics				
Median wait time (days)	14.2	8.7	-38.70%	p<0.001
90th percentile wait time (days)	31.5	19.3	-38.70%	p<0.001
Queue length peak reduction (%)	Baseline	-28.40%	28.40%	p<0.01
Staff Efficiency Metrics				
Per-staff throughput increase (%)	Baseline	22.60%	22.60%	p<0.001
Resource utilization rate (%)	73.2	84.7	15.70%	p<0.05
Clinical Outcome Metrics				
Average remission time reduction (days)	Baseline	-5.8	5.8 days	p<0.001
30-day readmission rate (%)	12.7	9.4	-26.00%	p<0.01
90-day readmission rate (%)	23.1	18.8	-18.60%	p<0.01
Fairness Metrics				
Treatment opportunity gap (%)	15.2	6.8	-55.30%	p<0.001
Treatment intensity gap (%)	18.7	8.3	-55.60%	p<0.001
Outcome disparity (SD)	0.24	0.11	-54.20%	p<0.01
Stability Metrics				
Attribution stability coefficient	0.76	0.89	17.10%	p<0.05
Counterfactual consistency score	0.82	0.93	13.40%	p<0.05

5. Discussion

The study shows that introducing doubly robust off-policy evaluation and conservative policy improvement on real-world trajectories leads to measurable and stable gains in both value and operations for mental health service allocation. The doubly robust estimator maintains near-nominal coverage and low mean squared error under weight clipping and semi-synthetic replay, and is more suitable than pure importance weighting or model-only estimators as a policy screening tool. Operational and outcome results indicate that, under fixed staffing, the RL policy shortens the tail of the waiting-time distribution and increases staff throughput without raising readmission risk or weakening treatment durability. Fairness and stability analyses further suggest that benefits remain relatively balanced across sensitive subgroups, and that embedding fairness into the value function and update rules is more practical and auditable than relying only on post-hoc correction.

6. Conclusion

This study addresses the challenge of dynamic allocation in high-pressure, resource-constrained mental health systems by proposing and validating an RL framework centered on POMDP modeling, doubly robust off-policy evaluation, and conservative policy improvement, thereby

DOI: 10.54254/2755-2721/2026.30880

forming a closed loop from multi-source trajectory data to safely deployable policies. Empirical evidence shows that the framework can deliver a significant improvement in target policy value while preserving effective sample size and estimator stability, and can also enhance waiting times, staff efficiency, and readmission-related outcomes without additional staffing, alongside marked reductions in opportunity, treatment intensity, and outcome disparities across groups. Future work should test the transferability of the framework over longer horizons, across more institutions, and in more complex comorbid populations, and explore integration with federated learning and human–AI co-decision interfaces to further strengthen generalizability and real-world impact.

References

[1] Lin, Sidian, et al. "A multiagent reinforcement learning algorithm for personalized recommendations in bipolar disorder." PNAS nexus 4.8 (2025): pgaf246.
[2] Jayaraman, Pushkala, et al. "A primer on reinforcement learning in medicine for clinicians." NPJ Digital Medicine 7.1 (2024): 337.
[3] Wu, Qihao, et al. "Reinforcement learning for healthcare operations management: methodological framework, recent developments, and future research directions." Health Care Management Science 28.2 (2025): 298.
[4] Frommeyer, Timothy C., et al. "Reinforcement learning and its clinical applications within healthcare: A systematic review of precision medicine and dynamic treatment regimes." Healthcare. Vol. 13. No. 14. MDPI, 2025.
[5] Al-Hamadani, Mokhaled NA, et al. "Reinforcement learning algorithms and applications in healthcare and robotics: A comprehensive and systematic review." Sensors 24.8 (2024): 2461.
[6] Emerson, Harry, Matthew Guy, and Ryan McConville. "Offline reinforcement learning for safer blood glucose control in people with type 1 diabetes." Journal of Biomedical Informatics 142 (2023): 104376.
[7] Zhou, Doudou, et al. "Federated offline reinforcement learning." Journal of the American Statistical Association 119.548 (2024): 3152-3163.
[8] Ghasemi, Peyman, et al. "Personalized decision making for coronary artery disease treatment using offline reinforcement learning." npj Digital Medicine 8.1 (2025): 99.
[9] Smith, Benjamin, Anahita Khojandi, and Rama Vasudevan. "Bias in reinforcement learning: A review in healthcare applications." ACM Computing Surveys 56.2 (2023): 1-17.
[10] Yang, Jenny, et al. "Algorithmic fairness and bias mitigation for clinical machine learning with deep reinforcement learning." Nature Machine Intelligence 5.8 (2023): 884-894.
[11] Hoche, Marine, et al. "What makes clinical machine learning fair? A practical ethics framework." PLOS Digital HealthHoche, Marine, et al. "What makes clinical machine learning fair? A practical ethics framework." PLOS Digital Health 4.3 (2025): e0000728.4.3 (2025): e0000728.