ALPACA: A Reinforcement Learning Environment for Medication Repurposing and Treatment Optimization in Alzheimer's Disease

Nolan Brady NOLAN.BRADY@COLORADO.EDU
University of Colorado Boulder, United States

Tom Yeh TOM.YEH@COLORADO.EDU
University of Colorado Boulder, United States

Abstract

Evaluating personalized, sequential treatment strategies for Alzheimer's disease (AD) using clinical trials is often impractical due to long disease horizons and substantial inter-patient heterogeneity. To address these constraints, we present the *Alzheimer's Learning Platform for Adaptive Care Agents* (ALPACA), an open-source, Gym-compatible reinforcement learning (RL) environment for systematically exploring personalized treatment strategies using existing therapies. ALPACA is powered by the Continuous Action-conditioned State Transitions (CAST) model trained on longitudinal trajectories from the Alzheimer's Disease Neuroimaging Initiative (ADNI), enabling medication-conditioned simulation of disease progression under alternative treatment decisions. We show that CAST autoregressively generates realistic medication-conditioned trajectories and that RL policies trained in ALPACA outperform no-treatment and behavior-cloned clinician baselines on memory-related outcomes. Interpretability analyses further indicated that the learned policies relied on clinically meaningful patient features when selecting actions. Overall, ALPACA provides a reusable in silico testbed for studying individualized sequential treatment decision-making for AD.

Keywords: Alzheimer's Disease, ADNI, Medical Simulation Environment, Disease Forecasting

Environment and Code Availability The ALPACA environment proposed in this paper is made available for use on GitHub [1] and as a pip package [2].

Institutional Review Board (IRB) This study used de-identified data from the ADNI. [3] Therefore, IRB approval was not required.

1. Introduction

The prevalence of AD increased by 147.9% worldwide between 1990 and 2019 Li et al. (2022). Despite this growth, identifying effective treatment regimens remains constrained by the slow, expensive, and ethically bound nature of clinical trials. These constraints are particularly limiting for personalized treatment strategies, in which clinicians must decide what to prescribe, when to initiate therapy, and how to adapt treatment as patient trajectories evolve over the years. Consequently, much of the AD research has focused on a narrow set of therapeutic targets, most notably amyloid-β and tau, while many medication repurposing and combination strategies remain comparatively underexplored.

In silico experimentation offers a potential path forward by shortening iteration cycles and enabling large-scale hypothesis generation but hinges on the availability of physiologically plausible simulations that respond meaningfully to intervention decisions. Mechanistic models, such as ordinary differential equations (ODEs), provide interpretability but are often vulnerable to misspecification. In contrast, much of the deep learning literature emphasizes trajectory reconstruction or disease state classification rather than closed-loop simulations. Although offline rein-

1. https://github.com/nolanrbrady/ALPACA-RL
2. https://pypi.org/project/ALPACA-DT-Sim/

3. Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). ADNI investigators contributed to the ADNI design and data collection but did not participate in this analysis. A complete list of ADNI investigators is available at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

forcement learning can optimize policies from retrospective data, it becomes unreliable for both policy improvement and evaluation when candidate actions fall outside the support of the original dataset, thus limiting its use cases Fujimoto et al. (2018).

These challenges are particularly pronounced in AD because of its underlying biology. Once believed to arise from a largely monolithic cause and therefore amenable to a single dominant treatment strategy, recent molecular and clinical stratification studies indicate the presence of biologically distinct AD subtypes with differential treatment sensitivities Tijms et al. (2024). Compounding this complexity, evidence suggests that treatment effects are also time-dependent, such that the timing of intervention can yield meaningfully different patient outcomes Amirrad et al. (2017); Winblad et al. (2006); Perneczky and Froelich (2025); Tarawneh and Pankratz (2024); Sims et al. (2023b). Together, these factors give rise to a combinatorial space of patient profiles, medication classes, and initiation windows that are infeasible to explore exhaustively through randomized clinical trials.

To address these challenges, we introduce ALPACA (*Alzheimer's Learning Platform for Adaptive Care Agents*), an in silico reinforcement learning environment with continuous-valued clinical states, multi-binary medication actions, and transitions generated by CAST, a medication-conditioned autoregressive forecasting model trained on longitudinal trajectories from the ADNI dataset. Figure 1 summarizes the ALPACA pipeline and CAST-based action-conditioned state transitions.
Our contributions are:

- **CAST Model.** A medication-conditioned autoregressive model for multivariate AD trajectory simulation (Section 3.1).
- **ALPACA environment.** A Gym-compatible RL environment with continuous clinical states and a 17-class multi-binary medication action space (Section 3.2).
- **Benchmark analysis.** Policy learning and comparison with clinician baselines, accompanied by interpretability analysis (Section 4).

2. Related Works

To contextualize ALPACA, we reviewed two complementary areas: (i) simulation-based reinforcement learning in offline medical settings and (ii) prior AD-focused simulation environments.

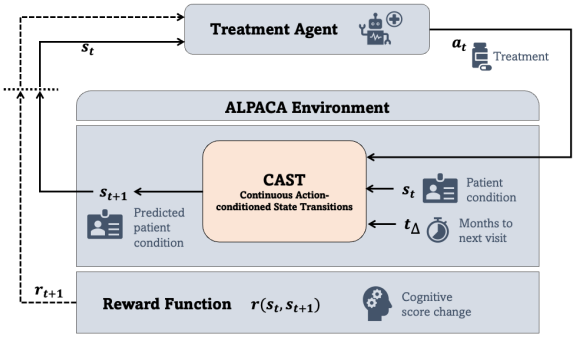


Figure 1: The ALPACA environment wraps the CAST model. Given a current state s_t and medical action a_t, the CAST model predicts the future state s_{t+1} (t_Δ months ahead), with r_{t+1} representing the resulting reward.

2.1. Simulation-based Reinforcement Learning Environments

In many medical settings, reinforcement learning is primarily applied offline because of ethical and practical barriers to online experimentation Jayaraman et al. (2024). Offline RL enables policy learning from static datasets Levine et al. (2020), but it raises challenges for both policy optimization and evaluation when the target policy deviates from the data-generating distribution.

In offline reinforcement learning, the extrapolation error is the primary failure mode. Fujimoto et al. showed that the distributional mismatch between a learned policy and a fixed batch of data can lead to unreliable value estimates for underrepresented or unseen actions Fujimoto et al. (2018). In clinical applications, constraining learned policies to support the observed clinician behavior can mitigate out-of-distribution recommendations and promote conservative improvements Huang et al. (2024). However, this constraint simultaneously limits a policy's ability to generalize beyond its historical practice. The second major bottleneck lies in the evaluation of offline policies. When the evaluation policy deviates from the behavior policy, distribution shifts can induce high-variance estimates in off-policy evaluation (OPE), undermining reliable performance assessment Uehara et al. (2022). These challenges are further exacerbated in long-horizon settings, where com-

pounding variance and bias make policy evaluation increasingly unstable Levine et al. (2020); Bossens and Thomas (2022). Together, these limitations pose significant challenges in verifying the efficacy and safety of learned policies, which is critical in medical decision-making contexts.

Offline model-based reinforcement learning (offline MBRL) partially mitigates these issues by learning a transition dynamics model from retrospective data and using simulated rollouts for policy optimization. Model-based Offline Policy Optimization (MOPO), while not strictly medical, demonstrates the method's ability to generalize past the bounds of static training data, where other offline RL policies cannot Yu et al. (2020). MOPO achieves this by incorporating a penalty based on model uncertainty to manage distribution shifts during optimization while allowing controlled generalization beyond strict behavior cloning Yu et al. (2020). The medical applications of offline MBRL include TR-GAN, which combines observational trajectories with simulated counterfactual rollouts for treatment recommendation Sun et al. (2022), and OMG-RL, which uses learned dynamics to support offline inverse reinforcement learning for dosing under a reward inferred from clinician behavior Lim and Lee (2024). ALPACA extends the offline MBRL paradigm to AD by providing a medication-conditioned transition model in a reusable and Gym-compatible environment.

2.2. Current State of AD Reinforcement Learning Environments

Reinforcement learning has been applied across clinical domains such as sepsis, oncology, diabetes, and neurodegenerative disease to optimize sequential treatment policies Luo et al. (2024); Ghaffari et al. (2016); Ahn and Park (2011); Oberst and Sontag (2019); Man et al. (2014); Wang et al. (2023); Bhattarai et al. (2023).

In AD, only a limited number of studies have directly addressed the environment gap. Bhattarai et al. construct an environment based on a discretized Markov decision process derived from ADNI, in which Mini-Mental State Exam (MMSE) based cognitive measures are mapped to a finite set of disease stages using decision trees Bhattarai et al. (2023). The action space consists of a discrete set of interventions including AD medications, antihypertensives, and supplements Bhattarai et al. (2023). Although this formulation aligns naturally with value-based methods such as Deep Q-learning, it compresses continuous clinical trajectories into categorical disease states and substantially constrains the space of possible treatment strategies.

In contrast, Saboo et al. developed a mechanistic simulation based on coupled differential equations modeling amyloid pathology, brain atrophy, and neural activity Saboo et al. (2021). Reinforcement learning was used in this setting to learn the patterns of information processing load across simulated brain regions. Consequently, the learned policy functions primarily as a model of brain responses to pathology rather than as a framework for optimizing sequential treatment decisions Saboo et al. (2021).

These efforts advance AD simulation; however, common limitations persist. Existing environments often rely on coarse state abstractions and restricted medication spaces, limiting their ability to evaluate diverse treatment sequences and repurpose hypotheses. ALPACA addresses these gaps by modeling continuous clinical states and using learned autoregressive, medication-conditioned transitions to simulate alternative treatment regimes in a multi-binary action space spanning 17 therapeutic classes.

3. ALPACA: Transition Model and Environment

First, we introduce the CAST model and discuss its role as a forecasting engine for the ALPACA environment. Finally, we cover environmental assumptions and reward function design.

3.1. CAST Model

3.1.1. Model Architecture

The CAST model employs a mixture-of-experts (MoE) transformer architecture designed to capture the heterogeneity of the progression of Alzheimer's disease Shazeer et al. (2017). Unlike standard transformers, the MoE design allows distinct experts to specialize in different medical scenarios, resulting in higher-fidelity trajectory predictions. The network is composed of three transformer layers with an embedding dimension of 256 and four attention heads. In each layer, only one of the eight experts is activated per visit to ensure computational efficiency. The model operates in an autoregressive manner with causal masking, taking 21 clinical variables, 17 multi-binary treatment actions, and a time-to-next-visit

feature as inputs. It predicts the patient's future state across both continuous biomarkers (e.g., tau, amyloid-β, and brain volume) and cognitive assessments (e.g., ADNI-Mem and ADNI-EF).

3.1.2. Training Data

Raw ADNI Dataset The data used in the preparation of this paper were obtained from the Alzheimer's Disease Neuroimaging Initiative database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by the Principal Investigator Michael W. Weiner, MD. The primary goal of the ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessments can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease.

Longitudinal Data Processing Raw longitudinal data from the ADNI dataset were transformed into a structured format that was optimized for autoregressive visit-to-visit state forecasting. Medication records were consolidated into therapeutic drug classes (e.g., AD treatments, statins, and antihypertensives), with activity windows determined by the start and end dates relative to each visit. The partitioning for drug-class mappings can be found in Appendix I.

Patient age was dynamically recalculated from baseline to ensure temporal consistency, and participants with fewer than three visits were excluded to provide a sufficient trajectory length for modeling. The full outline of the demographic information for the ADNI patient population used to train the CAST model is presented in Table 1.

Table 1: Demographic and Trajectory Summary

Characteristic	Full Dataset
Overview	
Visits	12,984
Unique subjects	1,905
Demographics	
Age (mean ± SD)	76.2 ± 7.4
Range (yrs)	55.0–103.1
Sex (Male / Female)	54.8% / 45.2%
Race (White)	93.0%
Race (Black / Asian / Other)	3.9% / 1.7% / 1.4%
Cognitive Scores	
ADNI MEM	0.38 ± 1.16
ADNI EF2	0.14 ± 1.02
Trajectory Lengths	
Mean visits per subject	6.82 ± 3.72
Median (IQR)	6 (4–9)
Range (min–max)	3–22
95th percentile	15
Temporal Variable	
Next visit interval (months)	8.6 ± 6.9
Visit Disease State	
Healthy / Impaired (MCI/AD)	71.4% / 28.6%

Patient demographic information for the ADNI data patients used in the creation of the forecasting model. Impaired visits in this table is measured as ADNI Mem score under -0.1 Crane et al. (2012). Distributions across splits were comparable.

Data Imputation Missing data across continuous biomarkers (tau, amyloid-β, neuroimaging volumes, ADNI-Mem, ADNI-EF, and age) were imputed using an iterative ExtraTrees-based method Pedregosa et al. (2011), which was chosen for its ability to capture non-linear relationships without assuming parametric distributions. The imputer and z-score scaler were fitted only to the training set and reused unchanged for validation and testing to avoid data leakage. Categorical variables were one-hot encoded using k–1 dummy coding to prevent multicollinearity.

Resulting Data After preprocessing, two datasets were generated: one for the CAST model and the other for the behavior cloning model. The CAST dataset was structured such that each sample represented a patient state–action pair, with the prediction target as the subsequent visit state, which supported the autoregressive loss calculation. A behavior cloning clinician dataset was generated to capture real-world treatment patterns by mapping patient states to prescribed medication regimens, enabling a comparison between the learned agent policies and a learned clinician policy from the dataset. Details on feature and action variables can be found in Appendix H.

Dataset Split To ensure independence across individuals, the data were split at the subject level into training (70%), validation (15%), and test (15%) sets and were kept the same for the clinician and CAST model datasets.

3.1.3. Training Protocol

The model was trained to forecast longitudinal patient trajectories using teacher forcing and the AdamW optimizer with an L2 weight decay (0.01). The learning rates were managed using a ReduceLROnPlateau scheduler Paszke et al. (2019), and regularization was enforced using a dropout rate of 0.3 in the transformer layer.

The composite loss function combined three terms: the mean squared error for continuous variables, binary cross-entropy for categorical outputs, and an auxiliary load-balancing loss (scaled by 0.005) to prevent expert collapse.

To ensure a robust evaluation, all experiments utilized pre-split training, validation, and test sets with subject identifiers excluded to prevent data from being leaked.

3.1.4. Forecasting Model Validation

We validated the CAST model using held-out data to assess its ability to predict the trajectories of unseen patients. Because most ADNI visits occurred at varying time intervals, the CAST model was trained and evaluated based on actual gaps between visits. Validation relied on two complementary tests: the Maximum Mean Discrepancy (MMD) and a Mantel test. The MMD test, which measures the kernel-based distributional similarity between the forecasted and ground truth trajectories, was computed using an RBF kernel (1,000 permutations) on z-scored transition vectors. The Mantel test complements the MMD by evaluating whether the pairwise similarity structure between time points (i.e., temporal adjacency in feature space) is preserved between the predicted and ground truth trajectories Panda et al. (2019); Mantel (1967). The group-level analysis for the Mantel test was split into two tests: a Fisher-Z one-sided t-test on the Mantel r values, and a permutation-based group test. The permutation test used 5,000 permutations, where only the time dimension of the forecasted series was permuted.

This suite of tests was selected because it provides a comprehensive characterization of the dynamic model performance and how it compares with real-world patients.

Evaluation Method The model was initialized with each test set subject's first observed state and generated subsequent transitions using only past predictions and ground truth actions in an autoregressive manner. This maximized the opportunity for model errors to compound and drift away from the ground truth, thereby faithfully evaluating the ability of the forecasting model to reconstruct the high-fidelity trajectories.

Results The trajectory reconstruction in the test set did not show statistically significant differences in either step-wise dynamics or long-range trends relative to the ground truth trajectories (at $\alpha = 0.05$). The short-range MMD, which measures the distribution of one-step visit-level transitions, failed to reject the difference between the forecasted and ground truth trajectories (MMD $= 4.65 \times 10^{-3}$, $p = 0.3916$). Similarly, the long-range MMD, which measures the overall drift from the first to the last visit, failed to reject the difference between the start and end state distributions (MMD $= 3.79 \times 10^{-3}$, $p = 1.00$). At the feature level, no individual feature showed a statistically significant deviation between the two distributions. The Fisher-Z one-sided t-test on Mantel r values showed strong preservation of interfeature dependency between the forecasted and ground truth trajectories in the test set ($r = 0.7104$, $CI = 0.7472 - 0.8102$, $p = 1.99 \times 10^{-4}$, n = 145). The permutation group test (5,000 within-subject timepoint shuffles) showed that the observed mean Mantel-like similarity significantly exceeded the shuffled null hypothesis.

These results demonstrate that the forecasting model captures physiologically valid transition dynamics, generalizes unobserved patient states, and provides a stable foundation for reinforcement learning experiments. The convergence of the MMD and Mantel results indicates a strong distributional and relational similarity between the forecasting model trajectories and ground truth, even in the presence of minor differences in data.

3.2. ALPACA Environment

The ALPACA environment wraps the CAST model as its core state transition function [4].

3.2.1. State Initialization and Population

To preserve patient privacy, ALPACA does not use any real patient's initial visit to seed an episode. Instead, we fit three separate Gaussian Mixture Models (GMMs) to the multivariate distribution of baseline states in the ADNI training set. In total, we fit one using all participants, one restricted to cognitively healthy participants, and one restricted to cognitively impaired participants. Consistent with prior work, impairment was defined as an ADNI-Mem score below -0.1 Crane et al. (2012).

Each GMM was trained independently, and for each model we selected the number of mixture com-

4. Due to the tight coupling between CAST and the ADNI dataset no other datasets are currently supported within the ALPACA environment.

ponents by fitting candidate models with 1–100 components and choosing the configuration with the lowest Bayesian Information Criterion (BIC). At the beginning of each episode, the environment samples a synthetic initial patient state from a single GMM corresponding to the user-selected cohort. This design allows cohort-specific initialization for healthy, impaired, or mixed-population simulations while maintaining the privacy of individual ADNI participants.

3.2.2. Action Space and Constraints

At each time step, the agent selects a subset of multi-binary actions that represent the combinations of medication classes. To maintain medical validity, the environment enforces two constraints:

1. If the agent selects "No Medication," it must not select any other treatments.
2. At least one valid action must be taken.

Violating these constraints results in early termination and a -10 reward, encouraging the agents to make explicit and interpretable decisions.

3.2.3. Transition Dynamics and Validity

At each step, ALPACA advances the environment by aggregating a patient's prior visits and autoregressively forecasting the next visit state. During inference, we fixed the transition interval to six months rather than using irregular visit spacing as was done in training. This simplifies downstream policy evaluation and avoids the need to dynamically compute the next visit time, which could introduce unnecessary bias. After prediction, the states are inverse-transformed using the original feature scaler and validated against the training distribution. Each feature must lie within three standard deviations of the training mean before the state is returned to the agent. If any feature violates this bound, the episode is terminated, and a neutral reward of zero is assigned. This termination rule is intended to enforce simulation validity by preventing policies from receiving reward from physiologically implausible CAST predictions, rather than to approximate clinical adverse-event modeling. We do not introduce adverse-event penalties because ALPACA is designed as a data-driven hypothesis-generation environment. Adding such penalties risks injecting external assumptions into the reward function and biasing policy exploration toward pre-specified clinical expectations.

3.2.4. Reward Function Formulation

The reward signal is centered on ADNI-Mem, a composite cognitive score derived from several clinically evaluated memory tests Crane et al. (2012). This metric was chosen for its strong correlation with neuropsychiatric symptom severity, capturing fundamental patient states without being explicitly included in the reward function De Vito et al. (2025).

The reward at each step was computed as the change in the ADNI-Mem between consecutive states, weighted by the standard error of the difference (SE). Given the normalized nature of the ADNI-Mem, we set the standard deviation to one. Because a precise test-retest reliability coefficient for the ADNI-Mem composite score has not been reported, we set $r_{xx} = 0.91$ as a high-reliability design parameter for reliable-change normalization. To verify this parameter decision we evaluate the sensitivity of policy learning to this choice through reward ablations over multiple reliability values in Appendix E.

The reward r_{t+1} can be formalized as:

$$r_{t+1} \;=\; \mathrm{clip}\!\left(10 \cdot \frac{M_{t+1} - M_t}{M_{\mathrm{diff}}},\; -10,\; 10\right) \qquad (1)$$

where:

- M_t is the ADNI-Mem score at timestep t,
- M_{t+1} is the ADNI-Mem value predicted for timestep $t+1$,
- $M_{\mathrm{diff}} = \sqrt{2(1 - r_{xx})} \cdot \sigma$ is the standard error of difference, with $\sigma = 1$ for z-scaled ADNI-Mem and r_{xx} the test–retest reliability coefficient,
- $\mathrm{clip}(x,\; a,\; b)$ bounds the value of x within the range $[a,\; b]$.

3.2.5. Episode Semantics

Each episode spans up to 22 time steps, representing 11 years of disease progression at six-month intervals. Patient age was incremented accordingly, and each treatment index corresponded to a specific therapeutic class (e.g., statins, antihypertensive agents, and metformin). These design choices ensure that ALPACA produces physiologically valid trajectories and enables an interpretable analysis of the learned policies.

4. Benchmark Policy Training

In this section, we propose four RL-based treatment benchmark policies (Proximal Policy Optimization (PPO), Advantage Actor–Critic (A2C), Soft Actor–Critic (SAC), and Branching Dueling Q-Network (BDQ)). We then compared the learned policies with a behavior-cloned clinician policy, optimal treatment heuristic policy, and no medication policy.

4.1. Policies

4.1.1. RL-based Treatment Policies

To evaluate ALPACA across diverse reinforcement learning paradigms, we benchmarked four agents: PPO, A2C, SAC, and BDQ. We used Stable Baselines3 Raffin et al. (2021) for PPO, A2C, and SAC and adapted the official BDQ implementation Tavakoli et al. (2017). Together, these methods span on-policy, off-policy, and value-based learning, providing a broad assessment of agent behavior in medical simulation settings. All models were trained for 500,000 time steps across four parallel environments, with input normalization applied to standardize the heterogeneous clinical feature scales.

PPO and A2C used standard two-layer multilayer perceptron (MLP) policies with 256 hidden units per layer. Because the SAC model is formulated for continuous action spaces, we implemented a wrapper that maps discrete medication actions to a continuous box space via dynamic thresholding. This preserves the action interpretability while enabling stable optimization. We additionally stabilized SAC using delayed learning starts and gradient clipping. For BDQ, we used a shared trunk with two 256-unit layers, followed by separate 128-unit branches for each action dimension and a 128-unit value head.

Our analysis focused on patient trajectories corresponding to MCI and AD. Although ALPACA also supports cognitively unimpaired profiles, we prioritized symptomatic patients to better align the evaluation with the clinical objective of treatment optimization.

4.1.2. Clinician Policy

To provide a real-world baseline for comparison with reinforcement learning agents, we trained a clinician policy Bayesian feedforward neural network (BFNN) using a behavioral cloning approach. This model approximates clinician decision-making by mapping patient states directly to treatment vectors observed in the ADNI dataset. The same preprocessing pipeline and subject splits were used as ALPACA for comparability, with the key distinction that the clinician model was trained to predict multi-binary treatment assignments at each visit, rather than forecasting future states. The input features included 12 continuous and nine categorical patient characteristics, and the output included 17 binary medication indicators that represented different therapeutic classes. During training, we applied a class-balanced binary cross-entropy (BCE) loss with per-action positive weights $(\text{neg}/\text{pos})^{0.55}$ to mitigate class imbalance, while preventing overly aggressive reweighting of rare medication actions. The Bayesian nature of the clinician model architecture was chosen because of its relative simplicity and ability to quantify predictive uncertainty in the evaluation metrics. Because this model serves as a baseline, incorporating uncertainty estimates provides additional context for the traditional performance metrics.

The behavior cloning performance was evaluated using a held-out test set. To assess robustness across multiple draws of patient states, we used Monte Carlo sampling and quantified predictive fidelity using exact-match accuracy and Hamming loss, as well as macro- and micro-averaged F1 scores to account for class imbalance. We also assessed probabilistic calibration using the Brier score and calibration error metrics, including the adaptive calibration error (ACE) and expected calibration error (ECE). All metrics were implemented with open-source Scikit-Learn utilities Pedregosa et al. (2011), except ACE and ECE, which followed the implementation described in Küppers et al. (2020)

4.1.3. Clinician Heuristics Policy

In addition to the behavior-cloned clinician policy, we implemented a heuristic clinician baseline intended to approximate consensus guideline-based treatment decisions for Alzheimer's disease. This policy follows a simple rule. At each step, it selects the "No Medication" action unless the patient's ADNI-Mem score falls below -0.1, in which case it selects "AD Treatment". We chose the -0.1 threshold based on the criterion described by Crane et al. for delineating MCI and AD onset using the ADNI-Mem composite score Crane et al. (2012). Clinical guidelines recommend initiating AD-specific therapies at the onset of MCI or mild AD Grossberg et al. (2019). Therefore, this heuristic maps our discrete action space onto an

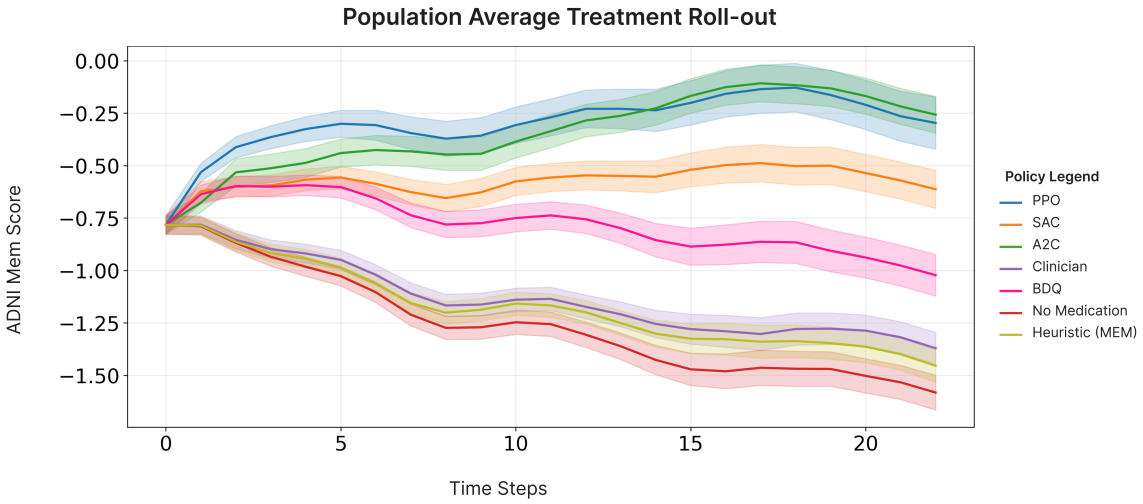


Figure 2: Policy performance over 1000 simulated rollouts (mean and 95% CI across patients).

idealized treatment strategy grounded in established clinical reasoning while reducing variability and noise relative to behavior-cloned policies trained on observational clinician actions.

4.2. Results

4.2.1. Behavior Cloning

We evaluated the clinician policy on the held-out test set using 50 Monte Carlo simulations. The model achieved an exact match across all 17 medications in 19.9% of cases, but individual medication predictions matched clinician choices in 91% of cases (Hamming loss = 0.090). This indicates that while the exact treatment combinations were difficult to reproduce, common prescription patterns were captured reliably.

Class imbalance in the dataset posed a significant challenge. The macro F1 score was low (0.104), reflecting the under-prediction of rare medications, whereas the micro F1 score was higher (0.487), indicating reasonable accuracy at the individual medication level despite skewed action frequencies.

In addition to the point prediction performance, the Bayesian modeling approach enables the assessment of the calibration and uncertainty. The model produced well-calibrated probabilities (Brier = 0.085, ECE = 0.107, and ACE = 0.144) and a strong predictive log-likelihood (−0.3). Variance decomposition showed that nearly all uncertainty was aleatoric (0.133) rather than epistemic (0.003), reflecting the variability among patients rather than model limitations.

Overall, the clinician policy effectively reproduced the dominant prescribing strategies observed in the ADNI while remaining conservative in the use of rare medications. Although it does not perfectly match all clinical actions, its fidelity and calibration make it a useful baseline for comparison with reinforcement learning agents.

4.2.2. Policy Evaluation Method

We evaluated all policies on identical cohorts by generating 1000 initial patient states from ALPACA's start-state model and rolling out each policy from the same starting point. All policies used a normalizing wrapper in StableBaselines3 to normalize the incoming state values. We used the default environment settings and a fixed seed (42) for reproducibility.

4.2.3. Policy Evaluation Results

Across matched simulated rollouts, the RL policies significantly outperformed both the no-treatment baseline and the behavior-cloned clinician policy (all $p < 0.001$), while the clinician policy also outperformed the no-treatment policy ($p < 0.001$). Statistical significance was assessed using a non-parametric paired Wilcoxon signed-rank test across matched initial states. PPO achieved the strongest average performance (Table 2, Figure 2).

Table 2: Performance of policies evaluated over 1000 simulated patient rollouts. Reported values are means calculated across rollouts.

Policy	Cumulative Reward	Final ADNI-Mem	ADNI-Mem Δ	Episode Length
Learned policies				
PPO	**3.38**	**−0.46**	**0.32**	12.62
A2C	−0.66	−0.59	0.19	9.93
SAC	0.18	−0.76	0.02	16.79
BDQ	−7.25	−0.98	−0.21	15.34
Baseline policies				
Clinician	−13.72	−1.36	−0.58	18.85
Heuristic	−15.10	−1.42	−0.64	19.09
No Medication	−17.81	−1.53	−0.76	18.85

Although trained policies improved group-level outcomes relative to clinician baselines, these effects were not uniform across individuals. Some impaired patients showed limited or no improvement relative to clinician policies, highlighting patient-level heterogeneity in simulated treatment response (Appendix Section G). To further characterize these results, we provide additional analyses of trajectory truncation and termination behavior in Appendix Section D. We also report model-free offline reinforcement learning benchmarks trained directly on ADNI trajectories in Appendix Section C.

4.2.4. SHAP Policy Evaluation

To evaluate whether the learned policies relied on clinically meaningful signals rather than environment-specific artifacts, we applied SHAP analysis to the policies trained in the ALPACA environment Lundberg and Lee (2017). The SHAP values quantify how individual patient features shift the probability of selecting an action. We focus on the *No Medication* and *AD Treatment* actions because they offer the least obscured insights into a policy's interpretation of disease severity and treatment needs. The results reported here used the aggregated SHAP values across all four trained policies (Figure 3).

No Medication The attributions for the No Medication action were consistent with clinical expectations. The probability of withholding treatment increased with higher memory scores and relatively preserved whole-brain volume. A higher hippocampal volume similarly increased the likelihood of taking no medical action, consistent with a lower disease burden Mortimer (1997); Stern et al. (2019). In contrast, elevated amyloid-β reduced the likelihood of selecting No Medication. Although the effect magnitudes varied by architecture, the aggregated SHAP patterns converged on these relationships, suggesting that the policies learned a reasonable criterion for when treatment was unnecessary.

AD Treatment The attributions for AD Treatment action were also biologically plausible. Higher cognitive performance (memory and executive function) reduced the likelihood of treatment, mirroring the No Medication findings and reflecting a reduced clinical need. In contrast, increased ventricular volume, a structural marker associated with cortical atrophy, increased the probability of selecting AD Treatment. Sex also contributed modestly, with female patients showing a higher tendency toward treatment selection, consistent with the higher prevalence and burden of AD reported in women than in men Li et al. (2022).

Tau and the treatment window Higher tau decreases attribution toward "AD Treatment" and increases attribution toward "No Medication" (Figure 3). Although this pattern may appear counterintuitive, elevated tau levels are typically associated with later-stage disease and reduced treatment responsiveness. The resulting action attributions suggest that the policy has learned an implicit treatment window, favoring intervention when pathology is more likely to remain modifiable (lower tau) and down-weighting treatment when tau is high and progression is less amenable to treatment. This is consistent with evidence that individuals with high tau levels derive reduced benefits from continued treat-

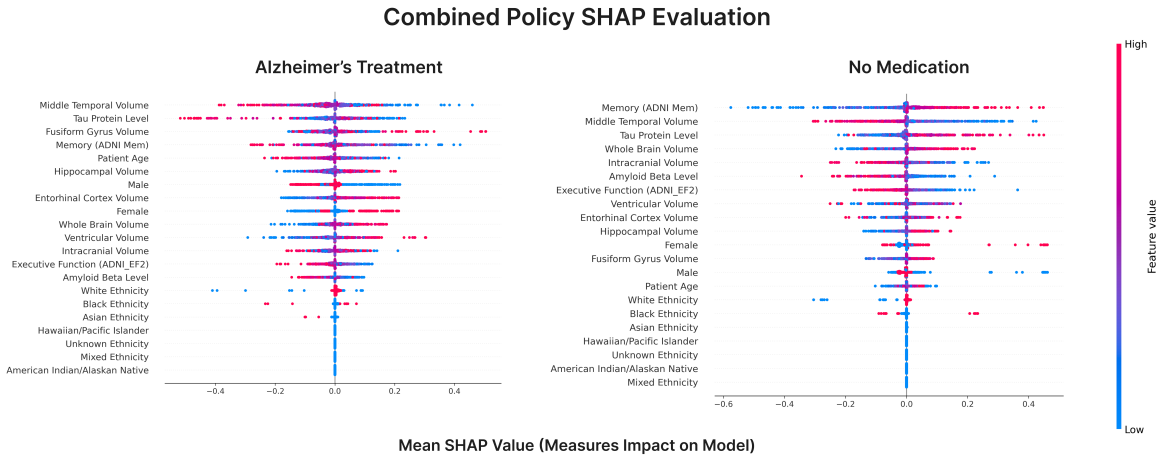


Figure 3: Aggregated SHAP values from all models (PPO, SAC, A2C, and BDQ) for the No Medication and AD Treatment actions. The red values indicate higher feature values, and higher SHAP scores indicate a greater propensity to act.

ment relative to those with lower tau levels Sims et al. (2023a).

Overall, the aggregated SHAP analysis suggests that the learned policies consistently rely on neurobiologically relevant features across architectures, reflecting established markers of Alzheimer's disease severity and progression. These patterns provide evidence that ALPACA supports clinically coherent policy learning.

5. Discussion

ALPACA is intended to support controlled and reproducible exploration of sequential treatment strategies in Alzheimer's disease, where clinical heterogeneity and long disease horizons make exhaustive empirical evaluation difficult. In this work, the CAST dynamics model generated medication-conditioned trajectories that preserved distributional and relational structure relative to held-out ADNI trajectories, while policy interpretability analyses suggested that learned treatment behaviors were shaped by clinically coherent markers of disease severity and progression. These results do not establish causal treatment effects, but they support the use of ALPACA as a stable in silico environment for studying personalized treatment strategies in controlled settings.

This framing is particularly relevant for Alzheimer's disease because medication effects may vary across biological subtypes, disease stages, comorbidity profiles, and treatment timing. Conventional trials often estimate average treatment effects across heterogeneous populations, which may obscure subgroup-specific responses and contribute to inconsistent findings across medication repurposing studies. ALPACA provides a complementary setting for exploring these possibilities at scale by simulating alternative medication regimens across diverse synthetic patient profiles. Rather than replacing clinical validation, we view these simulations as being helpful in identifying where treatment hypotheses appear robust, where they fail, and which patient subgroups may warrant more targeted downstream investigation.

ALPACA also provides a platform for studying interpretability in healthcare reinforcement learning. Post-hoc methods such as SHAP and LIME can provide useful feature-level summaries, but their utility is limited when clinical decisions depend on longitudinal trajectories. Because ALPACA exposes the full state, action, and rollout history of each simulated patient, it enables the development of interpretability methods that go beyond single-step feature attribution. The development of such methods could help disambiguate policies that rely on clinically meaningful longitudinal patterns from those that exploit artifacts of the learned simulator.

Finally, ALPACA may be useful for early-stage clinical trial planning and medication repurposing research. By organizing actions according to therapeu-

tic class and allowing policies to be evaluated across synthetic patient cohorts, the environment can help researchers probe various nuances related to the hypothesized treatment regime before committing to costly empirical studies. Its primary role is therefore not to recommend deployable treatment policies, but to provide a reproducible and low-risk testbed for refining hypotheses, stress-testing modeling assumptions, and improving the precision and interpretability of future neurodegenerative disease research.

6. Conclusion

We introduce ALPACA, an open-source, Gym-compatible reinforcement learning environment for Alzheimer's disease treatment research, built on a medication-conditioned forecasting model trained on longitudinal trajectories from ADNI. Policies trained in ALPACA improved memory-related outcomes relative to clinician-derived baselines while remaining consistent with biologically and clinically plausible treatment patterns. These results position ALPACA as an initial step toward clinically grounded simulation environments for Alzheimer's disease.

Despite the results, several limitations remain. ALPACA currently relies on discretized medication actions, coarse temporal resolution, and limited granularity in patient-state representations. In addition, although ADNI-Mem provides a more clinically meaningful and sensitive objective within ADNI, its limited adoption across Alzheimer's disease datasets compared to metrics like MMSE or MoCA makes comparison with future environments, datasets or models derived from other sources more difficult.

These limitations motivate several directions for future work. First, extending reinforcement learning methods such as ALPACA to leverage continuous action spaces and richer multimodal patient-state representations such as MRI and genomics would greatly improve downstream policy realism and clinical relevance. Second, important future work remains in evaluating how reward signal and reward function design shapes policy learning in Alzheimer's disease treatment. These contributions would allow for continued advancements in the safety and usability of in-silico treatment optimization.

Acknowledgments

We are very grateful to Yu-Yun Tseng for assistance with figures, and to both Yu-Chee Tseng and Yu-Yun Tseng for their valuable feedback on the manuscript.

The authors acknowledge the use of artificial intelligence (AI) tools for assistance with manuscript editing, code generation, and brainstorming during this research. All resulting text, code, figures, analyses, and claims were carefully reviewed and verified by the authors, who take full responsibility for the final paper.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association, Alzheimer's Drug Discovery Foundation, Araclon Biotech, and BioClinica Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

References

Inkyung Ahn and Jooyoung Park. Drug scheduling of cancer chemotherapy based on natural actor-critic approach. *Bio Systems*, 106 2-3:121–9, 2011. URL https://api.semanticscholar.org/CorpusID:20937389.

Farideh Amirrad, Emira Bousoik, Kiumars Shamloo, Hassan Al-Shiyab, Viet-Huong V. Nguyen, and Hamidreza Montazeri Aliabadi. Alzheimer's disease: Dawn of a new era? *Journal of Pharmacy & Pharmaceutical Sciences*, 20:184, July 2017. ISSN 1482-1826, 1482-1826. doi: 10.18433/J3VS8P.

Kritib Bhattarai, Sivaraman Rajaganapathy, Trisha Das, Yejin Kim, Yongbin Chen, Alzheimer's Disease Neuroimaging Initiative, Australian Imaging Biomarkers, Lifestyle Flagship Study of Ageing, Qiying Dai, Xiaoyang Li, Xiaoqian Jiang, and Nansu Zong. Using artificial intelligence to learn optimal regimen plan for alzheimer's disease. *Journal of the American Medical Informatics Association: JAMIA*, 30(10):1645–1656, 2023. ISSN 1527-974X. doi: 10.1093/jamia/ocad135.

David M. Bossens and Philip S. Thomas. Low variance off-policy evaluation with state-based importance sampling. *2024 IEEE Conference on Artificial Intelligence (CAI)*, pages 871–883, 2022. URL https://api.semanticscholar.org/CorpusID:254408540.

Paul K. Crane, Adam Carle, Laura E. Gibbons, Philip Insel, R. Scott Mackin, Alden Gross, Richard N. Jones, Shubhabrata Mukherjee, S. McKay Curtis, Danielle Harvey, Michael Weiner, Dan Mungas, and for the Alzheimer's Disease Neuroimaging Initiative. Development and assessment of a composite score for memory in the alzheimer's disease neuroimaging initiative (adni). *Brain Imaging and Behavior*, 6(4):502–516, December 2012. ISSN 1931-7565. doi: 10.1007/s11682-012-9186-z.

Alyssa N. De Vito, Zachary J. Kunicki, Hannah E. Joyce, Edward D. Huey, and Richard N. Jones. Parallel changes in cognition, neuropsychiatric symptoms, and amyloid in cognitively unimpaired older adults and those with mild cognitive impairment. *Alzheimer's & Dementia*, 21(2):e14568, February 2025. ISSN 1552-5260. doi: 10.1002/alz.14568.

Scott Fujimoto and Shixiang Shane Gu. A minimalist approach to offline reinforcement learning. *ArXiv*, abs/2106.06860, 2021. URL https://api.semanticscholar.org/CorpusID:235422620.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. In *International Conference on Machine Learning*, 2018. URL https://api.semanticscholar.org/CorpusID:54457299.

Ali Ghaffari, B. Bahmaie, and Mostafa Nazari. A mixed radiotherapy and chemotherapy model for treatment of cancer with metastasis. *Mathematical Methods in the Applied Sciences*, 39:4603 – 4617, 2016. URL https://api.semanticscholar.org/CorpusID:124440566.

George T. Grossberg, Gary Tong, Anna D. Burke, and Pierre N. Tariot. Present algorithms and future treatments for alzheimer's disease. *Journal of Alzheimer's Disease*, 67:1157 – 1171, 2019. URL https://api.semanticscholar.org/CorpusID:73419065.

Yong Huang, Rui Cao, Thomas Hughes, and Amir Rahmani. Smart pain relief: Harnessing conservative q learning for personalized and dynamic pain management. *Smart Health*, 2024. URL https://api.semanticscholar.org/CorpusID:273172715.

Pushkala Jayaraman, Jacob Desman, Moein Sabounchi, Girish N. Nadkarni, and Ankit Sakhuja. A primer on reinforcement learning in medicine for clinicians. *NPJ Digital Medicine*, 7, 2024. URL https://api.semanticscholar.org/CorpusID:274302999.

Ilya Kostrikov, Ashvin Nair, and Sergey Levine. Offline reinforcement learning with implicit q-learning. *ArXiv*, abs/2110.06169, 2021. URL https://api.semanticscholar.org/CorpusID:238634325.

Aviral Kumar, Aurick Zhou, G. Tucker, and Sergey Levine. Conservative q-learning for offline reinforcement learning. *ArXiv*, abs/2006.04779, 2020. URL https://api.semanticscholar.org/CorpusID:219530894.

Fabian Küppers, Jan Kronenberger, Amirhossein Shantia, and Anselm Haselhoff. Multivariate confidence calibration for object detection. In *The*

IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2020.

Sergey Levine, Aviral Kumar, G. Tucker, and Justin Fu. Offline reinforcement learning: Tutorial, review, and perspectives on open problems. *ArXiv*, abs/2005.01643, 2020. URL https://api.semanticscholar.org/CorpusID:218486979.

Xue Li, Xiaojin Feng, Xiaodong Sun, Ningning Hou, Fang Han, and Yongping Liu. Global, regional, and national burden of alzheimer's disease and other dementias, 1990–2019. *Frontiers in Aging Neuroscience*, 14, October 2022. ISSN 1663-4365. doi: 10.3389/fnagi.2022.937486. URL https://www.frontiersin.orghttps://www.frontiersin.org/journals/aging-neuroscience/articles/10.3389/fnagi.2022.937486/full.

Yooseok Lim and Sujee Lee. Omg-rl:offline model-based guided reward learning for heparin treatment. *ArXiv*, abs/2409.13299, 2024. URL https://api.semanticscholar.org/CorpusID:272770561.

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.

Zhiyao Luo, Mingcheng Zhu, Fenglin Liu, Jiali Li, Yangchen Pan, Jiandong Zhou, and Tingting Zhu. Dtr-bench: An in silico environment and benchmark platform for reinforcement learning based dynamic treatment regime. *ArXiv*, abs/2405.18610, 2024. URL https://api.semanticscholar.org/CorpusID:270094986.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc D. Breton, Boris P. Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator. *Journal of Diabetes Science and Technology*, 8:26 – 34, 2014. URL https://api.semanticscholar.org/CorpusID:33955269.

Nathan Mantel. The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27(2_Part_1):209–220, February 1967. ISSN 0008-5472.

James A. Mortimer. Brain reserve and the clinical expression of alzheimer's disease. *Geriatrics*, 52 Suppl 2:S50–3, 1997. URL https://api.semanticscholar.org/CorpusID:9810796.

Michael Oberst and David A. Sontag. Counterfactual off-policy evaluation with gumbel-max structural causal models. In *International Conference on Machine Learning*, 2019. URL https://api.semanticscholar.org/CorpusID:155092754.

Sambit Panda, Satish Palaniappan, Junhao Xiong, Eric W. Bridgeford, Ronak D. Mehta, Cencheng Shen, and Joshua T. Vogelstein. hyppo: A multivariate hypothesis testing python package. 2019. URL https://api.semanticscholar.org/CorpusID:195798646.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, Adam Lerer, James McDonnell, Zhichao Jia, Fan Zhu, Michael Liu, Xiaowei Deng, Arvind Mangalam, Bhargav Singh, Ye Fang, Honghao Lu, Tero Sourek, and Viktor Kang. PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32:8024–8035, 2019.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Robert Perneczky and Lutz Froelich. Clinically meaningful benefit and real-world evidence in alzheimer's disease research and care. *Alzheimer's & Dementia : Translational Research & Clinical Interventions*, 11, 2025. URL https://api.semanticscholar.org/CorpusID:278094428.

Antonin Raffin, Ashley Hill, Adam Gleave, Anssi Kanervisto, Maximilian Ernestus, and Noah Dormann. Stable-baselines3: Reliable reinforcement learning implementations. *Journal of machine learning research*, 22(268):1–8, 2021.

Krishnakant V. Saboo, A. Choudhary, Yurui Cao, G. Worrell, David T. Jones, and R. Iyer. Reinforcement learning based disease progression model for alzheimer's disease. *ArXiv*, June 2021. URL https://www.semanticscholar.org/paper/Reinforcement-Learning-based-Disease-Progression-Saboo-Choudhary/fe2c1b25d4e4bb16612a5381c55f51602943180a.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *CoRR*, abs/1701.06538, 2017. URL http://arxiv.org/abs/1701.06538.

John R. Sims, Jennifer A. Zimmer, Cynthia D. Evans, Ming Lu, Paul Ardayfio, JonDavid Sparks, Alette M. Wessels, Sergey Shcherbinin, Hong Wang, Emel Serap Monkul Nery, Emily C. Collins, Paul Solomon, Stephen Salloway, Liana G. Apostolova, Oskar Hansson, Craig Ritchie, Dawn A. Brooks, Mark Mintun, Daniel M. Skovronsky, and TRAILBLAZER-ALZ 2 Investigators. Donanemab in early symptomatic alzheimer disease: The trailblazer-alz 2 randomized clinical trial. *JAMA*, 330(6):512–527, August 2023a. ISSN 0098-7484. doi: 10.1001/jama.2023.13239.

John R. Sims, Jennifer A. Zimmer, Cynthia Duggan Evans, Ming ning Lu, Paul A. Ardayfio, Jondavid Sparks, Alette M. Wessels, Sergey Shcherbinin, Hong Wang, Emel Serap Monkul Nery, Emily C. Collins, Paul R. Solomon, Stephen Salloway, Liana G. Apostolova, Oskar Hansson, Craig W. Ritchie, Dawn A. Brooks, Mark Mintun, and Daniel M. Skovronsky. Donanemab in early symptomatic alzheimer disease: The trailblazer-alz 2 randomized clinical trial. *JAMA*, 2023b. URL https://api.semanticscholar.org/CorpusID:259946737.

Yaakov Stern, Carol A. Barnes, Cheryl L. Grady, Richard N. Jones, and Naftali Raz. Brain reserve, cognitive reserve, compensation, and maintenance: operationalization, validity, and mechanisms of cognitive resilience. *Neurobiology of Aging*, 83:124–129, 2019. URL https://api.semanticscholar.org/CorpusID:207973821.

Zhaohong Sun, Wei Dong, Haomin Li, and Zhengxing Huang. Adversarial reinforcement learning for dynamic treatment regimes. *Journal of biomedical informatics*, page 104244, 2022. URL https://api.semanticscholar.org/CorpusID:253652461.

Denis Tarasov, Alexander Nikulin, Dmitry Akimov, Vladislav Kurenkov, and Sergey Kolesnikov. CORL: Research-oriented deep offline reinforcement learning library. In *3rd Offline RL Workshop: Offline RL as a "Launchpad"*, 2022. URL https://openreview.net/forum?id=SyAS49bBcv.

Rawan Tarawneh and V. Shane Pankratz. The search for clarity regarding "clinically meaningful outcomes" in alzheimer disease clinical trials: Clarityad and beyond. *Alzheimer's Research & Therapy*, 16, 2024. URL https://api.semanticscholar.org/CorpusID:267701073.

Arash Tavakoli, Fabio Pardo, and Petar Kormushev. Action branching architectures for deep reinforcement learning. *ArXiv*, abs/1711.08946, 2017. URL https://api.semanticscholar.org/CorpusID:962757.

Betty M. Tijms, Ellen M. Vromen, Olav Mjaavatten, Henne Holstege, Lianne M. Reus, Sven van der Lee, Kirsten E. J. Wesenhagen, Luigi Lorenzini, Lisa Vermunt, Vikram Venkatraghavan, Niccoló Tesi, Jori Tomassen, Anouk den Braber, Julie Goossens, Eugeen Vanmechelen, Frederik Barkhof, Yolande A. L. Pijnenburg, Wiesje M. van der Flier, Charlotte E. Teunissen, Frode S. Berven, and Pieter Jelle Visser. Cerebrospinal fluid proteomics in patients with alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nature Aging*, 4(1):33–47, January 2024. ISSN 2662-8465. doi: 10.1038/s43587-023-00550-7.

Masatoshi Uehara, Chengchun Shi, and Nathan Kallus. A review of off-policy evaluation in reinforcement learning. *ArXiv*, abs/2212.06355, 2022. URL https://api.semanticscholar.org/CorpusID:254591267.

Guangyu Wang, Xiaohong Liu, Zhen Ying, Guoxing Yang, Zhiwei Chen, Zhiwen Liu, Min Zhang, Hongmei Yan, Yuxing Lu, Yuanxu Gao, Kanmin Xue, Xiaoying Li, and Ying Chen. Optimized glycemic control of type 2 diabetes with reinforcement learning: a proof-of-concept trial. *Nature Medicine*, 29:2633 – 2642, 2023. URL https://api.semanticscholar.org/CorpusID:261884154.

B. Winblad, A. Wimo, K. Engedal, H. Soininen, F. Verhey, G. Waldemar, A.-L. Wetterholm,

A. Haglund, R. Zhang, and R. Schindler. 3-year study of donepezil therapy in alzheimer's disease: Effects of early and continuous therapy. *Dementia and Geriatric Cognitive Disorders*, 21(5–6): 353–363, 2006. ISSN 1420-8008, 1421-9824. doi: 10.1159/000091790.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Y. Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *ArXiv*, abs/2005.13239, 2020. URL https://api.semanticscholar.org/CorpusID:218900501.

Appendix A. CAST Robustness to Action Perturbations

To assess whether CAST remains stable under deviations from observed clinician actions, we performed an action perturbation analysis during rollout. Specifically, each binary medication action was independently flipped with probability $p \in \{0.25, 0.5\}$. This allowed us to evaluate whether the model remains locally stable under perturbed action sequences rather than only under the ground-truth clinician actions used in the primary validation setting.

Across both autoregressive and teacher-forced evaluations, CAST remained stable under moderate perturbation. In the autoregressive setting, the mean Mantel correlation decreased only slightly from 0.6989 at baseline to 0.6924 under $p = 0.25$. In the teacher-forced setting, the mean Mantel correlation changed from 0.6819 at baseline to 0.6830 under $p = 0.25$. Mantel p-values remained highly significant throughout ($p = 1.99 \times 10^{-4}$ for all conditions). Short-horizon MMD increased only modestly under $p = 0.25$, with the associated p-values remaining non-significant or near-significant ($p = 0.078$ in the autoregressive setting and $p = 1.0$ in the teacher-forced setting).

Under stronger perturbation ($p = 0.5$), the performance degraded more noticeably. In the autoregressive setting, the mean Mantel correlation declined to 0.6814, short-horizon MMD increased to 1.14×10^{-2}, and the corresponding MMD p-value dropped below the significance threshold ($p = 0.001$). In the teacher-forced setting, the mean Mantel correlation decreased more modestly to 0.6769, while short-horizon MMD increased to 4.52×10^{-3} with an MMD p-value of 0.4625.

We interpret these findings as evidence of local robustness around the empirical action distribution rather than validation of arbitrary counterfactual action sequences. At the same time, the performance drop under stronger perturbation suggests that CAST is meaningfully sensitive to medication inputs rather than relying only on state history.

Table 3: Trajectory validation results under different probabilities of action perturbation for autoregressive and teacher-forced rollouts.

Rollout Type	Probability of Jitter	Real Mean Mantel R	Real Mean Mantel P	MMD Stat (Short)	MMD P-value (Short)	MMD Stat (Long)	MMD P-value (Long)
Autoregressive	Baseline ($p = 0$)	0.6989	1.99×10^{-4}	4.65×10^{-3}	0.3916	3.79×10^{-3}	1.0000
Autoregressive	$p = 0.25$	0.6924	1.99×10^{-4}	6.01×10^{-3}	0.0780	5.51×10^{-3}	1.0000
Autoregressive	$p = 0.5$	0.6814	1.99×10^{-4}	1.14×10^{-2}	0.0010	1.03×10^{-2}	0.8781
Teacher Forced	Baseline ($p = 0$)	0.6819	1.99×10^{-4}	2.02×10^{-4}	1.0000	2.69×10^{-3}	1.0000
Teacher Forced	$p = 0.25$	0.6830	1.99×10^{-4}	2.38×10^{-3}	1.0000	2.75×10^{-3}	1.0000
Teacher Forced	$p = 0.5$	0.6769	1.99×10^{-4}	4.52×10^{-3}	0.4625	3.87×10^{-3}	1.0000

Appendix B. Evaluation of Compounding Error in Trajectory Forecasting

To assess the impact of trajectory length on CAST performance, we calculated the normalized root mean squared error (RMSE) between predicted and ground-truth visits at each visit position and averaged the results across all test subjects. We evaluated CAST under both teacher-forced and autoregressive state prediction. In the teacher-forced setting, both ground-truth actions and ground-truth states were used to predict the next state. In the autoregressive setting, ground-truth actions were retained, but each next-state prediction was conditioned only on previously predicted states. Under autoregressive evaluation, CAST showed gradually increasing error, with RMSE rising from 0.22 at visit 1 to 0.58 at visit 19. Under teacher-forced evaluation, the model began with a similar RMSE of 0.22 at visit 1 and increased only modestly to 0.30 at visit 19. Visit 0 is the initial visit used to condition the patient in the CAST model and, as such, has an RMSE of 0 relative to baseline.

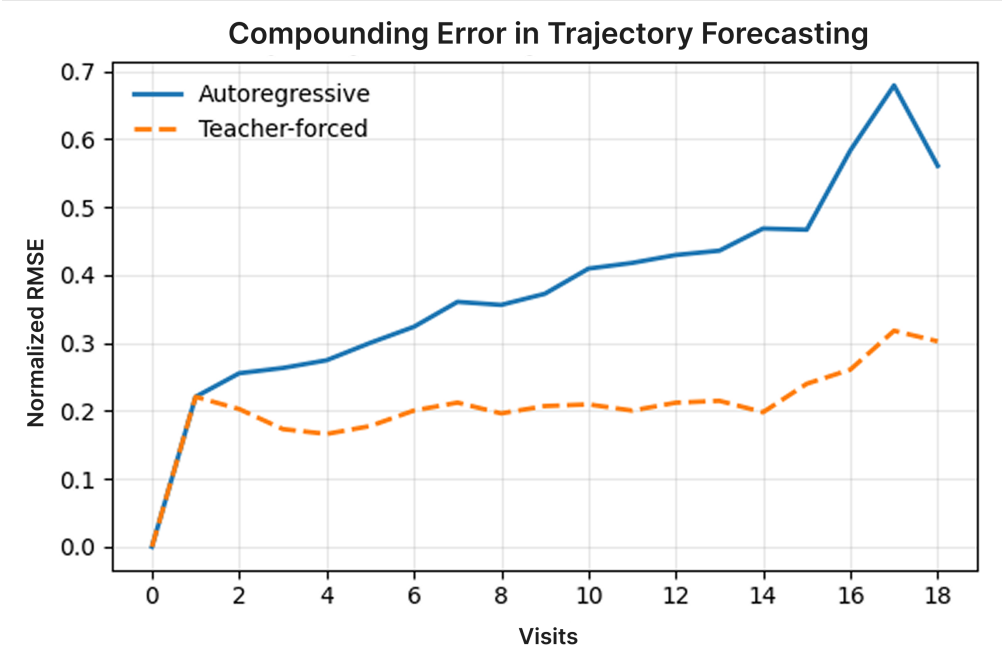


Figure 4: Comparison of the normalized root mean squared error accumulation between autoregressive and teacher-forced rollouts over unseen patient trajectories using the CAST model.

Appendix C. Model-free Offline Reinforcement Learning Baselines

To contextualize the policies trained directly within ALPACA, we additionally evaluated model-free offline reinforcement learning baselines trained from observed ADNI trajectories. Specifically, we evaluated Conservative Q-Learning (CQL), Implicit Q-Learning (IQL), Behavior Cloning (BC), and Twin Delayed Deep Deterministic Policy Gradient with Behavior Cloning (TD3-BC) Kumar et al. (2020); Kostrikov et al. (2021); Fujimoto and Gu (2021). These baselines provide a complementary comparison to the simulator-trained policies because they learn directly from retrospective trajectories rather than through closed-loop interaction with the ALPACA environment.

C.0.1. Policy Adaptations

The ALPACA action space is naturally multi-binary, with each action representing a combination of active medication classes. Because the CORL baselines assume continuous actions Tarasov et al. (2022), we adapted them to the finite action structure of ALPACA by representing observed medication combinations as categorical actions. Logged multi-binary medication vectors were mapped to discrete action IDs, and policies were restricted to medication combinations observed more than three times in the training set.

Under this formulation, policy networks output logits over supported action IDs, and critic networks estimate values over the same finite action catalog. This keeps both policy learning and fitted Q-evaluation (FQE) closer to the empirical action support of the ADNI trajectories.

Conservative Q-Learning For CQL, continuous action sampling was replaced with exact evaluation over the supported categorical action catalog. This allowed the conservative penalty and Bellman updates to be computed over medication combinations observed in the training data.

Implicit Q-Learning For IQL, the actor update was reformulated as categorical advantage-weighted behavior cloning. Specifically, the continuous-action behavior cloning objective was replaced with an advantage-weighted negative log-likelihood over logged action IDs.

TD3-BC For TD3-BC, actor outputs were reformulated as logits over the supported action catalog. Target actions were selected using the target actor's argmax action, and the mean-squared-error behavior cloning term was replaced with cross-entropy over logged action IDs. Because these changes alter the continuous-control structure of TD3-BC, we interpret this baseline as a categorical adaptation rather than a native implementation of the original algorithm.

Behavior Cloning For behavior cloning, the regression objective was replaced with cross-entropy classification over logged action IDs. As a result, the learned policy emits only medication combinations contained in the supported action catalog.

C.0.2. Policy Training and Evaluation

The offline RL policies were trained using the same subject-level train, validation, and test splits used for CAST. Unlike policies trained through interaction with ALPACA, these offline baselines learned only from observed ADNI trajectories and were restricted to the supported action catalog described above. This design was chosen to align learned policies with the empirical action support of the dataset and to improve the credibility of downstream off-policy evaluation.

Early stopping during policy training was performed using FQE. For each candidate policy checkpoint, an FQE model was fit on the training set, and policy value was estimated on the validation set. The best validation FQE value was then used to select the final checkpoint for each offline RL policy.

Final policy evaluation was also performed using FQE under the constrained ALPACA action catalog. The FQE evaluator was trained by Bellman regression under the target policy, and policy values were estimated from the fitted Q-function. In Table 4 we report two complementary FQE evaluations on the test set. First, in the held-out evaluation, the FQE evaluator was fit using the combined training and validation sets and then used to estimate policy performance on the held-out test set. This variant maximizes the amount of

data available for fitting the evaluator while preserving an independent test set for final evaluation. Second, we performed a five-fold cross-fitted sensitivity analysis, in which the FQE evaluator was trained on four folds and evaluated on the remaining held-out fold. This analysis provides a robustness check in which FQE is evaluated on data disjoint from the data used to fit the evaluator.

In addition to estimating scalar policy value, we fit two auxiliary FQE models to estimate downstream changes in ADNI-Mem and ADNI-EF2. Policy value was used for early stopping and primary model selection, while the ADNI-Mem and ADNI-EF2 estimates were used to provide additional clinical context for the learned policies.

C.0.3. Relationship to ALPACA Rollout Evaluation

The model-free offline RL experiments provide a complementary benchmark for ALPACA, rather than a duplicate evaluation of the simulator-trained policies. Because the two policy classes are trained under different regimes, they require different evaluation procedures. Policies trained within ALPACA are optimized through closed-loop interaction with the learned CAST transition model, so their performance is assessed through simulator rollouts. By contrast, the offline RL baselines are trained directly from observed ADNI trajectories, so their performance is assessed using fitted Q-evaluation (FQE) over logged transitions. This separation preserves the intended estimand in each setting.

Despite this difference, the offline RL results are directionally consistent with the simulator-based findings. Adapted TD3-BC achieves the strongest policy value in both the held-out and cross-fitted evaluations, while CQL and IQL achieve the strongest cross-fitted ADNI-Mem and ADNI-EF2 estimates, respectively. These results provide a support-constrained comparison showing that policies trained directly from observed ADNI trajectories recover broadly similar trends while remaining limited to empirically supported treatment combinations.

Table 4: Offline reinforcement learning policy performance under standard held-out and sensitivity evaluation.

	Standard Evaluation			Sensitivity Evaluation		
Algorithm	Policy Value	ADNI-Mem Δ	ADNI-EF2 Δ	Policy Value	ADNI-Mem Δ	ADNI-EF2 Δ
Learned policies						
Adapted TD3-BC	**-2.809**	**-0.168**	-0.306	**-4.179**	-0.239	-0.263
IQL	-4.193	-0.245	**-0.304**	-5.036	-0.273	**-0.206**
CQL	-5.583	-0.279	-0.344	-5.876	**-0.217**	-0.251
Baseline policies						
Behavior Cloning	-5.711	-0.304	-0.429	-4.796	-0.270	-0.274
No Medication	-5.863	-0.291	-0.530	-5.901	-0.221	-0.434

Appendix D. Impact of Trajectory Truncation on Policy Results

To compare termination behavior with downstream reward outcomes, we report rollout termination frequencies and completed-trajectory outcomes in Table 5. Despite baseline policies completing a larger fraction of rollouts than policies trained in ALPACA, this difference was driven primarily by constraint-violation terminations, which occurred only for learned policies. In contrast, aggregate out-of-bounds termination rates were broadly similar between learned and baseline policies, suggesting that out-of-bounds termination is not uniquely induced by policy optimization.

One possible explanation for the similar out-of-bounds rates is the use of GMM-based synthetic patient initialization. Because initial states are sampled from the impaired-patient GMM, some synthetic patients may initialize close to the boundary of one or more state features, increasing the likelihood of later exceeding the three-standard-deviation validity threshold. This effect may further be amplified in the impaired cohort because the validity bounds are computed from the broader training distribution, which includes both cognitively impaired and healthy participants.

Importantly, when the analysis is restricted to completed rollouts, the qualitative policy ranking remains consistent with the main evaluation. PPO achieves the largest mean ADNI-Mem improvement, followed by A2C and SAC, while the baseline policies show substantially worse completed-rollout outcomes. This suggests that the learned policies' memory-related gains are not solely an artifact of early termination.

Table 5: Rollout termination frequencies and completed-trajectory outcomes by policy.

Policy	Completed		Out-of-bounds		Constraint violation		Completed rollout outcomes		
	Count	Rate	Count	Rate	Count	Rate	Mean Reward	Final Cog.	ADNI-Mem Δ
Learned policies									
PPO	382	38.2%	248	24.8%	370	37.0%	10.45	-0.34	0.47
A2C	327	32.7%	179	17.9%	494	49.4%	9.12	-0.33	0.40
SAC	539	53.9%	461	46.1%	0	0.0%	2.30	-0.61	0.11
BDQ	477	47.7%	304	30.4%	219	21.9%	-6.62	-1.01	-0.28
Baseline policies									
Clinician	665	66.5%	335	33.5%	0	0.0%	-15.61	-1.43	-0.66
Heuristic (MEM)	680	68.0%	320	32.0%	0	0.0%	-17.02	-1.48	-0.72
No Medication	668	66.8%	332	33.2%	0	0.0%	-19.69	-1.60	-0.83

Appendix E. Reward Function Ablation Evaluation

To assess the stability of the proposed reward function, we performed an ablation study evaluating the downstream performance of policies trained under alternative reward formulations. The ablations focused on reward clipping, the test–retest reliability coefficient used for reliable-change normalization, and reward scaling. Because these reward variants differ in numerical scale, all evaluations are reported using the mean change in ADNI-Mem between the first and final visits across 500 simulated episodes. These were chosen due to their centrality to the ALPACA environment and their scale invariance to reward function changes such as scaling.

For each ablation, we first retrained each policy using the corresponding ablated reward function, referred to here as the native reward function. We then evaluated the resulting policies in two settings: an environment using the same native reward function used during training, and an environment using the reward function proposed in this work, referred to here as the proposed reward function. This evaluation design allows us to distinguish whether policies trained under ablated rewards improve the underlying ADNI-Mem outcome, rather than merely exploiting differences in reward scaling or shaping. The proposed evaluation results are shown in Table 6, and the native evaluation results are shown in Table 7.

The ablation study indicates that the proposed reward function provides a stable learning signal for optimizing simulated memory-related outcomes. Policies trained with the full reward formulation generally achieved stronger mean ADNI-Mem improvements than policies trained with raw, weakly scaled, or non-normalized reward variants. These results support the mathematical structure of the proposed reward function as an effective in silico optimization target. However, they should not be interpreted as evidence that ADNI-Mem alone is sufficient to capture optimal clinical outcomes. Rather, the ablation demonstrates that the proposed reward formulation facilitates training policies that improve the intended memory-related outcome within the ALPACA simulation environment.

Table 6: Reward function ablation results evaluated in the proposed ALPACA environment.

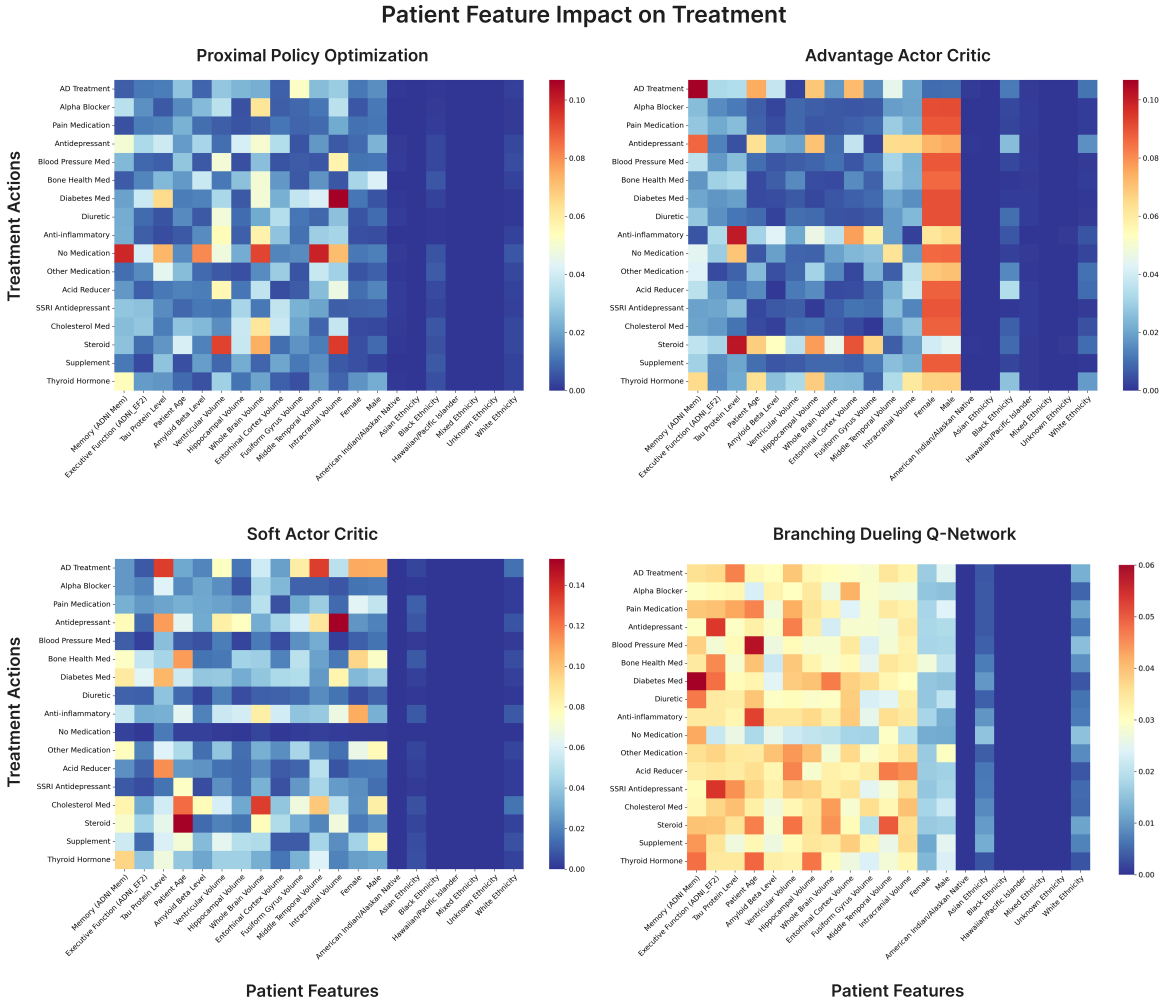
Training reward	**PPO**	**SAC**	**A2C**	**BDQ**	**Mean**
Proposed reward function	0.392	**0.083**	0.193	-0.262	**0.102**
No clipping	0.363	-0.400	0.136	-0.116	-0.004
No reliable-change normalization	0.295	-0.039	0.109	-0.380	-0.003
Raw reward	-0.291	-0.143	0.106	-0.608	-0.234
$r_{xx} = 0.50$	0.295	-0.039	0.109	-0.380	-0.003
$r_{xx} = 0.85$	0.379	0.005	0.156	-0.276	0.066
$r_{xx} = 0.99$	**0.413**	-0.149	**0.261**	**-0.159**	0.092
Scale = 1	-0.153	-0.055	0.053	-0.558	-0.178

Table 7: Reward function ablation results evaluated in each policy's native training environment.

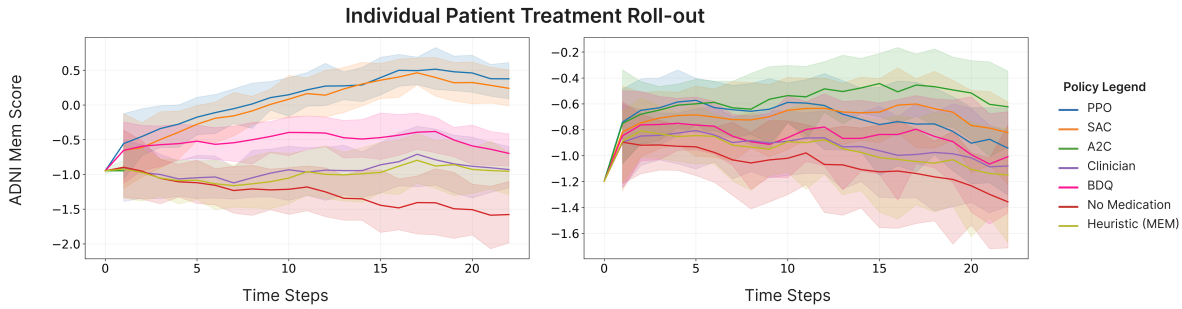
Training reward	**PPO**	**SAC**	**A2C**	**BDQ**	**Mean**
Proposed reward function	0.392	**0.079**	0.193	-0.262	**0.100**
No clipping	0.363	-0.409	0.136	**-0.116**	-0.006
No reliable-change normalization	0.295	-0.041	0.109	-0.380	-0.004
Raw reward	-0.291	-0.147	0.106	-0.608	-0.235
$r_{xx} = 0.50$	0.295	-0.041	0.109	-0.380	-0.004
$r_{xx} = 0.85$	0.379	0.026	0.156	-0.276	0.071
$r_{xx} = 0.99$	**0.413**	-0.136	**0.261**	-0.159	0.095
Scale = 1	-0.153	-0.045	0.053	-0.558	-0.176

Appendix F. Influence of Patient Feature on Treatment Action

These heatmaps were derived from the SHAP values calculated during inference as each policy treated patients within the ALPACA environment. The SHAP values measure the relative impact of each patient feature on treatment. Larger SHAP values indicate that a feature contributes more strongly to the policy's propensity to select the corresponding treatment action.



Appendix G. Individual Treatment Simulation



This figure shows two synthetic patients generated in the ALPACA environment, illustrating substantial variability in treatment trajectories among individuals with symptoms of cognitive impairment. Although policies perform well on average, these examples highlight patient-specific treatment outcome differences that may reflect underlying disease heterogeneity.

Appendix H. Patient Features and Actions

H.1. Patient Features

- **Cognitive Measures:**
  ADNI_MEM, ADNI_EF2
- **Biomarkers:**
  TAU_data, ABETA
- **Demographics:**
  subject_age, PTGENDER_Female, PTGENDER_Male
- **Race Indicators:**
  PTRACCAT_Am Indian/Alaskan, PTRACCAT_Asian, PTRACCAT_Black, PTRACCAT_Hawaiian/Other PI, PTRACCAT_More than one, PTRACCAT_Unknown, PTRACCAT_White
- **Structural MRI Volumes:**
  Ventricles, Hippocampus, WholeBrain, Entorhinal, Fusiform, MidTemp, ICV
- **Longitudinal Timing:**
  next_visit_months

H.2. Action Space

The medication action space consisted of 17 binary indicators denoting whether each medication class was active during a visit.

- AD Treatment_active
- Alpha Blocker_active
- Analgesic_active
- Antidepressant_active
- Antihypertensive_active
- Bone Health_active
- Diabetes Medication_active
- Diuretic_active
- NSAID_active
- No Medication_active
- Other_active
- PPI_active
- SSRI_active
- Statin_active
- Steroid_active
- Supplement_active
- Thyroid Hormone_active

Appendix I. Drug Class Mapping

The following mapping was used to consolidate individual medications into therapeutic classes for the action space.

Table 8: Drug–class mapping used in ALPACA.

Drug	**Mapped Class**
Aricept, Donepezil, Namenda, Exelon	AD Treatment
Lipitor, Simvastatin, Crestor, Zocor, Atorvastatin	Statin
Lisinopril, Atenolol, Amlodipine, Metoprolol, Norvasc, Losartan	Antihypertensive
Levothyroxine, Synthroid	Thyroid Hormone
Aspirin, Ibuprofen, Aleve, ASA	NSAID
Tylenol, Acetaminophen	Analgesic
Zoloft, Lexapro, Sertraline, Citalopram, Prozac	SSRI
Trazodone	Antidepressant
Metformin	Diabetes Medication
Vitamin D, Vitamin D3, Vitamin B12, Vitamin C, Vitamin E, Calcium, Multivitamin, Fish Oil	Supplement
Omeprazole, Prilosec	PPI
Hydrochlorothiazide	Diuretic
Fosamax	Bone Health
Prednisone, Prednisolone	Steroid
Flomax	Alpha Blocker
No medication	No Medication