Evaluating Early Treatment Initiation Regimes in Treatment-Resistant Depression: A Reinforcement Learning-Based Off-Policy Evaluation Approach

MSR55

You Wang[1 2] G. Caleb Alexander, MD, MS[1 2 3] Jason B. Gibbons, PhD[4]

1. Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health
2. Center for Drug Safety and Effectiveness, Johns Hopkins Bloomberg School of Public Health
3. Division of General Internal Medicine, Johns Hopkins Medicine
4. Division of Pharmacoepidemiology and Pharmacoeconomics, Department of Medicine, Brigham and Women's Hospital, Harvard Medical School

Contact: You Wang | ywan1033@jhu.edu

OBJECTIVES

This research aims to develop, pilot, and empirically evaluate a reinforcement learning (RL)-based framework for evaluating early treatment initiation strategies among individuals with treatment-resistant depression (TRD), using administrative claims data.

1. Evaluate fixed early treatment initiation strategies and a learned adaptive strategy against observed clinical practice
2. Assess the feasibility of offline RL in a psychiatric pharmacoepidemiologic setting
3. Validate model recommendations through concordance-outcome analysis

Study at a Glance

1,327,189 individuals with TRD
15,000,564 treatment transitions over 12 months
32 treatment combinations • **19** state features

METHODS

Design & Population

Retrospective cohort study using de-identified administrative claims from a large commercially insured U.S. population. Adults with TRD followed for 12 monthly decision points. Patient-level split: train 70%, validation 15%, test 15%.

MDP Formulation

States: 19 features — demographics, comorbidities, prior treatment (5 binary), treatment gap duration, utilization

Actions: 32 combinations — antidepressants, antipsychotics, psychotherapy, esketamine, ECT/TMS

Rewards: −10 suicide attempt/self-harm; −5 hospitalization, ED, ideation; −1 discontinuation (≥3 mo); 0 otherwise

Horizon: 12 monthly decision points. Discount γ = 0.99.

Model Architecture

Averaged Dueling Double Deep Q-Network (Averaged-DQN). Dueling architecture separates state value V(s) from action advantage A(s,a). Double Q-learning reduces overestimation. Averaged-DQN (K=20 snapshots) further stabilizes training by averaging Q-value estimates across recent network checkpoints.

Architecture: 3 shared layers (256→256→128), LayerNorm, ReLU, dropout(0.2), separate value (64→1) and advantage (64→32) streams. 500 epochs × 1,000 steps, batch 4,096.

Strategy Definitions

Learned adaptive: optimizes every monthly decision
Fixed initiation: initiate specified treatment at month 0, adapt optimally thereafter
Observed practice: actual clinician decisions (reference)

Strategy values = composite adverse outcome burden scores for **ranking only** — not treatment effect estimates.

Concordance-Outcome Analysis

Compared observed outcomes when clinician decisions coincidentally matched the learned strategy's recommendation (concordant) vs. when they differed (discordant). Sensitivity analyses restricted to actively treated months.

RESULTS

Table 1. Treatment Strategy Evaluation (Test Set, n = 199,079)

Strategy	Value	Benefit	Relative	95% CI	Freq. %
Observed practice	−0.919	*Ref.*	—	—	—
Learned adaptive	−0.831	**+0.088**	**+9.6%**	(0.088, 0.089)	—
Antidepressants	−0.842	+0.078	+8.5%	(0.077, 0.078)	32.5
Psychotherapy + AD	−0.854	+0.065	+7.1%	(0.064, 0.065)	4.3
Esketamine	−0.889	+0.030	+3.3%	(0.029, 0.030)	0.2
Psychotherapy	−0.953	−0.033	−3.6%	(−0.033, −0.032)	3.4
Antipsychotics	−1.044	−0.125	−13.6%	(−0.125, −0.124)	4.6
AD + Antipsychotics	−1.086	−0.167	−18.2%	(−0.167, −0.166)	9.1
No treatment	−1.463	−0.544	−59.2%	(−0.544, −0.542)	43.9

Strategy values are composite adverse outcome burden scores; less negative = more favorable. 95% CIs from B=200 individual-level bootstrap. Selected strategies shown; see eTable 1 for full 34-strategy ranking.

Figure 1. Training Dashboard (Averaged-DQN, 500 epochs; epoch 350 selected)

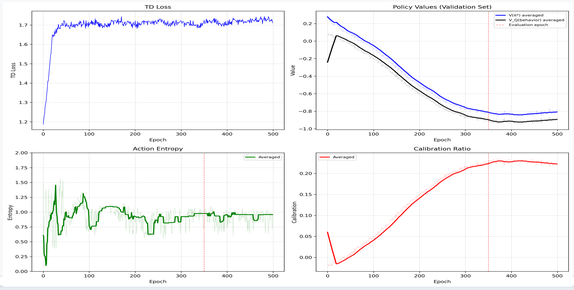


Figure 2. Treatment Strategy Advantages with 95% Confidence Intervals

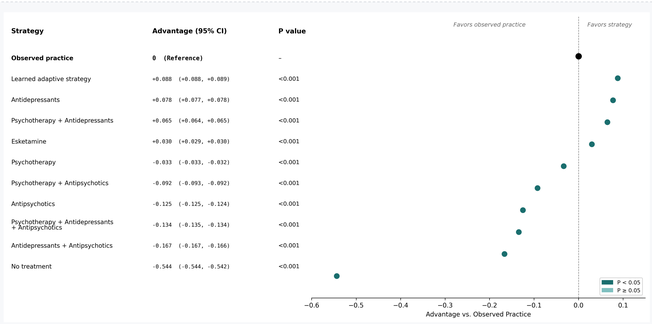


Figure 3. Strategy Values Across 12-Month Follow-Up

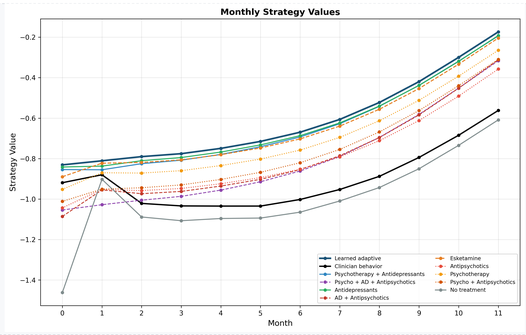
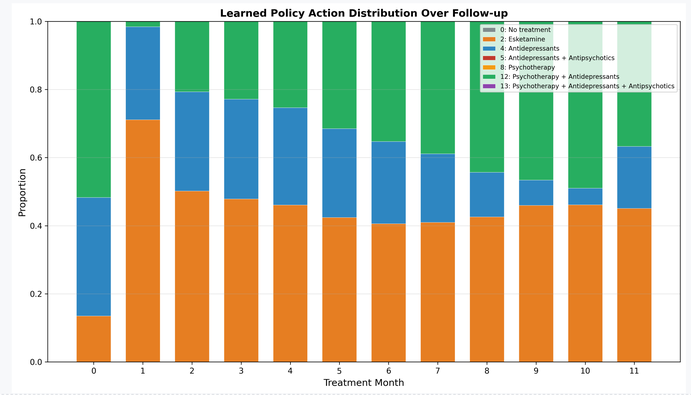


Figure 4. Treatment Recommendations of the Learned Adaptive Strategy Across Follow-Up



KEY FINDING

The learned adaptive strategy showed **+9.6%** relative benefit over observed practice. Among fixed strategies, antidepressants (+8.5%) and psychotherapy + antidepressants (+7.1%) ranked highest.

CONCORDANCE-OUTCOME ANALYSIS

Primary: Concordant months (10.0%) had better outcomes than discordant
(mean score −0.101 vs −0.408; difference +0.307; P < .001)

High severity: Concordance advantage strongest (diff +0.526; P < .001)

Sensitivity 1 (active treatment only): diff +0.039; P < .001
Sensitivity 2 (always-treated, 12/12 months): diff +0.012; P = .049

→ Learned strategy optimizes beyond treating the untreated

Concordance highest at month 0 (25.3%), declining in subsequent months — greatest model departure from clinicians occurs at treatment initiation.

At treatment initiation: learned strategy recommended psychotherapy + AD in 51.7%, AD in 34.8%, esketamine in 13.5% of individuals.
Across all months: esketamine 40.9%, AD 35.4%, psychotherapy + AD 23.7%. No treatment < 0.1% (vs. 43.9% observed).

CONCLUSIONS

Offline RL produced clinically plausible treatment strategy rankings for TRD from administrative claims data. Combination strategies — particularly psychotherapy + antidepressants — ranked most favorably; no treatment ranked last.

The concordance-outcome analysis confirmed that concordant treatment months were associated with better outcomes, with sensitivity analyses demonstrating optimization beyond treatment continuity alone.

Limitations: Claims lack symptom severity (Markov assumption likely violated); low-frequency treatments subject to positivity violations; reward weighting not empirically calibrated; direct Q-value estimation (fitted Q-evaluation a future direction).

Next steps:
→ Extension to EHR with PHQ-9 and repeated clinical measures
→ Fitted Q-evaluation for strategy-specific estimation
→ Time-varying concordance with clinical severity
→ Prospective validation of strategy recommendations

Contact: You Wang | ywan1033@jhu.edu | Epidemiology, BSPH, JHU