A Multi-Agent Reinforcement Learning Framework for Evaluating the U.S. 'Ending the HIV Epidemic' Plan

Dinesh Sharma
University of South Florida
Tampa, FL 33620
dsharma1@usf.edu

Ankit Shah
University of South Florida
Tampa, FL 33620
ankitshah@usf.edu

Chaitra Gopalappa
University of Massachusetts Amherst
Amherst, MA 01003
chaitrag@umass.edu

Abstract

Human immunodeficiency virus (HIV) is a major public health concern in the United States, with about 1.2 million people living with HIV and 35,000 newly infected each year. There are considerable geographical disparities in HIV burden and care access across the U.S. The 2019 'Ending the HIV Epidemic (EHE)' initiative aims to reduce new infections by 90% by 2030, by improving coverage of diagnoses, treatment, and prevention interventions and prioritizing jurisdictions with high HIV prevalence. Identifying optimal scale-up of intervention combinations will help inform resource allocation. Existing HIV decision analytic models either evaluate specific cities or the overall national population, thus overlooking jurisdictional interactions or differences. In this paper, we propose a multi-agent reinforcement learning (MARL) model, that enables jurisdiction-specific decision analyses but in an environment with cross-jurisdictional epidemiological interactions. In experimental analyses, conducted on jurisdictions within California and Florida, optimal policies from MARL were significantly different than those generated from single-agent RL, highlighting the influence of jurisdictional variations and interactions. By using comprehensive modeling of HIV and formulations of state space, action space, and reward functions, this work helps demonstrate the strengths and applicability of MARL for informing public health policies, and provides a framework for expanding to the national-level to inform the EHE.

Keywords Deep Reinforcement Learning · Multi-agent Reinforcement Learning · Proximal Policy Optimization · Disease Modeling · HIV Modeling

1 Introduction

Human immunodeficiency virus (HIV) poses a significant public health challenge in the United States (U.S.), with more than 1.2 million individuals living with HIV. In 2019, there were 34,800 new infections reported, equating to a rate of 12.6 infections per 100,000 people HIV.gov [2023a]. The primary treatment for HIV is antiretroviral therapy (ART), which involves the daily use of a combination of HIV medications, referred to as an HIV regimen, to manage the infection. While ART cannot cure HIV, it plays a crucial role in extending the lifespan of people living with HIV (PWH) and preventing the risk of HIV transmission. The purpose of ART is to achieve viral load suppression (VLS), defined as having fewer than 200 copies of HIV per milliliter of blood, making it undetectable through viral load testing HIV.gov [2023b]. Additionally, pre-exposure prophylaxis (PrEP) is a medication used by individuals at risk for HIV to prevent infection through sexual contact or injection drug use. PrEP effectively prevents HIV from establishing itself and spreading within the body HIV.gov [2023c].

Despite the availability of these preventive and treatment measures, reducing the incidence of new HIV infections remains a formidable challenge nationwide. Disparities in infection rates persist across different geographical areas and risk groups. Specifically, 57 jurisdictions, comprising counties and states, accounted for over 50% of all new HIV diagnoses reported in 2016 and 2017 HIV.gov [2020]. In response to this alarming situation, the U.S. Department of Health and Human Services launched the 'Ending the HIV Epidemic (EHE)' initiative in 2019, with the ambitious goal of eradicating the HIV epidemic in the U.S. by 2030. The EHE initiative aims to achieve two primary objectives: (i) reduce new HIV incidence by 75% in these EHE-designated jurisdictions by 2025 and (ii) decrease overall HIV incidence by 90% across all jurisdictions by 2030 HIV.gov [2020]. Additionally, the initiative targets achieving the 95-95-95 care continuum benchmarks, which entail ensuring that 95% of PWH are aware of their infections, 95% of those aware of their status are linked to care, and 95% of individuals in care attain VLS. Furthermore, the initiative aims to attain 50% PrEP coverage among eligible individuals by 2025 in EHE jurisdictions and by 2030 nationwide AHEAD [2022a].

To accomplish these ambitious targets, several essential actions have been identified: (i) prompt diagnosis (testing) of all individuals with HIV, (ii) rapid and effective treatment with ART to achieve VLS, and (iii) adequate prevention measures to reduce new HIV transmissions, including the widespread use of PrEP among high-risk groups. Given the limited resources, finding jurisdiction-specific optimal combinations of these actions, to minimize new HIV infections, would be key to optimal resource allocation.

The literature studies have not adequately addressed the disparities among all jurisdictions in the context of HIV. Recent literature has mainly focused on six specific cities: Atlanta (GA), Baltimore (MD), Los Angeles (CA), Miami (FL), New York City (NY), and Seattle (WA), which together account for 24.1% of the total population living with HIV in the U.S. Nosyk et al. [2020a]. Notably, Krebs et al. [2019] and Zang et al. [2020] have developed a framework to construct and calibrate a dynamic compartmental HIV transmission model exclusively for these six U.S. cities. Nosyk et al. [2019] projected HIV incidence trajectories for a decade (2020-2030) within these cities, highlighting the need for new strategies to achieve the EHE goal. Nosyk et al. [2020a] identified the optimal combination of strategies that offer both health benefits and cost-effectiveness within these six cities. Focusing on specific risk groups, Krebs et al. [2020] examined the cost-effectiveness of optimal strategies for persons who inject drugs (PWID) within the same six cities. Furthermore, Nosyk et al. [2020b] delved into the impact of racial and ethnic disparities in healthcare access on progress towards the EHE goal within these cities. Expanding beyond these cities, Fojo et al. [2021] employed a combination of HIV testing and PrEP coverage as intervention strategies to project HIV incidence in 32 priority metropolitan statistical areas. Additionally, Wheatley et al. [2022] assessed the cost-effectiveness of PrEP in reducing overall HIV incidence specifically in Atlanta, GA. Khatami and Gopalappa [2021] explored the optimal combination of HIV testing and retention-in-care strategies towards the EHE goal, examined at the national level. Meanwhile, Khurana et al. [2018] investigated the impact of PrEP on overall HIV incidence in the U.S. from 2016 to 2030. Recently, Khatami and Gopalappa [2021] introduced a national HIV elimination model. This model identified an optimal sequence of HIV-testing and retention-in-care rate combinations at 5-year intervals from 2015 to 2070, employing Q-learning, a classical reinforcement learning (RL) algorithm. Additionally, Tatapudi and Gopalappa [2022] developed a national HIV epidemic compartmental simulation model representing the U.S. population, comprising 96 jurisdictions. This model was designed to assess the impact of jurisdictional heterogeneity and mixing on national-level HIV prevention analyses.

The models in the literature either model each jurisdiction independently, disregarding any interactions between them, or collectively model the national epidemic ignoring jurisdictional differences. This significant gap in the literature leaves unaddressed the potential plans and necessary actions tailored to the stage of the HIV epidemic in each jurisdiction. All of these aspects are essential to get closer to the overarching EHE goal by 2030. To our knowledge, no prior research has explored dynamic jurisdictional interactions in these models. Capturing the interaction between various jurisdictions for HIV spread in terms of sexual partnership mixing among all the risk groups across jurisdictions can help showcase a more realistic impact of HIV transmission and effect of HIV epidemic on each jurisdiction from its respective neighboring areas. This aspect could make a substantial contribution toward realizing the EHE goal.

Deep reinforcement learning (DRL) offers a promising approach for deriving (near-)optimal policies in dynamic and stochastic environments, making it well-suited to tackle the complexity of disease dynamics and time-dependent decision-making in this research context. In this paper, we develop a policy-based DRL methodology to explore optimal combinations of actions (including testing, retention-in-care, and prevention rates) at the jurisdictional level while considering interactions. Our objective is to narrow the gap in achieving the EHE targets. We specifically formulate this research problem as a multi-agent reinforcement learning (MARL) problem, treating individual jurisdictions as independent entities. This approach empowers each jurisdiction to autonomously make decisions concerning preventive measures in pursuit of the EHE objectives, offering a decentralized strategy that accommodates diverse needs and circumstances within jurisdictions. By adopting this decentralized approach, we overcome limitations associated with national models, while including jurisdictional interactions, and facilitate tailored decision-making for each jurisdiction. To show the effectiveness of our approach, we conduct a numerical study focusing on the jurisdictions in two states, California and Florida. We compare our approach with a single-agent reinforcement learning (SARL) framework, where all jurisdictions within a state are collectively represented as a single agent.

The study makes several important contributions. The primary contribution lies in the development of a MARL framework aimed at reducing nationwide HIV infections. Our novel approach takes into consideration jurisdictional interactions when determining optimal intervention strategies. Another noteworthy contribution is the modeling of this public health challenge as a Markov decision process (MDP) and the meticulous construction of its key components. Specifically, the state within the MDP formulation encapsulates all necessary information in a concise format for decision-making in each jurisdiction. The reward signal is finely tuned using the new infection rate and the available budget for HIV care and treatment in each jurisdiction. Through our experimental findings, we demonstrate the effectiveness of our sequential decision-making framework for generating intervention policies across varying jurisdictional budgets. Our study provides valuable insights, including the potential for a significant reduction in HIV infections by 2030 through increased intervention efforts and a larger allocation of resources.

The rest of the paper is organized as follows: Section 2 discusses related literature. Section 3 offers an overview of our simulation model and proposed approaches. Section 4 presents computational results and associated discussions. Lastly, Section 5 provides the paper's conclusion.

2 Literature Review

In addition to the recent literature studies on the HIV public health challenge mentioned in Section 1, several other relevant studies deserve attention. Kok et al. [2015] developed an optimal resource allocation model using a system dynamics approach to the continuum of HIV care, aiming to minimize new HIV infections in Vancouver. Okosun et al. [2013] conducted an analysis to determine the conditions for optimal control of the disease, considering factors such as the effective use of condoms, treatment regimes, and screening of infectives. Gopalappa et al. [2012] and Lin et al. [2016] focused on assessing the cost-effectiveness of various interventions in HIV prevention. Additionally, several studies Lasry et al. [2011], Zaric and Brandeau [2001], Lasry et al. [2012], Yaylali et al. [2016], Juusola and Brandeau [2016], Gromov et al. [2018], Yaylali et al. [2018] developed optimization models for allocating HIV funds to enhance the effectiveness of HIV prevention efforts. In the context of another recent public health challenge, COVID-19, Matrajt et al. [2021], Yu et al. [2021], Han et al. [2021] utilized optimization models to determine vaccine prioritization strategies when vaccine supply is limited. Libotte et al. [2020] and Tsay et al. [2020] investigated optimal control strategies for managing the COVID-19 pandemic. Tang et al. (2022) formulated a bi-objective optimization problem for multi-period COVID-19 vaccination planning, aiming to minimize operational costs and the distance traveled by vaccine recipients. Finally, Rawson et al. [2020] employed an optimization approach to explore efficient strategies for ending COVID-19 lockdowns without overwhelming healthcare services.

Development and application of RL solution methodologies in the public health domain has experienced significant growth in recent times. For instance, Kwak et al. [2021] applied a DRL approach to determine optimal intervention strategies during the COVID-19 pandemic. This included identifying the optimal timing and intensity of lockdowns and travel restrictions for various countries and territories. Libin et al. [2021] employed DRL to model the spread of pandemic influenza in Great Britain, utilizing 379 patches to capture the infection process and develop mitigation policies. Awasthi et al. [2022] proposed a DRL model with a contextual bandits approach to optimize the distribution of COVID-19 vaccines across five different states in India. Furthermore, Allioui et al. [2022] introduced a DRL-based method for extracting masks from COVID-19 computed tomography (CT) images, enabling accurate clinical diagnoses by extracting visual features of COVID-19 infected areas. Jalali et al. [2021] combined RL with optimized convolutional neural networks (CNNs) to reduce ensemble classifier size and enhance model performance. Kompella et al. [2020] developed an agent-based pandemic simulator for COVID-19, employing DRL-based methodology to optimize mitigation policies. Kumar et al. [2021] proposed a model incorporating recurrent neural networks (RNNs) for forecasting COVID-19 cases and integrated DRL to optimize these values based on symptoms. Lastly, Bednarski et al. [2021] investigated the use of RL models to facilitate the redistribution of medical equipment, enhancing the public health response in preparation for future crises.

The primary objective of an RL agent is to make decisions that assess the impact of all actions in terms of short-term and long-term utility for stakeholders. This characteristic makes RL an ideal approach for addressing complex decision problems within public health, such as resource allocation during a pandemic, monitoring and testing strategies, and adaptive sampling for hidden populations Weltz et al. [2022]. Several DRL algorithms, including Deep Q-Network (DQN) Mnih et al. [2013], Dueling Double Deep Q-Network (D3QN) Wang et al. [2016], Soft Actor-Critic (SAC) Haarnoja et al. [2018], Proximal Policy Optimization (PPO) Schulman et al. [2017], have been applied to various public health problems Kwak et al. [2021] Kompella et al. [2020] Libin et al. [2021]. Notably, PPO has demonstrated strong performance in addressing problems with high-dimensional state spaces and continuous action spaces.

MARL is a subfield of RL where multiple agents interact in a shared environment, and each agent's actions not only affect its own rewards but also impact the rewards and states of other agents. MARL focuses on the study of how agents should learn to make decisions in such interactive and often competitive or cooperative environments. MARL algorithms find applications across various fields, including traffic control, resource allocation, robot path planning, production systems, and maintenance management Oroojlooy and Hajinezhad [2022].Chu et al. [2019] employed a multi-agent advantage actor-critic approach for adaptive traffic signal control within complex urban traffic networks. Their strategy involved forming a cooperative MARL system, with each intersection serving as a local RL agent equipped with its own local observations and partial knowledge of the overall system. Yang et al. [2020] integrated traffic signals and vehicle speeds to enhance urban traffic control performance. They utilized a multi-agent modified PPO method, making adjustments to the clip hyperparameter for improved learning. In the realm of wireless networks, Nasir and Guo [2019] developed a distributively executed dynamic power allocation model using multi-agent deep Q-learning. For large-scale fleet management challenges, Lin et al. [2018] applied contextual deep Q-learning and contextual multi-agent actor-critic algorithms. They treated each available vehicle as an agent and grouped vehicles in the same region at the same time as identical agents. Furthermore, Yu et al. [2020] employed a multi-agent actor-attention-critic mechanism to minimize heating, ventilation, and air conditioning costs in a multi-zone commercial building under dynamic pricing conditions. Their approach took into consideration random zone occupancy, thermal comfort, and indoor air quality comfort.

In this paper, we use a DRL approach to develop a policy-based methodology for optimizing actions in the dynamic and stochastic HIV disease environment. We treat individual jurisdictions as independent agents in a MARL framework to obtain a decentralized decision-making strategy to get closer to achieving the EHE targets, overcoming national model limitations, and accommodating diverse jurisdictional needs and constraints.

3 Methodology

The model framework comprises a compartmental simulation model for simulating disease progression, an MDP tailored for addressing the problem, and a DRL agent tasked with solving the resulting MDP. In this approach, we represent each jurisdiction as an individual agent and employ MARL to facilitate interaction among these agents within the environment. This enables each agent to learn its optimal policy while taking into account the dynamic changes in the environment caused by other agents operating within the same context. Figure 1 provides a schematic overview of our MARL framework where we use compartmental simulation model used in Tatapudi and Gopalappa [2022] as the simulator. In the SARL approach, we represent all jurisdictions as a single agent, with the accumulation of information from all jurisdictions forming the agent's state.

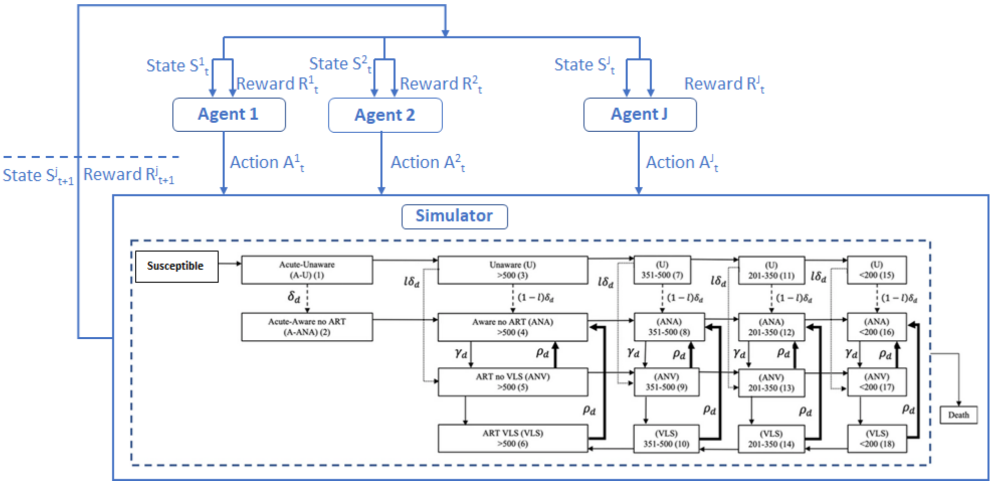


Figure 1: Schematic of the multi-agent reinforcement learning (MARL) framework with compartmental simulation from Tatapudi and Gopalappa [2022].

3.1 Simulation Model

The disease progression of HIV is tracked through the compartmental simulation model developed in Tatapudi and Gopalappa [2022]. An overview of this compartmental simulation is provided in the simulator of Figure 1. The model encompasses four stages within the care continuum: Unaware, Aware no ART, ART no VLS, and ART VLS. Each of these care continuum stages is further divided into five disease progression stages based on CD4 count: Acute, CD4 > 500, CD4 351-500, CD4 201-350, and CD4 < 200. Additionally, the model includes compartments for individuals who are susceptible to infection and those who have succumbed to the disease, resulting in a total of 22 compartments. In the model, δ_d represents the diagnostic rate, ρ_d represents the dropout rate, and γ_d represents the rate of entry to care in disease stage d. Furthermore, the parameter l signifies the proportion of individuals who are linked to care within three months of diagnosis.

Table A2 shows the various proportions of partnership mixing with the jurisdictional interaction category, as considered in Tatapudi and Gopalappa [2022]. The highest interaction takes place within the same jurisdiction for all risk groups, heterosexual males (HM), heterosexual females (HF), and men who have sex with men (MSM). The proportion of interaction in other jurisdictions in the same state is divided equally among all the jurisdictions in that state. The out of state interaction is divided equally among all the remaining states that are being modeled.

3.2 Markov Decision Process (MDP)

The MDP can be defined as a tuple $N, S, \mathbf{a}, P, R, \gamma$. Here, N represents the number of agents, where $N = 1$ corresponds to a single-agent MDP. S is the set of environmental states, denoted as o^j for each agent j to represent their local observations. For each agent j, $\mathbf{a}^j$ signifies their set of possible actions. Consequently, the total collection of actions, denoted as $\mathbf{a}$, is defined as the Cartesian product: $\mathbf{a} = \mathbf{a}^1 \times \mathbf{a}^2 \times \ldots \times \mathbf{a}^N$. The state transition function, P, is a mapping from the current state $s \in S$ and the joint action $\mathbf{a}$ to the next state $s' \in S$ at time $t + 1$. This is represented as $P : S \times \mathbf{a} \rightarrow \Delta(S)$, where $\Delta(S)$ denotes the probability distribution over states. The set of reward functions, R, includes individual reward functions $R^j : S \times \mathbf{a} \times S \rightarrow R$, where R^j signifies the reward for agent j when transitioning from state $(s, \mathbf{a})$ at time t to the next state s' at time $t + 1$. Lastly, the discount factor $\gamma \in [0, 1]$ is utilized to adjust for future rewards. It plays a crucial role in the MDP formulation by determining the weightage of future rewards relative to immediate rewards. The key elements of the MDP formulation are further described below.

3.2.1 State

State represents the information about a jurisdiction at a particular time. The state of a jurisdiction $j \in J$ at time t can be represented as,

$$s_t^j = [p_{k,j,t}, \mu_{u,k,j,t}, \mu_{a,k,j,t}, \mu_{ART,k,j,t}, \mu_{prep,k,j,t}; \forall k \in \{HM, HF, MSM\}] \tag{1}$$

where, $p_{k,j,t}$ denotes the proportion of people living with HIV (PWH), diagnosed or undiagnosed, $\mu_{u,k,j,t}$ denotes the proportion of PWH unaware of their infection, $\mu_{a,k,j,t}$ denotes the proportion of PWH aware of their infection but not on ART, $\mu_{ART,k,j,t}$ denotes the proportion of PWH aware of their infection and on ART or on VLS, and $\mu_{prep,k,j,t}$ denotes the proportion of people who have been prescribed PrEP among the people with PrEP indicators, for each risk group $k \in \{HM, HF, MSM\}$ in jurisdiction $j \in J$ at time t.

3.2.2 Action

The action $\mathbf{a}_t^j$ for agent $j \in J$ at time t represents a combination of three sets of measures: testing, retention-in-care and treatment, and prevention, aimed at controlling HIV infections. These actions are characterized by the changes in the following proportions: (i) the proportion of PWH who are unaware of their infection ($\mu_{u,k,j,t}$), (ii) the proportion of PWH who are aware of their condition and are on ART ($\mu_{ART,k,j,t}$), and (iii) the proportion of individuals taking PrEP as a preventive measure ($\mu_{prep,k,j,t}$). Note that, instead of directly formulating testing and retention-in-care using corresponding rates, we formulated as above because of its attractive mathematical properties that help efficiently constrain the number of action choices and thus improve the chance of convergence of an RL algorithm, as shown in Khatami and Gopalappa [2021].

$$\mathbf{a}_t^j = [a_{unaware,k,t}^j, a_{ART,k,t}^j, a_{prep,k,t}^j; \forall k \in \{HM, HF, MSM\}] \tag{2}$$

where, $a_{unaware,k,t}^j$ denotes the proportion decrease in $\mu_{u,k,j,t}$, $a_{ART,k,t}^j$ denotes the proportion increase in $\mu_{ART,k,j,t}$, and $a_{prep,k,t}^j$ denotes the proportion increase in $\mu_{prep,k,j,t}$, for each risk group $k \in \{HM, HF, MSM\}$ in jurisdiction $j \in J$ at time t.

3.2.3 State Transition Function

The state transition function calculates the probabilities of a system transitioning from state s_t to the next state s_{t+1} when an action $\mathbf{a}_t$ is taken. For this problem, the state transition probabilities are not available, hence we use the simulator to provide the Markov jumps from one state to another under the action taken by the agent(s).

3.2.4 Reward

Rewards reflect the impact of an action on the current state. Immediate rewards are computed as the negative of the total new infections, as our objective is to minimize this value over time. Notably, each jurisdiction, j, has a budget,

B^j, allocated towards reducing HIV infections RWHAP [2022]. In addition to the reward of lowering the infection count, the reward function also contains a penalty term, which is the difference in the cost allocated by the agent and the budget allocated for that jurisdiction. The rewards $R^j_{a^j_t}(s^j_t, s^j_{t+1})$ for an agent $j \in J$ transitioning from state s^j_t to the next state s^j_{t+1} under action $\mathbf{a}^j_t$ can be estimated as follows:

$$R^j_t = -\sum_k i_{k,j,t} - P^j_t \tag{3}$$

where, $i_{k,j,t}$ represents the new infections for risk groups $k \in \{HM, HF, MSM\}$ in jurisdiction j at time t. P^j_t is the penalty term, which is calculated as, $P^j_t = C^j_t - B^j_t$, the difference between the cost allocated by agent C^j_t and the budget B^j_t for jurisdiction j at time t. Notably, C^j_t is obtained using HIV related cost functions in Khatami and Gopalappa [2021] and estimation of those costs from the simulation model, and the budget B^j_t can be obtained from Ryan White HIV/AIDS Funding Program HRSA [2023].

3.3 Deep Reinforcement Learning (DRL) Solution Methodology

The optimal values for testing, treating, and PrEP can be obtained using the DRL solution methodology. In our MARL framework, each agent samples an action and interacts with the environment, represented by the simulator. Although each agent acts independently, their simultaneous interactions within the same environment can influence the actions taken by all of them. Consequently, each agent receives a reward from the simulator and transitions to the next state.

3.3.1 DRL Algorithm

Among the various DRL algorithms, PPO Schulman et al. [2017] is known for its stability and efficiency. It is an on-policy algorithm that can be used for environments with continuous action spaces and very large state spaces. Our research problem has the characteristics that are well suited for such an approach. PPO adopts the concepts from trust region policy optimization (TRPO) Schulman et al. [2015]. The objective in TRPO, called surrogate advantage, is calculated as:

$$L^{CPI}(\theta) = \hat{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_l}(a_t|s_t)}\hat{A}_t\right] = \hat{E}_t[r_t(\theta)\hat{A}_t] \tag{4}$$

The probability ratio $r_t(\theta)$ of two policies give a measure of how far the new policy π_θ is from the old policy π_{θ_k}. As having a large policy update at a time would destabilize the training, PPO-clip puts a limit on the extent of policy update with the following objective.

$$L^{CLIP}(\theta) = \hat{E}_t\left[\min(r_t(\theta)\hat{A}_t, clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t)\right], \tag{5}$$

where ϵ is the hyperparameter that clips the probability ratio and prevents r_t from moving outside the interval $[1-\epsilon, 1+\epsilon]$. The probability ratio is clipped with $1+\epsilon$ or $1-\epsilon$ depending on whether the advantage is positive or negative. The surrogate objective for PPO-clip can be generalized with the following equation:

$$L^{CLIP}_{(s,a,\theta_l,\theta)} = \min\left(\frac{\pi_\theta(a|s)}{\pi_{\theta_l}(a|s)}A^{\pi_{\theta_l}}(s,a), g(\epsilon, A^{\pi_{\theta_l}}(s,a))\right) \tag{6}$$

where,

$$g(\epsilon, A) = \begin{cases}(1+\epsilon)A & A \geq 0\\ (1-\epsilon)A & A < 0\end{cases} \tag{7}$$

3.3.2 Solution Approach

We are employing two approaches, SARL and MARL, in our experiments. These correspond to a single decision-making system and decentralized decision-making for each agent, respectively. In the SARL approach, we represent the overall group of jurisdictions by their cumulative value, which is then averaged over the number of jurisdictions. The first component of our state space, denoted as p_k and representing the proportion of people living with HIV, is calculated using the equation below:

$$p_k = \frac{\sum\limits_{j\in J} I_{k,j}}{\sum_{j\in J} N_{k,j}} \tag{8}$$

where $I_{k,j}$ represents the total number of PWH, and $N_{k,j}$ represents the total population for risk group $k \in \{HM, HF, MSM\}$ in jurisdiction j. The remaining components of the state space for SARL are computed in a manner analogous to the MARL formulation (1). The action and reward follow a similar structure as shown in equations (2) and (3) but with $j = 1$, meaning we obtain these values for a single cumulative jurisdiction. Notably, in MARL, the state, action, and reward for SARL are similar to equations (1), (2), and (3), respectively. Each agent maintains its own state representation, samples its own action, and receives a reward for each transition. Algorithm 1 shows the implementable steps of the MARL framework for (near-)optimal interventions in HIV epidemic. The SARL framework is implemented in a similar manner as shown in Algorithm 1, with the changes mentioned above.

Algorithm 1 MARL Framework for Optimal Interventions in the HIV Epidemic

```
Input: Initialize policy parameters $\theta_l^j$, initialize value function parameters $\phi_l^j$ for $j = 1, 2, ..., N$ agents, set episode length $T$ (here we used $T = 12$ years corresponding to year 2019 to 2030), set number of episodes $M$, set buffer size $B$ (here we used $B = 10$ episodes), set total timesteps $X = 0$, set $l = 0$.
Initialize replay buffer $\mathcal{D}_l^j$ for each agent $j$.
for episode $m = 0, 1, 2, ..., M$ do
    for year $t = 1, 2, ..., T$ do
        At $t = 1$, extract the initial state for each agent $j$ as:
            $s_t^j = [p_{k,j,t}, \mu_{u,k,j,t}, \mu_{a,k,j,t}, \mu_{ART,k,j,t}, \mu_{prep,k,j,t}; \forall k \in \{HM, HF, MSM\}].$
        Sample actions $\mathbf{a}_t^j$ comprising the change in proportions from each agent $j$ by running policy $\pi_l^j = \pi_{\theta_l^j}$
            $\mathbf{a}_t^j = [a_{unaware,k,t}^j, a_{ART,k,t}^j, a_{prep,k,t}^j; \forall k \in \{HM, HF, MSM\}].$
        Calculate the diagnostic and dropout rates using the change in unaware and ART proportions, respectively.
        Simulate the actions in the compartmental simulation model.
        Observe next state $s_{t+1}^j$ and reward $R_{t+1}^j$, and collect set of trajectories $\{\tau_t^j\}$, where
            $\tau_t^j = (s_t^j, a_t^j, R_{t+1}^j, done).$
        Add the trajectories to the replay buffers $\mathcal{D}_l^j$.
        $X \leftarrow X + 1$
        if ($X$ mod $B \times T$) = 0 then
            Compute advantage estimates, $A_t^j$ based on the current value function $V_{\phi_l^j}^j$.
            Update the policy by maximizing objective (via stochastic gradient ascent):
                $\theta_{l+1}^j = \arg\max_{\theta^j} \frac{1}{|\mathcal{D}_l^j| T} \sum_{\tau_t^j \in \mathcal{D}_l^j} \sum_{t=0}^{T} \min\left( \frac{\pi_{\theta^j}(a_t^j | s_t^j)}{\pi_{\theta_l^j}(a_t^j | s_t^j)} A^{\pi_{\theta_l^j}}(s_t^j, a_t^j), g(\epsilon, A^{\pi_{\theta_l^j}}(s_t^j, a_t^j)) \right).$
            Fit value function by regression on mean-squared error:
                $\phi_{l+1}^j = \arg\min_{\phi^j} \frac{1}{|\mathcal{D}_l^j| T} \sum_{\tau_t^j \in \mathcal{D}_l^j} \sum_{t=0}^{T} \left( V_{\phi_l^j}^j (s_t) - R_t^j \right)^2.$
            Clear replay buffer for each agent $j$.
            $l \leftarrow l + 1$
        end if
    end for
end for
```

4 Experiments and Results

4.1 Experimental Setup

We adopted the initial data for HIV infection in all jurisdictions for the year 2018 from Tatapudi and Gopalappa [2022]. This data encompasses all age groups from 13 to 100 and includes all 22 care continuum stages for three risk groups. The compartmental simulation model, used as the simulator in our study to determine transitions from time t to $t + 1$, was also sourced from the same paper. In our experiments, we selected two distinct groups of jurisdictions, one from California and the other from Florida. Each of these groups consists of eight jurisdictions corresponding to their respective states. In the SARL approach, the values of jurisdictions within a state were aggregated to create a single representation for the overall state. Consequently, a single set of actions was learned and implemented across all jurisdictions within a state. In contrast, in the MARL approach, each jurisdiction within a state was treated as an individual agent. These agents had the autonomy to sample actions independently, representing changes within their respective jurisdictions. All of these agents interacted with the same environment. The specific jurisdictions used in these groups are detailed below.

California: Alameda County (CA1), Los Angeles County (CA2), Orange County CA (CA3), Riverside County (CA4), Sacramento County (CA5), San Bernardino County (CA6), San Diego County (CA7), (rest of) California (CA8)

Florida: Broward County (FL1), Duval County (FL2), Hillsborough County (FL3), Miami-Dade County (FL4), Orange County FL (FL5), Palm Beach County (FL6), Pinellas County (FL7), (rest of) Florida (FL8)

The values of the parameters in Algorithm 1 were set as follows: $T = 12$, $M = 100,000$ and the optimizer was Adam. Figures A.2, A.3, A.4 and A.5 in the Appendix show the convergence in learning in MARL and SARL in California and Florida, respectively.

The budget allocation for each jurisdiction used in the experiment was obtained from Health Resource and Service Administration website. Ryan White HIV/AIDS Program (RWHAP) has allocated funds as EHE awards to 49 recipients that includes 39 metropolitan areas and 8 states HRSA [2023]. The total budget allocated to two states of interest, California and Florida was about $54.3 million and $48.4 million, respectively. While this budget from RWHAP has been used for our numerical study, it reflects only a fraction of the overall resources spent for HIV, and nationally, about 50% of people diagnosed with HIV have received some type of service from RWHAP. The total HIV related cost, which includes testing cost, retention-in-care cost and PrEP cost, was estimated from the simulator. The testing cost and retention-in-care cost were calculated using the cost functions defined in Khatami and Gopalappa [2021], whereas the annual PrEP related costs were taken from Sansom et al. [2021].

The feasibility of implementing an action was analyzed to set a constraint on the maximum possible action choices. Khatami and Gopalappa [2021] looked at the past trend in HIV proportion to set the feasible actions, specifically from 2010 to 2014, and found 2.3 % decrease in unaware proportion and 13.8 % increase in proportion of PWH in ART. Looking at the similar trend from 2015 to 2019, we found 1.6 % decrease in unaware proportion (14.9 % in 2015 to 13.3 % in 2019) CDC [2021] and an increase of 4.8 % in VLS proportion (59.8 % in 2015 to 64.6 % in 2019) CDC [2022] during that 5 years time interval. Thus, we set the feasible range of action choices as 0-0.5 % decrease in unaware proportion and 0-4 % increase in ART proportion each year. Additionally, the percentage of people with PrEP indicators that have been prescribed PrEP has gone from 13.2 % in 2017 to 30.1 % in 2021 which marks an increase of 16.9 % during that 5 years AHEAD [2022b]. Thus, the third action, increase in PrEP proportion, was given a range of 0-4 % each year.

4.2 Results

Figure 2 a) and b) depict the total new incidence in California and Florida from 2019 to 2030 for both solution approaches, SARL and MARL. In both states, MARL significantly outperforms SARL. In California, there is a notable 19% reduction in incidence when using the MARL approach, while SARL results in a 4.4% increase in total incidence from 2019 to 2030. A similar trend is observed for Florida, where MARL yields a 23% decrease in new incidence, while SARL results in a 0.6% increases. When comparing these incidence numbers for each jurisdiction within these states (results are available in the Appendix), a consistent pattern emerges in which the MARL approach outperforms the SARL approach across all jurisdictions. Figure 2 c) and d) display the differential costs incurred in California and Florida for both MARL and SARL. As larger jurisdictions have higher prevalence, the decisions are biased towards larger jurisdictions resulting in higher intervention rates even for jurisdictions with lower prevalence as well. Thus, it adds to the cost of the interventions, without much impact in incidence in the lower prevalence jurisdictions. Although the bias in policies here is towards larger jurisdictions, this may not always be true such as when larger jurisdictions don't have a higher prevalence. Nonetheless, the results highlight the advantage of multi-jurisdictional modeling compared to a single agent national model. On the otherhand, Figure A.1 a) shows a higher total incidence estimates in California in the scenario with mixing (interactions between jurisdictions) compared to the scenario that assumed jurisdictions are independent (no interactions with other jurisdictions), demonstrating that a single jurisdiction (or city) model too is insufficient as it can misrepresent the true epidemic.

Despite the superior performance of MARL, it is still a considerable distance from achieving the EHE targets. To move closer to the ambitious EHE initiative goal, a more aggressive policy approach is needed, involving increased

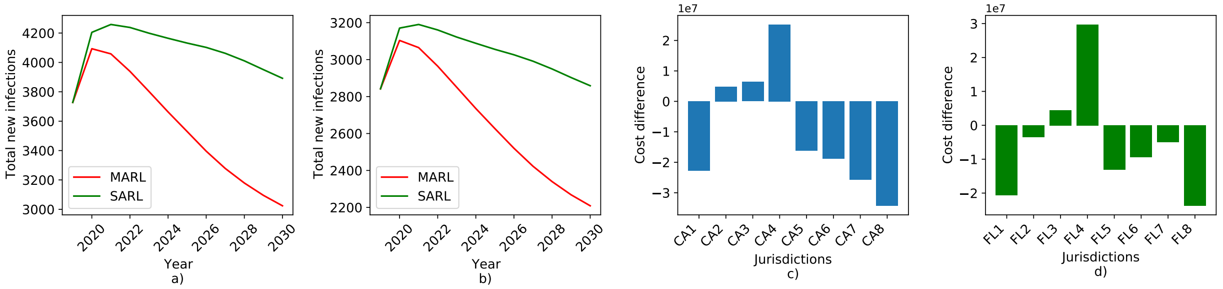


Figure 2: Comparing total new incidence from SARL and MARL in a) California and b) Florida, and cost difference (MARL minus SARL) in c) California and d) Florida.

intervention measures and a greater allocation of resources. To assess the effectiveness of these approaches in reducing HIV incidence, we modeled two different scenarios. The first scenario involves doubling the range of the action space, giving the agent more freedom to choose actions that result in higher levels of intervention. Specifically, we increased the action range to 1% decrease in unaware proportion, 8% increase in ART proportions and 8% increased in PrEP proportions. Figure 3 a) and c) illustrate that expanding the action range leads to a more significant reduction in the number of incidences by 2030, amounting to approximately a 34% reduction in California and 35% reduction in Florida. The second scenario involves increasing the budget allocation while maintaining the expanded action ranges. Figure 3 b) and d) illustrate the comparison of MARL in California and Florida with the initial budget and the increased budget, which is ten times the initial allocation. The new incidence trajectory demonstrates a reduction of approximately 58% from 2019 to 2030 in California while showing about 75% reduction in Florida in the same time interval. Although this level of reduction still falls short of the EHE goal, these scenarios clearly outline a viable pathway towards achieving those targets. The key takeaway is the need to allocate more resources and adopt more aggressive intervention measures in the pursuit of HIV elimination.

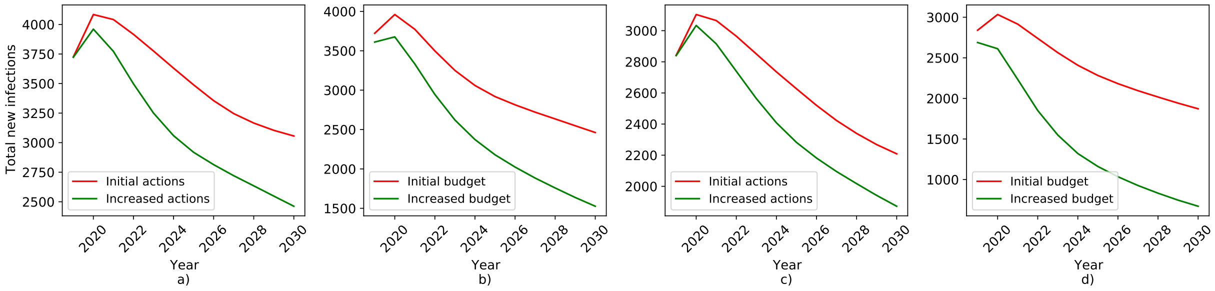


Figure 3: Comparing total new incidence from MARL with a) initial & increased actions and b) initial & increased budget in California, and c) initial & increased actions and d) initial & increased budget in Florida.

5 Conclusions

This paper explored the minimization of new HIV infections through a range of intervention measures, including testing, treatment, and prevention. To address this challenge, a MARL approach using a policy-based method is proposed. The MARL framework empowers each jurisdiction to make independent decisions tailored to its specific context. This approach is compared with the SARL approach, where all jurisdictions are treated as a single entity, and a centralized decision is made for all of them. Experiments demonstrate the superior performance of MARL compared to SARL, resulting in a significantly greater reduction in HIV incidence. This can be attributed to the independent decision making power of agents in MARL in light of disparity among various jurisdictions, compared to a single decision in SARL that overlooks the jurisdictional characteristics in HIV infections unique to each jurisdiction.

A Appendix

A.1 Hyperparameters

Table A1: List of hyperparameters used in PPO

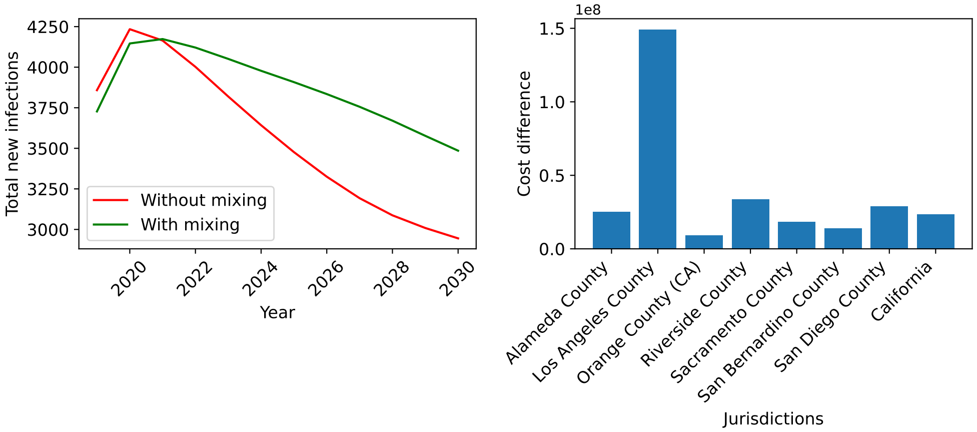
Hyperparameters	Value
Discount factor	0.99
Actor learning rate	0.0003
Critic learning rate	0.0003
Initial exploration	0.4
Final exploration	0.05
Decay rate	0.0046
Decay frequency	1000
PPO clip parameter	0.2
K epochs	20

A.2 Mixing Assumptions

Figure A.1 a) shows the incidence estimates considering mixing assumptions compared to the scenario without mixing where each jurisdiction is trained on its own. As the network is trained in the single agent and tested with mixing considered among all the jurisdictions, we can see individual jurisdictions model underestimates the total incidence numbers. Therefore, mixing assumptions across the interacting jurisdictions should be considered to get a better estimates of incidence in any jurisdiction. This also justifies the need of a combined modeling as a single jurisdiction (or city) model might not be able to capture the proper estimates of new infections.

Table A2: Sexual partnership mixing among jurisdictions Tatapudi and Gopalappa [2022]

Jurisdiction interaction	HM	HF	MSM
Same jurisdiction	57%	65%	47%
Other jurisdiction in same state	28%	23%	31%
Other states	14%	12%	22%



a)

b)

Figure A.1: Comparing a) new incidence with and without mixing and b) cost difference (without mixing - with mixing) in California.

A.3 Reward

Figures A.2 and A.4 show the learning curve for MARL and SARL approaches respectively for jurisdictions in California. Figures A.3 and A.5 show the learning curve for MARL and SARL approaches respectively for jurisdictions in Florida.

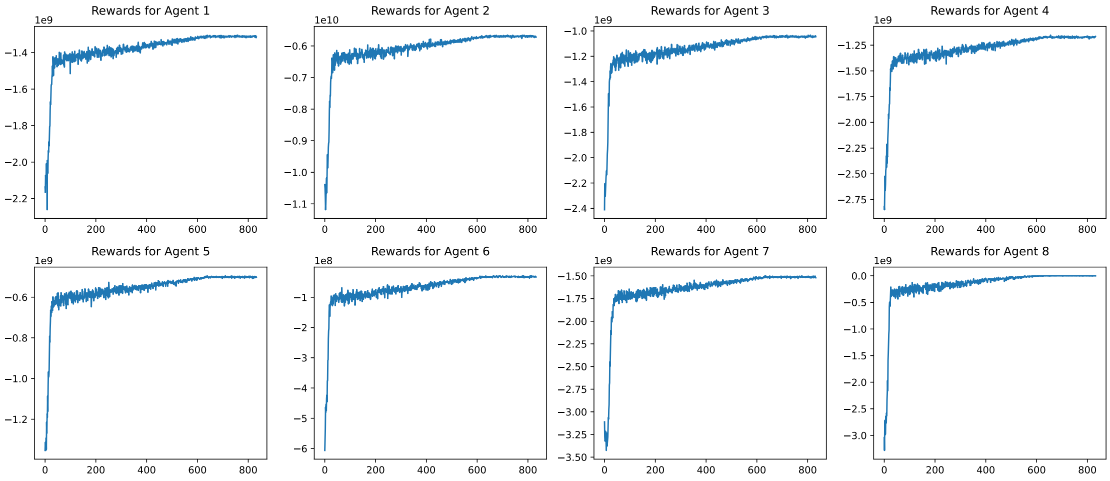


Figure A.2: Reward curve for MARL agents in California.

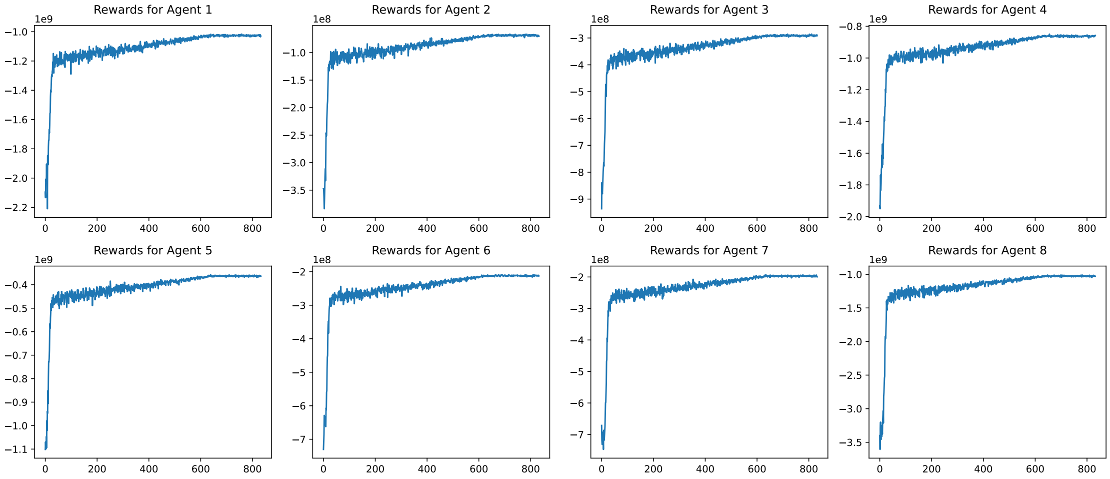


Figure A.3: Reward curve for MARL agents in Florida.

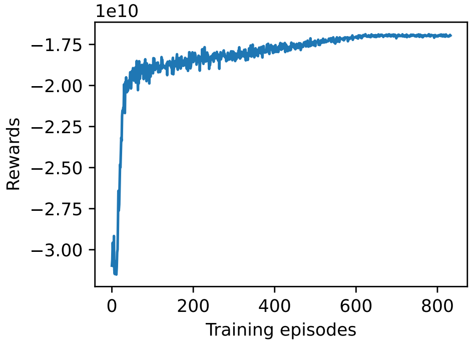


Figure A.4: Reward curve for SARL agent in California.

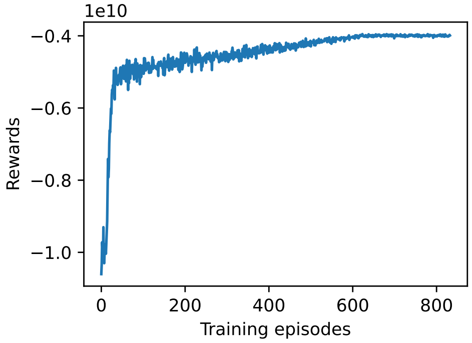


Figure A.5: Reward curve for SARL agent in Florida.

A.4 New incidence

Below is the overall incidence number for all the jurisdictions in each of the states, California and Florida.

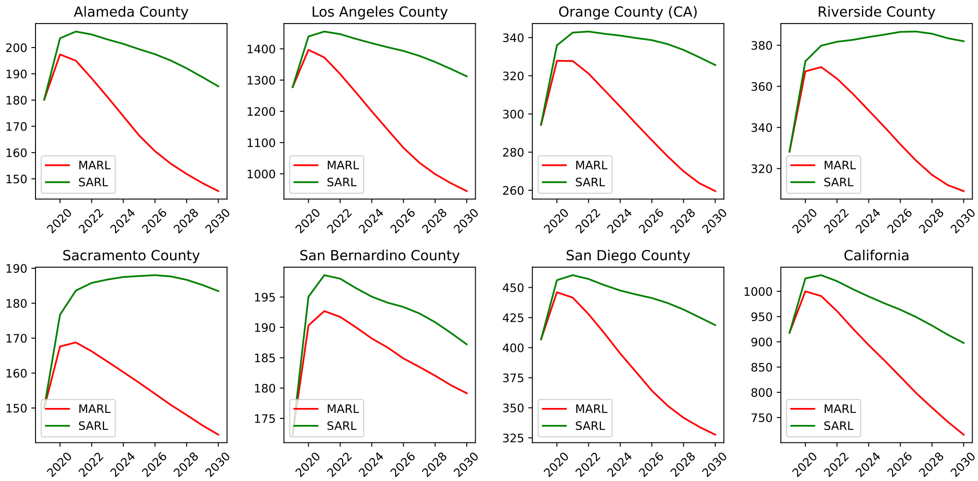


Figure A.6: Comparing new incidence with SARL and MARL in California.

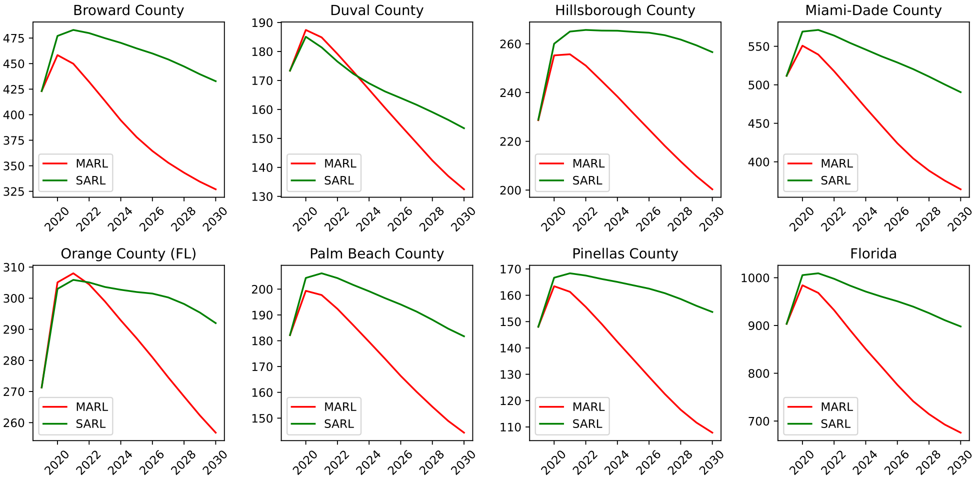


Figure A.7: Comparing new incidence with SARL and MARL in Florida.