Mathematical Biosciences and Engineering

MBE, 18(6): 7666–7684.
DOI: 10.3934/mbe.2021380
Received: 12 July 2021
Accepted: 25 August 2021
Published: 06 September 2021

http://www.aimspress.com/journal/MBE

Research article

A reinforcement learning model to inform optimal decision paths for HIV elimination

Seyedeh N. Khatami and Chaitra Gopalappa*

Mechanical and Industrial Engineering Department, University of Massachusetts Amherst, Amherst, MA 01003, USA

* **Correspondence:** Email: chaitrag@umass.edu; Tel: +1-4135452306.

Abstract: The 'Ending the HIV Epidemic (EHE)' national plan aims to reduce annual HIV incidence in the United States from 38,000 in 2015 to 9300 by 2025 and 3300 by 2030. Diagnosis and treatment are two most effective interventions, and thus, identifying corresponding optimal combinations of testing and retention-in-care rates would help inform implementation of relevant programs. Considering the dynamic and stochastic complexity of the disease and the time dynamics of decision-making, solving for optimal combinations using commonly used methods of parametric optimization or exhaustive evaluation of pre-selected options are infeasible. Reinforcement learning (RL), an artificial intelligence method, is ideal; however, training RL algorithms and ensuring convergence to optimality are computationally challenging for large-scale stochastic problems. We evaluate its feasibility in the context of the EHE goal. We trained an RL algorithm to identify a 'sequence' of combinations of HIV-testing and retention-in-care rates at 5-year intervals over 2015-2070 that optimally leads towards HIV elimination. We defined optimality as a sequence that maximizes quality-adjusted-life-years lived and minimizes HIV-testing and care-and-treatment costs. We show that solving for testing and retention-in-care rates through appropriate reformulation using proxy decision-metrics overcomes the computational challenges of RL. We used a stochastic agent-based simulation to train the RL algorithm. As there is variability in support-programs needed to address barriers to care-access, we evaluated the sensitivity of optimal decisions to three cost-functions. The model suggests to scale-up retention-in-care programs to achieve and maintain high annual retention-rates while initiating with a high testing-frequency but relaxing it over a 10-year period as incidence decreases. Results were mainly robust to the uncertainty in costs. However, testing and retention-in-care alone did not achieve the 2030 EHE targets, suggesting the need for additional interventions. The results from the model demonstrated convergence. RL is suitable for evaluating phased public health decisions for infectious disease control.

Keywords: decision-making in epidemics; agent-based simulation modeling; national HIV goals; reinforcement learning, ending the HIV epidemic; artificial intelligence in public health

Abbreviations: EHE: 'Ending the HIV Epidemic'; RL: Reinforcement learning; HIV: human immunodeficiency virus; US: United States; PWH: people living with HIV; ART: antiretroviral therapy; NHSS: National HIV Surveillance System; MDP: Markov decision process; PATH 2.0: Progression and Transmission of HIV/AIDS; HETs: heterosexuals; MSM: men who have sex with men; QALYs: quality-adjusted life-years; DP: dynamic programming

1. Introduction

The human immunodeficiency virus (HIV) continues to persist as a major public health issue in the United States (US), with about 1.2 million people living with HIV (PWH) as of 2015 and about 38,000 becoming newly infected each year [1]. The 2019 'Ending the HIV Epidemic (EHE)' US national strategic plan aims to reduce new infections by about 75% (to 9300 cases) by 2025 and by about 90% (to 3000 cases) by 2030 [2,3], by scaling-up four strategies, diagnose, treat, prevent, and respond [2]. Diagnoses followed by antiretroviral therapy (ART) treatment are key interventions as, in addition to being therapeutic, they can reduce HIV transmissions by up to 100% [4]. Since the implementation of the first national strategic plan in 2010, national guidelines have recommended at least annual testing for high-risk populations [5] and treatment initiation immediately upon diagnosis. However, estimates from the National HIV Surveillance System (NHSS) indicate that actual testing is less frequent than recommended, e.g., 3 to 5 years among those diagnosed with HIV in 2015 [6,7]. Further, though an estimated 70 to 80% of persons diagnosed with HIV were linked to care-and-treatment upon diagnosis, only 48% were on ART treatment in 2015, indicating high rates of care drop-out [8].

In this study, we develop a model to identify an optimal sequence of combinations of testing and retention-in-care rates at every 5-year interval from 2015 to 2070, to reduce HIV incidence. Identifying optimal testing rates, which is the inverse of how often to test, helps inform testing guidelines and implement social support and outreach programs to enable uptake [9]. Identifying optimal retention-in-care rates, which is the proportion of persons in care at the end of the year from among those in care at the beginning of that year, helps inform social service and support programs necessary to reduce the current high rates of care drop-out [10]. Identifying risk-group specific testing and retention-in-care rate combinations help direct resources to relevant support programs. We use non-linear cost functions to assign fixed and variable costs of testing and retention-in-care. To model the need for additional outreach and social support services to address barriers to testing and sustained care-and-treatment [11–13], we assume that the variable unit costs increase non-linearly with testing and retention-in-care rates [14,15]. As the type of service program-needs and its effectiveness vary by population [16], we utilize varying cost functions to generate the uncertainty range in optimal decisions.

Previous studies that have evaluated combinations of testing and retention-in-care rates have identified the most cost-effective combinations through either use of comparative analysis to evaluate a few pre-selected scenarios in stochastic simulation models [17–19], use of static parametric optimization techniques to evaluate non-dynamic decisions suitable for short-term decision-making [20,21], or use of dynamic optimal control techniques that evaluate dynamic decisions but using deterministic

differential equation-based models [20–22]. Our approach evaluates dynamic decision sequences in a stochastic and dynamic agent-based simulation environment through formulating the problem as a Markov decision process (MDP) and solving using reinforcement learning (RL), an area of artificial intelligence [23–25]. This methodology enables both simulating the dynamic changes in the epidemic over time while also evaluating the corresponding dynamic changes in decisions over time in a stochastic environment to identify the most optimal sequence choices to reduce new infections. We use a previously validated Progression and Transmission of HIV (PATH 2.0) model [26], a dynamic stochastic agent-based model, to simulate the epidemic and evaluate the decisions. Previous RL models in HIV have focused on patient-level clinical decisions such as optimal treatment protocols [27,28]. Recent literature has seen an emergence in the use of RL for public health decision-making related to the COVID-19 pandemic, but they predominantly use deterministic equation-based model environments [29–34]. While deterministic models are suitable for diseases that spread easily, agent-based models are more suitable for capturing the individual-level interactions for slow-spreading diseases such as HIV [35]. However, a combination of agent-based models and RL creates computational challenges. The number of iterations of RL training needed to ensure convergence increases exponentially with the size of the possible choices (225^{11} possible sequences in this application), and agent-based models are computationally expensive (e.g., each iteration of PATH 2.0 takes about 30 minutes). To overcome this challenge, we reformulate the decision variables to use proportions unaware and on ART as proxies, proving its mathematical viability, which reduces the number of choices to 36^{11}. To the best of our knowledge, this is the first model to evaluate the EHE goal of HIV elimination as a sequential decision-making problem in a stochastic dynamic environment and is naturally suited for informing the sequential goals of the US national plan.

2. Methodology

An MDP is a stochastic formulation of a decision-making problem, and RL is a machine learning methodology that uses 1) a simulation model to evaluate a policy (sequence of decisions) and 2) a control optimization algorithm to control the selection of policies to evaluate [36]. We used the Progression and Transmission of HIV/AIDS (PATH 2.0), a stochastic dynamic agent-based simulation model [26], and Q-learning RL algorithm [36,37]. We describe the MDP formulation in Section 2.1, and the RL algorithm in Section 2.22.

2.1. Mathematical formulation of the decision-making problem as an MDP

Let epidemic state at time t be a multivariate parameter $X_t = \left[\ p_i, \mu_{u,i}, \mu_{a,i}, \mu_{ART,i}; \forall i\right]_t$, where

- p_i is the HIV prevalence calculated as the number of people living with HIV (PWH) in risk group i divided by the total number of people in the population; we modeled two risk groups, heterosexuals (HETs) and men who have sex with men (MSM), thus $i \in \{$ HETs, MSM$\}$, and
- $\mu_{u,i}$, $\mu_{a,i}$, $\mu_{ART,i}$ are the proportions of PWH in risk group i that are unaware of their infection, aware of their infection but not on ART, and aware and on ART, respectively, and $\mu_{u,i} + \mu_{a,i} + \mu_{ART,i} = 1;\ \forall i$, thus representing all stages along the care-continuum.

Note that, for HIV prevalence p_i, we use the total population size as the denominator instead of the commonly used public health definition that uses the population size in that specific risk group i as

the denominator. This modification makes the state space mutually exclusive and collectively exhaustive, a necessary property for an MDP.

Let the intervention decision at time t be a multivariate parameter $D_t = [\delta_i, (1-\rho_i); \forall i]_t$, where δ_i is the diagnostic rate and $(1-\rho_i)$ is the retention-in-care rate in risk group i, where $i \in \{$ HETs, MSM$\}$.

Then, $\{X_t, D_t\colon\ t = 0\colon \mathrm{T} = 2015, 2020, 2025, \ldots, 2070\}$ is an MDP defined with a 4-tuple $\{\Omega, \mathrm{A}, P_a, R_a\}$, where,

- Ω is the state space, defined as the set of all possible states of the epidemic, i.e., $\Omega = \{[\ p_i, \mu_{u,i}, \mu_{a,i}, \mu_{ART,i}; \forall i \in \{\ \mathrm{HETs}, \mathrm{MSM}\}]\}$ and $X_t \in \Omega$,
- A is the action space, defined as the set of all possible decisions (referred to as actions in MDP terminology), i.e., $A = \{[\delta_i, (1-\rho_i), \forall i \in \{\mathrm{HETs}, \mathrm{MSM}\}]\}$, and $D_t \in \mathrm{A}$,
- P_a is the one-step transition probability matrix under action a, with element $P_a(x, x')$ being the probability that the epidemic transitions from state $X_t = x$ to $X_{t+1} = x'$ when action a is taken, and
- R_a is the immediate reward matrix under action a, with element $R_a(x, x')$ being the immediate reward (total benefits minus total costs) of taking action a when the epidemic is in state x and, as a result, it transitions to state x'; we model costs as intervention costs and benefits as the total quality-adjusted life-years (QALYs) lived in the population.

Note that the epidemic at any time t can be represented by one and only one *state*, and the probability of transitioning to an epidemic state x' at time $t+1$ is only dependent on the epidemic state x at time t, i.e., $Pr\{X_{t+1}|X_t, X_{t-1}, X_{(t-2)}, \ldots\} = Pr\{X_{t+1}|X_t\}$, thus satisfying the necessary Markov property for the MDP. Also, note that we use $t = [0\colon \mathrm{T}] = [2015, 2020, 2025, \ldots, 2070]$ to denote that we evaluate decisions at every five-year interval (consistent with the decisions made in the EHE national strategic plan). The initial year is $t = 0 = year\ 2015$, and thus, the first decision-making interval is for the period 2016 to 2020. We chose 2015 as the start year because, at the time of model development, the latest surveillance data available for HIV was 2016.

The objective is to identify the optimal decision $\boldsymbol{d} \in [d_1, .., d_T]$ (referred to as an optimal *policy* in MDP terminology) that maximizes the expected reward, i.e.,

$$\boldsymbol{d} = \underset{[d_1,..,d_T]\epsilon A^T}{arg\,max\,\mathbb{E}} \left[\sum_{t=1}^{t=T} \left(\gamma^t R_{a=d_t}(\ x, x')\right)\right] \tag{1}$$

where, γ is the discounting factor, and $\mathbb{E}[.]$ is the expected value. Thus, $\boldsymbol{d}$ is the sequence of optimal actions at 5-years intervals over the period 2016 to 2070. Conceptually, Eq (1) suggests that the decision d_t at every decision-making epoch t is evaluated not just based on its costs and impacts during the current epoch ($R_{a=d_t}(\ x, x')$), but is also based on the costs and impacts of decisions that would have to be made in all future decision epochs ($[d_{t+1}, d_{t+2}, \ldots d_T]$) to eliminate HIV, while also optimizing those future decisions. Intuitively, a policy that leads to zero new infections will be optimal if it has the lowest future costs and the highest benefits (QALYs), though it may have higher immediate costs. Under this objective function, it is necessary not to discount future costs and benefits as discounting would diminish the weight given to infections averted and costs prevented in the future. Thus, discounting may not identify strategies that lead to HIV elimination. Therefore, we set $\gamma = 1$.

The problem of solving for the optimal policy $\boldsymbol{d}$ can be visualized as a decision-tree (Appendix Section 1 Figure A1), with the *epidemic state* in 2015 being the start node, *actions* (a) being the decision nodes, *epidemic state* the decision transitions it to in next 5-year interval being the chance nodes (with the probability of transition defined by P_a and value of the outcome defined by R_a), the possible *epidemic states* in 2070 being the end nodes, and a *policy* being a path or a sequence of decisions in the decision-tree. Analytically estimating the outcome of each decision path is complex as the dynamics of the system makes estimating P_a and R_a complex. It is also sometimes impractical because of the large dimensions of the state space and action space, as will be seen in below formulation of the 4-tulpe $\{\Omega, \mathrm{A}, P_a, R_a\}$. Therefore, we solve for the optimal policy $\boldsymbol{d}$ using RL (see Section 2.2).

We next discuss the formulation of each element of the 4-tulpe $\{\Omega, \mathrm{A}, P_a, R_a\}$.

State space

We formulate the state space $\Omega = \{[\ p_i, \mu_{u,i}, \mu_{a,i}, \mu_{ART,i}; \forall i \in \{\ \mathrm{HETs}, \mathrm{MSM}\}]\}$ as a finite state space by discretizing each of its elements as follows,

$$p_{HET} \in \begin{Bmatrix}[0,0.0005), [0.0005,0.0015), [0.0015,0.0025), [0.0025,0.0035), \\ [0.0035,0.0045), [0.0045,0.0055), \geq 0.0055\ ]\end{Bmatrix},$$

$$p_{MSM} \in \begin{Bmatrix}[0,0.0005), [0.0005,0.0015), [0.0015,0.0025), [0.0025,0.0035), \dots, \\ [0.0085,0.0095), [0.0095,0.015), \geq 0.015\ ]\end{Bmatrix},$$

$$\mu_{u,\mathrm{HET}} \in \{[10\%, 11.25\%), [11.25\%, 13.75\%), [13.75\%, 16.25\%), < 10\% \cup \geq 16.25\%\ ]\},$$

$$\mu_{u,\mathrm{MSM}} \in \{[10\%, 11.25\%), [11.25\%, 13.75\%), [13.75\%, 16.25\%), [16.25\%, 18.75\%), < 10\% \cup \geq 18.75\%\ ]\},$$

$$\mu\_(ART, i) \in \{[85\%, 95\%), [75\%, 85\%), [65\%, 75\%), [55\%, 65\%), [45\%, 55\%), \geq 95\% \cup < 45\%\},$$

$$\mu_{a,i} = \ 1 - \mu_{u,i} - \mu_{ART,i},$$

thus, making the state a categorical value instead of a numerical value, e.g., the epidemic is assigned the state

$$[[0,0.0005), [0.0085,0.0095), [11.25\%, 13.75\%), [10\%, 11.25\%), [45\%, 55\%), [55\%, 65\%)]$$

when

$$p_{HET} \in [0,0.0005),\ p_{MSM} \in [0.0085,0.0095),\ \mu_{u,HET} \in\ [11.25\%, 13.75\%),$$

$$\mu_{u,MSM} \in [10\%, 11.25\%),\ \mu_{ART,HET} = [45\%, 55\%),\ \mu_{ART,MSM} = [55\%, 65\%).$$

Note that we exclude $\mu_{a,i}$ in the notation as it is redundant and can be calculated as

$$\mu_{a,i} = \ 1 - \mu_{u,i} - \mu_{ART,i}.$$

Further, note that any state with at least one element in the open range (last element in each of the discretized ranges of $p_i, \mu_{u,i}, \mu_{a,i}, \mu_{ART,i}$) is either not desirable, i.e., takes the epidemic to a worse

state than current, or exceeds the feasibility constraint (see Constraints below), and thus, should be avoided. Therefore, we combine those into a single state, say Δ. The epidemic is assigned the state Δ when $p_{HET} \geq 0.0055$ or $p_{MSM} \geq 0.015$ or $\mu_{u,HET} < 10\%$ or $\mu_{u,HET} \geq 16.25\%$ or $\mu_{u,MSM} < 10\%$ or $\mu_{u,MSM} \geq 18.75\%$ or $\mu_{ART,HET} < 45\%$ or $\mu_{ART,MSM} < 45\%$ or $\mu_{ART,HET} \geq 95\,\%$ or $\mu_{ART,MSM} \geq 95\%$.

As state Δ is representative of all epidemic states we want to avoid, we associate it with a very large cost, such that any action that would take the epidemic to that state would have a high negative reward and thus be marked as a bad decision. The upper bounds on $p_i, \mu_{u,i}$ and lower bounds on $\mu_{ART,i}$ are set to values in 2015 to indirectly constrain the decisions to lead to a better epidemic state. Considering all the possible combinations of the discretized values, noted above for $p_{HET}, p_{MSM}, \mu_{u,HET}, \mu_{u,MSM}, \mu_{ART,HET}, \mu_{ART,MSM}$, there are a total of 16,500 ($= 5 \times 11 \times 3 \times 4 \times 5 \times 5$) *states* excluding Δ. Hence, the final size of the state space is $|\Omega| = 16{,}500 + 1$.

Action space

Instead of directly formulating an action as a combination of diagnostic rate (δ_i) and retention-in-care rate $(1 - \rho_i)$, which is the decision of interest here, we formulate it using changes in proportions unaware and on ART as a proxy, i.e., instead of using action space as $\mathrm{A} = \{[\delta_{HET}, \delta_{MSM}, (1 - \rho_{HET}), (1 - \rho_{MSM})]\}$ we use a proxy as

$$A = \{[a_{unaware,HET}, a_{unaware,MSM}, a_{ART,HET}, a_{ART,MSM}]\},$$

where, $a_{unaware,i}$ is the percentage decrement in $\mu_{u,i}$ (the proportion unaware in risk group i), and $a_{ART,i}$ is the percentage increment in $\mu_{ART,i}$ (the proportion on ART in risk group i).

We formulated the action space 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 the RL algorithm. We discuss these mathematical properties through four Remarks as follows.

Remark 1: Given the system state x at time $\mathrm{t} - 1$, $(X_{t-1} = x)$, corresponding to every action $a_{unaware,i}$, there is a unique diagnostic rate (δ_i) and corresponding to every action $a_{ART,i}$, there is a unique retention-in-care rate $(1 - \rho_i)$.

Remark 2: From a public health perspective, all actions that transition the epidemic state to a higher proportion unaware ($\mu_{u,i}$) or to a lower proportion on ART ($\mu_{ART,i}$), compared to its current state, are undesirable and should not be selected.

Remark 3: Setting the action space to use $a_{unaware,i}$ and $a_{ART,i}$ instead of diagnostic rate and retention-in-care rate, respectively, helps efficiently control the number of possible actions (interventions) and thus is more computationally efficient.

Remark 4: Setting the action space to use $a_{unaware,i}$ and $a_{ART,i}$ instead of *changes* in the diagnostic rate and retention-in-care rate, respectively, over two consecutive decision-making time-steps, also helps efficiently control the number of possible actions and thus is more computationally efficient.

We support Remarks 1 to 4 through proofs in the Appendix Section 3. Briefly, for any given epidemic state, corresponding to every combination of $a_{unaware,i}$ and $a_{ART,i}$, there is a unique combination of diagnostic and retention-in-care rates, which essentially implies that our formulation of *action* as $\mathrm{A} = \{[a_{unaware,i}, a_{ART,i}; \forall \mathrm{i}]\}$ would yield the same results as the more direct metrics of $\mathrm{A} = \{[\delta_i, (1 - \rho_i); \forall i]\}$ (Remark 1). In fact, for evaluating the proxy action in the simulation, we first

estimate the diagnostic and retention-in-care rates (see Appendix Section 2) and use that as input to the simulation. We use this estimation method, which generates functional expressions between δ_i and $a_{unaware,i}$ and between $(1-\rho_i)$ and $a_{ART,i}$ through the representation of the system as a differential equations model (see Appendix Section 2), followed by showing that the functional expressions are bijection functions (see Appendix Section 3) to prove Remark 1. Note that the proxy action space elements ($a_{unaware,i}$ and $a_{ART,i}$) directly modify part of the state space elements ($\mu_{u,i}$ and $\mu_{ART,i}$, respectively), which gives the flexibility to, at any decision time-step, not choose actions that would take the epidemic to a state worse than the current (Remark 2), and thus, also avoid transitioning into the undesirable state Δ. If on the other hand, we choose δ_i and $(1-\rho_i)$ as elements of the action space, we prove that we have to run the simulation model (~30 mins per run) to evaluate what state that action would transition the system into (could be better than current, worse than current, or Δ) thus requiring many more evaluations (Remark 3). Remark 4 has a similar purpose as Remark 3 except that it evaluates the use of changes in the diagnostic rate and retention-in-care rate over two consecutive decision-making time-steps, as the proxies $a_{unaware,i}$ and $a_{ART,i}$ are also decrements or increments (of $\mu_{u,i}$ or $\mu_{ART,i}$, respectively) over two consecutive decision-making time-steps.

We assumed two possible choices for decrements in $\mu_{u,i}$, decrease by 0 or 2.5%, and three possible choices for increments in $\mu_{ART,i}$, increase by 0, 10, or 20%, each relative to their values at the time of decision-making. That is, we formulated the possible action choices as

$$a_{unaware,i} \in \{0\%, -2.5\%\} \text{ and } a_{ART,i} \in \{0\%, 10\%, 20\%\}, \ \forall i \in \{\text{HETs}, \text{MSM}\}$$

resulting in 36 possible *actions* (2×3 for HETs $\times$ 2×3 for MSM) to choose from at every 5-year decision interval between 2015 and 2070, and thus, 36^{11} possible decision sequences. On the other hand, if we had directly formulated an action as changes in testing and retention-in-care rates, we would in the least have 225 action choices, and thus 225^{11} possible sequences (see Appendix Section 3). The size of the action space can exponentially increase the number of RL iterations for convergence (see convergence discussion in next section) and thus becomes infeasible to model or guarantee convergence.

For public health decision-making and implementation, testing rates and retention-in-care rates are more meaningful. Therefore, in Results, we present both metrics, the changes in proportions unaware and on ART, and the direct metrics of testing rate, estimated from the simulation as the inverse of the time from infection to diagnosis (a proxy for how often to test), and retention-in-care rate, estimated from the simulation as the proportion retained in care for the entire year among those in care at the beginning of that year.

Transition probabilities and immediate rewards

Generating the full one-step transition probability matrices P_a and reward matrices R_a is infeasible considering the size of the state space and action space. Therefore, we used PATH 2.0 [26], discussed later under Simulator, to simulate *actions* and stochastic *transitions*, track corresponding *states* it transitions to, and estimate *immediate rewards.* We estimated the *immediate rewards* as benefits minus costs. We model benefits as the total population quality-adjusted-life-years (QALYs) lived converted to a monetary value by multiplying with the US gross domestic product (GDP) per capita of $54,000 to denote the economic value added for every QALY lived [9,38,39]. Costs include total population costs for HIV testing, care, and treatment. Specifically, we estimated $R_a(x, x') = (c_l L_t - C_t)$, where,

c_l = cost per QALY lived, a health utility measure to control for the willingness to pay for every QALY lived; here we assumed it is equal to \$54,000, the GDP per capita in the US in the year 2015,

L_t = sum of QALYs of all people in the population at decision-making epoch t,

C_t = sum of HIV-related costs and intervention costs at decision-epoch t.

We used the PATH 2.0 simulation model for the estimation of C_t and L_t. We present the estimation of the unit-costs corresponding to the interventions in Appendix Section 4.

Constraints

We set the following constraints, which can be interpreted as cost or feasibility constraints:

a) The maximum possible decrement in proportion unaware and the maximum possible increment in proportion on ART, achievable in a 5-year interval, were set at 2.5 and 20%, respectively, as evident from the choice of *actions*. For reference as to the feasibility of achieving these maximum scale-ups, between 2010 and 2015, there was about a 2.3% decrease in proportion unaware (17.3% in 2010 to 15% in 2014) and a 13.8% increase in proportion on ART (46% in 2010 to 59.8% in 2015 [40,41]).

b) The maximum proportion aware and maximum proportion on ART were set at about 90 and 95% respectively, which are the targets typically aimed in national and global strategic plans [2,42].

2.2. Reinforcement learning algorithm to identify optimal policy

Reinforcement learning (RL) and dynamic programming (DP) are commonly used algorithms to solve MDP problems. Applying DP guarantees convergence to the optimal solution; however, it requires estimating, under each action, a probability matrix of transitions between all states. In the case of large-scale problems such as our current application to HIV, estimating the transition probability matrices for all states and actions is computationally infeasible. This curse of dimensionality makes DP suitable for only small-scale problems [36]. Therefore, we use Q-learning, a machine learning control optimization algorithm that uses an iterative feedback and control process to identify the optimal policy. Thus, it does not require a priori knowledge of transition probability matrices and is known to converge to near-optimal solutions [36].

Q-learning algorithm

In this study, we use PATH 2.0 ('simulator') to simulate a specific *action* and train the Q-learning algorithm. The simulator returns as 'feedback' to the Q-learning algorithm ('optimizer'), the *immediate reward* of the action simulated, and the epidemic *state* it transitions to at the end of the 5-year period. The optimizer tracks the *total reward* for the period 2015 to 2070 by summing the *immediate reward* of each 5-year *action* while also tracking the epidemic states visited. It then controls what action should be next taken by observing the total reward of previous actions and sending that decision back to the simulator [36,43,44]. By repeating this iterative process a large number of times (training the Q-learning algorithm), the optimizer learns to pick a better action each time to eventually find the optimal decision.

The optimizer tracks total reward as Q-values,

$$Q_{k+1}(x,t,a) = (1-\alpha)Q_k(x,t,a) + \alpha\left[R_a(x,x') + \gamma \max_{b\in A(x',t+1)} Q_k(x',t+1,b)\right] \quad (2)$$

at every decision epoch t, given system state as x, selects the action $(a(x,t))$ to simulate using a

decaying epsilon greedy method,

$$a(x,t) = \begin{cases} random\ selection\ from\ action\ space\ A\ with\ probability\ \epsilon_t \\ arg\max_{b \in A(x,t)} Q_k(x,t,b)\ \ with\ probability\ (1-\epsilon_t) \end{cases} \quad (3)$$

where k is the iteration of the Q-learning algorithm, α is a learning rate, γ is the discounting factor, which was set to 1 here, and

$$\epsilon_t = \begin{cases} \frac{0.85}{k} + 0.049 & k \leq 4000 \\ \frac{0.85}{k-3500} + 0.049 & k > 4000 \end{cases}.$$

Typically, ϵ_t is set to decrease as k increases, thus balancing more exploration of random actions in the beginning and exploitation of the greedy actions in future iterations. We additionally defined ϵ_t to explore at a higher rate when $k > 4000$ to test for convergence. We initialize the Q-values to some constant (C), i.e., $Q_k(x,t,a) = C, \forall (x,t,a)$, for $k = 0$. We then iterate over k and all decision epochs within each iteration to update $Q_{k+1}(x,t,a)$ using immediate rewards $R_a(x,x')$ returned by the simulator at every decision epoch t after simulating action a. The algorithm is known to converge to near-optimal solutions when k becomes a sufficiently large value [36]. That is, the optimal action $(say\ a^*(x,t))$ to be taken at time "t" when the system is in state "x" is defined as $a^*(x,t) \in arg\max_{b \in A(x,t)} Q_{k_{max}}(x,t,b)$. The schematic of the above iterative training process and summary of the above training steps of the algorithm are shown in Appendix Section 5 (Figure A3 and Table A4 respectively).

The optimal policy $\boldsymbol{d} = [d_0, d_1, \ldots d_T]$ would then be identified by also using PATH 2.0 simulation along with trained Q-values ($Q_{k_{max}}$). Specifically, we set $d_0 = a^*(x_0, 0)$, where x_0 =epidemic state in the year 2015, simulate a^*, and say, the epidemic transitions to state x_1 in 2020, then set $d_1 = a^*(x_1, 1)$, simulate that action, and continue this iterative process until T. As the simulation is stochastic, we repeat this process 100 times to generate an uncertainty range.

Simulator- Progression and Transmission of HIV (PATH 2.0) model

PATH 2.0 is an agent-based stochastic simulation model that individually tracks HIV-infected persons by simulating HIV disease progression through a health-state transition model and sexual transmissions of HIV through a novel dynamic transmission model. PATH 2.0 is calibrated to be representative of the US HIV epidemic and has been validated to accurately simulate the epidemic for the years 2010 through 2015. Details of the model, its validation, and its adoption to inform HIV-related decisions in the US are presented elsewhere [26,38].

We used the PATH 2.0 model to simulate an *action,* $a(x,t)$, selected by the Q-learning algorithm at a decision epoch t given system state x, simulate *state transitions,* and calculate the corresponding *immediate reward* $R_a(x,x')$ by assigning QALYs (L_t) and costs (C_t) to every person in the simulation. Specifically, for the selected proxy action ($a = \left[a_{unaware,i}, a_{ART,i}; \forall \mathrm{i}\right]$), we estimate diagnostic and retention-in-care rates (Appendix Section 2), and use PATH 2.0 to simulate diagnosis, care-and-treatment, and based on the care status of an infected person, simulate transmissions to their

susceptible partners, thus transitioning the epidemic to a new state x'. We assign a QALY of 1 per year if the person is healthy, between 0 and 1 if HIV-infected (varying based on disease stage), and 0 if deceased [26,38]. Based on estimated diagnostic and retention-in-care rates and the effectiveness of the intervention programs, we estimate the number of persons intervened for each program and corresponding costs, which are discussed in more detail in Appendix Sections 2 and 4, respectively. Briefly, in estimating testing and retention-in-care costs, we made the following assumptions based on data from intervention programs [9,39,45]. HIV testing programs can be conducted in clinical or non-clinical settings, each having its own fixed and variable costs [9,39]. Some people get tested voluntarily and incur only the cost of testing, while some get tested as a result of an outreach intervention and thus incur additional outreach costs [9], which we modeled as a non-linear function of the number of people outreached [15,46]. In accordance with current CDC recommendations, we assumed only persons with high-risk are recommended for regular testing and intervened through outreach programs and that 6% of heterosexual females, 10% of heterosexual males, and all MSM are high-risk populations [47–49]. We assumed a non-linear variable cost function for retention-in-care to model the additional support programs needed to retain a larger proportion of people in care. Details of intervention costs are included in Appendix Section 4.

In summary, we assumed that the first decision-making year is 2015 for the period 2016 to 2020, and decisions are made at every 5-year interval and solved for a decision sequence that optimally reaches close to zero new infections by 2070, i.e., in the Q-learning algorithm $t \in \{2015, 2020, \dots, 2070\}$. We used 2015 as the initial year as per the latest data available at the time of model development [26, 38]. However, the time-step of the simulation is monthly. Every iteration (k) of the Q-learning algorithm consists of simulating the PATH 2.0 model from 2015 to 2070 in monthly time-steps within a feedback and control loop to update Q-values and determine actions to be simulated every 5-years. Repeating this process for a large number of iterations, the optimizer learns to pick a better action at each iteration eventually *converging* to an optimal *policy*. The model is coded in NetLogo 6.0.2 software [50].

Evaluating convergence of Q-learning algorithm to optimal policy

An algorithm has converged if it has reached a local optima through the iterative search process, i.e., successfully solved for an optimal combination of testing and retention-in-care rates. If the number of iterations is not sufficiently large, there is a risk that the algorithm is terminated before convergence. The ideal number is typically determined through experimentation. Further, there could be multiple local optima, i.e., multiple policies could yield similar total rewards, and because of the stochastic nature of the epidemic system, the optimal policy could be a range rather than a point estimate. Therefore, we ran the model for varying number of iterations, 2000, 3000, 4000, and 5000, and compared the corresponding total rewards (Appendix Section 6, Figures A4 to A13), to ensure convergence and obtain the uncertainty range in optimal policies. The relative difference in the average costs and QALYs between the varying iterations were at most 2% in each cost function evaluated (see cost functions in Uncertainty Analysis Section 3), suggesting convergence. The corresponding optimal policies differed slightly, more so in future years than earlier years, suggesting stochastic uncertainty as the model projects further into the future. Therefore, in Results, we present the range of optimal policies across these iterations as the uncertainty range.

3. Uncertainty analysis

We modeled two types of uncertainty:

1) The inherent stochasticity in the epidemic system is modeled through: a) the use of PATH 2.0, which is a stochastic simulation model where input parameters are drawn from probability distributions and events simulated using stochastic functions; b) the use of MDP with Q-learning, which is a stochastic control optimization method and; c) the use of varying numbers of MDP iterations (2000 to 5000), and simulating the optimal policy from each iteration a 100 times to generate the average values for output metrics.

2) The uncertainty in intervention costs is modeled by using three different cost functions, each with varying assumptions for the following four unit-costs a) the fixed cost per clinic for a retention-in-care program, b) the variable cost per person for a retention-in-care outreach program, c) the marginal increase in variable cost for a retention-in-care outreach program, and d) the marginal increase in variable cost for a testing outreach program [49]. These four parameters were chosen as they are related to support programs, which tend to have more variability as the type of support needed varies by individual-needs. As the model attempts to find the optimal balance in testing and retention-in-care rates, we selected a median and alternating bounds of the above four unit cost range to generate the overall range in optimal policy from uncertainty in costs. Therefore, we have three cost-functions as follows:

Median (Median Testing and Retention-in-care Costs): Uses the median values for all four parameters.

LTHR (Low Testing High Retention in Care Costs): Uses the lowest value for the testing costs and the highest value for the retention-in-care costs.

HTLR (High Testing Low Retention in Care Costs): Uses the highest values for the testing costs and the lowest values for the retention-in-care costs.

In summary, for each cost-function assumption, we trained the Q-learning algorithm with multiple stopping conditions (2000, 3000, 4000, and 5000 iterations). Using the trained Q-values ($Q_{k_{max}}$), for each cost function and stopping condition pair, we simulated 100 runs and extracted the average values (over the 100 runs) of the optimal policy and corresponding impacts generated, specifically, values for the testing rate, retention-in-care rate, proportion of people with HIV (PWH) aware of their infection, proportion of PWH on ART, number of new infections, number of PWH, and incremental total costs, which are useful metrics from a public health perspective.

4. Results

For the end of 2015, the PATH 2.0 simulation model estimated an annual testing rate of 0.26 for high-risk heterosexuals and 0.4 for MSM, i.e., an average time from infection to diagnosis of 3.8 years for heterosexuals and 2.5 years for MSM. These results closely match the CDC estimates for the time from infection to diagnosis in 2015 [6], estimated using data from the NHSS [7]. For the end of 2015, the model estimated an annual retention-in-care rate of 86% for heterosexuals and 91% for MSM. Figure 1a,b presents the optimal policy for the period 2016 to 2070, specifically the optimal combination of testing (bottom) and retention-in-care (top) rates over time for heterosexuals (and MSM). Figure 1c,d shows the corresponding proportions aware (top) and on ART (bottom) for

heterosexuals (and MSM). The figures present the uncertainty range (shaded bands) for each of the three cost function assumptions (Median: blue bands, LTHR: red bands, HTLR: green bands).

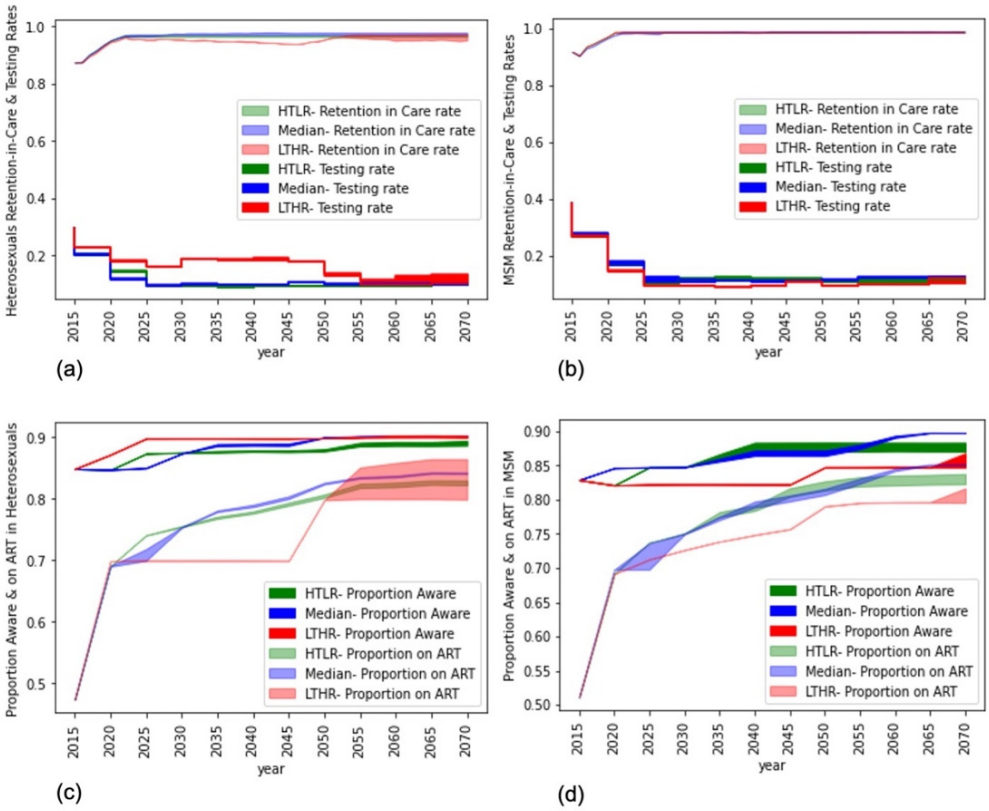

Figure 1. Optimal combinations of testing (top lines) and retention-in-care (bottom lines) rates for heterosexuals (1a) and MSM (1b) and corresponding proportion aware (top lines) and proportion on ART (bottom lines) for heterosexuals (1c) and MSM (1d) for three cost functions of HTLR (green), Median (Blue), and LTHR (Red). Results are an average of over 100 simulation runs of the optimal policy. The shaded region is the uncertainty around the optimal policy generated by the reinforcement learning algorithm.

For the period 2016 to 2020, under all three cost function assumptions, the model suggests a testing rate of 0.2 for high-risk heterosexuals (Figure 1a) and 0.3 for MSM (Figure 1b), equivalent to testing once every 5 and 3.5 years, respectively. Over the period 2016 to 2020, under all three cost function assumptions, the model suggests aggressive retention-in-care programs to gradually increase annual retention-rates from 86 to 94% for HETs (Figure 1a) and from 91 to 96% for MSM (Figure 1b). The uncertainty bands for this period, under all three cost function assumptions, for both testing and retention-in-care rates are narrow, suggesting a high chance that the algorithm has converged. Achieving the above testing and retention-in-care rates corresponded to about 85% of all heterosexual PWH and 82% of all MSM PWH being aware of their infection by the end of 2020, and about 70% of all heterosexual PWH and 70% of all MSM PWH on ART by the end of 2020 (Figure 1c,d).

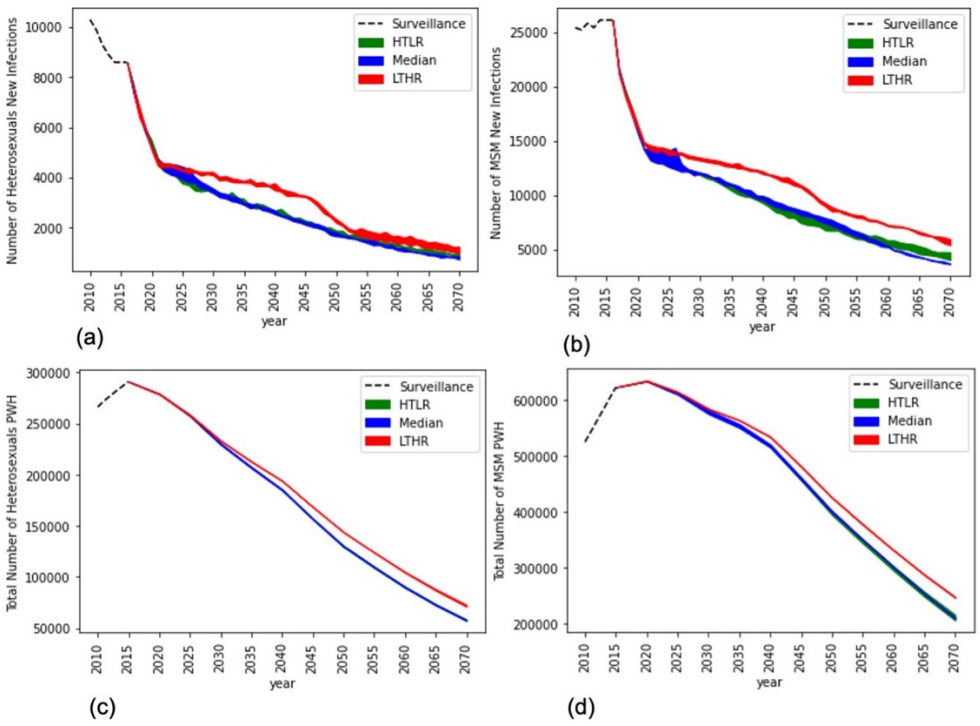

Figure 2. The number of new infections for heterosexuals (1a) and MSM (1b) and the number of people living with HIV for heterosexuals (1c) and MSM (1d) for the HTLR (green), Median (Blue), LTHR (Red) cost function assumptions. Results are an average of over 100 simulation runs of the optimal policy. The shaded region indicates the uncertainty range in the optimal policy.

Implementing the combination of testing and retention-in-care rates for the period 2016 to 2020 generated a 50% reduction in annual new infections among heterosexuals (from 9000 in 2016 to 4500 by the end of 2020) and a 42% reduction in annual new infections among MSM (from 26,000 in 2016 to 15,000 by the end of 2020) which is a significant reduction compared to trends over the previous 5-year period (Figure 2a,b). There was a modest decline in the number of heterosexual PWH, breaking the previous trend of growth (Figure 2c). There was modest-growth in the number of MSM PWH, which is a shift from the previous high-growth rate but suggests that the number of PWH will continue to increase for a short period before declining (Figure 2d). The annual cost of HIV increased by about 22% over this period, suggesting a high initial investment to achieve the above reduction in new infections (Figure 3).

For the period 2021 to 2025, compared to the previous 5-year period, the model suggests relaxing the frequency of testing while modestly increasing and maintaining a high retention-in-care rate for both heterosexuals and MSM and under all three cost function assumptions. Specifically, for heterosexuals, the model suggests testing rates of 0.14, 0.11, and 0.18 under the HTLR, Median, and LTHR cost assumptions, equivalent to a testing frequency of every 7, 9, and 5.5 years, respectively (Figure 1a). For MSM, the model suggests a testing rate of around 0.15 under all three cost function assumptions, equivalent to a testing frequency of every 6.5 years (Figure 1b). The model suggests scaling-up retention-in-care programs to increase annual retention-rates from 94 to 96% for heterosexuals (Figure 1a) and from 96 to 98% for MSM (Figure 1b) over the period 2021 to 2025. The reduction in new infections was modest compared to the previous 5-year period for both heterosexuals

(Figure 2a) and MSM (Figure 2b), but the number of PWH declined at a faster rate compared to the previous 5-year period for heterosexuals (Figure 2c), and for the first time there was a reduction in the number of MSM PWH (Figure 2d).

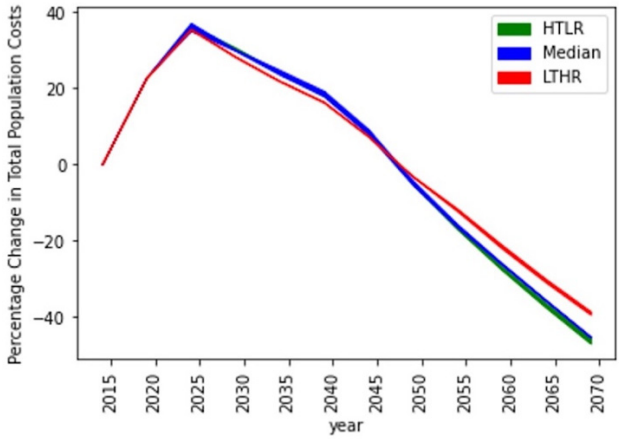

Figure 3. Changes in total population costs corresponding to the optimal policy under the three cost function assumptions, HTLR (green), Median (Blue), LTHR (Red).

For the period 2026 to 2030, the testing rate reduced to 0.1 (test once every 10 years or less) under Median and HTLR cost functions in heterosexuals (Figure 1a) and under all cost functions for MSM (Figure 1b) and continued to remain at those values for the remaining duration of the simulation (up to 2070). Under the LTHR cost function for heterosexuals, the testing rate continued to remain at 0.18 (test once every 5.5 years) until 2050 and then reduced to 0.1 for the remaining duration of the simulation. The corresponding proportion aware and on ART gradually increased to 80% (Figure 1c,d), and the corresponding number of new infections (Figure 2a,b) and PWH (Figure 2c,d) gradually decreased over the period 2026 to 2070 for both heterosexuals and MSM. For heterosexuals (Figure 2a), under the three cost function assumptions, the number of new infections reduced to about 3200 to 4000 cases per year by 2030 (53 to 62% reduction compared to 2015) and further reduced to 750 to 1200 cases per year by 2070 (86 to 91% reduction compared to 2015). For MSM (Figure 2c), under the three cost function assumptions, the number of new infections reduced to about 11,000 to 14,000 cases per year by 2030 (46 to 58% reduction compared to 2015), and further reduced to 3500 to 6000 cases per year by 2070 (80 to 86% reduction compared to 2015).

Comparing across the cost function assumptions, for heterosexuals, the optimal rates were generally intuitive, with the highest testing and lowest retention-in-care in LTHR and lowest testing and highest retention-in-care rates in HTLR, though the differences in retention-care rates were modest (Figure 1a). For MSM, however, the optimal annual retention-in-care rates were similar in all three cost functions, and the results were counter-intuitive for testing rate as the model suggested a slightly lower testing rate in LTHR compared to Median and HTLR. That is, some of the testing resources were shifted to retention-in-care programs to offset its higher costs while maintaining annual retention-rates at the level of Median and HTLR (Figure 1b). And as a result, the proportion of MSM on ART in

LTHR, though lower than in Median and HTLR, was higher than that of heterosexuals on ART in LTHR (Figure 1d), which suggests the higher significance of retention-in-care programs to ensure sustained care-and-treatment, compared to testing. The optimality of this counter-intuitive strategy in MSM was evaluated by a counterfactual simulation run using the optimal LTHR strategy of heterosexuals for both heterosexuals and MSM. The number of new infections, PWH, and costs were higher in the counterfactual simulation, confirming the optimality of the strategy. Details of this run and the results are presented in Appendix Section 7.

5. Discussion and conclusions

This paper proposes a methodology for phased-decision-making, which is typical in public health for epidemic control. We modeled the question of 'how to optimally reach HIV elimination in the US' as a sequential decision-making problem by formulating it as an MDP and solving it using a Q-learning algorithm. Compared to the most commonly used approach of explicitly evaluating a few pre-selected scenarios, our approach enables selecting an optimal from all possible choices (36^{11} possible choices) through probabilistic projections of the decisions and the epidemic. In identifying the optimal sequence of combinations of testing and retention-in-care rates, we took a societal perspective to evaluate costs and QALYs. Though the stochastic dynamic sequential decision-making models are very attractive for evaluations of phased decision-making, they are computationally expensive to solve as the state space and action space of epidemic control problems are typically large, giving rise to issues of non-convergence and thus limiting its applicability. In this study, we reformulated the action space by taking indirect metrics that significantly reduced the size of the action space, thus leading to a successful application. Though applied to HIV, the proposed approach can be used for other infectious diseases as typically testing and treatment are key methods to control spread.

Our analysis is subject to limitations. We constrained the maximum intervention scale-up to one time and two times the maximum scale-up achieved in the past five years for proportion aware and proportion on ART, respectively, to balance feasibility and aggressive scale-ups. We also did not consider the availability of new interventions in the future, such as vaccines. We did not consider preventive interventions such as pre-exposure prophylaxis or changes in sexual behaviors in future generations. We did not evaluate interventions specific to people who inject drugs. We did not consider changes in demographics over time. Limitations to any model-based analysis, such as any drawbacks from the quality of data and unavailability of data leading to parameter estimations rather than the use of actual data, also apply to the PATH 2.0 simulation model and are discussed elsewhere [26]. Data on costs of support services and outreach are limited, and thus, we made assumptions on its variability to evaluate its sensitivity to optimal decisions. The model is limited to current testing and treatment technologies, i.e., we did not consider availability of a cure or significant improvements in costs of testing and treatment. Availability of a cure can dramatically change the timeline for HIV elimination due to reduction in transmissions, and reductions in costs could lead to better trade-offs with GDP and thus changes in the optimal decision. However, note that, as the model outcome is in favor of allocating resources to treatment over testing though costs of treatment are much higher, we can expect this outcome to remain in the event of significant reductions in costs of treatment relative to costs of testing.

Despite these limitations, we believe the approach used in this paper in evaluating phased-decisions related to the two most effective interventions is very timely in light of the 'Ending the HIV

Epidemic' federal plan for an HIV elimination objective [2].

The optimal policies generally suggest more frequent testing for the first 10 years and less frequently after as the number of new infections decreases and the proportion aware increases. It suggests more frequent testing for MSM than heterosexuals for the first 10 years, which supports the risk-based testing proposed under current CDC guidelines [51], and similar testing rates for both risk groups thereafter. It suggests implementing retention-in-care programs to gradually increase annual retention-in-care rates to 95% within the first 10 years and maintain it at that rate thereafter. Generally, the model suggested aiming for a higher retention-in-care rate than the testing rate, suggesting prioritization of spending on retention-in-care programs. Optimal decisions were robust to uncertainty in costs, under the range assumed, except for MSM under LTHR. In this scenario, the model suggested taking advantage of the lower testing costs and maintaining a higher testing rate for a longer duration than in the Median and HTLR cost functions.

While testing and retention-in-care rates help inform intervention programs, the corresponding targets for proportions aware and on-ART serve as control measures, i.e., help direct resources to relevant support programs and risk groups in cases where actual proportions (that are tracked through NHSS) fall short of the target proportions. Proportions aware and on-ART are leading indicators in the EHE, and thus, actual proportions are regularly tracked as part of the NHSS.

The federal plan for 'Ending the HIV Epidemic' aims for a 75% reduction in annual new infections by 2025 (to 9300 new cases per year) and a 90% reduction by 2030 (to 3000 new cases per year) [2,3]. Our results indicate that testing and retention-in-care programs alone would be insufficient to reach this goal optimally. The optimal strategy would achieve a significant reduction in annual incidence by 2030, reducing to about 14,200 to 18,000 annual new infections, which is a 50 to 60% reduction from 2015, but then gradually decrease at a slower rate thereafter, reaching about 4250 to 7200 annual new infections by 2070, which is an 86 to 91% reduction from 2015. To further accelerate the reduction in new infections, other novel interventions that are recently emerging, such as pre-exposure prophylaxis for HIV-negative individuals and more targeted testing through identification of transmission clusters, should be explored [2].

Acknowledgments

Financial support for this study was provided by a grant from the National Institute of Allergy and Infectious Diseases of the National Institutes of Health under Award Number R01AI127236. The funding agreement ensured the authors' independence in designing the study, interpreting the data, writing, and publishing the report. We also acknowledge Dr. Paul Farnham, Dr. Ram Shrestha, Dr. Zihao Li, and Dr. Stephanie Sansom, from the Centers for Disease Control and Prevention for their inputs on this work.

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. *Centers for Disease Control and Prevention*, Estimated HIV incidence and prevalence in the United States, 2010–2015, 2020. Available from: https://www.cdc.gov/hiv/pdf/library/reports/surveillance/cdc-hiv-surveillance-supplemental-report-vol-23-1.pdf.
2. *HIV. gov*., What is 'Ending the HIV Epidemic: A Plan for America'?, 2019. Available from: https://www.hiv.gov/federal-response/ending-the-hiv-epidemic/overview.
3. *America's HIV Epidemic Analysis Dashboard,* 2020. Available from: https://ahead.hiv.gov/indicators/incidence/. Published 2020. Accessed December 2020.
4. *Centers for Disease Control and Prevention*, Evidence of HIV Treatment and Viral Suppression in Preventing the Sexual Transmission of HIV, 2020. *A*vailable from: https://www.cdc.gov/hiv/pdf/risk/art/cdc-hiv-art-viral-suppression.pdf.
5. B. M. Branson, H. H. Handsfield, M. A. Lampe, R. S. Janssen, A. W. Taylor, S. B. Lyss, et al., Revised recommendations for HIV testing of adults, adolescents, and pregnant women in health-care settings, *Morb. Mortal Wkly. Rep.*, **55 (**2006), 1–CE.
6. A. F. Dailey, B. E. Hoots, H. I. Hall, R. Song, H. Demorah. Vital signs: human immunodeficiency virus testing and diagnosis delays-United States, *Morb. Mortal Wkly. Rep.*, **66** (2017), 1300.
7. *Centers for Disease Control and Prevention*, HIV Surveillance Report, 2015. Available from: http://www.cdc.gov/hiv/library/reports/hiv-surveillance.html.
8. *U.S. Department of Health & Human Services*, 2017 National HIV/AIDS Strategy (NHAS) Progress Report Released, 2018. Available from: https://www.hiv.gov/blog/2017-national-hivaids-strategy-nhas-progress-report-released.
9. R. K. Shrestha, H. A. Clark, S. L. Sansom, B. Song, H. Buckendahl, C. B. Calhoun, et al., Cost-effectiveness of finding new HIV diagnoses using rapid HIV testing in community-based organizations, *Public Health Rep.*, **123** (2008), 94–100.
10. R. K. Shrestha, L. Gardner, G. Marks, J. Craw, M. Faye, T. P. Giordano, et al., Estimating the cost of increasing retention in care for HIV-infected patients: results of the CDC/HRSA retention in care trial, *J. Acquir. Immune Defic. Syndr.*, **68** (2015), 345.
11. M. L. G. Buot, J. P. Docena, B. K. Ratemo, M. J. Bittner, J. T. Burlew, A. R. Nuritdinov, et al., Beyond race and place: distal sociological determinants of HIV disparities, *PloS One*, **9** (2014), e91711.
12. J. McMahon , C. Wanke, N. Terrin, S. Skinner, T. Knox, Poverty, hunger, education, and residential status impact survival in HIV, *AIDS Behav.*, **15** (2011), 1503–1511.
13. *HIV AND AIDS SOCIAL ISSUES*, 2016. Available from: https://www.avert.org/professionals/hiv-social-issues.
14. M. L. Brandeau, S. Z. Gregory, Optimal investment in HIV prevention programs: more is not always better, *Health Care Manag. Sci.*, **12** (2009), 27.
15. L. Guinness, L. Kumaranayake, K. Hanson, A cost function for HIV prevention services: is there a'u'--shape?, *Cost Eff. Resour. Allocation*, **5** (2007), 1–12.
16. J. A. Pellowski, S. C. Kalichman , K. A. Matthews and N. Adler. A pandemic of the poor: Social disadvantage and the US HIV epidemic, *Am. Psychol.*, **68** (2013), 197.
17. C. Gopalappa, P. G. Farnham, A. B. Hutchinson, S. L. Sansom, Cost effectiveness of the National HIV/AIDS Strategy goal of increasing linkage to care for HIV-infected persons, *J. Acquir. Immune Defic. Syndr.*, **61** (2012), 99–105.

18. F. Lin, P. G. Farnham, R. K. Shrestha, J. Mermin, S. L. Sansom, Cost effectiveness of HIV prevention interventions in the US, *Am. J. Prev. Med.*, **50** (2016), 699–708.
19. R. A. Bonacci, D. R. Holtgrave, US HIV incidence and transmission goals, 2020 and 2025, *Am. J. Prev. Med.*, **53** (2017), 275–281.
20. A. L. Avancena, D. W. Hutton, Optimization models for HIV/AIDS resource allocation: A systematic review, *Value Health*, 2020.
21. S. Kok, A. R. Rutherford, R. Gustafson, R. Barrios, J. S. Montaner, K. Vasarhelyi, Optimizing an HIV testing program using a system dynamics model of the continuum of care, *Health Care Manag. Sci.*, **18** (2015), 334–362.
22. K. O. Okosun, O. Makinde, I. Takaidza, Impact of optimal control on the treatment of HIV/AIDS and screening of unaware infectives, *Appl. Math. Model.*, **37** (2013), 3802–3820.
23. L. N. Steimle, D. L. Kaufman, B. T. Denton, Multi-model Markov decision processes: A new method for mitigating parameter ambiguity, *Optim. Online*, 2018.
24. S. M. Shechter, M. D. Bailey, A. J. Schaefer, M. S. Roberts, The optimal time to initiate HIV therapy under ordered health states, *Oper. Res.*, **56** (2008), 20–33.
25. J. E. Mason, B. T. Denton, N. D. Shah, S. A. Smith, Optimizing the simultaneous management of blood pressure and cholesterol for type 2 diabetes patients, *Eur. J. Oper. Res.*, **233** (2014), 727–738.
26. C. Gopalappa, P. G. Farnham, Y. H. Chen, S. L. Sansom, Progression and transmission of HIV/AIDS (PATH 2.0) a new, agent-based model to estimate HIV transmissions in the United States, *Med. Decis. Making*, **37** (2017), 224–233.
27. C. Yu, Y. Dong, G. Ren, Incorporating causal factors into reinforcement learning for dynamic treatment regimes in HIV, *BMC Med. Inf. Decis. Making*, **19** (2019), 19–29.
28. S. Parbhoo, *A reinforcement learning design for HIV clinical trials*, 2014.
29. V. Kompella, R. Capobianco, S. Jong, J. Browne, S. Fox, L. Meyers, et al., Reinforcement learning for optimization of COVID-19 mitigation policies, preprint, arXiv: 2010.10560.
30. R. Padmanabhan, N. Meskin, T. Khattab, M. Shraim, M. Al-Hitmi, Reinforcement learning-based decision support system for COVID-19, *Biomed. Signal Process. Control*, (2021), 102676.
31. M. I. Uddin, S. A. Ali Shah, M. A. Al-Khasawneh, A. A. Alarood, E. Alsolami, Optimal policy learning for COVID-19 prevention using reinforcement learning, *J. Inf. Sci.*, 2020.
32. H. Khadilkar, T. Ganu, D. P. Seetharam, Optimising lockdown policies for epidemic control using reinforcement learning, *Trans. Indian Natl. Acad. Eng.*, **5** (2020), 129–132.
33. R. Wan, X. Zhang, R. Song, Multi-objective reinforcement learning for infectious disease control with application to COVID-19 spread, preprint, arXiv: 2009.04607.
34. M. Arango, L. Pelov, Covid-19 pandemic cyclic lockdown optimization using reinforcement learning, preprint, arXiv: 2009.04647.
35. T. Smieszek, L. Fiebig, R. W. Scholz, Models of epidemics: when contact repetition and clustering should be included, *Theor. Biol. Med. Model.*, **6** (2009), 1–15.
36. A. Gosavi, Simulation-based optimization: Parametric optimization techniques and reinforcement learning, *Interfaces*, **35** (2005), 535.
37. R. S. Sutton, A. G. Barto, *Reinforcement Learning: An Introduction*, MIT Press, 2018.
38. Z. Li, D. W. Purcell, S. L. Sansom, D. Hayes, I. Hall, Vital signs: HIV transmission along the continuum of care-United States, 2016, *Morb. Mortal. Wkly. Rep.*, **68** (2019), 267–272.

39. E. U. Jacobson, K. A. Hicks, E. L. Tucker, P. G. Farnham, S. L. Sansom, Effects of reaching national goals on HIV incidence, by race and ethnicity, in the United States, *J. Public Health Manag. Pract.*, **24** (2018), E1–E8.
40. *U.S. Department of Health & Human Services*, 2017 National HIV/AIDS Strategy (NHAS) Progress Report Released, 2018. Available from: https://www.hiv.gov/blog/2017-national-hivaids-strategy-nhas-progress-report-released.
41. *Centers for Disease Control and Prevention*, HIV Prevention Progress Report, 2019. Available from:https://www.cdc.gov/hiv/pdf/policies/progressreports/cdc-hiv-preventionprogressreport.pdf.
42. *UNAIDS*, 90-90-90 An ambitious treatment target to help end the AIDS epidemic, 2014. Available from: https://www.unaids.org/sites/default/files/media_asset/90-90-90_en.pdf.
43. O. Gottesman, F. Johansson, J. Meier, J. Dent, D. Lee, S. Srinivasan, et al., Evaluating reinforcement learning algorithms in observational health settings, preprint, arXiv:1805.12298.
44. C. Kreatsoulas, S. Subramanian, Machine learning in social epidemiology: Learning from experience, *SSM- Popul. Health*, **4** (2018), 347.
45. E. M. Gardner, M. P. McLees, J. F. Steiner, C. del Rio, W. J. Burman, The spectrum of engagement in HIV care and its relevance to test-and-treat strategies for prevention of HIV infection, *Clin. Infect. Dis.*, **52** (2011), 793–800.
46. L. Kumaranayake, The economics of scaling up: cost estimation for HIV/AIDS interventions, *Aids*, **22** (2008), S23–S33.
47. A. Lansky, J. Christopher, O. Emeka, S. Catlainn, M. P. Joyce, E. DiNenno, et al., Estimating the number of heterosexual persons in the United States to calculate national rates of HIV infection, *PloS One*, **10** (2015), e0133543.
48. A. Chandra, V. G. Billioux, C. Copen, C. Sionean, HIV risk-related behaviors in the United States household population aged 15-44 years: data from the National Survey of Family Growth, 2002 and 2006-2010, *Natl. Health Stat. Rep.*, **46** (2012), 1–19.
49. N. Khurana, E. Yaylali, P. G. Farnham, K. A. Hicks, B. T. Allaire, E. Jacobson, et al., Impact of improved HIV care and treatment on PrEP effectiveness in the United States, 2016–2020, *J. Acquir. Immune Defic. Syndr.*, **78** (2018), 399–405.
50. U. Wilensky, *NetLogo*, Evanston, IL: Center for Connected Learning and Computer-Based Modeling, Northwestern University, 1999.
51. *Centers for Disease Control and Prevention*, Recommendations for HIV screening of gay, bisexual, and other men who have sex with men-United States, 2017, *MMWR Morb. Mortal Wkly. Rep.*, **66** (2017), 830.

AIMS Press