Computational Approaches to Optimize HIV Treatment Strategies Using Synthetic Data

Marie Huynh
Dept. of Biomedical Data Science
Stanford University

Marc Nasser
Dept. of Electrical Engineering
Stanford University

***Abstract*—Human Immunodeficiency Virus (HIV) is a chronic viral infection that attacks the body's immune system, specifically targeting CD4 cells (T cells). The current clinical standard for managing HIV involves the lifelong use of antiretroviral therapy (ART) combinations that target various stages of the virus's life cycle. The vast number of possible drug combinations, the extensive treatment histories, the rapid mutations of the virus, and adherence differences make it difficult to manually identify the most effective therapy, leading to variability in patient outcomes. Effective HIV management requires clinicians to adjust ART regimens continually to address these challenges. Leveraging a synthetic dataset from Health Gym AI [2], we developed two approaches—Q-learning and a closed-form Markov Decision Process (MDP)—to optimize HIV treatment strategies. Drawing inspiration from approaches that combine kernel-based and model-based reinforcement learning for HIV therapy [4], we aim to maximize health outcomes by minimizing viral load and enhancing immune function (CD4 count) throughout treatment. Both approaches were evaluated for their ability to recommend effective treatment strategies under varying conditions. This work highlights the potential of reinforcement learning and MDP-based methods in supporting dynamic and individualized HIV therapy planning.**

I. Introduction

In 2023, 39.9 million people worldwide were living with HIV, highlighting the ongoing global health challenge posed by the virus despite advances in treatment and prevention strategies [5]. The current standard of care for HIV treatment is antiretroviral therapy (ART) which involves taking a combination of HIV medicines–HIV treatment regimen–every day. [3] Nevertheless, the management of HIV treatment presents several significant challenges. First, there is substantial variability in patient responses to antiretroviral therapy (ART), driven by factors such as age, genetic differences, and underlying physiological conditions. Additionally, HIV's high mutation rate enables rapid adaptation, producing new strains that may evade existing therapies. This evolutionary capacity of the virus introduces uncertainty in predicting long-term treatment outcomes, as previously effective regimens can lose efficacy due to emergent resistance. Another critical challenge is patient adherence; irregular medication intake or therapy discontinuation can lead to viral rebound and accelerate the development of drug resistance. Adherence patterns vary widely among individuals, complicating efforts to model and incorporate adherence behavior into treatment optimization accurately. Moreover, ART typically involves drug combinations targeting different aspects of the viral life cycle, and the interactions between these drugs can result in highly variable effects. For patients with extensive treatment histories, the cumulative impact of prior regimens adds yet another layer of complexity, making it difficult to predict the efficacy of new treatment plans. Finally, privacy concerns and a lack of consent for sharing health information with public health agencies pose barriers to comprehensive data collection and integration [1].

Selecting the most effective HIV therapy can thus be framed as a sequential decision-making process with long-term consequences. To address these challenges, recent research has explored computational approaches to optimize treatment strategies. Parbhoo, Sonali, et al. developed a mixture-of-experts framework combining a Bayesian Partially Observable Markov Decision Process (POMDP) and a history alignment kernel to optimize HIV therapy selection. It leverages patient-specific data to optimize long-term treatment outcomes, demonstrating superior performance in tailoring therapies based on individual patient histories compared to existing approaches. [4] Additionally, the integration of reinforcement learning techniques, such as prioritized experience replay in double deep Q-networks (PER-DDQN), has been shown to improve treatment strategy efficiency and reduce overall treatment duration by dynamically learning from patient-specific data while mitigating computational overhead. [6] Additionally, Yu, et al. proposed a Causal Policy Gradient (CPG) algorithm that integrates causal relationships into reinforcement learning, improving the efficiency and interpretability of dynamic treatment regimes by explicitly modeling how treatment actions influence outcomes, thereby refining the learning process and resulting strategies. [7]

II. Problem Statement

A successful HIV treatment is determined by the improvement of a patient's health. We define a patient's health status with two indicators: Cluster of Differentiation 4 (CD4) count and Viral Load (VL). The goal is to take the series of treatments that increase CD4 count and decrease the Viral Load. However, our goal is not to prescribe treatments that only yield immediate patient health improvements. Instead, we aim to select treatments that also contribute to long-term health improvements, ensuring the patient is healthier in the months ahead. The uncertainty arises from the fact that a treatment does not guarantee a specific outcome. The patient's response

to treatment is inherently probabilistic, influenced by unique human biology and other aforementioned individual factors.

This makes our problem a sequential decision-making challenge, where we aim to identify the best treatment for a given health status by considering and predicting future health statuses and their treatment options.

III. Dataset

A. Data Source and Type

To answer our problem statement, we aim to leverage a synthetic dataset from Health Gym AI [2], which contains time-series data on viral load, CD4 counts, demographics, and drug combinations. This dataset–created using Generative Adversarial Networks (GANs)–simulates real-world HIV treatment and mirrors the distributions, correlations, and trends observed in real datasets. It addresses critical challenges in healthcare machine learning by providing accessible, de-identified data that maintains privacy while enabling reproducible research. The authors validate the datasets' realism through qualitative and statistical tests, ensure low disclosure risks, and demonstrate its utility by training RL agents to optimize treatment strategies, achieving outcomes comparable to those trained on real-world data.

B. Data Variables

Our data contains records of 8916 patients over 60 months. The main variables of our dataset are summarized in Table I. This dataset captures key clinical features essential for understanding and modeling the progression of HIV infection and the impact of antiretroviral therapies. Left untreated, HIV can lead to Acquired Immunodeficiency Syndrome (AIDS), a condition in which the immune system becomes severely compromised, increasing susceptibility to opportunistic infections. Thus, two types of markers are of interest: (1) HIV Viral load (VL), which is the primary clinical marker used to monitor disease progression, and (2) the CD4 Counts (both absolute and relative) which are the primary clinical markers used to monitor immune system health.

Variable	Type	Description
Viral Load (VL)	Numeric	Reflects how much HIV virus is in one's body
Absolute Count for CD4 (CD4)	Numeric	Number of CD4 cells per μL, indicating immune health
Relative Count for CD4 (Rel CD4)	Numeric	Percentage of CD4 cells relative to total lymphocytes
Gender	Binary	Patient gender, encoded as 0 for female, 1 for male
Ethnicity	Categorical	Ethnic background, with categories
Base Drug Combination	Categorical	Type of antiretroviral combination, e.g., FTC + TDF, 3TC + ABC

TABLE I: Variables of interest in the Health Gym Synthetic HIV Dataset.

IV. Methods

We are modeling our problem as a Markov Decision Process (MDP). We start by defining the states, actions, transitions, and the reward function.

A. Model Variables

1) States: The states are currently defined by two metrics: the CD4 count and the Viral Load (VL). As the CD4 count and VL are both continuous values, we decided to discretize the state space by considering two factors: (1) the number of samples in each state should be similar to one another to avoid over-representation or under-representation of some states and (2) the number of bins/states created yields to the best results. We bin these metrics into 8 bins each, presented in Tables III and IV. This results in 64 unique states. The justification of this decision is later presented in the section **IV. D. Model Hyper-parameters Choices**. We also explore adding a history buffer to our state space to take into account previous states in our decision-making. We thus add the previous VL, the previous CD4, and the previous action to the state space with a parameter history length. For state space constraints, we were only able to experiment with a length of 1 which did not yield any good results.

2) Actions: The actions involve selecting different combinations of drugs, with each action corresponding to a unique drug combination. While there are 96 possible actions based on the 3 most relevant drug categories (with 6, 4, and 4 choices, respectively), our dataset includes only 44 drug combinations used. As our dataset size is extremely large, it is safe to assume that there are known drug combinations patients shouldn't take. This significantly reduces the action space to 44 actions.

3) Transitions: The transitions represent the probability of moving to a new health state given the current state and proposed treatment (action). They are obtained using a maximum likelihood estimation (MLE) approach. Given our large sample size, we avoided using a Dirichlet distribution, as actions that were not sampled for a given state were likely excluded for valid reasons. Thus, there was no need to assume a prior for such cases. Implementing a transition model is justified as our sample size is very large, and our action and state spaces are not very large.

4) Reward Function: We define our rewards based on the patient's state: $R(s) = 2 * \text{VLscore}(s) + \text{CD4score}(s)$. Since $VLscore(s)$ is negative, and $CD4score(s)$ is positive (cf. Tables III and IV), our reward function penalizes VL and encourages CD4 count.

A state with a higher CD4 count and a lower VL results in a higher reward; therefore, our model aims to take actions that maximize the likelihood of achieving the best possible subsequent states. We set the VL coefficient to be higher because reducing VL is prioritized over increasing CD4, and HIV treatments generally have greater control over lowering VL than improving CD4 levels.

B. Approach and Justifications

1) Exact Value Iteration on a Discretized State-Space MDP: An exact solution on a discretized state-space MDP was the most effective approach for this problem. The relatively small action and state spaces, combined with the completeness of the state space (a stationary state space where the set of states

S is equal to the set of next states S'), made it feasible to apply value iteration. This method allowed us to solve the Bellman equation iteratively, offering the best chance of achieving optimality:

$$V(s) = \max_a \left[R(s,a) + \gamma \sum_{s'} P(s' \mid s,a) V(s') \right],$$

where $V(s)$ represents the value of state s, $R(s,a)$ is the immediate reward for taking action a in state s, $P(s' \mid s,a)$ is the probability of transitioning to state s' from s after taking action a, and γ is the discount factor.

We aim to consider both long-term and short-term utilities, with a slightly higher importance on short-term outcomes, as getting better in the near future is important and is the stepping-stone to allow us to think long-term. Therefore, we focus on the near future rather than very distant outcomes (years), prioritizing treatments that are effective over the next few months. This approach also allows us to quickly identify and address poor short-term effects within a few months. To account for this in our infinite-horizon problem (treatment never ends), we introduce a discount factor γ with a value of 0.4, emphasizing immediate outcomes over long-term effects. This approach is fast and very effective due to the completeness of our space's state space and the ease of creating a transition model.

2) Q-Learning: We also implemented a Q-learning approach to optimize HIV treatment strategies. Q-learning is a model-free reinforcement learning algorithm that aims to learn the optimal policy by iteratively updating the Q-value function:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right],$$

where $Q(s,a)$ is the Q-value for taking action a in state s, α is the learning rate, r is the reward received after transitioning to the next state s', and γ is the discount factor that models the importance of future rewards. In our implementation, we attempted to incorporate a history buffer into the state space by augmenting each state with the previous viral load (VL), CD4 count, and action. This addition increased the dimensionality of the state space, making it significantly larger and resulting in the loss of the stationary state space property, as the augmented state S was no longer equivalent to the original state S' ($S \neq S'$). The non-stationarity of the state space posed additional challenges, as Q-learning assumes a stationary environment for effective learning. To encourage exploration in this expanded state space, we experimented with an exploration bonus into the Q-learning updates. This bonus, added to the reward as $\frac{1}{\sqrt{\text{visit count}+1}}$ prioritizes less-explored state-action pairs. This approach helped mitigate the imbalance in state-action pair exploration, especially in the larger augmented state space. Despite these enhancements, the performance of the Q-learning model with a history buffer was inferior to our former approach without history. The larger state space and non-stationarity made convergence slower and less reliable. As a result, we ultimately disregarded the inclusion of history in the state space and reverted to a simpler representation for our final implementation.

3) MDP over POMDP: We chose to use an MDP instead of a POMDP because our dataset does not allow us to infer the true health status (state) of a patient from their viral load and CD4 levels (observations). By assuming that the observations fully determine the states, we simplified the problem by treating the observations as the states, making an MDP a more suitable framework for our approach.

C. Policy Evaluation

We evaluate our policy through two methods to ensure its effectiveness, both using rollouts.

1) Policy Rollout - Random Policy: We perform multiple policy rollouts, treating each state in our state space as a starting point. Using a depth of 60, we follow our policy and transition model to simulate state transitions and collect the cumulative reward. Similarly, we perform rollouts using a random policy instead of our own. We then compare the cumulative rewards of the two policies. If our policy is effective, it should significantly outperform the random policy, as discussed in Results and Discussion.

2) Policy Evaluation - Evaluation of CD4 and VL changes in Rollouts: Our second evaluation method provides more interpretability. We perform rollouts with depths going from 1 to 60, with each rollout considering all possible starting states, and calculate the scores for each rollout as follows. We track four counts for each rollout based on the end state of the rollout:

- Best scenario: The number of times the end state resulted in an increase in CD4 and a decrease in VL compared to the start state.
- Good scenario: The number of times the end state resulted in a decrease in CD4 and a decrease in VL compared to the start state.
- Bad scenario: The number of times the end state resulted in an increase in CD4 and an increase in VL compared to the start state.
- Worst scenario: The number of times the end state resulted in a decrease in CD4 and an increase in VL compared to the start state.

We visualize these outcomes using a scatter plot of outcome counts for each scenario versus rollout depth, as detailed in the Results and Discussion section. This approach provides a clear and interpretable measure of policy performance over time.

D. Model Hyper-parameters Choices

Our model includes several customizable hyperparameters, such as:

1) State space size.
2) Action space size.
3) Discount factor γ value.
4) Reward function coefficients for VL and CD4.

To determine the optimal values for each hyperparameter, we used the same approach: selecting the configuration that

produced the best results based on our evaluation criteria (**IV. C. Policy Evaluation - CD4 and VL Outcome Analysis**). Rather than repeating the process for each hyperparameter, we demonstrate the methodology by illustrating how we selected the optimal state space size. We discretized the state space by dividing CD4 and VL levels into bins, ensuring a uniform number of samples in each bin. The desired number of samples per bin determined the total number of states. We generated a policy for each state space configuration tested, performed rollouts starting from each state, and computed rollout scores as previously described. These scores were averaged over all rollouts, and we compared the ratio:

$$\frac{\text{Best Outcome Count}}{\text{Possible Outcomes Count}}$$

The state space configuration with the highest outcome ratio was selected as optimal. The results of our experiments are summarized in Table II:

Number of States	**625**	**144**	**64**	**49**
$\frac{\text{Best Outcome Count}}{\text{Possible Outcomes Count}}$	0.57	0.625	0.87	0.83

TABLE II: Comparison of Best Outcomes to Total Possible Outcomes for Different State Space Sizes

As 64 states yielded the best results, we divided CD4 and VL into 8 bins each. Analytically, we can interpret these results as follows: the model didn't perform very well with a high number of states because it led to an over-representation of the state space, resulting in a limited number of actions for each state. This increases the odds of not taking the optimal action for that state.

When the state space size is too small, states are underrepresented, making it hard to identify changes like an increase in CD4 or a decrease in VL, and the model is unable to distinguish between states that require different actions, leading to suboptimal actions. Additionally, the large value ranges within each state can mask these changes, incorrectly interpreting them as no change, which is medically inaccurate. The same procedure was applied to determine the optimal values for the remaining hyperparameters, including action space size, discount factor, and reward function coefficients for VL and CD4.

CD4 Range	**Score**
0–201	0.1
201–279	0.2286
279–360	0.3571
360–465	0.4857
465–607	0.6143
607–840	0.7429
840–1253	0.8714
1253–∞	1.0

TABLE III: Score for each CD4 bin used in the reward function.

VL Range	**Score**
0–9	-0.1
9–16	-0.2286
16–30	-0.3571
30–54	-0.4857
54–100	-0.6143
100–209	-0.7429
209–724	-0.8714
724–∞	-1.0

TABLE IV: Score for each VL bin used in the reward function.

V. Results and Discussion

A. Model choices

As discussed in previous sections, we adopted an exact-solution approach using a discretized state-space MDP. This method involved strategically discretizing the state space and leveraging its completeness and stationary property to define a well-structured transition function. The abundance of samples in our dataset ensured that the transition function was well-defined, enabling the use of value iteration to solve the problem efficiently. The final hyperparameter values were selected empirically, as detailed in **IV. D. Model Hyperparameters Choices**. These include a discount factor of 0.4, reward function coefficients of 2 for VLscore(s) and 1 for CD4score(s), a state space size of 64, and 96 possible actions (of which 44 are utilized).

B. Performance

Using the policy evaluation methods described in Policy Evaluation, we obtained the following results:

For our first evaluation method, where we compare cumulative rewards against a random policy, the results are presented in Table V.

Policy	**Rollout Cumulative Reward (depth=60)**
Closed-Form MDP Policy	**26.0020**
Q-learning	-0.31
Q-learning with history	-16.71
Random Baseline	-56.9327

TABLE V: Comparison of Cumulative Rewards between Our Policy and Random Baseline.

Our results indicate that the Closed-Form MDP policy significantly outperforms the random baseline, achieving an average cumulative reward of 26.002. This suggests that our policy consistently selects actions leading to favorable outcomes, such as increasing CD4 counts or decreasing viral load (VL). In contrast, the random baseline achieves a cumulative reward of -56.933, demonstrating that random action selection frequently results in undesirable outcomes, such as decreasing CD4 or increasing VL. The Q-Learning approach underperformed compared to the Closed-Form MDP policy, with cumulative rewards of -0.31 (without history) and -16.71 (with history). These results highlight the challenges posed by larger, non-stationary state spaces when using Q-learning.

Our second method of evaluation (Policy Evaluation - Evaluation of CD4 and VL changes in Rollouts) led to the plot shown in Figure 1.

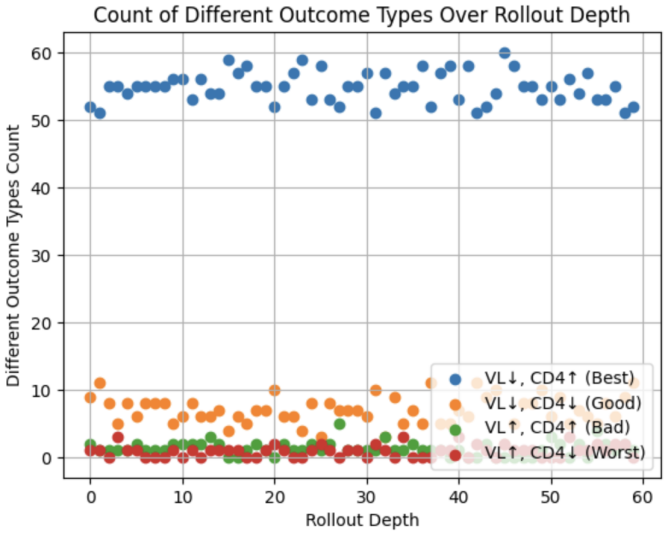

Fig. 1: Outcome Types Count v/s Rollout Depth

- *Best Outcomes Dominate Across Rollouts*: Figure 1 shows that the number of times the action leads to the best outcome (VL↓, CD4↑) consistently exceeds the other outcomes, regardless of the rollout depth. This indicates that the policy is effective at favoring the best outcomes for both the short term and the long term.
- *Rollouts With Depth 1-2 Shows Lower Ratios for Best Outcomes*: At rollout depths of 1-2, the best outcomes ratio is smaller than deeper rollouts. This is expected because of the non-zero discount factor which accounts for future rewards. As rollouts with shallow depths consider very short rewards, they may not showcase favorable outcomes since our policy may decrease CD4 before increasing it again, but the shallow depth cuts off before the increase. So for depths 1-2, the policy more frequently selects actions that do not have favorable end-state outcomes since the end-state is immediate to the start-state.
- *Consistent Ratios for Higher Depths*: Beyond a small rollout depth, the ratio of best outcomes stabilizes across rollouts. This is because the discount factor is small, causing the policy to prioritize only the next few steps. As a result, for a given state, the decisions focus on nearby future states, reflecting the diminishing importance of rewards further into the future, leading to outcomes that are indifferent to bigger rollout depths.

Overall, our model is a success, and we are choosing actions that will lead to better outcomes in both the short term and long term, achieving our goal.

VI. Conclusion and Future Work

In this project, we successfully developed and evaluated a policy for optimizing HIV treatment strategies using a Markov Decision Process (MDP) framework. Our approach involved discretizing the state space and leveraging its stationary properties to construct a well-defined transition model, allowing us to employ value iteration to compute an optimal policy. Our resulting policy demonstrated superior performance compared to a random baseline and other reinforcement learning methods, including Q-learning. Evaluation through cumulative rewards and clinical outcome-based metrics, such as changes in CD4 counts and viral load, highlighted the efficacy of the policy in achieving favorable outcomes both in the short term and in the long term. We highlighted significant trade-offs in hyperparameter tuning, such as balancing the granularity of the state space to optimize model performance. Furthermore, insights from our evaluation metrics, including the dominance of best outcomes in rollouts, further validated the practical applicability of our approach in dynamic HIV treatment planning. Despite the success of our best approach, several limitations still need to be addressed. For instance, while we assumed that observations directly represent states, real-world clinical data often includes latent health states that cannot be fully observed. Our current model would benefit from access to more diverse data and improved proxies for latent variables, such as adherence patterns or biomarkers that indicate underlying health states. We could address that limitation by extending the model to a Partially Observable Markov Decision Process (POMDP), which could better capture the uncertainty inherent in clinical settings. Patient-specific factors could further enhance the model's ability to personalize treatment recommendations. Another promising direction is further exploring how to better leverage patient history within the model. For example, incorporating temporal patterns of CD4 and viral load changes or past treatments could provide richer insights into long-term outcomes, even if doing so introduces additional complexity to the state space. Addressing these gaps can help us build a more robust, comprehensive, and clinically applicable solution.

References

[1] Juli M Bollinger, Gail Geller, Elizabeth May, Janesse Brewer, Leslie Meltzer Henry, and Jeremy Sugarman. Brief report: challenges in obtaining the informed perspectives of stakeholders regarding hiv molecular epidemiology. *JAIDS Journal of Acquired Immune Deficiency Syndromes*, 93(2):87–91, 2023.

[2] Nicholas I-Hsien Kuo, Mark N Polizzotto, Simon Finfer, Federico Garcia, Anders Sönnerborg, Maurizio Zazzi, Michael Böhm, Rolf Kaiser, Louisa Jorm, and Sebastiano Barbieri. The health gym: synthetic health-related datasets for the development of reinforcement learning algorithms. *Scientific data*, 9(1):693, 2022.

[3] National Institutes of Health. Hiv treatment basics, n.d. Accessed: 2024-12-06.

[4] Sonali Parbhoo, Jasmina Bogojeska, Volker Roth, and Finale Doshi-Velez. Combining kernel and model based reinforcement learning for hiv therapy selection.

[5] UNAids. Global hiv aids statistics — fact sheet, 2023. Accessed: 2024-10-24.

[6] Changyeon Yoon, Jaemoo Choi, Hee-Dae Kwon, and Myungjoo Kang. Optimal sti controls for hiv patients based on an efficient deep q learning method. *Journal of Theoretical Biology*, 594:111914, 2024.

[7] Chao Yu, Yinzhao Dong, Jiming Liu, and Guoqi Ren. Incorporating causal factors into reinforcement learning for dynamic treatment regimes in hiv. *BMC medical informatics and decision making*, 19:19–29, 2019.